ULB Faculté des Sciences

Causal and predictive modeling of customer churn

Lessons learned from empirical and theoretical research

Thesis presented by Théo Verhelst

in fulfilment of the requirements of the PhD Degree in Computer Science

Academic year 2023-2024

Supervisor: Professor Gianluca Bontempi

Machine Learning Group

Thesis jury

Matthieu Defrance (Université Libre de Bruxelles), chair
Maarten Jansen (Université Libre de Bruxelles), secretary
Catharina Olsen (Vrije Universiteit Brussel)
Szymon Jaroszewicz (Polish Academy of Sciences)
Wouter Verbeke (Katholieke Universiteit Leuven)

orange™ innoviris .brussels we fund your future MLG MACHINE LEARNING GROUP

Causal and predictive modeling of customer churn

Lessons learned from empirical and theoretical research

Théo Verhelst

Supervisor
Gianluca Bontempi

Thesis jury
Matthieu Defrance, chair
Maarten Jansen, secretary
Catharina Olsen
Szymon Jaroszewicz
Wouter Verbeke

ULB

The ladder in the cover illustration represents *Pearl's causal hierarchy*, described in Section 2.2.1. The shadow of the ladder symbolizes the data as the only source of information available to us, which is a deformed representation of the reality of the world.

Declaration of authorship

This thesis was written under the supervision of Prof. Gianluca Bontempi (Université Libre de Bruxelles, Belgium). The members of the jury are:

- Prof. Matthieu Defrance, chair (Université Libre de Bruxelles, Belgium)
- Prof. Maarten Jansen, secretary (Université Libre de Bruxelles, Belgium)
- Prof. Catharina Olsen (Vrije Universiteit Brussel, Belgium)
- Prof. Szymon Jaroszewicz (Polish Academy of Sciences, Poland)
- Prof. Wouter Verbeke (Katholieke Universiteit Leuven, Belgium)
- Prof. Gianluca Bontempi (Université Libre de Bruxelles, Belgium)

I declare that this thesis does not include any content that has been acknowledged for the award of any other degree or diploma at any university or equivalent institution. To the best of my knowledge, this thesis does not include material that has been previously published or authored by another individual, except in cases where proper acknowledgment is provided within the thesis text.

Théo Verhelst

Summary

Customer churn is an important concern for large companies, especially in the telecommunications sector. Customer retention campaigns are often used to mitigate churn, but targeting the right customers based on their historical profiles presents an important challenge. Companies usually have recourse to two data-driven approaches: churn prediction and uplift modeling. In churn prediction, customers are selected on the basis of their propensity to churn in the near future. In uplift modeling, only customers who react positively to the campaign are considered. Uplift modeling is used in various other domains, such as marketing, healthcare, and finance. Despite the theoretical appeal of uplift modeling, its added value with respect to conventional machine learning approaches has rarely been quantified in the literature.

This doctoral thesis is the result of a collaborative research project between the Machine Learning Group (ULB) and Orange Belgium, funded by Innoviris. This collaboration offers a unique research opportunity to assess the added value of causal-oriented strategies to address customer churn in the telecommunication sector. Following the introduction, we give the necessary background in probability theory, causality theory, and machine learning, and we describe the state of the art in uplift modeling and counterfactual identification. Then, we present the contributions of this thesis:

- An empirical comparison of various predictive and causal models for selecting customers in churn prevention campaigns. We perform several benchmarks of different state-of-the-art approaches on real-world datasets and in live campaigns with our industrial partner, we propose a new approach that exploits domain knowledge to improve predictions, and we make available the first public churn dataset for uplift modeling, whose unique characteristics make it more challenging than the few other public uplift datasets.
- Counterfactual identification allows one to classify the different behaviors of customers in response to a marketing incentive. This can be used to establish profiles of customers sensitive to the campaign, and subsequently improve marketing operations. We derive novel bounds and point estimators on the probability of counterfactual statements based on uplift models.
- A comprehensive comparison of predictive and uplift modeling, starting from firm theoretical foundations and highlighting the parameters that influence the performance of both approaches. In particular, we provide a new formulation of the measure of profit, a formal proof of the convergence of the uplift curve to the measure of profit, and an illustration, through simulations, of the conditions under which predictive approaches still outperform uplift modeling.

Our theoretical and empirical assessments of uplift modeling suggest that it often fails to deliver the anticipated advantages over predictive modeling, especially in scenarios such as customer churn within the telecom sector, characterized by class imbalance, limited separability, and cost-benefit considerations. These results are broadly aligned with the practical experience of our industrial partner and with the existing scientific literature. Our counterfactual probability estimators allow us to characterize customers at a level inaccessible to conventional predictive modeling, revealing new insights on the behavior and preferences of customers.

Résumé

L'attrition de la clientèle est une préoccupation importante pour de nombreuses entreprises, notamment dans le secteur des télécommunications. Des campagnes de fidélisation sont souvent utilisées pour réduire le taux de désabonnement, mais cibler les bons clients en fonction de leur profil représente un défi majeur. Les entreprises ont généralement recours à l'une de deux approches : la prédiction de *churn* (attrition) et la modélisation de l'*uplift*. Dans la prédiction de churn, les clients sont sélectionnés sur la base de leur propension estimée à se désabonner dans un avenir proche. Dans la modélisation de l'uplift, seuls les clients qui réagissent positivement à la campagne sont pris en compte. Les prédictions de ces deux approches sont basées sur les caractéristiques des clients. La modélisation de l'uplift est aussi utilisée dans d'autres domaines tels que le marketing, la médecine et la finance. Malgré son attrait théorique, la valeur ajoutée de la modélisation de l'uplift par rapport à l'approche plus conventionnelle de prédiction de churn a rarement été quantifiée dans la littérature.

Cette thèse doctorale est le résultat d'un projet de recherche collaborative entre le Machine Learning Group (ULB) et Orange Belgique, financé par Innoviris. Cette collaboration offre une opportunité unique pour évaluer la valeur ajoutée de stratégies causales pour prévenir l'attrition de la clientèle dans le secteur des télécommunications. Après l'introduction, nous présentons les base théoriques nécessaires en théorie des probabilités, théorie de la causalité et apprentissage automatique, et nous décrivons l'état de l'art en matière de modélisation de l'uplift et d'identification contrefactuelle. Nous présentons ensuite les contributions de cette thèse :

- Une comparaison empirique de divers modèles prédictifs et causaux pour la sélection des clients dans les campagnes de prévention du désabonnement. Nous comparons différentes approches de pointe sur des jeux de données réels et dans des campagnes de rétention avec notre partenaire industriel, nous proposons une nouvelle approche qui exploite la connaissance du domaine pour améliorer les prédictions, et nous rendons public un jeu de données de churn pour la modélisation de l'uplift, dont les caractéristiques uniques le rendent plus difficile que les quelques autres jeux de données d'uplift publics.
- L'identification contrefactuelle permet de classer les différents comportements des clients en réponse à une incitation marketing. Elle peut être utilisée pour établir des profils de clients réagissant positivement à la campagne et, par la suite, améliorer les opérations de marketing. Nous dérivons de nouvelles bornes et plusieurs estimateurs ponctuels de la probabilité de propositions contrefactuelles basées sur des modèles de l'uplift.
- Une comparaison de l'approche prédictive et de la modélisation de l'uplift à partir de fondements théoriques, mettant en évidence les paramètres influençant la performance des deux approches. En particulier, nous donnons une nouvelle formulation de la mesure de profit, une preuve formelle de la convergence de la courbe d'uplift vers la mesure de profit, et une illustration, par des simulations, des conditions dans lesquelles l'approche prédictive reste plus performante que la modélisation de l'uplift.

Nos évaluations théoriques et empiriques de la modélisation de l'uplift suggèrent que cette dernière n'apporte souvent pas les avantages escomptés par rapport à la modélisation prédictive, en particulier dans des scénarios tels que la prédiction d'attrition de clientèle dans le secteur des télécommunications, caractérisée par

un déséquilibre entre les classes, une séparabilité des classes limitée, et des considérations de coût-bénéfice. Ces résultats sont largement conformes à l'expérience pratique de notre partenaire industriel et à la littérature scientifique existante. Nos estimateurs de probabilités contrefactuelles nous permettent de caractériser les clients à un niveau inaccessible à la modélisation prédictive conventionnelle, révélant de nouvelles perspectives sur le comportement et les préférences des clients.

Acknowledgments

When the time came to find a subject for my master's thesis in 2017, I contacted Prof. Gianluca Bontempi, in the hope that his expertise would drive me forward and possibly lead to a Ph.D. after my master's degree. I have not regretted this decision; Gianluca not only offered me unique research opportunities, but also fostered an environment conducive to learning and growth. Notably, he offered a precious guidance and structured supervision in the initial stages of research, gradually giving more autonomy as I reached the conclusion of the project, which can be seen in the diminishing number of revisions he made to my manuscripts. I can only be thankful for this invaluable guidance.

The initial stages of the project were also under the supervision of Olivier Caelen. Having worked both in academia and industry, his mentoring was crucial in helping me understand the relationship between academic research and practice in business, and in motivating me to do 10-km runs. I am also grateful for the help, support, and supervision provided by Denis, Jeevan, Jean-Christophe, and the other people at Orange.

On the university side, I am glad to have worked, and shared coffee and beers with amazing colleagues such as Jacopo, Gian Marco, Bertrand, and others. And, of course, Robin, Cédric and Antoine; long discussions at the Gauguin were a crucial part of my development as a researcher. Special mention to Robin for the math support and the willingness to listen to whatever I had on my mind at any time.

I am very grateful to the members of the jury for accepting to read and evaluate my thesis. Also, although they have already been mentioned, I have to thank again Robin, Gianluca, Jacopo, and Gian Marco, who have dedicated their time to read this thesis and gave me insightful feedback. Your help was very much appreciated. Thank you also to Ronnie Raeymaekers and the people at Innoviris for making this Ph.D. project possible.

Finalement, merci à mes chers parents et frères, sœurs, et adelphes, pour leur affection et soutien qui dure depuis aussi longtemps que je m'en souvienne. Merci à Coralie d'être à mes côtés, de me soutenir, de m'écouter, de me comprendre, et, plus généralement, d'exister.

Contents

Index of notation

General

$P(\cdot)$	Probability measure, 158
$\boldsymbol{v}$	Random variable (bold font), 158
v	Realization of $\boldsymbol{v}$, 158
$\mathscr{V}$	Domain of $\boldsymbol{v}$, 158
$f_{\boldsymbol{v}}(\cdot)$	Probability density function of $\boldsymbol{v}$, 160
$\mathbb{E}[\boldsymbol{v}]$	Expected value of $\boldsymbol{v}$, 16
$\mathrm{Var}(\boldsymbol{v})$	Variance of $\boldsymbol{v}$, 17
$M_n(\boldsymbol{v})$	nth central moment of $\boldsymbol{v}$, 17
$R_n(\boldsymbol{v})$	nth raw (non-central) moment of $\boldsymbol{v}$, 17
$H(\boldsymbol{v})$	Entropy of $\boldsymbol{v}$, 18
$H(\boldsymbol{v}, \boldsymbol{w})$	Joint entropy of $\boldsymbol{v}$ and $\boldsymbol{w}$, 19
$H(\boldsymbol{v} \mid \boldsymbol{w})$	Conditional entropy of $\boldsymbol{v}$ given $\boldsymbol{w}$, 19
$D(P \parallel Q)$	Relative entropy between probability distributions P and Q, 19
$I(\boldsymbol{v}; \boldsymbol{w})$	Mutual information between $\boldsymbol{v}$ and $\boldsymbol{w}$, 20
$\mathbb{I}[\cdot]$	Iverson bracket, equals one when the expression between brackets is true, zero otherwise, 40
$\boldsymbol{v} \perp \boldsymbol{w} \mid \boldsymbol{x}$	Variables $\boldsymbol{v}$ and $\boldsymbol{w}$ are conditionally independent given $\boldsymbol{x}$, 162

Machine learning

$\boldsymbol{y}$	Outcome indicator with domain $\mathscr{Y} = \{0, 1\}$, 48
$\boldsymbol{t}$	Treatment indicator with domain $\mathscr{T} = \{0, 1\}$, 48
$\boldsymbol{r}$	Reach indicator with domain $\mathscr{R} = \{0, 1\}$, 81
$\boldsymbol{x}$	Feature vector with domain $\mathscr{X} \subseteq \mathbb{R}^n$, 48
$\mathrm{do}(\boldsymbol{t} = t)$	Causal intervention $\boldsymbol{t} = t$ 32
$S_t = P(\boldsymbol{y}_t = 1)$	Probability of the outcome $\boldsymbol{y} = 1$ under $\mathrm{do}(\boldsymbol{t} = t)$, 48
α	$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0)$, probability of being a *sure thing* customer, 60
β	$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0)$, probability of being a *persuadable* customer, 60
γ	$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1)$, probability of being a *do-not-disturb* customer, 60
δ	$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1)$, probability of being a *lost cause* customer, 60
μ	Vector of counterfactual probabilities, $\mu = [\alpha, \beta, \gamma, \delta]$, 123
$\mu_1, \mu_2, \mu_3, \mu_4$	Another notation for $\alpha, \beta, \gamma, \delta$; 123
$S_t(x)$	$P(\boldsymbol{y}_t = 1 \mid \boldsymbol{x} = x)$ for $t = 0, 1$; 48
$\alpha(x), \beta(x), \ldots$	$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x), P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x), \ldots$, 60
U	Uplift, defined as $U = S_0 - S_1$, 48
$\boldsymbol{D}$	Training set or test set, defined as $\boldsymbol{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{t}^{(i)})\}_{i=1}^{N}$, 54
$\mathscr{M}(x, \boldsymbol{D})$	Prediction for features x of model $\mathscr{M}$ trained on set $\boldsymbol{D}$, 89
$\tau(D_{\mathrm{tr}}) \in \mathbb{R}$	Classification threshold for a model trained on a training set D_{tr}, 89
$\rho \in [0, 1]$	Prescription rate, 89

Graph theory

$\text{PA}_G(\boldsymbol{v})$	The set of parents of vertex $\boldsymbol{v}$ in a DAG G, 28
$\text{CH}_G(\boldsymbol{v})$	The set of children of vertex $\boldsymbol{v}$ in a DAG G, 28
$\text{AN}_G(\boldsymbol{v})$	The set of ancestors of vertex $\boldsymbol{v}$ in a DAG G, 28
$\text{DE}_G(\boldsymbol{v})$	The set of descendants of vertex $\boldsymbol{v}$ in a DAG G, 28
$G_{\overline{\boldsymbol{v}}}$	A graph G but with all edges going into $\boldsymbol{v}$ removed, 32
$\boldsymbol{v} \perp_G \boldsymbol{w} \mid \boldsymbol{x}$	Nodes $\boldsymbol{v}$ and $\boldsymbol{w}$ are d-separated by $\boldsymbol{x}$ in the graph G, 28

Probability distributions and special functions

$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and standard deviation σ, 102
$\text{Bern}(p)$	Bernoulli distribution with probability p, 21
$\text{Bin}(n, p)$	Binomial distribution with n trials and probability p, 170
$\text{Cat}(p_1, \ldots, p_k)$	Categorical distribution with event probabilities $p_1, \ldots, p_k$; 124
$\text{Beta}(a, b)$	Beta distribution with parameters a, b; 21
$\text{Dir}(m)$	Dirichlet distribution with parameter vector $m = [a, b, c, d]$, 23
M	Sum of the parameters of the Dirichlet distribution, 106
m_1, m_2, m_3, m_4	Another notation for a, b, c, d; 24
$\text{BB}(m)$	Bivariate beta distribution from Olkin and Trikalinos (2015) with parameter vector $m = [a, b, c, d]$, 123
$\text{GD}(a_1, \ldots, b_3)$	Generalized Dirichlet distribution with parameters $a_1, a_2, a_3, b_1, b_2, b_3$; 128
$\Gamma(\cdot)$	Gamma function, a generalization of the factorial function, 22
$\psi(\cdot)$	Digamma function, a generalization of the harmonic numbers, 23
$\text{B}(a, b)$	Bivariate beta function, 22
$\text{B}(m)$	Multivariate beta function, for an input vector $m = [m_1, \ldots, m_k]$, 24
$x^{\overline{n}}$	Rising factorial, $x^{\overline{n}} = x(x+1)\ldots(x+n-1)$, 23
$x_{\overline{n}}$	Harmonic difference, $x_{\overline{n}} = \sum_{i=0}^{n-1}(x+i)^{-1}$, 185
$\left[{a \atop b}\right]$	Unsigned Stirling number of the first kind, 193

Part I

Overview

Anything that happens, happens.
Anything that, in happening, causes something else to happen, causes something else to happen.
Anything that, in happening, causes itself to happen again, happens again.
It doesn't necessarily do it in chronological order, though.

Mostly Harmless, Douglas Adams

1

Introduction

Machine learning is reshaping the contemporary technological landscape. It enables systems to learn and improve from data without explicit programming. The rapidly advancing capabilities of machine learning systems not only capture attention, but also drive innovation, spark discussions, and change the way we interact with technology. Larger and more complex models, fueled by increased computational power and vast datasets, have accomplished unprecedented achievements, notably in the processing and generation of text, images, and videos. State-of-the-art large language models contain more than a hundred billion parameters (Brown et al., 2020), allowing them to comprehend and generate human-like text on a wide range of topics. Generative models have demonstrated astonishing proficiency in creating realistic images and videos (Ramesh et al., 2021).

While these two applications have received a lot of media attention in recent years, machine learning is used in a much wider range of domains, such as healthcare, content recommendation, online advertisement, energy management, fraud detection, scientific discovery, etc. Patient diagnosis can be improved by leveraging large medical databases and electronic medical records with machine learning (Mintz and Brodie, 2019). Social media, e-commerce, and entertainment websites use recommender systems to help the user find new and relevant items (Portugal, Alencar, and Cowan, 2018). Current power systems are using machine learning to address new issues such as dynamic resource allocation and incorporating renewable energy sources and large-scale real-time sensor data (Ibrahim, Dong, and Yang, 2020).

These recent advances in machine learning are made possible by increased computational power and the development of new learning algorithms, but also, and more crucially, by the increasing pervasiveness of data in various sectors. We are now living in the so-called *big data* era. More and more systems generate data, such as sensors on various devices, social media interactions, online transactions, and digital communications. This surge in data production has reached unprecedented scales, as shown in Fig. 1.1, and is often characterized by the three Vs: volume, velocity, and variety. The sheer volume of data is massive, generated at an accelerating pace (velocity), and exhibits diverse formats and structures (variety). This abundance of data is a goldmine for machine learning algorithms, providing the raw material for training and refining models to extract meaningful insights, make predictions, and automate decision-making processes across diverse domains.

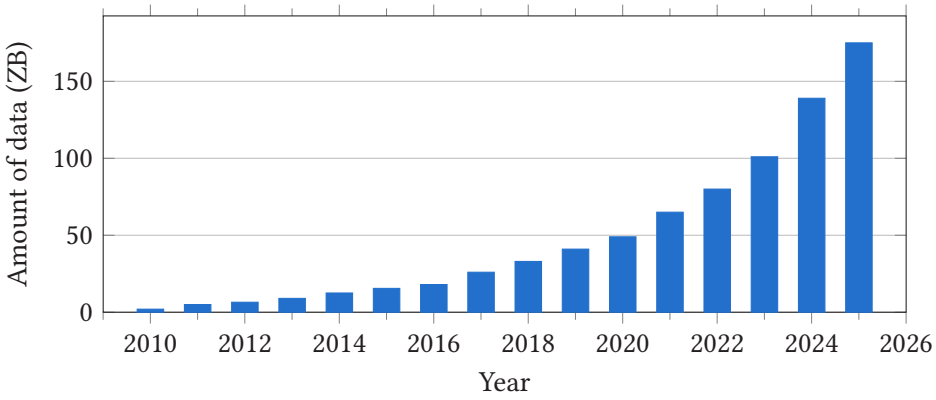


Figure 1.1 Amount of data data captured, created, and replicated worldwide, in zettabytes (1 ZB = 1 trillion GB = 10^{21} bytes). Values after 2018 are estimated by extrapolation. Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere.

However, traditional machine learning faces significant challenges when dealing with big data, primarily due to the need for distributed computation. As datasets grow in size, the computational demands increase, often surpassing the capabilities of a single machine. Distributing computations across multiple machines becomes essential, introducing complexities related to communication, synchronization, and load balancing.

Another issue faced by machine learning in the big data era is that they often operate in the realm of association rather than causation. Although associations can reveal intricate patterns and relationships, they may not necessarily represent true cause-and-effect dynamics. Understanding causality is essential for informed decision-making, understanding and justifying predictions, and, more broadly, making intuitive sense of the world. There is a need to develop methodologies and models that go beyond mere correlations, enabling the extraction of causal insights from large datasets.

1.1 Machine learning in business analytics

The development of machine learning has led to a significant shift in business practices, with the integration of artificial intelligence becoming commonplace among most large companies. It empowers organizations to extract valuable insights, predict trends, and optimize various processes. For example, financial companies can detect fraudulent behavior among millions of credit card transactions (Dal Pozzolo, Caelen, Le Borgne, et al., 2014), and telecom companies can predict which customers are likely to *churn*, that is, to stop their subscription (Jain, Khunteta, and Srivastava, 2021). However, these tasks have characteristics quite different from the assumptions underlying the most common machine learning models. For example, there are much fewer fraudulent transactions than genuine transactions. This characteristic is called *class imbalance*. Although it is desirable from a customer and business perspective, it poses a significant challenge for machine learning algorithms, which often operate optimally under the assumption of a balanced dataset. The inherent class imbalance in such tasks can lead to biased models that prioritize accuracy but may struggle to identify the minority class. In some settings, the data labels (i.e., fraudulent vs. genuine, or churner vs. non-churner) may

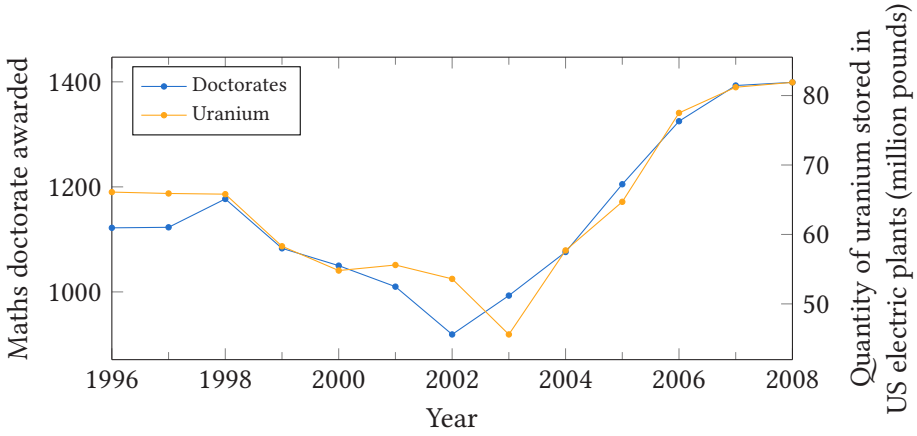


Figure 1.2 The number of math doctorates awarded in the US correlates with the quantity of uranium stored in US electric plants. Sources: Fiegener (2010) and United States Census Bureau (2011), idea from tylervigen.com.

not even be available, and specific strategies must be adopted, such as transfer learning (Lebichot et al., 2021).

Another defining characteristic of such settings in business application is the inherent difficulty in determining the outcome of interest from the available data. This is called low *class separability*. Customer behavior is influenced by a large number of factors, most of which cannot be measured or influenced. Customers might decide to churn when they move to a new place, or when they hear about another company through advertisements. New privacy policies restrict the scope of data collection and storage permissible for companies and, as a result, machine learning models which predict customer behavior must deal with significant uncertainty.

1.2 The limits of predictive analytics

The belief that sheer volume of data and powerful machine learning models always lead to meaningful insights can have unexpected and disappointing results. As the adage says, "correlation does not imply causation". With a significant amount of data, spurious correlations will most certainly emerge, linking variables that are causally unrelated. This is demonstrated in Fig. 1.2, where we can observe a striking correlation between the number of maths doctorates awarded in the US and the total quantity of uranium stored in US power plants between 1996 and 2008. This example comes from Tyler Vigen, who listed several other spurious correlations on his website.[1] This clearly demonstrates that detecting an association[2] between events is never sufficient to prove the existence of a causal link between these events.

Despite this fundamental limitation, the achievements of machine learning in detecting subtle patterns in large datasets has led the data science community to develop even larger and more data-intensive algorithms, without attempting to validate the causal nature of the associations exploited by these models. Judea Pearl, who developed the theory of causality based on structural models, writes:

[1] https://tylervigen.com/spurious-correlations, last accessed 2023-12-12.

[2] The word *correlation* refers to a linear relationship between variables, whereas *association* refers to any kind of relationship between variables, linear or otherwise.

Figure 1.3 A deer swimming in a lake. An image recognition model based on spurious associations might classify it as a sea animal. Image by Misserion on Flickr, under the CC BY-2.0 license.

> Machine learning as it is currently practiced cannot yield the kind of understanding that intelligent decision making requires.
>
> *Data versus Science: Contesting the Soul of Data-Science*, Pearl (2020)

This oversight may be harmless for some tasks. For example, a model classifying images of deer and dolphins might predict "dolphin" if the background is blue water, and "deer" if the background contains trees, without even taking into account the visual characteristics of the animal. If only the model accuracy is considered, this approach might be sufficient. However, Ming, Yin, and Yixuan Li (2022) show that the validity of such a model is restricted to the specific distribution represented by the training set; a deer swimming in a lake (Fig. 1.3) might be classified as "dolphin" with a high degree of confidence. In the majority of real-world problems, a high accuracy for in-distribution samples (that is, samples coming from a distribution similar to that of the training set) is not sufficient.

This issue can be avoided by imposing an *inductive bias* during the learning phase. In the example of animal picture classification, this might consist in ensuring that the predictions are based on the visual characteristics of the animal rather than on the background. More generally, an inductive bias restricts the space of possible functions that a model can explore in order to accomplish its objective. This is a critical element in settings where machine learning models are used to make decisions with real-world consequences, such as granting loans or evaluating criminal offenses (Chakraborty, Majumder, and Menzies, 2021). Careless modeling has been shown to be biased with respect to protected individual characteristics such as gender and race (Buolamwini and Gebru, 2018). Causality theory provides essential tools to formalize the notions of generalization, bias, and fairness in the context of automated decision-making (Jalaldoust and Bareinboim, 2023; Plečko and Bareinboim, 2022).

Counterfactual reasoning is another facet of causal analysis, and refers to potential events in scenarios that contradict the observed facts. An example of a counterfactual question is "Would I still have a headache if I did not take this pill?" Counterfactual statements are used in a variety of applications, such as articulating the notion of fairness (as discussed above), finding which patient will benefit from a given treatment, or finding the right customers to call in phone marketing campaigns. Counterfactual reasoning represents the highest of the three layers of Pearl's *causal hierarchy*, described

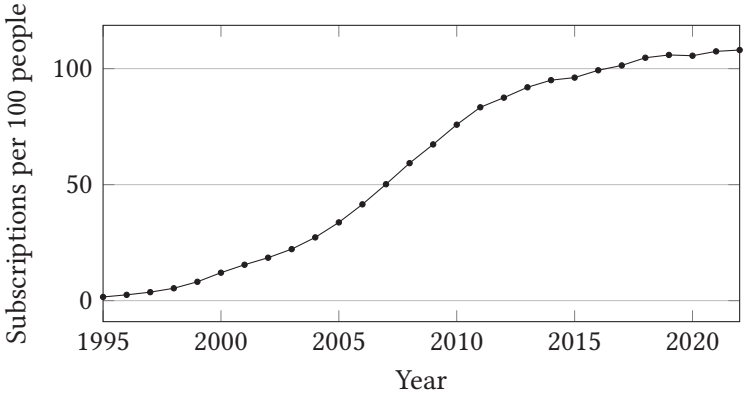


Figure 1.4 Number of mobile cellular subscriptions worldwide per year, per 100 people. Source: International Telecommunication Union (ITU) World Telecommunication / ICT Indicators Database, under the CC BY-4.0 license.

in detail in Section 2.2.1, the first and second layers representing, respectively, observation (e.g., I see that people who take the pill do not have a headache) and interventions (e.g., when I give a pill to someone, their headache diminishes). Although counterfactual reasoning permeates our intuition of causality, its mathematical definition is not trivial. Differentiating the three layers of this hierarchy represents one of the biggest achievements of causality theory (Bareinboim et al., 2020).

Generally, causal analysis aims to either infer the causal mechanisms underlying the observed data or to inform other systems (such as machine learning models) about these mechanisms. The literature on inferring causal mechanisms can be further divided into two subfields: *causal discovery* and *causal inference*. Causal discovery is concerned with determining the cause-and-effect relationships between variables and, as such, provides general insights about the system under study. Causal inference estimates the magnitude of the effect of one variable on another, given that we have established their causal relationship. This represents a more fine-grained knowledge of the system. In this thesis, we will focus on causal inference, more specifically, on the effect of a direct marketing action on the propensity of a customer to churn in the telecommunication industry.

1.3 Customer churn in telecom

Telecom companies, operating in a highly competitive and saturated market, need to develop innovative strategies to maintain a competitive edge. The global number of mobile phone users has consistently risen since the early 21st century, reaching 8.6 billion mobile phone subscriptions worldwide in 2022. The number of subscriptions to mobile services has exceeded the global population, as shown in Fig. 1.4. As highlighted by Jain, Khunteta, and Srivastava (2021), it is more expensive to acquire new customers than retaining existing ones. Consequently, companies have transitioned from a sales-oriented to a customer-oriented marketing approach. By fostering customer relationships grounded in trust and commitment, telecom companies can mitigate incentives for client churn, thereby enhancing benefits through subsequent customer lifetime value.

One of the various marketing processes used by telecom companies to improve

customer relationships is to conduct retention campaigns. This traditionally consists in selecting clients according to some simple statistical criteria and offering them a promotion or advantage. Typical promotions include a reduced invoice, free calls or SMS, or increased data limits. However, due to the simple nature of this statistical analysis, it is plausible that the customers thus reached might never have planned to churn in the first place. Although this is not an issue for the customer, it would be far more beneficial for the telecommunication company to contact only risky customers, to ensure that efforts are focused on customers who would otherwise churn if no action were taken. The problem of detecting potential churn can be addressed with machine learning, by collecting data about customers and using this information to infer typical patterns exhibited by risky clients. Today, most major telecommunication companies adopt this data-driven approach, and a large body of literature is devoted to customer churn prediction with machine learning (Idris and Khan, 2014; Mitrović et al., 2018; Óskarsdóttir, Bravo, et al., 2017; Óskarsdóttir, Van Calster, et al., 2018; Verbeke, Martens, and Baesens, 2014; Zhu, Baesens, and Broucke, 2017).

This thesis is the result of a research collaboration with Orange Belgium, a major telecommunication company in Belgium. As such, we focus on the issue of customer churn from the perspective of Orange Belgium. The company's experts categorize the reasons for customer churn into four categories.

Bill shock This occurs when a customer has an unusually large service usage, which results in an important "out-of-bundle" amount (i.e., the client is charged much more than usual), itself leading some customer to churn. This scenario is well understood and verified in practice. It is believed to be the most important cause of churn. Retention campaigns may even focus exclusively on this category by filtering out customers with a low out-of-bundle amount.

Customer dissatisfaction Multiple factors can influence customer satisfaction, including service quality and network quality. A customer who has numerous connection outages during phone calls or is unable to properly use online services will be more likely to look for other service providers.

Wrong positioning Choosing the right tariff plan suited to one's service usage habits is sometimes difficult. On the one hand, if not enough call time is provisioned, an "out-of-bundle" amount is likely to be charged at the end of the month. On the other hand, an expensive tariff plan results in a high fixed cost for the customer. When the needs of a customer do not correspond to the chosen tariff plan, we say that the customer is wrongly positioned. Wrong positioning results in most cases in a higher bill than expected and is a significant cause of churn.

Churn due to a move Subscriptions are often sold in bundles, comprising a mobile phone subscription, a landline phone, a television subscription, and an internet connection. In this case, the subscription is tied to the address of the customer. When the client moves to another place, it is quite common for them to also change to another telecommunication service provider. Therefore, this is a significant cause of churn, albeit of a different nature from the other scenarios discussed above.

The pipeline for a typical customer retention campaign is shown in Figure 1.5. First, a predictive model is trained on historical data from previous campaigns to predict

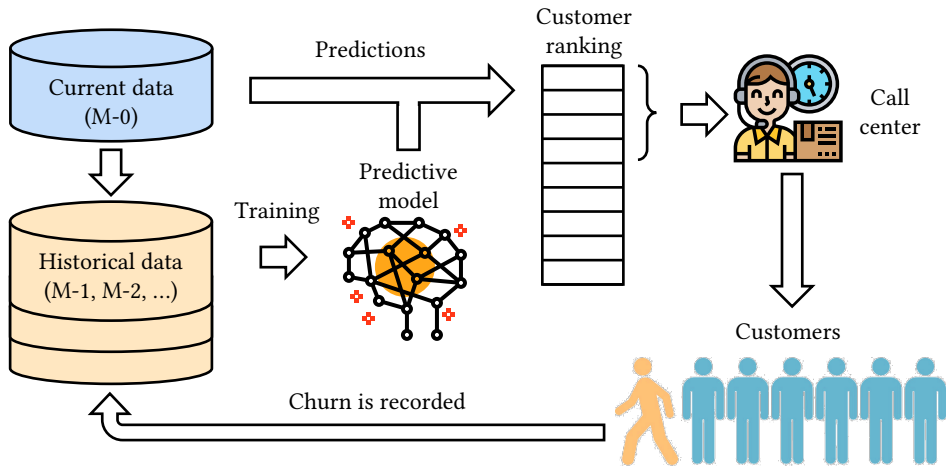


Figure 1.5 Overview of the pipeline for customer retention. Icons by eucalyp on iconfinder.com, licensed under CC BY 3.0, adapted for style.

which customers churned by detecting association patterns between customer characteristics and the churn outcome. Then, this model predicts a score for each of the current customers and ranks them accordingly. The list of customers with the highest scores is randomly divided into a target group and a control group, and the target group is sent to a call center. The call center contacts each of them individually and the reaction of the customer is recorded and added to the historical data set for training future models. The control group is used as a baseline to measure the effect of the campaign. If the proportion of customers who subsequently churned is significantly lower in the target group than in the control group, we can assume that the campaign successfully convinced some customers to stay.

The data used to predict customer churn consists of a monthly summary of the customer's activity, with a few hundreds of features grouped in 5 different categories:

- Service usage metadata: duration of calls, mobile data usage, etc.
- Subscription: tariff plan, cable connection, etc.
- Revenues: price of tariff plan, out-of-bundle amount, etc.
- Customer hardware: type of phone, number of devices, etc.
- Socio-demographics: age, region of residence, etc.

The customer ranking provided by the machine learning model is based on the probability of churn. However, this approach disregards the causal aspect of the problem. Targeting high-risk customers is not necessarily the best strategy: for instance, some customers slightly less inclined to churn could be far more receptive to retention offers, and focusing the campaign on these customers could be more effective. This idea is exploited by uplift models (Devriendt, Berrevoets, and Verbeke, 2021; Gutierrez and Gérardy, 2016). Instead of estimating the probability of churn based only on input features, uplift modeling focuses on estimating how much this probability changes when the marketing action is performed. In recent years, many machine learning-based uplift models have been developed, such as S-learner, T-learner, and X-learner (Künzel et al., 2019; W. Zhang, J. Li, and L. Liu, 2021).

The term *uplift* is used mainly in business settings where data from large campaigns are available, while in other fields the same quantity is called *conditional average treatment effect* (CATE), or *heterogeneous treatment effect* (Gutierrez and Gérardy, 2016). These fields usually assume having only access to passive observations, for example, data on the effect of a learning program on student dropout when the students decide by themselves to take up the program.

Beyond uplift modeling, the possible behavior of a customer can be summarized in terms of *counterfactual* statements (Devriendt, Berrevoets, and Verbeke, 2021). As mentioned in Section 1.2, counterfactual statements refer to potential events in situations that contradict the observed facts, such as "Given that a customer was not called and churned, would they still have churned if we had called them?" More generally, we can distinguish four types of customer based on counterfactual events:

- Sure thing: the customer does not churn regardless of the action.
- Persuadable: the customer churns only if not contacted.
- Do-not-disturb: the customer churns only if contacted.
- Lost cause: the customer churns regardless of the action.

Ideally, marketing actions should only target persuadable customers. However, we can observe only one of the two potential outcomes (this is known as the *fundamental problem of causal inference* (Holland, 1986)), and, generally, it is difficult to determine with certainty who are the persuadable customers. We can, however, estimate the probability of each customer to belong to each category.

Counterfactuals and uplift are closely related, yet formally distinct notions. The counterfactual distribution describes the probability of each possible combination of realized and hypothetical outcomes, while the uplift describes the change in the probability of the outcome due to treatment. While the counterfactual distribution is more informative, it is also more difficult to estimate than the uplift. A. Li and Pearl (2019) mention that the similarity between these two notions can lead to confusion, especially since they collapse under the assumption of monotonicity (the absence of negative causal effects). Estimating the counterfactual distribution serves several purposes.

- We can establish a profile of each customer category (*persuadable*, *do-not-disturb*, etc.) based on their characteristics such as age, spending habits, subscription, and more. This process can reveal significant business insights and offer new perspectives for future marketing strategies.
- The number of *persuadable* and *do-not-disturb* customers offer an important way to understand the efficacy of a churn prevention campaign by indicating, respectively, how many customers have been convinced to stay thanks to the campaign, and how many customers churned because of the campaign. If we consider only the campaign uplift, we have access to only the difference of these two numbers, hence we cannot estimate separately the positive and negative impact of the campaign.
- More generally, counterfactual probabilities are used in various domains such as algorithmic fairness (Plečko and Bareinboim, 2022, 2023), healthcare, or the legal domain (Balke and Pearl, 1994).

1.4 Motivation and aims

This doctoral thesis takes place in the context of a collaboration between the Machine Learning Group (MLG) from the Université Libre de Bruxelles (ULB) and Orange Belgium, a major telecommunication company in Belgium. It is funded by Innoviris, the public organization that supports and funds research and innovation in the Brussels-Capital Region. The objective of this research project is to assess how causal approaches to analytics can help mitigate the problem of customer churn for telecommunication companies, with a focus on the Brussels-Capital Region.

A unique opportunity offered by this research collaboration is the possibility of experimentation in real-world direct marketing campaigns. Orange Belgium conducts direct marketing campaigns at regular intervals in a variety of use cases, such as *up-sell*, *cross-sell*, *migration*, and also *churn prevention*. In this thesis, we focus on churn prevention, but most of the considerations we make are relevant to other use cases. As mentioned in the previous sections, the traditional approach used by practitioners at Orange Belgium consists in targeting customers with the highest propensity to churn. This approach is at odds with the modern literature on custom targeting and causal inference, which instead suggests the use of causal approaches such as uplift modeling.

In this thesis, we seek to evaluate various causally informed methods for understanding and mitigating customer churn, using large amounts of data, and validating these methods with direct marketing campaigns. After the initial review of the literature at the beginning of the research project, it appeared that uplift modeling was theoretically the most appropriate approach to address the issue of customer churn. Yet, initial empirical results did not seem to suggest that uplift modeling brings a significant improvement over the approach previously used. Most of the contributions of this thesis constitute an investigation of this discrepancy from a theoretical and practical perspective. We also propose new ways to estimate counterfactual probabilities, which allow us to characterize the causal nature of customer behavior. Our results on uplift modeling and counterfactual inference apply naturally to a much larger range of domains than churn prediction.

1.5 Thesis contributions

The contributions of this thesis are as follows:

- The publication of the first public churn dataset with anonymized customer data from Orange Belgium, allowing the research community to evaluate new uplift strategies on challenging and realistic data (Section 4.1).
- A benchmark of various uplift models on two churn datasets and two other publicly available datasets (Section 4.2).
- The development of several strategies to integrate reach information into uplift modeling (Section 4.4).
- The comparison of uplift and predictive modeling in a series of real customer retention campaigns (Section 4.3).
- A new formulation of the measure of profit for uplift models, focusing on individual cost sensitivity (Section 5.2.2), and an empirical estimator of this measure (Section 5.2.5).

- A proof that the uplift curve (an evaluation curve often used in the uplift literature) is an estimator of our proposed measure of profit, highlighting the strict conditions necessary for the validity of the uplift curve (Section 5.2.4).
- A demonstration through theoretical analysis and simulations of the conditions under which the predictive approach outperforms uplift modeling (Section 5.3).
- A set of original bounds and point estimators on the probability of counterfactuals, derived from the scores estimated by an uplift model (Sections 6.2 and 6.3).
- Point estimators of the probability of counterfactuals based on bivariate distributions fitted using uplift scores (Section 6.4), showing a large improvement over the state of the art.
- An evaluation of the proposed counterfactual estimators with two different simulations (Section 6.5) and on a real-world dataset (Section 6.6).
- A characterization of different customer types using our counterfactual estimators and other customer descriptive features, giving new insights on the reaction of customers to churn campaigns (Section 6.6.3).

1.6 Activities summary

In this section, we summarize the communications and activities carried out during the research project, in terms of publication of articles (Section 1.6.1), presentations (Section 1.6.2), various other research activities (Section 1.6.3), and publication of code (Section 1.6.4).

1.6.1 Publications

The following publications were written while completing the requirements for the Doctor of Philosophy degree. The works are listed by chapter. Two of the publications of Chapter 6 are expected to be submitted for publication early 2024.

- Chapter 4:
  - Théo Verhelst, Jeevan Shrestha, et al. (2021). "Predicting reach to find persuadable customers: Improving uplift models for churn prevention". In: *Discovery science*. Ed. by Carlos Soares and Luis Torgo. Cham: Springer International Publishing, pp. 44–54. ISBN: 978-3-030-88942-5
  - Théo Verhelst, Denis Mercier, et al. (2023a). "A churn prediction dataset from the telecom sector: a new benchmark for uplift modeling". In: *ECML PKDD 2023 Workshops - Workshop on Uplift Modeling and Causal Machine Learning for Operational Decision Making*
- Chapter 5:
  - Théo Verhelst, Wouter Verbeke, et al. (2023). "Uplift vs. Predictive Modeling: a Theoretical Analysis". In: *Submitted to Journal of Machine Learning Research*
- Chapter 6:

  - Théo Verhelst, Denis Mercier, et al. (Mar. 2023b). "Partial counterfactual identification and uplift modeling: theoretical results and real-world assessment". en. In: *Machine Learning*. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-023-06317-w. URL: https://link.springer.com/10.1007/s10994-023-06317-w (visited on 05/03/2023)
  - Théo Verhelst and Gianluca Bontempi (2024). "Identifying counterfactual probabilities using bivariate distributions and uplift modeling". In: *to be submitted*
  - Théo Verhelst, Mercier Denis, et al. (2024). "Customer segmentation from counterfactual probabilities: new insights for the telecom sector". In: *to be submitted*

Moreover, the following publications provided an entry point in the research field of predictive analytics, churn prediction and causal inference:

- Théo Verhelst, Olivier Caelen, et al. (2020). "Understanding Telecom Customer Churn with Machine Learning: From Prediction to Causal Inference". In: *Artificial Intelligence and Machine Learning*. Ed. by Bart Bogaerts et al. ISSN: 16130073. Springer International Publishing, pp. 182–200. ISBN: 978-3-030-65154-1
- Bertrand Lebichot et al. (2021). "Transfer Learning Strategies for Credit Card Fraud Detection". In: *IEEE Access* 9, pp. 114754–114766. DOI: 10.1109/ACCESS.2021.3104472

The first one, based upon the candidate's master thesis, was presented at the Benelearn/BNAIC 2019 conference and published in the conference's post-proceedings.

1.6.2 Presentations

The content of this thesis has been presented at the following international conferences:

- 31st Benelux Conference on Artificial Intelligence (BNAIC 2019) and the 28th Belgian Dutch Conference on Machine Learning (Benelearn 2019), 6th to 8th of November 2019, Brussels, Belgium.
- 24th International Conference on Discovery Science (DS 2021), 11th to 13th of October 2021, online (planned to be in Halifax, Canada).
- Fundamental Challenges in Causality, 9th to 12th of May 2023, Grenoble, France (poster presentation).
- ECML PKDD 2023 - Workshop on Uplift Modeling and Causal Machine Learning for Operational Decision Making, 18th to 22nd of September 2023, Turin, Italy.

Moreover, advancements of the work presented in this thesis was periodically presented during bi-monthly meetings with the data science teams of Orange Belgium and Orange Spain.

1.6.3 Research activities

Beyond active participation in the aforementioned conferences, the candidate also attended the following activities:

- Hackaton "CodeVsCovid19", 27th to 30th of March 2020, online.
- Course "Causal Diagrams: Draw Your Assumptions Before Your Conclusions", July 2020, online, at edx.org.
- Course "Academic Writing: The Research Article" by G. Lucy, September to December 2020, Brussels, Belgium.
- Conference "Causal Data Science Meeting", 11th and 12th of November 2020, online.
- Summer School "DeepLearn Summer 2021" from the 26th to the 30th of July 2021, Las Palmas de Gran Canaria, Spain.
- Conference "Microsoft Research Summit", 19th to 21st of October 2021, online.
- Conference "Causal Data Science Meeting", 15th and 16th of November 2021, online
- FARI Brussels Conference, 5th and 6th of July 2022, Brussels, Belgium
- 39th International Conference on Machine Learning (ICML), 17th to 23rd of July 2022, Baltimore, USA.

We had the opportunity to invite various researchers to give seminars at the Maching Learning Group in the context of the research project supporting this doctoral thesis:

- Vincent Lemaire from Orange Lab (France) presented "FEARS: a FEature And Representation Selection approach for time series classification" on the 31st of January 2020.
- Lê Hoang Nguyen from École Polytechnique Fédérale of Lausanne (Switzerland) presented "The security of collaborative learning" on the 13th of October 2022.
- Wouter Verbeke from KU Leuven (Belgium) presented "An introduction to causal machine learning for operational decision making" on the 30th of March 2023.

1.6.4 Code availability

To ensure the reproducibility of our results, the code of the contributions of this thesis are published on the online platform GitHub:

- The code for the paper "Partial counterfactual identification and uplift modeling" is available at https://github.com/TheoVerhelst/Counterfactual-uplift-bounds
- The code for the paper "Uplift vs. Predictive Modeling: a Theoretical Analysis" is available at https://github.com/TheoVerhelst/Uplift-Predictive-Paper
- The code for the paper "A churn prediction dataset from the telecom sector: a new benchmark for uplift modeling" is available at https://github.com/TheoVerhelst/Churn-Uplift-Dataset-Paper

2

Background

This thesis addresses questions at the intersection of machine learning and causal inference. Although the history of these fields dates back to the twentieth century, they are nowadays increasingly active fields of research. In this section, we lay the foundations of the key concepts used in our work, starting with probability theory (Section 2.1) Pearl's theory of causality (Section 2.2), to finally introduce machine learning (Section 2.3).

2.1 Probability theory

Uncertainty is prevalent in the physical world. In quantum mechanics, for example, uncertainty is a fundamental component of the theory. Measuring the spin of an electron will either indicate *up* or *down*, and, in some situations, the outcome of the measurement cannot be predicted with certainty, even with the best measure instruments[1]. In the rest of science, uncertainty is due to limitations in our ability to model reality and make inferences from it. Probability theory is one of the many mathematical models of partial information and uncertainty (Parsons and Hunter, 1998). It provides a representation of the partial knowledge of a system, a set of inference rules to update this representation in the presence of data, and a way to generate data as if it were coming from a system compatible with the current state of our knowledge.

In this section, we define the mathematical notions related to probability theory that are used throughout this thesis. We assume the reader to be familiar with the concept of probability measure, discrete and absolutely continuous random variables, probability mass function, probability density function (pdf), joint and marginal distributions, and conditional independence. See Appendix A for a detailed introduction and a definition of these concepts.

2.1.1 Distribution summaries

It is often desirable to summarize a probability distribution into a compact numerical representation. This is useful to obtain a more concise representation of the data or

[1]There is a longstanding debate on whether this uncertainty is a fundamental aspect of nature, or if it is due to an incomplete formulation of the theory. Physicists showed that any theory in which some hidden variables predetermine the outcome of a quantum experiment must be nonlocal (Aspect, Grangier, and Roger, 1982; Bell, 1964), that is, involving a *spooky action at a distance* (Einstein et al., 1969).

when reporting the results of an experiment. This is also useful for more technical tasks, such as fitting the parameters of a distribution using the method of moments.

Expected value

The *expected value*, also called *mean*, of a distribution is the central tendency of that distribution or, loosely speaking, the average outcome that we can expect when collecting observations from the random variable.

Definition 2.1 (Expected value)**.** The *expected value* of a discrete random variable $\boldsymbol{x}$ is defined as

$$\mathbb{E}[\boldsymbol{x}] = \sum_{x \in \mathcal{X}} x P(\boldsymbol{x} = x) \tag{2.1}$$

and the expected value of an absolutely continuous random variable $\boldsymbol{x}$ with probability density $f_{\boldsymbol{x}}$ is defined as

$$\mathbb{E}[\boldsymbol{x}] = \int_{\mathcal{X}} x f_{\boldsymbol{x}}(x) \, \mathrm{d}x. \tag{2.2}$$

When the random variable to be summed or integrated over is not easily identified in the expression between brackets, we indicate it in subscript as $E_{\boldsymbol{x}}[\cdot]$. The expected value may possibly be infinite (i.e., either ∞ or $-\infty$), or be undefined.

As an example, consider the game of roulette in which the ball falls either into a red slot with probability $18/37$, into a black slot with probability $18/37$, or into slot 0, which always results in a loss, with probability $1/37$. Assume that a player bets \$1 on the red color. Then they will win \$2 with probability $18/37$, and loose their bet with probability $19/37$. The expected value of the gain, noted $\boldsymbol{x}$, is therefore

$$\mathbb{E}[\boldsymbol{x}] = \frac{18}{37}\,\$2 + \frac{19}{37}\,\$0 \approx \$0.973.$$

This indicates that betting on color earns on average 97.3% of the money put on the table.

An important property of the expected value operator $\mathbb{E}[\cdot]$ is its *linearity*: for any random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, and for any two real numbers a and b, we have

$$\mathbb{E}[a\boldsymbol{x} + b\boldsymbol{y}] = a\mathbb{E}[\boldsymbol{x}] + b\mathbb{E}[\boldsymbol{y}]. \tag{2.3}$$

This follows directly from its definition in terms of a sum (for discrete random variables) or an integral (for absolutely continuous random variables), which are themselves linear operators. We will often use this property in our mathematical developments.

2.1.2 Conditional expected value

The expected value operator can be generalized to conditional distributions. This represent the mean of a random variable given that we know the value of another random variable.

Definition 2.2. The *conditional expected value* of a discrete random variable $\boldsymbol{x}$ given the realization of another random variable $\boldsymbol{y} = y$ is defined as

$$\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y} = y] = \sum_{x \in \mathcal{X}} x P(\boldsymbol{x} = x \mid \boldsymbol{y} = y) \tag{2.4}$$

and the conditional expected value of an absolutely continuous random variable $\boldsymbol{x}$ with a conditional probability density $f_{\boldsymbol{x}|\boldsymbol{y}=y}$ is defined as

$$\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y} = y] = \int_{\mathcal{X}} x f_{\boldsymbol{x}|\boldsymbol{y}=y}(x)\,\mathrm{d}x. \tag{2.5}$$

It follows from this definition and the definition of conditional probability density function (Definition A.8) that the conditional expected value of an absolutely continuous random variable can also be expressed as

$$\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y} = y] = \int_{\mathcal{X}} x \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x, y)}{f_{\boldsymbol{y}}(y)}\,\mathrm{d}x = \frac{1}{f_{\boldsymbol{y}}(y)} \int_{\mathcal{X}} x f_{\boldsymbol{x},\boldsymbol{y}}(x, y)\,\mathrm{d}x. \tag{2.6}$$

Variance

The variance is another important quantity, which describes the tendency of the realizations to be close to the expected value. For example, the distribution of the throws of a high-level darts player will be much tighter than that of a beginner player, which is formally expressed as having a lower variance.

Definition 2.3 (Variance)**.** The *variance* of a random variable $\boldsymbol{x}$ is defined as

$$\mathrm{Var}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^2]. \tag{2.7}$$

This definition supposes that the random variable $\boldsymbol{x}$ has an expected value $\mathbb{E}[\boldsymbol{x}]$, and that the square of the difference between $\boldsymbol{x}$ and $\mathbb{E}[\boldsymbol{x}]$ is also a random variable that has an expected value. Some distributions do not respect these conditions, such as the Student t-distribution, a generalization of the standard normal distribution. Depending on the value of its parameter, its variance can be infinite, or its mean and variance can be undefined.

Higher moments

The notion of expected value and variance can be generalized to higher orders, leading to the notion of distribution *moments*. Moments come in two varieties: central moments and raw moments (also called non-central moments). The expected value is the first non-central moment, while the variance is the second central moment. Other moments, such as skewness and kurtosis, are sometimes used in the literature to describe distributions beyond their mean and variance. In this thesis, however, higher-order moments will only be used in the context of an estimation technique called the *method of moments*, hence we will not discuss the specific meaning of skewness or kurtosis.

Definition 2.4 (Moments)**.** The n-th *central moment* of a random variable $\boldsymbol{x}$ is defined as

$$M_n(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^n] = \begin{cases} \sum_{x \in \mathcal{X}} (x - \mathbb{E}[\boldsymbol{x}])^n P(\boldsymbol{x} = x) & \text{if } \boldsymbol{x} \text{ is discrete,} \\ \int_{\mathcal{X}} (x - \mathbb{E}[\boldsymbol{x}])^n f_{\boldsymbol{x}}(x)\,\mathrm{d}x & \text{if } \boldsymbol{x} \text{ is absolutely continuous.} \end{cases} \tag{2.8}$$

The n-th *raw moment* of a random variable $\boldsymbol{x}$ is defined as

$$R_n(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{x}^n] = \begin{cases} \sum_{x \in \mathcal{X}} x^n P(\boldsymbol{x} = x) & \text{if } \boldsymbol{x} \text{ is discrete,} \\ \int_{\mathcal{X}} x^n f_{\boldsymbol{x}}(x)\,\mathrm{d}x & \text{if } \boldsymbol{x} \text{ is absolutely continuous.} \end{cases} \tag{2.9}$$

One can see that by setting $n = 1$, the raw moment corresponds to the expected value in Definition 2.1, while with $n = 2$, the central moment corresponds to the variance in Definition 2.3. While higher-order central moments are easier to interpret, the simpler analytical form of raw moments lend themselves more easily to computations and mathematical derivations.

2.1.3 Information theory

Information theory is a discipline used at the intersection of mathematics, computer science, and engineering. It was first developed by Shannon (1948). In this work, we use two notions of information theory, namely entropy and mutual information. This section is inspired by the excellent textbook by T. M. Cover and Thomas (1991).

Entropy

Entropy, in information theory[2], is a measure of the quantity of uncertainty embedded in a probability distribution. As an example, consider the random experiment of a coin toss. This experiment has two outcomes, heads and tails, and when the coin is fair (i.e., each outcome is equally likely), the entropy is maximum. If the coin is biased towards one side, then the entropy is lower, because we expect the heavier side to come up with a higher probability, hence the outcome is less uncertain.

Definition 2.5 (Entropy)**.** The *entropy* $H(\boldsymbol{x})$ of a discrete random variable $\boldsymbol{x}$ is defined as

$$H(\boldsymbol{x}) = -\sum_{x \in \mathcal{X}} P(x) \log P(x). \tag{2.10}$$

In this definition, the base of the logarithm can be chosen arbitrarily, but a base of two is commonly used. In the coin toss example, if the coin is fair, assuming a logarithm in base two, the entropy is

$$H(\boldsymbol{x}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1.$$

If the coin has a probability 75% of landing on one side, the entropy becomes

$$H(\boldsymbol{x}) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \approx 0.8113.$$

Note that we do not provide a definition for the entropy of an absolutely continuous random variable. It would be tempting to replace the sum in Eq. (2.10) by an integral, much like in Definitions 2.1 and 2.4. However, the resulting quantity, called differential entropy, does not possess some of the mathematical properties of entropy. Alternative definitions exist, but we will not discuss them in this thesis.

Joint and conditional entropy

We can naturally extend the notion of entropy to the bivariate case:

[2]The notion of entropy also exists in statistical mechanics, which is a closely related but nonetheless distinct concept.

Definition 2.6 (Joint entropy)**.** The *joint entropy* $H(\boldsymbol{x}, \boldsymbol{y})$ of two discrete random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$H(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log P(x, y). \tag{2.11}$$

Also, we can extend the definition of univariate entropy to the case where the distribution is conditioned on another random variable. This represents the uncertainty remaining in the first variable, given that we know the value of the second.

Definition 2.7 (Conditional entropy)**.** The *conditional entropy* $H(\boldsymbol{y} \mid \boldsymbol{x})$ of a discrete random variable $\boldsymbol{y}$ given a random variable $\boldsymbol{x}$ is defined as

$$H(\boldsymbol{y} \mid \boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{x}} \left[\sum_{y \in \mathcal{Y}} P(y \mid \boldsymbol{x}) \log P(y \mid \boldsymbol{x}) \right]. \tag{2.12}$$

When $\boldsymbol{x}$ is also discrete, this reduces to

$$H(\boldsymbol{y} \mid \boldsymbol{x}) = -\sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} P(y \mid x) \log P(y \mid x) \tag{2.13}$$

$$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log P(y \mid x). \tag{2.14}$$

When $\boldsymbol{x}$ is absolutely continuous with probability density $f_{\boldsymbol{x}}$, this reduces to

$$H(\boldsymbol{y} \mid \boldsymbol{x}) = \int_{\mathcal{X}} f_{\boldsymbol{x}}(x) H(\boldsymbol{y} \mid \boldsymbol{x} = x) \, \mathrm{d}x \tag{2.15}$$

$$= -\int_{\mathcal{X}} f_{\boldsymbol{x}}(x) \sum_{y \in y} P(y \mid x) \log P(y \mid x) \, \mathrm{d}x. \tag{2.16}$$

Relative entropy

The relative entropy, also called the Kullback-Leibler (KL) divergence, measures the distance between two probability distributions. Here, by probability distributions, we mean probability mass functions for discrete variables and probability density functions for absolutely continuous random variables. It is close to zero when the two random variables have very similar distributions and can be arbitrarily large otherwise. This notion is used, for example, to measure the goodness-of-fit of a candidate distribution with respect to a reference distribution.

Definition 2.8. The *relative entropy* or *Kullback-Leibler divergence* between two probability mass functions P and Q with the same domain $\mathcal{X}$ is

$$D(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \tag{2.17}$$

The relative entropy between two probability density functions f and g with the same domain $\mathcal{X}$ is

$$D(f \parallel g) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \, \mathrm{d}x. \tag{2.18}$$

We use the convention that $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{a} = 0$ and $a \log \frac{a}{0} = +\infty$.

Mutual information

Mutual information represents the quantity of information that one variable possesses about another. We have seen in Appendix A.4.2 that observing one variable can change the distribution of another, leading to the conditional distribution. Mutual information measures the extent to which this conditional distribution differs from the marginal distribution. It has important uses in machine learning, for example in the task of feature selection. Feature selection consists in determining which variables (i.e., features) should be considered or discarded when building a model to predict the value of a target variable. Variables with a low mutual information with the target variable can usually be discarded.

Definition 2.9 (Mutual information)**.** The *mutual information* $I(\boldsymbol{x}, \boldsymbol{y})$ between two discrete random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$I(\boldsymbol{x}, \boldsymbol{y}) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}. \tag{2.19}$$

The mutual information between two absolutely continuous random variables with density $f_{\boldsymbol{x},\boldsymbol{y}}$ is defined as

$$I(\boldsymbol{x}, \boldsymbol{y}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{\boldsymbol{x},\boldsymbol{y}}(x, y) \log \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x, y)}{f_{\boldsymbol{x}}(x) f_{\boldsymbol{y}}(y)} \, \mathrm{d}x \, \mathrm{d}y. \tag{2.20}$$

Note that there exists a unique definition that generalizes these two cases and every other case as well: the mutual information between discrete and absolutely continuous variables, and between random variable that are neither discrete nor absolutely continuous (T. M. Cover and Thomas, 1991, p. 252). However, Definition 2.9 is sufficient for our purpose.

In the case of discrete variables, we can show the following identities:

$$\begin{aligned} I(\boldsymbol{x}, \boldsymbol{y}) &= H(\boldsymbol{y}) - H(\boldsymbol{y} \mid \boldsymbol{x}) && (2.21)\\ &= H(\boldsymbol{x}) - H(\boldsymbol{x} \mid \boldsymbol{y}) && (2.22)\\ &= H(\boldsymbol{x}) + H(\boldsymbol{y}) - H(\boldsymbol{x}, \boldsymbol{y}). && (2.23) \end{aligned}$$

These identities provide some intuition on the mutual information: it can be seen as the reduction in uncertainty in one variable due to the observation of the other. Fig. 2.1 provides a schematic representation of the relationship between entropy, conditional entropy, joint entropy and mutual information.

2.1.4 Convergence of random variables

In various applications of statistics and machine learning, we are interested in the behavior of an estimator as the number of data samples increases. This can be formalized with the notion of convergence of a sequence of random variables. For example, let $\boldsymbol{x}_n$ denote a measure of the performance of a machine learning model trained with n data samples. This quantity is a random variable, since the training set is assumed to be random as well. We want to know whether the sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ converges to some fixed value or distribution as n gets very large. There are three main definitions of convergence of a sequence of random variables: convergence in *distribution*, convergence in *probability*, and *almost sure* convergence. In this thesis, we only use the notion of convergence in probability; therefore, we will define only this one.

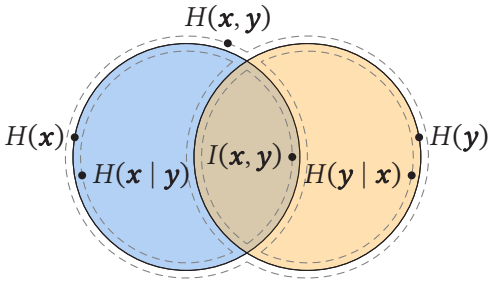


Figure 2.1 Schematic representation of the relationship between entropy and other quantities as a Venn diagram, where $H(\boldsymbol{y})$ and $H(\boldsymbol{x})$ symbolize sets and Eqs. (2.21) to (2.23) represent set relationships of union and difference. Diagram from (Verhelst, 2018).

Definition 2.10. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be an infinite sequence of absolutely continuous random variables with domain $\mathbb{R}$. This sequence *converges in probability* to a random variable $\boldsymbol{x}$, noted

$$\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x} \quad \text{in probability,} \tag{2.24}$$

if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|\boldsymbol{x}_n - \boldsymbol{x}| \geq \varepsilon\right) = 0. \tag{2.25}$$

Intuitively, this definition requires that the probability that $\boldsymbol{x}_n$ and $\boldsymbol{x}$ differ by a significant amount must decrease down to zero as $\boldsymbol{n}$ increases.

2.1.5 Families of probability distributions

Most random distributions in practice are defined in terms of one or more parameters. For example, the Bernoulli distribution, representing a binary experiment such as tossing a coin, is parameterized by a number p between 0 and 1, indicating the probability of, say, obtaining a tail. When the parameters are not fixed, we call this a *family of distributions*. In this section, we describe three families of distributions: the Bernoulli distributions, the beta distributions, and the Dirichlet distributions. We calculate the value of some of the notions defined in the previous sections. This section also serves as a reference point when these families of distributions are used in this thesis. To simplify the language, we shall thereafter refer to families of distributions as distributions.

Bernoulli distribution

The simplest distribution is the *Bernoulli* distribution, whose domain is the binary set $\{0, 1\}$. The parameter p indicates the probability of the realization 1. When a random variable $\boldsymbol{x}$ is distributed according to a Bernoulli distribution, we note

$$\boldsymbol{x} \sim \text{Bern}(p), \quad P(\boldsymbol{x} = 1) = p.$$

Now, let us compute the different quantities defined in the previous sections. The expected value of a Bernoulli-distributed random variable is

$$\mathbb{E}[\boldsymbol{x}] = 1p + 0(1 - p) = p. \tag{2.26}$$

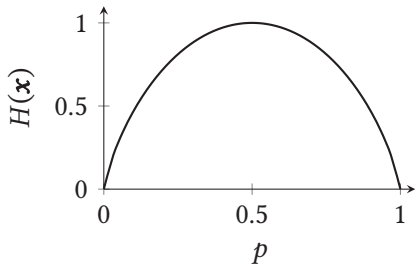


Figure 2.2 Entropy of a Bernoulli-distributed random variable as a function of p.

Its variance is

$$\mathrm{Var}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - p)^2] = \mathbb{E}[\boldsymbol{x}^2] + p^2 - 2\mathbb{E}[\boldsymbol{x}]p = p(1-p) \tag{2.27}$$

where we used the linearity of the expected value. The higher-order raw moments are all equal to p, since $1^n = 1$ for any n and $0^n = 0$ for any $n > 0$. The higher-order central moments are

$$\mathbb{E}[(\boldsymbol{x} - p)^n] = p(1-p)^n + (1-p)(-p)^n. \tag{2.28}$$

Finally, the entropy is

$$H(\boldsymbol{x}) = -p \log p - (1-p)\log(1-p). \tag{2.29}$$

The entropy as a function of p is depicted in Fig. 2.2.

Beta distribution

A beta-distributed random variable is absolutely continuous and takes its values in $[0, 1]$. It is parameterized by two positive values, usually noted a and b, which, loosely speaking, dictate how much of the probability density is concentrated close to respectively 0 and 1. The fact that a random variable $\boldsymbol{x}$ is beta-distributed is noted $\boldsymbol{x} \sim \mathrm{Beta}(a, b)$. Its probability density function is

$$f_{\boldsymbol{x}}(x) = \frac{1}{\mathrm{B}(a,b)} x^{a-1}(1-x)^{b-1} \tag{2.30}$$

where the term $\mathrm{B}(a, b)$ ensures that the integral of the pdf over $[0, 1]$ is equal to one. This term is the *beta function*, defined as

$$\mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{2.31}$$

where the term $\Gamma(\cdot)$ is the *gamma function*, a generalization of the factorial operator to the complex numbers.[3] Fig. 2.3 shows the pdf of two different beta distributions.

The derivation of the expected values, variance, moments and entropy of the beta distribution is slightly more complex than in the case of the Bernoulli distribution. For the sake of brevity, we will limit ourselves to give the formulas without showing their derivations. The expected value is

$$\mathbb{E}[\boldsymbol{x}] = \frac{a}{a+b}, \tag{2.32}$$

[3] See the Wikipedia article on the gamma function for more details.

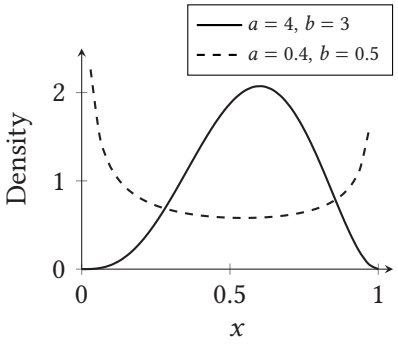


Figure 2.3 Probability density function of two beta-distributed random variables.

the variance is

$$\mathrm{Var}(\boldsymbol{x}) = \frac{ab}{(a+b)^2(a+b+1)}, \tag{2.33}$$

and the entropy is

$$H(\boldsymbol{x}) = \log(\mathrm{B}(a,b)) - (a-1)\psi(a) - (b-1)\psi(b) + (a+b-2)\psi(a+b) \tag{2.34}$$

where $\psi(a)$ is the *digamma* function, defined as

$$\psi(a) = \frac{\mathrm{d}\log\Gamma(a)}{\mathrm{d}a}. \tag{2.35}$$

We will not use the central moments of the beta distribution in this thesis, but the raw moments are frequently used in our simulation experiments. In fact, we give a generalization of the raw moments: for any $r, s > 0$, we can show that

$$\mathbb{E}[\boldsymbol{x}^r(1-\boldsymbol{x})^s] = \frac{\mathrm{B}(a+r, b+s)}{\mathrm{B}(a,b)}. \tag{2.36}$$

A key property of the gamma function is that, for any real x and any integer n, we have $\Gamma(x+n) = x^{\overline{n}}\Gamma(x)$ where $x^{\overline{n}} = x(x+1)\dots(x+n-1)$ is called the *rising factorial*. This allows for an alternate expression for the moments of the beta distribution, given that r and s are integers:

$$\mathbb{E}[\boldsymbol{x}^r(1-\boldsymbol{x})^s] = \frac{\mathrm{B}(a+r,b+s)}{\mathrm{B}(a,b)} = \frac{\Gamma(a+r)\Gamma(b+s)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+r+s)} = \frac{a^{\overline{r}}b^{\overline{s}}}{(a+b)^{\overline{r+s}}}. \tag{2.37}$$

This proves particularly useful in numerical computations since the rising factorial is much faster to compute and less affected by numerical instability than the gamma function. Finally, when $s = 0$, we have the raw moments $\mathbb{E}[\boldsymbol{x}^r] = a^{\overline{r}}/(a+b)^{\overline{r}}$, which itself reduces to $\mathbb{E}[\boldsymbol{x}] = a/(a+b)$ as in Eq. (2.32) when $r = 1$.

Dirichlet distribution

The Dirichlet distribution is a generalization of the beta distribution to multiple dimensions. We note the number of dimensions as an integer d, and the random vector $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_d]$. The domain of the Dirichlet distribution is the *unit simplex*, that is, the set of d-dimensional vectors x with positive values summing up to one:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid x_i \geq 0 \text{ for all } i \in \{1,\dots,d\} \text{ and } \sum_{i=1}^{d} x_i = 1 \right\}. \tag{2.38}$$

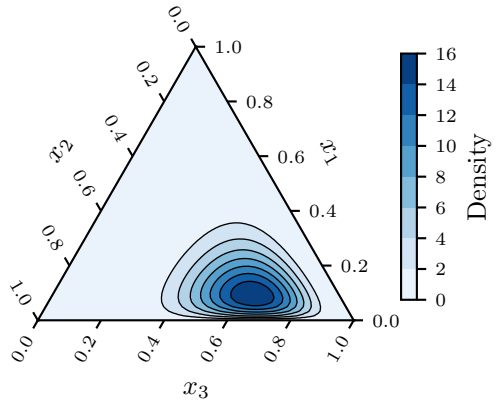


Figure 2.4 Probability density function of a 3-dimensional Dirichlet distribution with parameter vector $m = [2, 4, 8]$. The triangular grid conveniently represents the three-dimensional simplex. Plot drawn using the Python package `mpltern`.

The Dirichlet distribution is parameterized by a positive vector $m = [m_1, \dots, m_d]$, and we define that a random vector $\boldsymbol{x}$ follows a Dirichlet distribution, noted $\boldsymbol{x} \sim \mathrm{Dir}(m)$, when its probability density function is

$$f_{\boldsymbol{x}}(x) = \frac{1}{\mathrm{B}(m)} \prod_{i=1}^{d} x_i^{m_i - 1} \tag{2.39}$$

where the term $\mathrm{B}(m)$ is a generalization of the beta function to vector inputs:

$$\mathrm{B}(m) = \frac{\prod_{i=1}^{d} \Gamma(m_i)}{\Gamma\left(\sum_{i=1}^{d} m_i\right)}. \tag{2.40}$$

Fig. 2.4 shows the probability density function of a Dirichlet distribution with three dimensions on a triangular grid. One can see that the Dirichlet distribution is indeed a generalization of the beta distribution: if $\boldsymbol{y} \sim \mathrm{Beta}(a, b)$, then the vector $[\boldsymbol{y}, 1 - \boldsymbol{y}]$ follows a Dirichlet distribution with parameter vector $[a, b]$.

The expected value of a component $\boldsymbol{x}_j$ of the random vector $\boldsymbol{x}$ is

$$\mathbb{E}[\boldsymbol{x}_j] = \frac{m_j}{M} \quad \text{with } M = \sum_{i=1}^{d} m_i. \tag{2.41}$$

The variance is

$$\mathrm{Var}(\boldsymbol{x}_j) = \frac{m_j(M - m_j)}{M^2(M + 1)}. \tag{2.42}$$

More generally, the raw moments of the Dirichlet distribution are, for a positive real vector $a \in \mathbb{R}^d_{\geq 0}$,

$$\mathbb{E}\left[\prod_{i=1}^{d} \boldsymbol{x}_i^{a_i}\right] = \frac{\mathrm{B}(m + a)}{\mathrm{B}(m)}. \tag{2.43}$$

Similarly to the case of the moments of the beta distribution in Eq. (2.37), when the components of a are integers, we can use the properties of the gamma function to

express the raw moments in terms of rising factorials:

$$\mathbb{E}\left[\prod_{i=1}^{d} \boldsymbol{x}_i^{a_i}\right] = \frac{\prod_{i=1} m_i^{\overline{a_i}}}{M^{\overline{A}}} \quad \text{where } A = \sum_{i=1}^{d} a_i. \tag{2.44}$$

We do not provide an expression for the central moments of the Dirichlet distribution, since they are not used in this thesis. Lastly, the entropy of the random vector $\boldsymbol{x}$ is

$$H(\boldsymbol{x}) = \log \mathrm{B}(m) + (A-d)\psi(A) - \sum_{i=1}^{d}(a_i - 1)\psi(a_i). \tag{2.45}$$

2.2 Causality theory

Most scientific inquiries are inherently causal in nature, aiming to uncover the cause-and-effect relationships that underlie various phenomena. Whether it is estimating the efficacy of a new drug, assessing the impact of an educational policy, or unraveling the formation mechanisms of supermassive black holes at the center of galaxies, understanding causality is at the heart of scientific investigations.

Traditional statistics, based on the theory of probability presented in Section 2.1, has been a valuable tool in many scientific fields to draw associations and identify patterns in data. However, it has inherent limitations when it comes to establishing causal relationships. One of the main challenges with traditional statistical methods is that they primarily focus on associations or correlations between variables. These methods are excellent at revealing patterns, but they can't discern whether one variable is causing changes in another or if those changes are coincidental or influenced by other hidden factors.

Judea Pearl's theory of causality (Pearl, 2009), emerged as a response to these limitations. Pearl recognized the need for a more rigorous framework to address causal questions. His work was motivated, in part, by the complexities involved in social sciences, and particularly in econometrics, where randomizing variables for controlled experiments, as often done in natural sciences, is not always feasible or ethical.

Pearl's theory of causality introduced a paradigm shift by providing a structured framework for reasoning about causation. By representing causal relationships using directed acyclic graphs (DAGs), Pearl's approach enabled researchers to make causal inferences, even in observational data, where controlled experiments are impractical. He also emphasized the importance of counterfactuals, which involve comparing what happened with what might have happened in the absence of a specific cause, to understand causation. It is worth noting that this is not the only mathematical formulation of causality. The most widely used alternative formulation is Rubin's causal model (Donald B Rubin, 2005, 1974). It is based on the notions of *intervention* and *potential outcome*. There is a longstanding debate between proponents of both approaches (Weinberger, 2021), and since the formal equivalence of the two frameworks has been proven (Pearl, 2009, Sec. 7.4.4), the essence of the debate relates to the way assumptions are encoded, and the focus on either individual treatment effect in Rubin's framework or structural knowledge in Pearl's framework.

In the next section, we present an overview of Pearl's causal framework through the *causal hierarchy*, then we develop its mathematical formalism in the following four sections. The mathematical definitions are based upon Pearl (2009), Bareinboim et al.

(2020) and Peters, Janzing, and Schölkopf (2017). For a less technical introduction to causality, we refer the interested reader to the classic *The book of why: the new science of cause and effect* by Pearl and Mackenzie (2018), or *Why: A guide to finding and using causes* by Kleinberg (2015).

2.2.1 Pearl's causal hierarchy

In Pearl's framework, formal knowledge and its associated mathematical representation are organized in three layers of increasing abstraction (Bareinboim et al., 2020), which highlight the roles of *observation*, *intervention*, and *imagination*. The data at one layer is almost always not sufficient to estimate probabilities relating to higher layers.

1. The *observational* layer represents all knowledge derived from observing the world: information such as correlation and association between variables, conditional probabilities learned with statistical or machine-learning models, or even more complex knowledge representations such as graphical Bayesian models (Pearl, 1988).

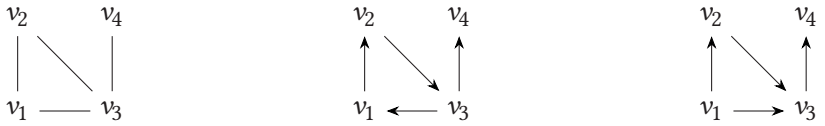
2. The *interventional* layer represents all knowledge related to the consequences of experiments, manipulations, and interventions in a system. Typical examples of interventional knowledge include the results of randomized experiments in pharmaceutical research or A / B testing in online marketing. This layer is dedicated to understanding the cause-and-effect relationships that emerge when specific interventions are implemented and enables the direct assessment of how these interventions influence the system's behavior.

3. The *counterfactual* layer represents knowledge relating to alternate states of the world that could have occurred if different choices or actions had been made. Given that we observe a specific outcome, would the outcome be different had we chosen another course of action? Counterfactuals also formalize notions such as the *probability of causation*.

Although the difference between the observational layer and the two other layers is fairly intuitive[4], the nuance between interventions and counterfactuals might be less clear. Observational knowledge can be gathered through experiments, whereas counterfactual knowledge cannot be deduced from observations or experiments alone.

To illustrate, let us consider the example of a telecom company attempting to prevent customer churn with marketing emails. The average causal effect of a marketing email on customer churn can be estimated with a randomized campaign. A *target group* is created by randomly selecting a subset of customers, and each of these customers receives an email from the company. The remaining customers constitute the *control group*, and they do not receive any email as part of the campaign. The difference in churn rate between the control and target group is called the *average causal effect*, and belongs to the interventional layer. Now, consider customers who have churned and are in the target group. What is the probability that they would have churned, had they not been called? This is a counterfactual probability and, without further assumption,

[4]We will avoid discussing the justification of separating observation from intervention as a fundamental aspect of knowledge representation, since this is the subject of a decades-long debate which is beyond the scope of this work.



(a) An undirected graph. **(b)** A directed graph. **(c)** A directed acyclic graph.

Figure 2.5 Examples of different types of graphs. Notice that (b) contains a cycle between v_1, v_2 and v_3.

this information cannot be computed from the available experimental data, since this involves estimating the probability of events that are, by definition, not observed.

Each layer is distinct from the ones below it and expresses information not available in lower layers. This has been proven by Bareinboim et al. (2020) by inscribing the three layers in a formal language and proving that, in a measure-theoretic sense, the information expressed in a given layer is almost always insufficient to answer queries in a higher layer. This implies, for example, that mere correlations are not sufficient to estimate causal effects, or that causal effects are not sufficient to estimate counterfactuals.

2.2.2 Directed acyclic graphs and independence

The most important feature of Pearl's causal framework is the representation of knowledge with *causal graphs*. A causal graph indicates direct cause-effect relationships between random variables. In this section, we provide the mathematical definition of graphs in general, which are used in various fields of science and mathematics. We then state important results relating probability distributions and graphs of random variables.

Definition 2.11 (Graph)**.** A *graph* G is a pair (V, E) where V is a finite set of vertices, and E is a set of pairs of vertices, called edges. If the pairs in E are ordered, we say that the edges are *oriented*, we note edges as $v_1 \to v_2$, and we say that the graph is *directed*. Otherwise, we say that the graph is *undirected*, and we note edges as $v_1 - v_2$.

Fig. 2.5 shows various examples of graphs. In this definition, vertices can be any kind of mathematical object, such as numbers, functions, or random variables. In causality theory, we are interested in a certain class of graphs, called *directed acyclic graphs*, or DAGs for short. *Acyclic* means that the graph does not contain loops or cycles, that is, following edges cannot lead to an already visited vertex. To define this last property, we first define the notion of *path*.

Definition 2.12 (Path)**.** A *directed path* in a graph $G = (V, E)$ is a sequence of n vertices $v_1, \dots, v_n$ in V such that $v_i \neq v_j$ for all distinct $i, j = 1, \dots, n$, and such that for all $i = 1, \dots, n-1$, there is an edge $v_i \to v_{i+1}$ in E. An *undirected path* in a graph $G = (V, E)$ is a sequence of n vertices $v_1, \dots, v_n$ in V such that $v_i \neq v_j$ for all distinct $i, j = 1, \dots, n$, and such that, for all $i = 1, \dots, n-1$, there is an edge $v_i \to v_{i+1}$ or an edge $v_{i+1} \to v_i$ in E.

Definition 2.13 (Directed acyclic graph)**.** A *directed acyclic graph* (DAG) is a graph $G = (V, E)$ such that, for all $v_1, v_2 \in V$, if there is a directed path from v_1 to v_2, then there is no directed path from v_2 to v_1.

The notion of *d-separation* is the graphical equivalent of conditional independence between random variables. The *d* in d-separation stands for *dependency*, as it establishes a criterion to determine whether two sets of variables are independent when separated by another set of variables. It is fundamental in causal inference, as it systematically determines the graphs compatible with the observed data, and the conditional independence between variables that should be observed in any data compatible with a given graph. The formal definition of d-separation is not intuitive, thus we will provide an example following the definition. First, we need to define some additional graph terminology.

Definition 2.14 (Graph relationships)**.** Let $G = (V, E)$ be a directed acyclic graph. For any vertex $v \in V$, we define the following sets.

- The *parents* of v, noted $\mathrm{PA}_G(v)$, is the set of vertices u such that there exists an edge $u \to v$.
- The *children* of v, noted $\mathrm{CH}_G(v)$, is the set of vertices w such that there exists an edge $v \to w$.
- The *ancestors* of v, noted $\mathrm{AN}_G(v)$; is the set of vertices u such that there exists a directed path from u to v.
- The *descendants* of v, noted $\mathrm{DE}_G(v)$, is the set of vertices w such that there exists a directed path from v to w.

Let u, v, w be three vertices in V. We give special names to the following edge configurations:

- The configuration $u \to v \to w$ is called a *chain*.
- The configuration $u \leftarrow v \to w$ is called a *fork*.
- The configuration $u \to v \leftarrow w$ is called a *collider*.

Note that the set of ancestors of v contains v, using the trivial path that contains only one vertex, v. By the same argument, v is also its own descendant, although it is not its own parent or child.

Definition 2.15 (d-separation)**.** An undirected path p is *blocked* by a set of vertices $Z \subseteq V$ if and only if

(i) p contains a chain $u \to z \to v$ or a fork $u \leftarrow z \to v$ for some $z \in Z$, or

(ii) p contains a collider $u \to z \leftarrow v$ such that neither z nor any of its descendants are in Z.

Vertex sets $X \subseteq V$ and $Y \subseteq V$ are d-separated by Z in graph G if every path from a node in X to a node in Y is blocked by Z. This is noted $X \perp_G Y \mid Z$. When Z is the empty set, we write $X \perp_G Y$.

To illustrate this notion, consider Fig. 2.6. In this fictional example, customers decide to churn based only on the amount to pay on the invoice. This invoice is in turn determined by the data usage (and indirectly by the age of the customer), and also by the location: a customer in a less populated area pays more because establishing the

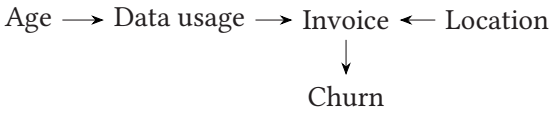


Figure 2.6 Fictional causal graph illustrating the notion of d-separation. Age and invoice are d-separated given data usage, whereas data usage and location are no longer d-separated once we know the invoice amount or the churn variable. Example from (Verhelst, 2018).

connectivity is more expensive for the operator. In this configuration, the age and the invoice are d-separated by the data usage, since knowing the data usage of a client removes any information the age may bring about the invoice amount. This illustrates the first property of a blocked path in Definition 2.15. Furthermore, data usage and location are d-separated given the empty set, since there is a collider (data usage → invoice ← location) between them. This illustrates the second property of a blocked path in Definition 2.15. However, if we know the invoice variable, data usage and location are no longer d-separated, as knowing the invoice amount allows one to *explain away* one variable with the other. If the client has a large invoice amount and lives in a populated area, that must mean that their data usage was probably high. This shows how a variable can open a path in a collider, as mentioned in the second property in Definition 2.15. Note that knowing the churn variable instead of the invoice would have the same effect since if the customer decides to churn, then the invoice amount is probably high, which brings us back to the previous case.

As mentioned before, Pearl's framework represents the structure of causal relationships between random variables with DAGs. In the following, we will therefore assume that the vertices of the graph G are random variables. We note the set of vertices as $\boldsymbol{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_d\}$, and the domain of $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_d)$ is noted $\mathscr{V}$. We assume that $\boldsymbol{V}$ has a joint probability $P(\boldsymbol{v}_1 = v_1, \dots, \boldsymbol{v}_d = v_d)$.

We mentioned earlier that the d-separation between nodes in a graph is the graph-theoretic equivalent of the notion of conditional independence between random variables. We now formalize this equivalence. First, we need to generalize the notion of conditional independence (Definition A.10) to sets of random variables.

Definition 2.16 (Conditional independence)**.** Let $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n), \boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_m)$ and $\boldsymbol{Z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_\ell)$ be three tuples of random variables. We note the domain of $\boldsymbol{X}$ as $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_n$, and, similarly, the domains of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are noted respectively $\mathscr{Y}$ and $\mathscr{Z}$. We say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *independent given* $\boldsymbol{Z}$, which we write $\boldsymbol{X} \perp_P \boldsymbol{Y} \mid \boldsymbol{Z}$, if

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z) \quad \text{whenever } P(Z) > 0$$

for any values $X \in \mathscr{X}, Y \in \mathscr{Y}, Z \in \mathscr{Z}$. When $\boldsymbol{Z}$ is the empty tuple, we write $\boldsymbol{X} \perp_P \boldsymbol{Y}$.

Note that here the notation $\perp$ receives a subscript P to differentiate it with the notion of d-separation. So far, we have not explained how graphs and probability distributions are related. The following definition indicates the property that we expect from a graph to represent a probability distribution. Intuitively, the probability distribution of a random variable should depend only on its parents in the graph.

Definition 2.17 (Markov compatibility, Pearl, 2009, Def. 1.2.2)**.** Let P be a probability measure on $\boldsymbol{V}$ and let $G = (\boldsymbol{V}, E)$. We say that G represents P, that G and P are

compatible, or that P is *Markov relative* to G, if

$$P(v_1, \dots, v_d) = \prod_{i=1}^{d} P(v_i \mid \mathrm{PA}_G(v_i)) \quad \forall (v_1, \dots, v_d) \in \mathscr{V}$$

where $\mathrm{PA}_G(v_i)$ denotes the realization of $\mathrm{PA}_G(\boldsymbol{v}_i)$ in $v_1, \dots, v_d$:

$$\mathrm{PA}_G(v_i) = \left(v_j \mid (\boldsymbol{v}_j \to \boldsymbol{v}_i) \in E\right)_{j=1}^{d}. \tag{2.46}$$

Finally, the following theorem shows the equivalence between d-separation and conditional independence. This theorem is the key element that enables causal inference methods to deduce structural causal knowledge (in the form of DAGs) from statistics computed with observed data (derived from the probability distribution).

Theorem 2.1 (Pearl, 2009, Thm. 1.2.5)**.** *Let $G = (\boldsymbol{V}, E)$, and let $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ be three tuples with elements in $\boldsymbol{V}$. The following two statements are true.*

(i) If $\boldsymbol{X} \perp_G \boldsymbol{Y} \mid \boldsymbol{Z}$, then $\boldsymbol{X} \perp_P \boldsymbol{Y} \mid \boldsymbol{Z}$ for all probability measures P compatible with G.

(ii) If $\boldsymbol{X} \perp_P \boldsymbol{Y} \mid \boldsymbol{Z}$ for all probability measures P compatible with G, then $\boldsymbol{X} \perp_G \boldsymbol{Y} \mid \boldsymbol{Z}$.

The first property can be used to validate a causal graph G, given some data coming from P: d-separation between vertices in G implies that the corresponding random variables will be conditionally independent in any data compatible with G. If we observe some dependency, then we can reject this causal graph. The second property indicates that some conditional independence must be true in all probability measures P to be reflected in a graph G. This is less convenient, because in practice we have access to only one set of data coming from one probability measure P. This is why it is common to make the assumption of *faithfulness*: a probability measure P is faithful to G if any conditional independence in P implies the corresponding d-separation in G. This rules out parallel causal pathways canceling out each other and resulting in the absence of dependency in observed data.

2.2.3 Causal models

In the previous section, we defined graphs and their relationship with probability distributions, suggesting the potential to model real-world processes with these two concepts. We could be tempted to think that associating a probability distribution and a compatible graph (see Definition 2.17) would be all that is needed to infer causation from data and to represent the three layers of the causal hierarchy presented in Section 2.2.1. Such a mathematical construct is named a *Bayesian network*. Furthermore, if the edges of the graph represent genuine causal relationships, then it is called a *causal Bayesian network*. An important achievement of Pearl and his collaborators is to realize that such a model is incomplete because it is not sufficient to compute counterfactuals, the highest layer of the causal hierarchy.

From an epistemological standpoint, this affirmation is not meaningful on its own: after all, counterfactuals could be a made-up construct with no bearing on reality, and the fact that Bayesian networks are unable to express them would be unavoidable. To make an analogy, this would be like adding a new rule to the rules of basketball, specifying that doing a back-flip earns more points, and then claiming that the best basketball players are bad because they cannot do back-flips. This is not reasonable, unless we are

ready to justify in which sense back-flips (respectively, counterfactuals) are an essential aspect of basketball (respectively, causality theory).

The legitimacy of counterfactuals as an important concept to take into account in causality theory is justified by two facts. First, it is ubiquitous in human reasoning. Imagining counterfactual scenarios is pervasive in the way we approach problems and we reason about the consequences of our actions. Second, the concept is a primitive component of Rubin's causal model, indicating that even these two competing views on the formalization of causality agree that counterfactuals should be present in the theory. If Pearl's framework is to succeed in providing the scientific community with a framework applicable even in the fields relying heavily on Rubin's framework, then it must include the notion of counterfactuals.

In this section, we define *causal models*, which are mathematical objects sufficiently expressive to encompass all three layers of the causal hierarchy. The essential aspect that differentiates causal models from Bayesian networks is that random variables are not defined by their conditional probability given their parents, but as the result of a deterministic function of their parents and some unobserved noise. Any Bayesian network can be easily transformed into an equivalent causal model (although this transformation is not unique), and, moreover, using deterministic functions and unobserved noise provides the machinery necessary to compute counterfactual probabilities.

Definition 2.18 (Causal model)**.** A *causal model*, also called *structural equation model* (SEM), or *structural causal model* (SCM) is a tuple $M = (P, \boldsymbol{U}, \boldsymbol{V}, G, F)$ where

(i) $\boldsymbol{U} = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_n)$ is a sequence of random variables called *exogenous variables* or *unobserved variables*, with domains $\mathcal{U}_1, \dots, \mathcal{U}_n$. The domain of $\boldsymbol{U}$ is noted $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_n$.

(ii) P is a probability measure on $\boldsymbol{U}$.

(iii) $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_d)$ is a sequence of random variables called *endogenous variables* or *observed variables*, with domains $\mathcal{V}_1, \dots, \mathcal{V}_d$.

(iv) G is a directed acyclic graph with vertices $\boldsymbol{V} \cup \boldsymbol{U}$ such that there is no directed path from $\boldsymbol{V}$ to $\boldsymbol{U}$.[5]

(v) F is a sequence $(f_1, \dots, f_d)$ where f_i is a function from the domain of $\mathrm{PA}_G(\boldsymbol{v}_i)$ to $\mathcal{V}_i$.[6]

(vi) All $\boldsymbol{v}_i \in \boldsymbol{V}$ are defined as

$$\boldsymbol{v}_i = f_i(\mathrm{PA}_G(\boldsymbol{v}_i)). \tag{2.47}$$

The distinction between the unobserved variables $\boldsymbol{U}$ with a probability distribution P, and the observed variables $\boldsymbol{V}$ as the result of deterministic functions F, suggests a two-phases procedure to generate data from a causal model:

1. First, sample a realization U from the probability distribution P.

[5]The absence of directed path from $\boldsymbol{V}$ to $\boldsymbol{U}$ is a consequence of the fact that we define F in terms of G. We could avoid this constraint by defining G in terms of F, however, such a definition introduces more technical details irrelevant to the intuition of causal models.

[6]The functions f_i must be measurable, in a measure-theoretic sense, to ensure that $f_i(\mathrm{PA}_G(\boldsymbol{v}_i))$ is a random variable.

2. Then, evaluate each function f_i and assign the result to the corresponding $\boldsymbol{v}_i$, starting with the members of $\boldsymbol{V}$ that have parents only in $\boldsymbol{U}$. More precisely, the functions f_i are evaluated in a topological order of G.

This two-step process is what allows causal models to compute counterfactuals, as we will show in Section 3.2. In the rest of this section, we illustrate the expressive power of causal models by explaining their relation to each layer of the causal hierarchy.

We should clarify a slight abuse of notation in Definition 2.18. We consider the input of f_i, that is, $\mathrm{PA}_G(\boldsymbol{v}_i)$, to be a tuple rather than a set. This implies that the domain of f_i is the Cartesian product of the domain of the parents of $\boldsymbol{v}_i$. Furthermore, if $\boldsymbol{v}_i$ has no parent in G, it should be a constant random variable (since it is not affected by exogenous variables in $\boldsymbol{U}$). In this case, the domain of f_i is the empty Cartesian product, which contains only the empty tuple. The value assigned by f_i to the empty tuple is the constant value of $\boldsymbol{v}_i$. This technical aspect is used in the following definitions.

Observational layer

Since P defines a probability distribution on $\boldsymbol{U}$, and since any member of $\boldsymbol{V}$ is ultimately a deterministic function of (a subset of) $\boldsymbol{U}$, we can deduce that a causal model entails a probability distribution on the endogenous variables $\boldsymbol{V}$. Therefore, any probability distribution, which is within the realm of the observational layer, can be modeled with a causal model.

Interventions

Interventions formalize the notion of modification of a system, and allow one to estimate their impact on the remaining variables. In general, any part of a causal model can be changed (removing or adding edges, changing the distribution, etc.), but in practice we are more interested in interventions where one variable is forced to take a specific value. This is represented as *graph surgery*, where the edges going into a given variable are removed and that random variable is set to a specific value.

Definition 2.19 (Intervention)**.** Let $M = (P, \boldsymbol{U}, \boldsymbol{V}, G, F)$ be a causal model. For any random variable $\boldsymbol{x} \in \boldsymbol{V}$ with domain $\mathcal{X}$, and for any $x \in \mathcal{X}$, an *intervention* $\mathrm{do}(\boldsymbol{x} = x)$ defines a new causal model $M_x = (P, \boldsymbol{U}, \boldsymbol{V}, G_{\overline{\boldsymbol{x}}}, F_x)$ where

- $G_{\overline{\boldsymbol{x}}}$ is G but with all edges going into $\boldsymbol{x}$ removed.
- F_x is F but where the function giving the value of $\boldsymbol{x}$ is replaced by the constant function $(\,) \mapsto x$, where $(\,)$ is the empty tuple.[7]

This new causal model M_x induces a new probability distribution over $\boldsymbol{V}$. The probability of a variable $\boldsymbol{y} \in \boldsymbol{V}$ to take its value in B under M_x is noted $P(\boldsymbol{y} \in B \mid \mathrm{do}(\boldsymbol{x} = x))$, or $P(\boldsymbol{y} \in B \mid \mathrm{do}(x))$ when the variable under intervention is clear from the context.

Counterfactuals

The motivating application of counterfactuals is to estimate the impact of a hypothetical intervention, given that the observed course of events is different from this intervention. In the notation developed so far, this would be to estimate the probability

[7] Since $\boldsymbol{x}$ has no parents anymore, the domain of its function is the empty Cartesian product, whose unique member is the empty tuple.

distribution of some random variable $\boldsymbol{y}$ under the intervention $\text{do}(\boldsymbol{x} = x)$ given that we have observed $\boldsymbol{x} = x'$. It appears that the $\text{do}(\cdot)$ notation is insufficient to express such a probability[8]. The notion of potential outcomes provides the necessary machinery to define counterfactual expressions. We define potential outcomes as the value of the random variables $\boldsymbol{V}$ in the causal model, given a realization of the unobserved variables $\boldsymbol{U} = U$. We can compute this realization using the two-step procedure mentioned after Definition 2.18. The following two definitions formally express this procedure.

Definition 2.20 (Potential outcome)**.** Given a model $M = (P, \boldsymbol{U}, \boldsymbol{V}, G, F)$ and a variable $\boldsymbol{y} \in \boldsymbol{V} \cup \boldsymbol{U}$ with domain $\mathscr{Y}$, the *potential outcome function* of $\boldsymbol{y}$, or *potential response function* of $\boldsymbol{y}$, is a function from $\mathscr{U}$ to $\mathscr{Y}$, noted $\boldsymbol{y}_M(U)$, and defined as the value given by M to $\boldsymbol{y}$ when we set $\boldsymbol{U} = U$, for $U \in \mathscr{U}$. It is recursively defined as

$$\boldsymbol{y}_M(U) = \begin{cases} y & \text{if } \boldsymbol{y} \in \boldsymbol{U} \text{ with value } y \text{ in } U, \\ f_{\boldsymbol{y}}(\boldsymbol{z}_M^{(1)}(U), \dots, \boldsymbol{z}_M^{(m)}(U)) & \text{if } \boldsymbol{y} \in \boldsymbol{V} \text{ with } \text{PA}_G(\boldsymbol{y}) = (\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(m)}). \end{cases}$$

The *potential outcome* (or *potential response*) $\boldsymbol{y}_M$ is a random variable defined as $\boldsymbol{y}_M = \boldsymbol{y}_M(\boldsymbol{U})$. When considering a model M_x under intervention $\text{do}(\boldsymbol{x} = x)$, we note $\boldsymbol{y}_{M_x} = \boldsymbol{y}_{\boldsymbol{x}=x}$, or $\boldsymbol{y}_x$ when it is clear from the context that we intervene on $\boldsymbol{x}$. We can see from Definition 2.18 that $\boldsymbol{y}(\boldsymbol{U})$ is the same as $\boldsymbol{y}$.

The following definition is adapted from (Bareinboim et al., 2020, Def. 7), where we generalized their definition to encompass any type of random variable rather than only discrete variables.

Definition 2.21 (Counterfactuals)**.** A *counterfactual* expression (or just *counterfactual* for short) is a logical expression that involves any number of potential outcomes, possibly subject to different interventions. Let $M_1, \dots, M_m$ be a sequence of models derived by interventions from a base model $M = (P, \boldsymbol{U}, \boldsymbol{V}, G, F)$ as in Definition 2.19, and possibly including M. Let $\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(m)}$ be a sequence of random variables in $\boldsymbol{V}$, with domains $\mathscr{Y}_1, \dots, \mathscr{Y}_m$. A counterfactual expression is a logical expression of the form

$$\boldsymbol{y}_{M_1}^{(1)} \in A_1 \wedge \dots \wedge \boldsymbol{y}_{M_m}^{(m)} \in A_m$$

where $A_i \subseteq \mathscr{Y}_i$ for all $i = 1, \dots, m$. Its probability can be computed as

$$P\left(\boldsymbol{y}_{M_1}^{(1)} \in A_1, \dots, \boldsymbol{y}_{M_m}^{(m)} \in A_m\right) = P(\boldsymbol{U} \in \Lambda) \tag{2.48}$$

where Λ is the subset of $\mathscr{U}$ satisfying the counterfactual expression:

$$\Lambda = \left\{U \in \mathscr{U} \mid \boldsymbol{y}_{M_1}^{(1)}(U) \in A_1, \dots, \boldsymbol{y}_{M_m}^{(m)}(U) \in A_m\right\}. \tag{2.49}$$

As an example, let $\boldsymbol{t}$ be the indicator variable of a medical treatment ($\boldsymbol{t} = 1$ for treatment, $\boldsymbol{t} = 0$ for no treatment), $\boldsymbol{y}$ be the outcome of the treatment ($\boldsymbol{y} = 1$ for survival, $\boldsymbol{y} = 0$ for death) and $\boldsymbol{x}$ be the age of the patient (in years). Suppose that we gave the treatment to a patient aged 60 who subsequently died. An interesting question to answer is "Would this patient have survived if we had avoided giving this treatment?". This question is formalized as the counterfactual probability

$$P(\boldsymbol{y}_{\boldsymbol{t}=0} = 1 \mid \boldsymbol{y} = 0, \boldsymbol{t} = 1, \boldsymbol{x} = 60).$$

[8] We could be tempted to write $P(\boldsymbol{y} = y \mid \text{do}(\boldsymbol{x} = x), \boldsymbol{x} = x')$, however, in the model under intervention $\text{do}(\boldsymbol{x} = x)$, the event $\boldsymbol{x} = x'$ has probability zero, therefore $P(\boldsymbol{y} = y \mid \text{do}(\boldsymbol{x} = x), \boldsymbol{x} = x')$ is undefined.

We can use the definition of conditional probabilities to convert this expression into a ratio of counterfactual probabilities as in Definition 2.21:

$$P(\boldsymbol{y}_{\boldsymbol{t}=0} = 1 \mid \boldsymbol{y} = 0, \boldsymbol{t} = 1, \boldsymbol{x} = 60) = \frac{P(\boldsymbol{y}_{\boldsymbol{t}=0} = 1, \boldsymbol{y} = 0, \boldsymbol{t} = 1, \boldsymbol{x} = 60)}{P(\boldsymbol{y} = 0, \boldsymbol{t} = 1, \boldsymbol{x} = 60)}.$$

In practice, it is often difficult to compute counterfactual probabilities by evaluating Eqs. (2.48) and (2.49). In fact, the full causal model $M = (P, \boldsymbol{U}, \boldsymbol{V}, G, F)$ is typically unknown, and its formal definition serves more as a theoretical basis for developing causal inference methods than as an object to be fully determined from experiments. While the distribution of the observed variables $\boldsymbol{V}$ and the causal graph G can be reasonably inferred from data and expert knowledge, it is often impossible to specify the background variables $\boldsymbol{U}$ and the functional dependencies in F. However, to make Definition 2.21 more concrete, we provide a detailed explanation of the computation of counterfactuals from a fully defined causal model, as well as a numerical example in a simple system, in Appendix B. In Chapter 6, we explore various approaches to estimate the probability of counterfactuals from different sets of assumptions.

2.3 Machine learning

Machine learning is the cornerstone of modern data-driven decision-making. As data sources continue to expand in size and complexity, machine learning procedures have become indispensable tools for extracting valuable insights, making predictions, and automating decision processes. This section provides an overview of the machine learning process, from data preparation to model training and evaluation. We specifically address the aspects pertinent to the prediction of customer churn in telecom. This section is inspired by (Bontempi, 2017; De Stefani, 2022).

Fig. 2.7 depicts a representation of the different steps of the machine learning procedure. We give a brief summary of the different steps:

- *Problem formulation*: Defining the nature of the problem, such as the dependency to be modeled, the corresponding outcome variable, and the features to be considered.
- *Experimental design*: Specifying how the data will be collected.
- *Data collection*: Gathering relevant data for the problem following the experimental design
- *Data preprocessing*: Cleaning, transforming, and preparing the data, handling missing values and outliers.
- *Learning phase*: Choosing a model and training it with data, possibly repeating this process until satisfactory performances are reached. This consists of three sub-steps:
  - *Model definition*: Choosing the class of machine learning model and its hyperparameters.
  - *Parameter learning*: Training the model by adjusting its parameters using the training data.

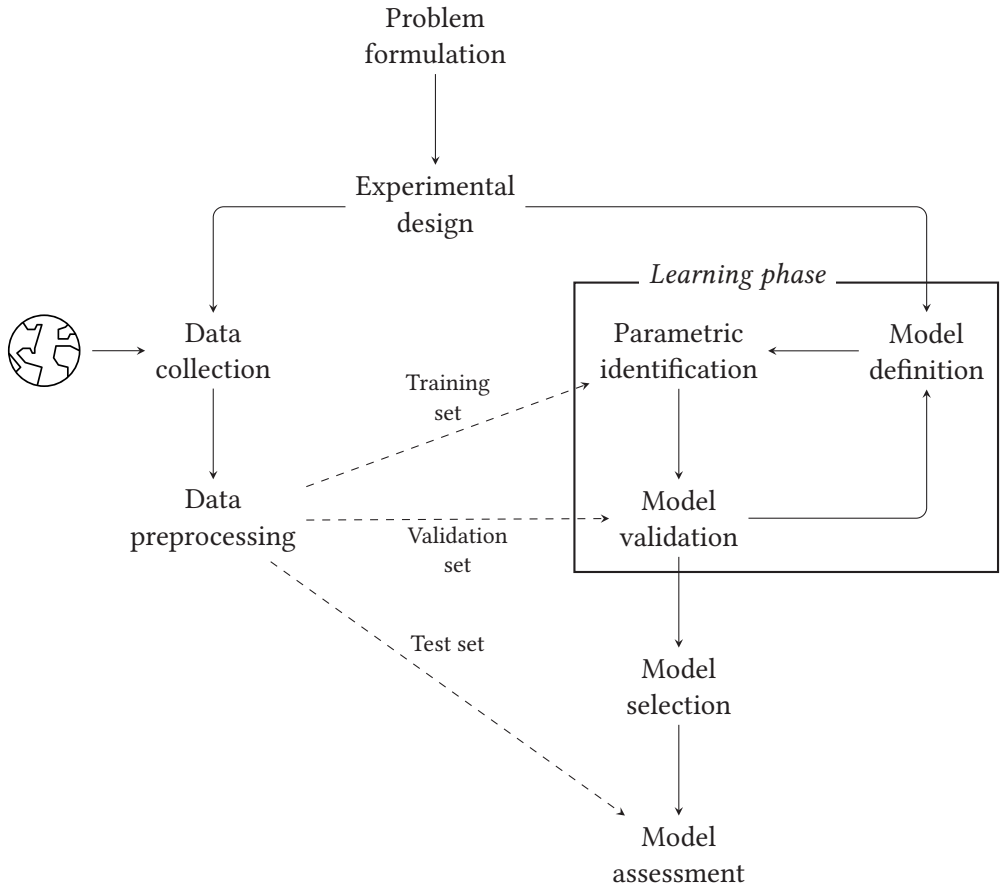


Figure 2.7 The machine learning procedure, from problem formulation to model assessment.

  - *Model validation*: Assessing the model performance on a validation dataset to compare different model classes or hyperparameter values.

- *Model selection*: Choosing the best performing model from different candidate models.

- *Model assessment*: Evaluating the generalization performance of the model on an independent test dataset to estimate its real-world performance accurately.

We go through each step of this procedure in Sections 2.3.1 to 2.3.5. Since model validation during the learning phase and the final model assessment are based on the same procedures and concepts, both are reviewed in the same subsection. Then, in Section 2.3.6, we review other important concepts that do not pertain to a specific step in the machine learning procedure. Finally, in Section 2.3.7, we illustrate the notions presented here on the example of customer churn prediction by telecom companies.

2.3.1 Problem formulation

The problem formulation step is an important phase of the machine learning process in which the practitioner defines the practical aspects of the problem, such as the input and output variables. They give the qualitative objective of the application setting (e.g.,

reducing customer churn) a quantitative and measurable definition (e.g., minimize the number of customers ending their subscription during the two months following a marketing action). The modeler enumerates the set of data that are available and relevant to the problem, and the broad class of machine learning procedures to be considered (supervised learning, unsupervised learning, reinforcement learning, causal inference, causal discovery, etc.).

2.3.2 Experimental design

In this phase, the practitioner determines how the data samples will be collected and what part of the input space should be targeted, to ensure that the training data are representative of the setting in which the machine learning model is intended to be applied. Let us take as an example the application setting of this thesis, where direct marketing retention campaigns are repeated every month to address customer churn. The experimental design consists of specifying what will be used as training and test data, which customers are going to be contacted, which communication channel is going to be used, etc. Also, in this example, it is important to create a control group to assess the causal impact of the campaign and to consider other business-related aspects such as avoiding contacting the same customers in consecutive campaigns. Note that, in some cases, the data come from a preexisting dataset (for example, publicly available online), which does not offer the possibility to influence the experimental design underlying these data.

2.3.3 Data preprocessing

Once the data have been collected according to the experimental design (or from an online source), it is very often in a form that is not compatible with the input expected by the learning algorithm. This incompatibility can be due to the representation of the data (e.g., unstructured data need to be formatted in tabular form) or to its statistical characteristics (e.g., some models perform better if the data are scaled to the unit interval). We briefly discuss the most common preprocessing steps.

Missing values

Some parts of the dataset might be missing, due to errors in the measurement process or the software that collected the data, or for other reasons. There exist a number of approaches to address this issue. The simplest is to discard the data samples containing missing values. If the number of missing values is relatively low compared to the number of samples, this can have only a minor impact on the performances. If this approach is not feasible (due to a large number of missing values or a low number of samples in the first place), then the practitioner must impute the missing values. This can be done with domain knowledge, for example, the information present in other features might be sufficient to infer a reasonable value for the missing feature. Or, if the data are following a temporal order, the missing value of a feature can be predicted from the past values. Lastly, domain-agnostic methods have been developed that are based on statistical properties of the data and assumptions on the process causing the missing values (Emmanuel et al., 2021; Ghahramani and Jordan, 1993; Hand, 1981)

Outliers

Some aspects of the data collection process can lead to erroneous values that are not typical in the statistical distribution of the rest of the data samples. As an example, consider data encoded on an accounting keyboard, where there is a *000* key right next to the *0* key. Pressing by mistake the *000* key instead of *0* scales the number by a factor of 100. Outliers can be detected using simple thresholds or statistical methods. They can be handled in the same way as missing values, or by choosing a learning algorithm robust to outliers (Huber, 2004).

Feature transformation

Some models take as input only certain types of feature or work better when the data distribution is within a certain scale.

For example, neural networks take as input numerical variables and, therefore, categorical variables must be encoded in some way. A common approach is to use one-hot encoding, in which a set of binary variables indicates the value taken by the categorical feature (Johannemann et al., 2019). Another solution applicable in binary classification settings is to replace the categorical value by the probability that the target is positive in the subgroups defined by the categorical values (Micci-Barreca, 2001).

Some learning algorithms benefit from scaling the data to a standard range. This can be achieved by either transforming the features so that their maximum and minimum values are, respectively, zero and one. It is clear that removing outliers beforehand is crucial. Another solution is to map the data to a standard normal distribution by subtracting the mean and dividing by the standard deviation of each feature.

Feature selection

Lastly, feature selection is often used to improve the performance and interpretability of the model. This consists in selecting only the features relevant for the prediction of the target variable. This is especially important for some learning algorithms designed for problems involving a small number of features. This also has implications in causal analysis, where learning to predict the target variable using only causes of this variable has interesting generalization properties (Schölkopf et al., 2012). Several methods exist for feature selection, such as filter methods and wrapper methods (Jović, Brkić, and Bogunović, 2015), or causal approaches (Bontempi and Flauder, 2015; Bontempi and Meyer, 2010).

2.3.4 Learning phase

The learning phase consists of a feedback loop composed of three steps: model definition, parametric identification, and model validation. A machine learning model contains a number of parameters, some of which must be determined manually by the practitioner (called the *hyperparameters*), while the other parameters are tuned during the parametric identification step. For example, the number and width of the hidden layers of a typical feedforward neural network must be set manually, while the weight of each link between the neurons is automatically adjusted in the learning phase. The performance of the trained model is then estimated on a validation set during the model validation step, and the whole learning phase can be repeated to evaluate the performance of different architectures and hyperparameter values.

To make the following discussion more precise, we introduce the mathematical notation of the machine learning procedure. The vector of feature is noted $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_d]$, and the target variable is noted $\boldsymbol{y}$. The domain of $\boldsymbol{x}$, noted $\mathcal{X}$, is usually a subset of the real vectors $\mathbb{R}^d$. The domain of $\boldsymbol{y}$, noted $\mathcal{Y}$, is either a subset of $\mathbb{R}$ in regression tasks or a finite set in classification tasks (e.g., $\mathcal{Y} = \{0, 1\}$ in the binary classification task). We assume the existence of a data-generating process that determines the value of $\boldsymbol{y}$ from the value of $\boldsymbol{x}$, and some unknown noise factor $\boldsymbol{w}$, in the form

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{w}). \tag{2.50}$$

This setting is very general, but some models make more restrictive assumptions on the nature of this process. For example, the assumption of *homoscedasticity* specifies that the variance of $\boldsymbol{w}$ is the same for all samples. Settings that do not conform to this assumption are named *heteroscedastic*. Another typical assumption is that the samples in the dataset $\boldsymbol{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{N}$ are *independent* and *identically distributed* (iid). While these assumptions have proven useful in developing powerful predictive models, the validity of these hypotheses should be assessed, and the impact of their possible violation should be taken into account when training a new model. For example, in direct marketing campaigns, a customer can discuss their positive or negative experience following a marketing call or email to their friends or family. This is called the *spillover effect*, and it leads to the data samples not being independent and identically distributed. This also occurs in medical studies where an individual influences the decision of other individuals to take the treatment.

Model definition

The phase of model definition involves selecting the type of model to address the specified problem. A wide range of models is found in the literature, such as linear models, decision trees, neural networks, physics-based models, etc. The choice should be based on the characteristics of the dataset and the requirements of the task. Model definition also includes configuring the model's architecture (this is called *structural identification*) and determining factors like the number of layers in a neural network or the depth of a decision tree (this is called *hyperparameter tuning*).

In mathematical terms, the practitioner determines the hypothesis function $\mathscr{M}_\theta(\boldsymbol{x})$, which is parameterized by a vector $\theta \in \Theta$, where Θ is, for example, a subset of $\mathbb{R}^m$ for some integer m. The hypothesis $\mathscr{M}_\theta$ is constructed in a way that allows one to find the optimal value of θ given a training dataset. The method used to find the optimal parameters and the definition of "optimal" in this context is the subject of the following two sections.

Parametric identification

This step is the core of the machine learning procedure, where the model tunes the values of θ that most closely fit the training set $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$. This is done by minimizing the *empirical risk* R_{emp}, which is defined as

$$R_{\text{emp}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} L(y^{(i)}, \mathscr{M}_\theta(x^{(i)})) \tag{2.51}$$

where the function L is called the *loss function*. Therefore, the objective of parametric identification is to find the optimal $\theta^* \in \Theta$ that satisfies the principle of *empirical risk*

minimization (ERM):

$$\theta^* = \arg\min_{\theta \in \Theta} R_{\text{emp}}(\theta). \tag{2.52}$$

This principle expresses the fact that we do not know the *true risk*, which is the loss function evaluated in the real setting where the model will be used in practice. The empirical risk minimization principle consists in estimating this risk using a training set instead.

The choice of loss function L depends on domain knowledge and the type of hypothesis $\mathscr{M}$ considered. In regression settings, the *mean squared error* (MSE) is often used, and it is defined as

$$L(y, \mathscr{M}_\theta(x)) = (y - \mathscr{M}_\theta(x))^2. \tag{2.53}$$

In binary classification settings, the model $\mathscr{M}_\theta$ typically estimates the probability that the target $\boldsymbol{y}$ is positive, given the features $\boldsymbol{x}$: $\mathscr{M}_\theta(x) \approx P(\boldsymbol{y} = 1 \mid \boldsymbol{x} = x)$. In this setting, the *cross-entropy* is commonly used as a loss function, and it is defined as

$$L(y, \mathscr{M}_\theta(x)) = -y \log \mathscr{M}_\theta(x) - (1 - y) \log(1 - \mathscr{M}_\theta(x)). \tag{2.54}$$

We can show that minimizing this quantity is equivalent to minimizing the likelihood of the parameters θ.

Once the loss function is defined, various approaches can be adopted to find the optimal parameters. In simple settings such as linear regression, an analytical solution for θ^* can be found. However, in more complex settings, the absence of an analytical solution leads to the use of other approaches. *Gradient-based methods* seek to minimize the empirical risk by solving the equation

$$\nabla R_{\text{emp}}(\theta) = 0 \tag{2.55}$$

where ∇ is the gradient of the empirical risk with respect to θ. The gradient descent method and the Newton method are two gradient-based methods that start with an initial solution θ_0, and repeatedly update the solution by following the first- or second-order gradient of the loss function, until a fixed point is found in the solution space.

Decision tree (Breiman et al., 1984) is an important algorithm that is not based on gradient and instead recursively partitions the feature space into smaller regions (represented by leaves in a tree), until the subset of samples in each region is sufficiently homogeneous (or another stopping criterion is met). In this case, the loss is used to quantify the goodness of fit of a given leaf. Although the performance of decision trees is limited in real-world applications, it has been used with success in ensemble methods such as random forests (Breiman, 2001) or gradient boosting (Friedman, 2001).

Finally, other approaches that are not gradient-based include support vector machines (Cortes and Vapnik, 1995), naive Bayes (Rish et al., 2001), nearest neighbors (T. Cover and Hart, 1967), genetic algorithms (Goldberg, 1989), simulated annealing (Kirkpatrick, Gelatt Jr, and Vecchi, 1983), particle swarm optimization (Kennedy and Eberhart, 1995), etc.

Model validation and assessment

Once a candidate model has been found by minimizing the empirical risk, its predictive performance should be evaluated on another dataset. This is to avoid the issue of *overfitting*, where the model has learned to predict particularly well the value of the

target variable in the training set, but in a way that does not generalize well on another dataset, even identically distributed. This is particularly important if the training set is small or if the model has a large number of parameters. It is common to use two held-out datasets, the *validation set* and the *test set*. The validation set is used to compare the performance of different model architectures or hyperparameter values during the model selection step. Therefore, it is used as many times as the learning phase is repeated. The test set is used after the best model has been selected (which is the subject of the next section) and serves as a final benchmark. Using a test set rather than the validation set avoids overfitting the validation set when evaluating a large number of alternative models during the learning phase. In some settings, the test set is notably different from the training and validation sets, for example, in the context of distribution shift, where the model is used in a setting with a data distribution different from that of the training samples. We should mention k-fold cross-validation as an alternative to a simple held-out validation set, which possesses more interesting statistical properties (Bontempi, 2017, Sec. 7.11.1). It consists of separating the dataset into k equal sized subsets called *folds*, training a model using the last $k-1$ folds, and using the first fold as a validation set. Then, the process is repeated using the second fold as validation set, etc. The average prediction error in all k folds is a more representative estimation of the generalization error of the model than when using a single validation set.

In this thesis, our main focus is on binary classification, so we will look at metrics commonly used to evaluate the effectiveness of a binary classification model. Assume that a model $\mathscr{M}_\theta$ predicts a score $\mathscr{M}_\theta(x^{(i)})$ for each sample $x^{(i)}$ in the test set $D_{\text{te}} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N_{\text{te}}}$. We use a threshold $\tau \in \mathbb{R}$ to determine the predicted label $\hat{y}^{(i)}$ as follows:

$$\hat{y}^{(i)} = \mathbb{I}[\mathscr{M}_\theta(x^{(i)}) \geq \tau] \tag{2.56}$$

where $\mathbb{I}[\cdot]$ is the *Iverson bracket*, equal to one when the expression between brackets is true, zero otherwise. Then, from the combined value of the true label y and the predicted label $\hat{y}$, we define the *confusion matrix* (Fawcett, 2006)

$$\text{CM} = \begin{array}{c} \begin{array}{cc} y=0 & y=1 \end{array} \\ \left[\begin{array}{cc} \text{TN} & \text{FN} \\ \text{FP} & \text{TP} \end{array}\right] \begin{array}{l} \hat{y}=0 \\ \hat{y}=1 \end{array} \end{array} \tag{2.57}$$

where each element of the matrix indicates the corresponding number of samples in the test set (therefore, we have $\text{TN} + \text{FN} + \text{TP} + \text{FP} = N_{\text{te}}$). From the confusion matrix, several evaluation metrics are defined:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{N_{\text{te}}} \tag{2.58}$$

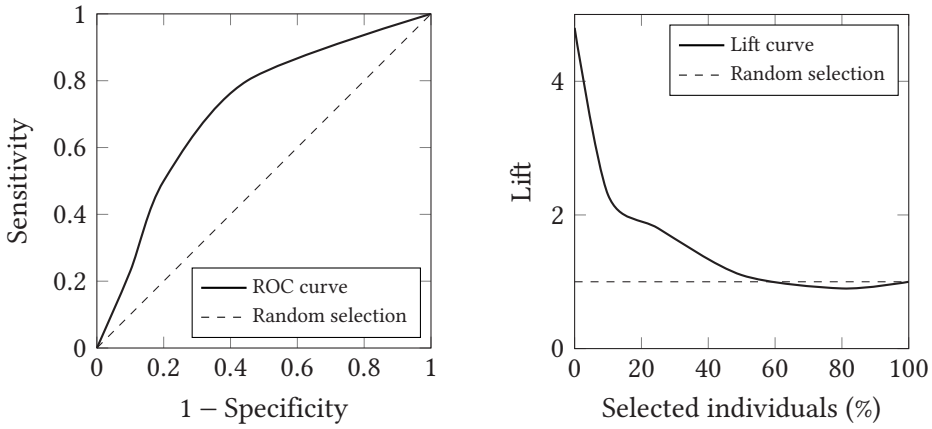
$$\text{Sensitiviy} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.59}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.60}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.61}$$

$$\text{F1-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2.62}$$

Accuracy is the simplest to interpret, it measures the ratio of correct classifications. *Sensitivity* is the proportion of samples correctly predicted to be positive among the



(a) Example of a ROC curve.

(b) Example of a lift curve.

Figure 2.8 Examples of evaluation curves for binary classification problems.

samples actually positive. It is also named *recall*, and *true positive rate* (TPR). *Specificity* measures the proportion of samples correctly predicted to be negative among the samples actually negative. *Precision* is the number of true positives among the samples predicted to be positive. Finally, the *F1-score*, is the harmonic mean of precision and sensitivity.

The confusion matrix and all performance metrics described above depend on a specific threshold τ. It is not always clear how to determine the ideal threshold, and different models might perform better or worse depending on the threshold. Two common metrics, the ROC curve and the lift curve, are parametric curves that represent the performance of a model as a function of the threshold τ.[9] The ROC curve is widely used in machine learning in general (Fawcett, 2006), while the lift curve is more commonly used in churn prediction and online retail (Verbeke, Martens, and Baesens, 2014; Zhu, Baesens, and Broucke, 2017). When comparing multiple models, it is often useful to compute the area under these curves to obtain a single quantitative measure of performance that does not rely on the choice of τ.

To make the following definitions easier, we now note the evaluation metrics defined above as functions of the threshold τ. The ROC curve is defined as the parametric curve

$$\mathrm{ROC}(\tau) = (1 - \mathrm{Specificity}(\tau), \mathrm{Sensitivity}(\tau)). \tag{2.63}$$

An example ROC curve is shown in Fig. 2.8a. A ROC curve approaching the upper left corner indicates that the model is simultaneously sensitive (i.e., it correctly detects positive outcomes) and specific (i.e., it avoids labeling negative instances as positive). The diagonal line serves as a baseline and indicates the expected performance of a random classifier.

The lift curve is defined as

$$\mathrm{Lift}(\tau) = \frac{\mathrm{Precision}(\tau)}{\mathrm{Precision}(1)} \tag{2.64}$$

[9] The evaluation measures presented above (accuracy, sensitivity, etc.) can also be evaluated for different thresholds τ, resulting in parametric curves. However, their typical usage in the literature assumes a fixed threshold, while the ROC and lift curves are more commonly computed for all possibles thresholds.

where Precision(1) designate the precision resulting from classifying all instances as positive, which boils down to the proportion of positive outcomes. An example lift curve is shown in Fig. 2.8b. The dashed horizontal line indicates the rate of positive outcomes in the overall population. The lift curve indicates the expected outcome rate if we select only a portion of the population according to the ranking provided by the model. If the lift curve is above the dashed line, e.g., when selecting the top 20% in Fig. 2.8b, this indicates that this subset of individual displays a higher probability of positive outcomes than the general population. A good model is able to rank first individuals with a high probability of a positive outcome, resulting in a higher lift curve.

2.3.5 Model selection

In the last step of the machine learning procedure, the practitioner selects the final model from the set of models trained and evaluated during the learning phase. This choice is dictated by different factors, the most important of which is often performance on the validation set. But other factors are being increasingly considered, such as interpretability (Moraffah et al., 2020), fairness (Mehrabi et al., 2021), or energy consumption (García-Martín et al., 2019). When these additional factors are not taken into account, the obvious strategy is to select the model with the best performance among the alternatives. This approach is called the *winner-takes-all* approach. However, it is well known that combining the predictions of different models can lead to better predictions than any individual model (Perrone and Cooper, 1995). This is called *ensemble learning*. For example, the random forest algorithm (Breiman, 2001) trains an ensemble of weak learners with low bias and high variance (see Section 2.3.6 for a definition of these concepts), so that their average predictions have low bias and low variance. Ensemble learning has many more advantages, as described by Sagi and Rokach (2018). Ensembles of machine learning models have been used successfully in a number of applications, such as healthcare (Livieris et al., 2019), fraud detection (Lebichot et al., 2021), or sentiment analysis (Basiri et al., 2021).

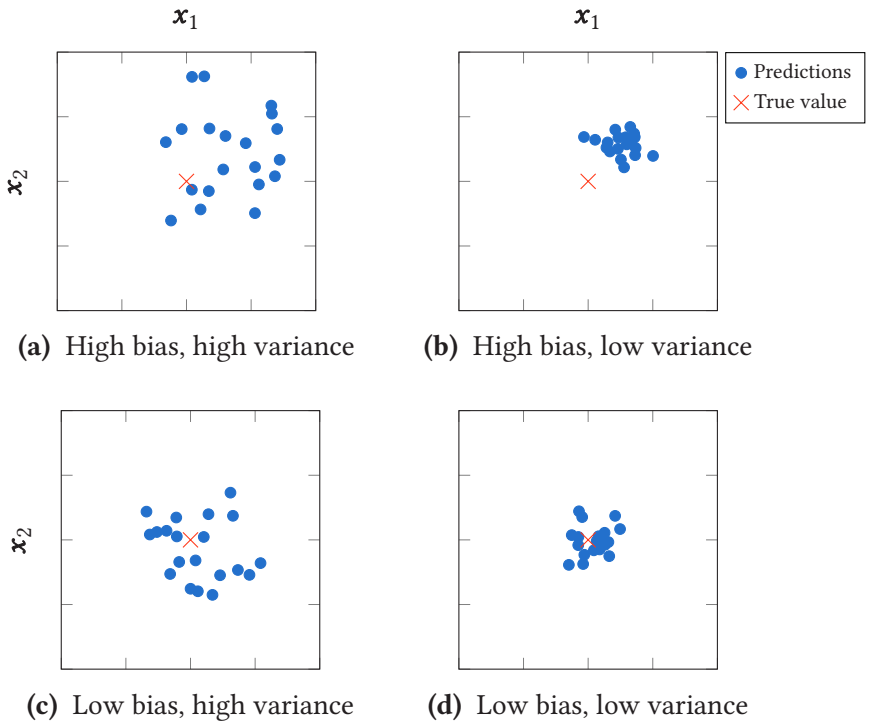
2.3.6 Other concepts

In this section, we review important concepts that do not pertain to a specific step in the machine learning procedure.

Bias and variance of a model

The concepts of *bias* and *variance* are fundamental to understand the generalization abilities of a machine learning model. These concepts, also used in statistics, represent two different aspects of the quality of the predictions of a model. The *bias* indicates the general tendency of the model to predict *around* the correct value, or, in contrast, the tendency of the model to predict a value that is systematically offset from the true value by a fixed amount. The *variance* quantifies the spread of the distribution of predicted values, regardless of whether these predictions are biased or not. The two notions are illustrated in Fig. 2.9.

These notions can be defined mathematically using the notions of expected value (Definition 2.1) and variance of a distribution (Definition 2.3). The notion of distribution variance is closely linked but distinct from the notion of variance of an estimator. Using the notation developed in Section 2.3.4, let us suppose that we aim to model the relationship between a target variable $\boldsymbol{y}$ and a set of covariates $\boldsymbol{x}$, where $\boldsymbol{y}$ follows



(a) High bias, high variance **(b)** High bias, low variance

(c) Low bias, high variance **(d)** Low bias, low variance

Figure 2.9 Illustration of the concept of bias and variance. The model has a high bias (first row) or a low bias (second row), and a high variance (first column) or a low variance (second column).

a data-generating process $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{w})$ for some unknown noise factor $\boldsymbol{w}$. Given an observation $\boldsymbol{x} = x$, we want to estimate

$$\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{x} = x] = \mathbb{E}[f(x, \boldsymbol{w})]. \tag{2.65}$$

For simplicity, we note $S(x) = \mathbb{E}[\boldsymbol{y} \mid \boldsymbol{x} = x]$. Let us assume that we train a model $\mathscr{M}$ on a dataset D_{tr}. In Section 2.3.4, we noted the parameters θ of the model after training as $\mathscr{M}_\theta$. Here, the value of θ is irrelevant, but the training set is an important part of the definitions. Hence, we note the prediction of $\mathscr{M}$ trained on D_{tr} for an observation $\boldsymbol{x} = x$ as $\mathscr{M}(x, D_{\text{tr}})$. Since the training set is the result of a random process, we note it as a random variable $\boldsymbol{D}_{\text{tr}}$. The model $\mathscr{M}$ can be considered as a function that takes as input a feature vector and a training set. Applying this function to a fixed x and a random training set $\boldsymbol{D}_{\text{tr}}$, we obtain the *sampling distribution* $\mathscr{M}(x, \boldsymbol{D}_{\text{tr}})$. While the estimand $S(x)$ is fixed, its estimation $\mathscr{M}(x, \boldsymbol{D}_{\text{tr}})$ is a random variable, due to the random nature of the training set. This allows us to define the concept of bias and variance of an estimator as follows.

Definition 2.22 (Bias of an estimator)**.** The bias of an estimator $\mathscr{M}(x, \boldsymbol{D}_{\text{tr}})$ of $S(x)$ is

$$\text{Bias}(\mathscr{M}, x) = \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[\mathscr{M}(x, \boldsymbol{D}_{\text{tr}})] - S(x). \tag{2.66}$$

When $\text{Bias}(\mathscr{M}, x) = 0$, we say that the estimator is *unbiased* (for $\boldsymbol{x} = x$).

Definition 2.23 (Variance of an estimator)**.** The variance of an estimator $\mathscr{M}(x, \boldsymbol{D}_{\text{tr}})$ of $S(x)$ is

$$\text{Var}(\mathscr{M}, x) = \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}\left[\left(\mathscr{M}(x, \boldsymbol{D}_{\text{tr}}) - \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[\mathscr{M}(x, \boldsymbol{D}_{\text{tr}})]\right)^2\right]. \tag{2.67}$$

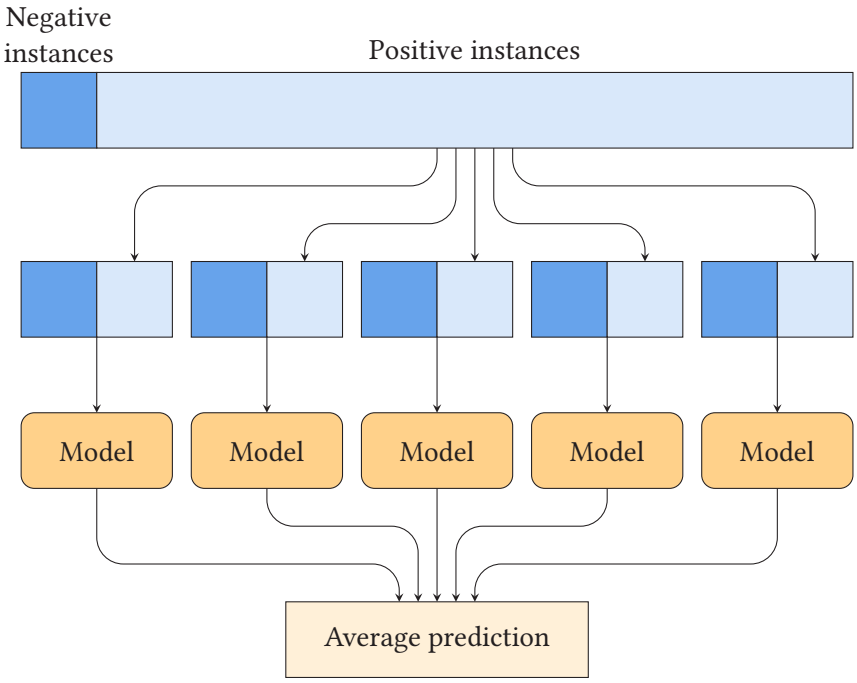


Figure 2.10 Overview of the EasyEnsemble methodology.

Class imbalance

Class imbalance designates the difference in the proportion of occurrence of the two outcomes of the target variable. This is an important notion in churn prediction, since most customers do not churn. Several methodologies exist to address class imbalance (Batista, Prati, and Monard, 2004), which modify the machine learning procedure described above in different ways. Some intervene during the data preprocessing step, others during the learning phase, and others during the model selection step. EasyEnsemble belongs to the latter category. In this thesis, we use the *EasyEnsemble* methodology (X.-Y. Liu, Wu, and Zhou, 2009), which consists in randomly selecting several subsets of negative instances (which are assumed to be more numerous), and pairing each of these subsets with the whole set of positive instances, in such a way that these new datasets are balanced. Then, a different model is trained on each of these balanced datasets. The prediction of all the base models is averaged to obtain the final prediction. This procedure is depicted in Fig. 2.10.

2.3.7 The example of churn prediction

We now illustrate the machine learning procedure presented in the previous sections on the example of customer churn prediction by telecom companies. This description is based on the practical experience of data scientists working at Orange Belgium. This serves as an illustration of the machine learning procedure rather than a rigorous state of the art of churn prediction. We recommend to the reader interested in the current trends in churn prediction the reviews by Coussement, Lessmann, and Verstraeten (2017), Geiler, Affeldt, and Nadif (2022), Jain, Khunteta, and Srivastava (2021), Tianyuan and Moro (2021), and Zhu, Baesens, and Broucke (2017).

Problem formulation Customer churn is addressed by conducting a direct market-

ing campaign using phone calls. Phone operators either suggest a tariff plan that might be more suited to the usage of the customer, or propose them a promotional offer. Then, we consider that a customer churns if they cancel their subscription in the 2 months following the phone call (or in the 2 months following the campaign if the customer is in the control group). If the customer churns later, the churn is not attributed to the campaign. To determine which customers should be called, we train a machine learning model to predict the probability of churn of all customers based on their characteristics and usage patterns.

Experimental design Historical data from previous campaigns is used to train a machine learning model to predict the churn outcome based on customer features. Then, a number of customers with the highest risk of churn (predicted from the current customer data) are forwarded to the marketing team, which is responsible for creating a control group and ensuring that customers are not called repeatedly for different campaigns.

Data preprocessing Data preprocessing is an important step in modern businesses where the same customer data is used for a variety of tasks. As such, the data must include as many relevant features as possible, possibly transforming existing ones into a form that is more easily leveraged by learning algorithms. This includes also the other steps mentioned in Section 2.3.3.

Model definition In this phase, we determine the most relevant class of machine learning models for churn prediction. Churn prediction is characterized by a low class separability (i.e., positive and negative outcomes are difficult to discriminate) and class imbalance (very few churners compared to non-churners), but a reasonable number of features (a few hundreds) and number of samples (at most a few millions), which is in the realm of conventional machine learning. Various benchmarks and internal experiments have shown that gradient boosting and random forest display the best performance for the task of churn prediction.

Parameter learning, model validation and selection Given that all the practical aspects of the machine learning procedure specific to churn have been defined in the previous steps, the following three steps (parameter learning, model validation and selection) are similar to any other conventional binary classification problem in machine learning. Typical evaluation metrics for churn prediction include the lift curve and the ROC curve, presented in Section 2.3.4.

Model assessment While the efficacy of the machine learning model used to predict churn is assessed on a test set with a ROC curve or a lift curve, as discussed in Section 2.3.4, we also compare the churn rate in the control group and in the target group in the two months following the campaign. This indicates the efficacy of the campaign in reducing churn, and this metric is more relevant to other stakeholders in the company than the predictive performance of the machine learning model.

The metric mentioned in the model assessment step is called the campaign *uplift*. Since uplift represents an important business objective, a dedicated approach called *uplift modeling* is now favored over conventional churn prediction (Devriendt, Berrevoets, and Verbeke, 2021). This is the subject of the next section.

3

State of the art

In this chapter, we introduce the two fields of research most closely linked to our contributions, that is, uplift modeling (Section 3.1) and counterfactual identification (Section 3.2), and we review the state of the art of both fields.

3.1 Uplift modeling

The concept of uplift, also known as conditional average treatment effect (Gutierrez and Gérardy, 2016) and heterogeneous treatment effect[1] (Athey and G. Imbens, 2016), has emerged as a important tool within the field of data-driven decision-making. It allows one to understand the magnitude of the causal effect of an action on an outcome. Uplift modeling is used primarily in scenarios where the objective is not just predicting individual outcomes, but identifying the specific individuals who are most likely to benefit from a particular intervention or treatment. This perspective distinguishes it from traditional predictive modeling, which focuses on estimating the likelihood of an outcome for an individual based on input features alone. More precisely, uplift models quantitatively estimate the impact of the intervention on the probability distribution of the outcome for each individual.

This section describes the state of the art in uplift modeling, with a focus on the aspects relevant to this thesis. In Section 3.1.1, we explain the concept of uplift modeling in mathematical terms and outline the assumptions most commonly used. In Section 3.1.2, we cover the most common approaches for learning uplift. We outline the evaluation used in the uplift literature in Section 3.1.3. Finally, in Section 3.1.4 we present the body of research that compares the performance of uplift modeling with respect to traditional predictive modeling, an issue that is receiving increasing attention in the literature. We refer the reader to the book by Michel, Schnakenburg, and Von Martens (2019) for an introduction to more technical aspects of uplift modeling, such as feature selection and transformation, software implementation, a more detailed discussion of the different uplift models, and other technical considerations.

[1]The term *uplift* is commonly employed in the context of marketing or customer management, whereas *conditional average treatment effect* and *heterogeneous treatment effect* are more typical in medical science, social science, and econometrics literature.

Table 3.1 The four categories of customers for churn prevention in terms of potential outcomes.

	$\boldsymbol{y}_0 = 0$	$\boldsymbol{y}_0 = 1$
$\boldsymbol{y}_1 = 0$	Sure thing	Persuadable
$\boldsymbol{y}_1 = 1$	Do-not-disturb	Lost cause

3.1.1 Problem formulation

Notation

In uplift modeling, the random variable $\boldsymbol{y}$ represents the binary outcome of interest, $\boldsymbol{t}$ is the binary treatment, and $\boldsymbol{x}$ is the random vector of the features. We note their respective domains $\mathcal{Y} = \{0, 1\}$, $\mathcal{T} = \{0, 1\}$, and $\mathcal{X} \subseteq \mathbb{R}^n$. Furthermore, we assume that $\boldsymbol{x}$ is continuous and has a probability density function $f_{\boldsymbol{x}}$. In the example of churn prevention with marketing campaigns, $\boldsymbol{y}$ is the churn outcome indicator ($\boldsymbol{y} = 1$ when the customer churns and $\boldsymbol{y} = 0$ when they stay), $\boldsymbol{t}$ indicates whether the customer was contacted during the campaign, and $\boldsymbol{x}$ is a set of descriptors used to characterize the customer in terms of usage patterns, demographics, etc.

Following Definition 2.20, we note the potential outcome of $\boldsymbol{y}$ under the intervention $\text{do}(\boldsymbol{t} = t)$, for $t = 0, 1$, as $\boldsymbol{y}_t$. This is a random variable that indicates the customer's churn outcome assuming that treatment $\boldsymbol{t} = t$ was applied, regardless of whether the treatment was in reality t or $1 - t$. The objective of uplift modeling is to find customers who are most sensitive to the treatment, that is, customers for whom the potential outcomes $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are the most likely to be different. In fact, we can classify customers into four categories, depending on the four possible values of the two potential outcomes $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$. This classification is represented in Table 3.1 for the case of prediction of customer churn. The ideal category of customer to target are the persuadable customers, since the other customers are either not influenced by the action or are negatively affected. Since we cannot observe both $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ (an issue called the *fundamental problem of causal inference*, Holland 1986), we cannot directly learn to predict the customer category from data. Instead, uplift modeling aims to estimate the difference in the probability of a positive outcome under the treatment scenario (customer is contacted) and the no-treatment scenario (customer is not contacted). Customers maximizing this difference are the most likely to generate a profit increase when contacted.

Since we will often refer to the conditional probability distribution of $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ given the observation $\boldsymbol{x} = x$, we use the following notation:

$$S_0(x) = P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x) \tag{3.1}$$

$$S_1(x) = P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x). \tag{3.2}$$

The uplift for an individual $\boldsymbol{x} = x$ is defined as

$$U(x) = S_0(x) - S_1(x). \tag{3.3}$$

We also note the marginal probabilities (i.e., without conditioning on $\boldsymbol{x} = x$) and the marginal uplift as

$$S_0 = P(\boldsymbol{y}_0 = 1) \tag{3.4}$$

$$S_1 = P(\boldsymbol{y}_1 = 1) \tag{3.5}$$

$$U = S_0 - S_1. \tag{3.6}$$

S_0 and S_1 can be estimated with the proportion of positive outcomes (e.g., churners) in the control and target groups respectively. The uplift U, called *average treatment effect* (ATE) in the literature (Pearl, 2017), can be interpreted as the causal effect of the campaign on the overall population, or the expected causal effect on a randomly selected individual, since it is the expected value of $U(\boldsymbol{x})$ over the distribution of $\boldsymbol{x}$:

$$U = P(\boldsymbol{y}_0 = 1) - P(\boldsymbol{y}_1 = 1) \tag{3.7}$$

$$= \int (P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x) - P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x)) f_{\boldsymbol{x}}(x)\,\mathrm{d}x \tag{3.8}$$

$$= \mathbb{E}[S_0(\boldsymbol{x}) - S_1(\boldsymbol{x})] = \mathbb{E}[U(\boldsymbol{x})]. \tag{3.9}$$

Note that, for example, in the literature pertaining to retail or online advertisements, the uplift is defined as $U = S_1 - S_0$, and similarly $U(x) = S_1(x) - S_0(x)$. This choice depends on whether the probability of the (positive) outcome $\boldsymbol{y} = 1$ should be minimized (e.g., in churn prevention) or maximized (e.g., in sales). The uplift is then defined so that a positive uplift corresponds to a beneficial outcome. Since we apply our results primarily to churn prevention, we use the convention $U = S_0 - S_1$.

Assumptions

We assume that the random variables $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{t}$ are part of a causal model $M = (P, \boldsymbol{W}, \boldsymbol{V}, G, F)$ (see Definition 2.18), of which we only know $\boldsymbol{V} = \{\boldsymbol{y}, \boldsymbol{t}\} \cup \boldsymbol{x}$. The latent variables $\boldsymbol{W}$[2] and the functions F are unknown. We do not know the graph G either; however, we assume that the treatment $\boldsymbol{t}$ does not influence the features $\boldsymbol{x}$ (i.e., there is no path from $\boldsymbol{t}$ to any member of $\boldsymbol{x}$ in G). This assumption is often made, but rarely stated in the literature, and it implies that $\boldsymbol{x}_{\mathrm{do}(\boldsymbol{t}=t)} = \boldsymbol{x}$ for any t. The probabilities $S_0(x)$ and $S_1(x)$ cannot be estimated directly from data without further assumptions, because they involve the potential outcomes $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$, while real-world data consist of samples of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$. Some uplift models are designed for experimental data, while others are designed for observational data. This distinction has an impact on how $S_0(x)$ and $S_1(x)$ can be estimated.

Definition 3.1 (Experimental data)**.** A dataset $\boldsymbol{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{t}^{(i)})\}_{i=1}^{N}$ of iid samples of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$ is said to be *experimental data* if $\boldsymbol{t}$ has no parent in the causal graph G and if $0 < P(\boldsymbol{t} = 1) < 1$. We also say that this is a *randomized controlled trial*. Otherwise, the dataset is said to be *observational data*.

The requirement that $\boldsymbol{t}$ has no parent in G is usually implemented by assigning the treatment at random, which corresponds to the graph surgery mentioned in Definition 2.19. Experimental data implies that

$$\boldsymbol{y}_t \perp \boldsymbol{t} \quad \text{for all } t \in \{0, 1\}. \tag{3.10}$$

This translates the intuition that the potential outcome of $\boldsymbol{y}$, had the treatment $\boldsymbol{t} = t$ been given, does not depend on the actual assignment of treatment. This precludes unobserved confounders that determine both the assignment of treatment and the outcome. The fact that Eq. (3.10) holds with experimental data can be proven from the

[2] we use $\boldsymbol{W}$ instead of $\boldsymbol{U}$ as in Definition 2.18 for the latent variables to avoid confusion with the notation for the uplift $U(x)$.

rules of *do-calculus* (Pearl, 2009, Thm. 3.4.1), or directly from Definition 2.20, however, we will accept this result without proof for the sake of brevity.

Experimental data is not always available because randomizing the treatment assignment can be costly or unethical (e.g., forcing people to smoke to study the impact of smoking on lung cancer), but learning causal effects in an observational setting is still possible. The required assumption is called *unconfoundedness*[3] (Pearl, 2009, Def. 9.2.9).

Definition 3.2 (Unconfoundedness)**.** A variable $\boldsymbol{y}$ is *unconfounded* with respect to $\boldsymbol{t}$ given $\boldsymbol{x}$ if, for all $y \in \mathcal{Y}$, $t \in \mathcal{T}$ and $x\mathcal{X}$,

$$P(\boldsymbol{y}_t = y \mid \boldsymbol{x} = x) = P(\boldsymbol{y} = y \mid \boldsymbol{t} = t, \boldsymbol{x} = x) \tag{3.11}$$

Or, alternatively, if for all $x \in \mathcal{X}$,

$$\boldsymbol{y}_t \perp \boldsymbol{t} \mid \boldsymbol{x} = x. \tag{3.12}$$

In this context, the vector of features $\boldsymbol{x}$ is called the *adjustment set.* We also say that $\boldsymbol{x}$ satisfies the *rule 2 of do-calculus* (Pearl, 2009, Thm. 3.4.1), or the *back-door criterion*[4] (Pearl, 2009, Def. 3.3.1).

Any dataset where the treatment is randomized (i.e., experimental data) is unconfounded, but observational data can also be unconfounded given the right adjustment set $\boldsymbol{x}$. The equivalence of the two expressions in Definition 3.2 follows from

$$\begin{aligned} P(\boldsymbol{y}_t = y \mid \boldsymbol{x} = x) &= P(\boldsymbol{y}_t = y \mid \boldsymbol{t} = t, \boldsymbol{x} = x) && \text{(by Eq. 3.12)} \\ &= P(\boldsymbol{y} = y \mid \boldsymbol{t} = t, \boldsymbol{x} = x). && \text{(by Definition 2.20)} \end{aligned}$$

This last equality follows from Definition 2.20, however we omit the details of this derivation for brevity. Unconfoundedness allows the estimation of the scores $S_0(x)$ and $S_1(x)$ from data, since, in this case, we have

$$S_0(x) = P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x) = P(\boldsymbol{y} = 1 \mid \boldsymbol{t} = 0, \boldsymbol{x} = x), \tag{3.13}$$

and similarly for $S_1(x)$.

3.1.2 Uplift models

In this section, we describe the uplift models that are used most often in the literature. For a more extensive comparison and description of existing uplift models, see the review by Kayaalp (2017), the book by Michel, Schnakenburg, and Von Martens (2019, Ch. 3), or the extensive benchmarks by Devriendt, Moldovan, and Verbeke (2018) and Rößler and Schoder (2022). In this section, we denote that $\mathscr{M}(x)$ is a model trained to predict some quantity $g(x)$ as $\mathscr{M}(x) \approx g(x)$.

[3] Also called *ignorability* by Rosenbaum and Donald B Rubin (1983), *exogeneity* by Pearl (2009), or *conditional independence assumption* by Gutierrez and Gérardy (2016).

[4] The back-door criterion is equivalent to the rule 2 of do-calculus with the additional assumption that there is no causal path from $\boldsymbol{t}$ to $\boldsymbol{x}$, which is always the case in our setting.

T-learner

This approach, called T-learner in Künzel et al. (2019), is often used as a baseline, and consists in modeling the probabilities $S_0(x)$ and $S_1(x)$ separately with two independent predictive models:

$$\mathscr{M}_0(x) \approx S_0(x) \quad \text{and} \quad \mathscr{M}_1(x) \approx S_1(x).$$

The two probability estimators $\mathscr{M}_0$ and $\mathscr{M}_1$ can be, for example, logistic regression, or any conventional machine learning model. The uplift is then estimated as

$$U(x) \approx \mathscr{M}_0(x) - \mathscr{M}_1(x). \tag{3.14}$$

The two predictors are trained, respectively, on the control and target group datasets. Since they are trained independently, they can have different probability calibrations, and, also, the variance of the uplift estimator is greater than the variance of the individual predictors $\mathscr{M}_0(x), \mathscr{M}_1(x)$, due to the subtraction in Eq. (3.14). See Radcliffe and Surry (2011) for an illustration of these problems.

S-learner

A related approach, called the S-learner in Künzel et al. (2019), is based on a single predictive model. The treatment indicator $\boldsymbol{t}$ is included in the vector of input features of the predictive model, along with $\boldsymbol{x}$:

$$\mathscr{M}(x, 0) \approx S_0(x) \quad \text{and} \quad \mathscr{M}(x, 1) \approx S_1(x).$$

Here, the notation $\mathscr{M}(x, t)$ indicates that the value of the treatment t is added as an input feature of the model, and should not be confused with the notation $\mathscr{M}(x, D_{\text{tr}})$ that indicates that the $\mathscr{M}$ is trained on a training set D_{tr}.

The customers are then ranked according to their estimated uplift, again obtained by subtraction:

$$U(x) \approx \mathscr{M}(x, 0) - \mathscr{M}(x, 1). \tag{3.15}$$

This approach avoids the problems of the T-learner by reducing the independence of the estimations of the treated and control probabilities. The drawback is that the treatment $\boldsymbol{t}$ is considered a priori as important as any other feature in $\boldsymbol{x}$. If $\boldsymbol{t}$ is considered not informative by the learner (e.g., as a result of a feature selection step), the model will be ineffective in modeling uplift.

Modified target variable

Presented in Jaskowski and Jaroszewicz (2012), this approach defines a new target variable $\boldsymbol{z}$ as

$$\boldsymbol{z} = \boldsymbol{y}(1 - \boldsymbol{t}) + (1 - \boldsymbol{y})\boldsymbol{t}.$$

In the setting of customer churn, the variable $\boldsymbol{z}$ is equal to one when the customer is in the control group and churns or is in the target group and does not churn. A customer for which $\boldsymbol{z} = 1$ is either a *persuadable*, a *sure thing*, or a *lost cause* (positive or null uplift). Similarly, a customer for which $\boldsymbol{z} = 0$ is either a *do-not-disturb*, a *sure thing*, or a *lost cause* (negative or null uplift). Jaskowski and Jaroszewicz (2012) showed that in the case of a balanced randomized experiment (i.e., such that $P(\boldsymbol{t} = 1) = P(\boldsymbol{t} = 0) = 0.5$), we have

$$U(x) = 2P(\boldsymbol{z} \mid \boldsymbol{x} = x) - 1. \tag{3.16}$$

Any model of the conditional probability of $\boldsymbol{z}$ is thus able to predict the uplift. This model assumes that $P(\boldsymbol{t} = 1) = P(\boldsymbol{t} = 0) = 0.5$, which is not always the case in real-world applications.

In a variation of the modified target approach, called *Lai's generalization* (Kane, V. S. Lo, and Zheng, 2014), a classification model learns to predict the joint value of the outcome and the treatment, i.e., the new target variable is

$$\boldsymbol{z} = \begin{cases} 0 \text{ if } \boldsymbol{y} = 0 \text{ and } \boldsymbol{t} = 0 \\ 1 \text{ if } \boldsymbol{y} = 0 \text{ and } \boldsymbol{t} = 1 \\ 2 \text{ if } \boldsymbol{y} = 1 \text{ and } \boldsymbol{t} = 0 \\ 3 \text{ if } \boldsymbol{y} = 1 \text{ and } \boldsymbol{t} = 1. \end{cases}$$

Then the uplift is estimated as

$$U(x) \approx \frac{P(\boldsymbol{z} = 0)}{P(\boldsymbol{t} = 0)} + \frac{P(\boldsymbol{z} = 1)}{P(\boldsymbol{t} = 1)} - \frac{P(\boldsymbol{z} = 2)}{P(\boldsymbol{t} = 0)} - \frac{P(\boldsymbol{z} = 3)}{P(\boldsymbol{t} = 1)}. \tag{3.17}$$

This has the advantage of being valid even with an imbalanced probability of treatment, but suffers from volatile performance (Devriendt, Moldovan, and Verbeke, 2018).

Tree-based models

This approach builds a model that predicts uplift directly, often inspired by a conventional machine learning algorithm. Several strategies have been proposed in the literature. For example, Athey and G. Imbens (2016) propose a modification of the classification and regression tree (CART) algorithm to learn uplift. Rzepakowski and Jaroszewicz (2012) use an information-theoretic criterion to build an uplift classification tree, which is naturally extended to a random forest (Breiman, 2001) by Guelman, Guillén, and Pérez-Marín (2015). More precisely, the split criterion when constructing a tree node is chosen to maximize the divergence of the target class distribution between the treated and control populations. The distribution divergence is estimated using information-theoretic measures such as the Kullback-Leibler distance (Csiszár and Shields, 2004), the Euclidian distance (L. Lee, 2000), or the chi-squared divergence (Jaroszewicz and Simovici, 2001). Other approaches exist, such as the *difference in difference* approach (B. Hansotia and Brad Rukstales, 2002), or the *Bayesian additive regression trees* (Hahn, Murray, and Carvalho, 2020).

X-learner

Künzel et al. (2019) propose a new algorithm for uplift modeling called *X-learner*. Like the T-learner or the modified target approach, this algorithm relies on a base learner, which can be any conventional machine learning model. X-learner is able to use information from the control group to improve the estimator for the target group, and vice versa. First, two models $\mathscr{M}_0(x)$ and $\mathscr{M}_1(x)$ are trained to predict, respectively, $S_0(x)$ and $S_1(x)$. This step is identical to the T-learner. Then, two models $\mathscr{M}'_0(x)$ and $\mathscr{M}'_1(x)$ are trained to predict, respectively, $P(\mathscr{M}_1(\boldsymbol{x}) - \boldsymbol{y}_0 \mid \boldsymbol{x} = x)$ and $P(\boldsymbol{y}_1 - \mathscr{M}_0(\boldsymbol{x}) \mid \boldsymbol{x} = x)$. The final prediction is a weighted average of the two final models:

$$U(x) \approx g(x)\mathscr{M}'_0(x) + (1 - g(x))\mathscr{M}'_1(x)$$

where $g(x)$ is a weighting function. By using parametric weighting, the average model gives more importance to the predictions of $\mathscr{M}'_1$ or $\mathscr{M}'_0$ where those models are more

confident. The X-learner approach can be used with observational data by using a propensity score as the weighting function $g(x)$, that is, by fitting

$$g(x) \approx P(\boldsymbol{t} = 1 \mid \boldsymbol{x} = x).$$

Note that this approach was originally designed for a continuous target $\boldsymbol{y}$, and thus may perform poorly when applied to a binary target. No assumption is required for the validity of this model, but Künzel et al. (2019) indicate that the X-learner does not perform well when the uplift is close to zero. Also, this algorithm is designed to take advantage of highly unbalanced target and control groups. Künzel et al. (2019) provide theoretical conditions that guarantee a fast convergence rate.

Other approaches

Shalit, Johansson, and Sontag (2017) propose a neural network architecture designed for uplift modeling, inspired from the domain adaptation literature (Ganin et al., 2016). The probabilities of the outcomes $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are learned by two separate heads of the neural network sharing a common base, which leverage the advantages of both the S-learner (the predictors are calibrated similarly, and all the data is used to learn both scores) and the T-learner (the impact of the treatment is taken into account even with numerous features).

Louizos et al. (2017) present a new approach to compute causal effects from observational data, called the *causal effect variational autoencoder* (CEVAE). It is an adaptation of the variational autoencoder (Kingma and Welling, 2013), which learns a latent representation of the input features using variational inference. Variational autoencoders are widely used in various domains, such as image recognition or time series prediction, due to their predictive power and the few assumptions they make on the data-generating process. CEVAE builds upon this method by representing the causal effect and the (possibly hidden) confounders as a latent space. They achieve state-of-the-art performance while being more robust than other uplift models.

Lastly, Zaniewicz and Jaroszewicz (2013) adapt the *support vector machine* (SVM), a well-known machine learning algorithm (Cortes and Vapnik, 1995), to predict uplift. They outperform the baseline methods and show performance similar to that of other state-of-the-art uplift models on their benchmark.

3.1.3 Performance evaluation

Various evaluation measures exist to quantify the performance of uplift models. All of them measure, in some way, the ability of the evaluated model to rank individuals according to their uplift. Since we do not have access to the true uplift of a given individual, these curves rely on the randomization of the treatment assignment and on the law of large numbers to estimate the average uplift of a subset of the population. The two most common evaluation measures are the *uplift curve* and the *Qini curve*. These two curves have been defined in different ways in the literature; see (Devriendt, Van Belle, et al., 2020) for a comparison. We will focus on the most common definitions. Other measures have been discussed in the literature, but they have not yet gained broad acceptance. Among those, we present the profit measure by Verbeke, Olaya, Guerry, et al. (2022), the regret measure by Fernández-Loria and Provost (2022b), and we mention a few other related measures.

Uplift curve

To compute the uplift curve, the uplift model $\mathscr{M}$ is used to predict a score for each sample in a test data set $D_{\text{te}} = \{(x^{(i)}, y^{(i)}, t^{(i)})\}_{i=1}^{N}$. The curve is then estimated by comparing the outcome rate in the control and target groups among the individuals with the highest scores. In the framework established by Devriendt, Van Belle, et al. (2020), this definition corresponds to the absolute, joint uplift curve (Eq. (8) in their paper). Intuitively, the uplift curve indicates the number of additional favorable outcomes that can be attributed to the causal effect of the action, as a function of the number of targeted individuals.

Definition 3.3 (Uplift curve)**.** Let $D_{\text{te}} = \{(x^{(i)}, y^{(i)}, t^{(i)})\}_{i=1}^{N}$ be a data set of N iid realizations of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$, where the treatment assignment $\boldsymbol{t}$ is randomized. Let $\mathscr{M}$ be a model trained on a data set D_{tr}, and let D_{te} be sorted in decreasing order according to $\mathscr{M}$: for any $i < j$, we have $\mathscr{M}(x^{(i)}, D_{\text{tr}}) \geq \mathscr{M}(x^{(j)}, D_{\text{tr}})$. The *uplift curve* is defined for $k \in \{1, \dots, N\}$ as

$$\text{Uplift}(k, D_{\text{tr}}, D_{\text{te}}) = \left(\frac{r_0(k)}{n_0(k)} - \frac{r_1(k)}{n_1(k)}\right) k \tag{3.18}$$

where the following notation is used, with $t = 0, 1$:

$$r_t(k) = \sum_{i=1}^{k} \mathbb{I}[y^{(i)} = 1 \text{ and } t^{(i)} = t] \quad \text{and} \quad n_t(k) = \sum_{i=1}^{k} \mathbb{I}[t^{(i)} = t]. \tag{3.19}$$

In the case where $r_t(k) = n_t(k) = 0$, the quotient $r_t(k)/n_t(k)$ is defined as 0.

In this definition, the quantity $n_t(k)$ represents the number of individuals with treatment $\boldsymbol{t} = t$ among the k individuals with the highest scores. The quantity $r_t(k)$ represents the number of positive outcomes (that is, such that $\boldsymbol{y} = 1$) with treatment $\boldsymbol{t} = t$ among the k individuals with the highest scores. An example of uplift curve is given in Fig. 3.1a. Note that the uplift curve reaches the population uplift $UN = (S_0 - S_1)N$ when $k = N$. Also, for some values of $k < N$, the uplift curve may be higher than UN. This indicates that some of the customers lower in the ranking are *do-not-disturb* customers (see Table 3.1), whose negative reaction to the campaign create a negative slope in the curve.

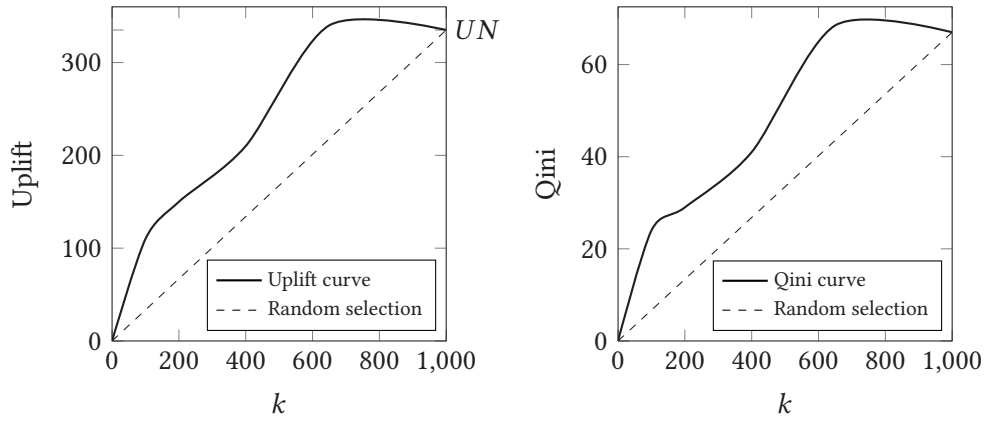
Qini curve

The Qini curve is also commonly used in the literature. We present the joint, absolute definition of the Qini curve in the framework presented by Devriendt, Van Belle, et al. (2020). Intuitively, this curve indicates the number of additional positive outcomes that can be attributed to the causal effect of the action, if we were to conduct a new randomized campaign (i.e., with target and control groups), as a function of the number of individuals included in this new campaign.

Definition 3.4 (Qini curve)**.** Let $D_{\text{tr}}, D_{\text{te}}$ and $\mathscr{M}$ be a training set, a test set, and a model as defined in Definition 3.3. The *Qini curve* is defined for $k \in \{1, \dots, N\}$ as

$$\text{Qini}(k, D_{\text{tr}}, D_{\text{te}}) = r_0(k) - r_1(k)\frac{n_0(k)}{n_1(k)} \tag{3.20}$$

where $r_0(k), r_1(k), n_0(k)$ and $n_1(k)$ are defined as in Definition 3.3. When $n_1(k) = 0$, the quotient $n_0(k)/n_1(k)$ is defined as 0.



(a) Example of uplift curve.

(b) Example of Qini curve.

Figure 3.1 Example of uplift and Qini curves for a population of $N = 1000$ individuals and a treatment rate of 20%. The dashed lines represent the curves resulting from selecting individuals at random. **(a)** The average uplift is 3.35%, therefore, the right end of the uplift curve reaches $UN = 335$. The uplift curve can go higher than the average uplift, reaching 340 when selecting 65% of the individuals. **(b)** The Qini curve is, on average, proportional to the uplift curve by a factor equal to the proportion of individuals assigned to the control group, $P(\boldsymbol{t} = 1)$.

It is easy to see that

$$\text{Qini}(k, D_{\text{tr}}, D_{\text{te}}) = \frac{n_0(k)}{k}\text{Uplift}(k, D_{\text{tr}}, D_{\text{te}}). \tag{3.21}$$

This means that, on average, the uplift curve and the Qini curve are related by a factor equal to the proportion of individuals assigned to the control group. An example of Qini curve is shown in Fig. 3.1b.

Measure of profit

As illustrated by Gubela and Lessmann (2021), the conventional uplift curve presented in Definition 3.3 does not take into account the cost and benefits associated with each individual and with the action $\boldsymbol{t} = 1$. The most general measure, in our opinion, for evaluating the performance of uplift models while taking into account costs and benefits was proposed by Verbeke, Olaya, Guerry, et al. (2022). Its generality stems from the fact that it is not tied to a specific operational setting (e.g., churn prediction or online retail). This is achieved by defining a cost-benefit matrix (see below) that can effectively represent the diverse range of settings characterized by cost sensitivity.

Let $F_{yt}^{D_{\text{tr}}}$ be the cumulative distribution function of the score from a model $\mathscr{M}$ trained on a data set D_{tr}, conditional on a particular realization of the potential outcome $\boldsymbol{y}_t = y$:

$$F_{yt}^{D_{\text{tr}}}(\tau) = P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau \mid \boldsymbol{y}_t = y). \tag{3.22}$$

In this expression, the probability is taken over the distribution of $\boldsymbol{x}$. The condition $\boldsymbol{y}_t = y$ indicates that we consider individuals who, when applying the treatment $\boldsymbol{t} = t$, have an outcome $\boldsymbol{y}_t = y$. Then, we define the *causal confusion matrix* $\text{CF}(\tau, D_{\text{tr}})$ for a

threshold τ as

$$\mathrm{CF}(\tau, D_{\mathrm{tr}}) = \begin{array}{c} \begin{matrix} \boldsymbol{t}=0 & \boldsymbol{t}=1 \end{matrix} \\ \begin{bmatrix} (1-S_0)F_{00}^{D_{\mathrm{tr}}}(\tau) & (1-S_1)(1-F_{01}^{D_{\mathrm{tr}}}(\tau)) \\ S_0 F_{10}^{D_{\mathrm{tr}}}(\tau) & S_1(1-F_{11}^{D_{\mathrm{tr}}}(\tau)) \end{bmatrix} \begin{matrix} \boldsymbol{y}=0 \\ \boldsymbol{y}=1 \end{matrix} \end{array} \tag{3.23}$$

Intuitively, this matrix indicates the expected proportion of positive and negative outcomes in the population if we use a threshold τ to determine which individuals should be targeted. The rows of the matrix indicate positive and negative outcomes, while the columns specify whether these outcomes belong to the targeted or non-targeted groups. As discussed by Verbeke, Olaya, Berrevoets, et al. (2021), the performance of a model should be measured relative to a baseline scenario, rather than in absolute terms. That is because, even when no action is carried out, an outcome will always occur, and therefore, the success of an action should be compared with the outcome resulting from the absence of action. This consideration leads us to define the *causal effect matrix* E as

$$\mathrm{E}(\tau, D_{\mathrm{tr}}) = \mathrm{CF}(\tau, D_{\mathrm{tr}}) - \mathrm{CF}(\infty, D_{\mathrm{tr}}) \tag{3.24}$$

with $\mathrm{CF}(\infty, D_{\mathrm{tr}})$ defined as

$$\mathrm{CF}(\infty, D_{\mathrm{tr}}) = \begin{bmatrix} 1-S_0 & 0 \\ S_0 & 0 \end{bmatrix}. \tag{3.25}$$

The matrix $\mathrm{CF}(\infty, D_{\mathrm{tr}})$ indicates the outcome distribution assuming that nobody is targeted, which corresponds to the outcome distribution in the control group. Finally, we define a cost-benefit matrix CB that expresses the sum of the costs and benefits for the two possible actions ($\boldsymbol{t}=0$ or $\boldsymbol{t}=1$) and the two possible outcomes ($\boldsymbol{y}=0$ or $\boldsymbol{y}=1$). Here, the cost-benefit matrix is the same for all individuals. It is noted

$$\mathrm{CB} = \begin{array}{c} \begin{matrix} \boldsymbol{t}=0 & \boldsymbol{t}=1 \end{matrix} \\ \begin{bmatrix} \mathrm{CB}_{00} & \mathrm{CB}_{01} \\ \mathrm{CB}_{10} & \mathrm{CB}_{11} \end{bmatrix} \begin{matrix} \boldsymbol{y}=0 \\ \boldsymbol{y}=1 \end{matrix} \end{array} \tag{3.26}$$

For ease of notation in the following definition, we note the sum of the elements in the componentwise product of two matrices A and B (also called the Frobenius inner product) as $A \oplus B$:

$$A \oplus B = \sum_{ij} A_{ij} B_{ij}. \tag{3.27}$$

This operation satisfies the same axioms as the inner product between two vectors.

Definition 3.5 (Verbeke, Olaya, Guerry, et al., 2022)**.** The measure of *causal profit* $\mathrm{CP}(\tau, D_{\mathrm{tr}})$, for a threshold τ, a training set D_{tr}, and a constant cost-benefit matrix CB, is defined as

$$\mathrm{CP}(\tau, D_{\mathrm{tr}}) = \mathrm{E}(\tau, D_{\mathrm{tr}}) \oplus \mathrm{CB}\,. \tag{3.28}$$

Measure of regret

Fernández-Loria and Provost (2022b) discuss the conceptual differences between causal effect estimation (i.e., uplift modeling) and causal classification (i.e., selecting persuadable customers). In particular, they express the objective of causal classification to be the minimization of the expected difference between the best potential outcome and

the potential outcome induced by the evaluated model. If we assume that all individuals with positive uplift should be targeted, then the ideal treatment can be expressed as $\boldsymbol{t}^* = \mathbb{I}[U(\boldsymbol{x}) > 0]$, and the corresponding potential outcome is noted $\boldsymbol{y}_{\boldsymbol{t}^*}$. Similarly, using a model $\mathscr{M}$ as a decision rule with a threshold of zero results in a treatment assignment $\boldsymbol{t}_{\mathscr{M}}(D_{\text{tr}}) = \mathbb{I}[\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) > 0]$, and the corresponding potential outcome is $\boldsymbol{y}_{\boldsymbol{t}_{\mathscr{M}}}(D_{\text{tr}})$. The evaluation measure proposed by Fernández-Loria and Provost (2022b) is expressed as

$$\text{Regret}(\mathscr{M}, D_{\text{tr}}) = \mathbb{E}[\boldsymbol{y}_{\boldsymbol{t}_{\mathscr{M}}}(D_{\text{tr}}) - \boldsymbol{y}_{\boldsymbol{t}^*}]. \tag{3.29}$$

One can see this approach as the converse of that adopted by Verbeke, Olaya, Guerry, et al. (2022) presented in the previous section: rather than comparing the factual outcome with the outcome resulting from a baseline scenario (such as taking no action), they compare the factual outcome with the best potential outcome that could be taken. However, the baseline profit and the profit generated from the best potential outcomes are independent of the prediction model $\mathscr{M}$, therefore their values are irrelevant to optimize $\mathscr{M}$.

Other measures

A. Li and Pearl (2019) define the benefit in terms of the counterfactual category of the customer: persuadable, sure thing, lost cause, or do-not-disturb (see Table 3.1). They allow for arbitrary costs for the four different counterfactual categories, and one can view the theoretical framework of Verbeke, Olaya, Guerry, et al. (2022) as a basis to determine the cost coefficients used by A. Li and Pearl (2019). Gubela, Lessmann, and Jaroszewicz (2020) provide another measure of the profit of a marketing campaign. The formula they propose is tailored to the specific aspects of customer retention (cost of contacting a customer, cost of the incentive, etc.), but does not consider benefits that can vary across individuals. Haupt and Lessmann (2022) provide a cost-sensitive measure of the profit generated by individuals in the context of customer targeting. They also discuss how to incorporate cost sensitivity into the uplift modeling framework. Lastly, Gubela and Lessmann (2021) propose value-driven evaluation metrics for marketing campaigns, taking into account the trade-off between maximizing uplift and maximizing revenues. Their metric aggregates the estimated uplift and the expected value of individuals into a score used to either rank individuals or evaluate a ranking model.

3.1.4 Predictive versus uplift modeling

Uplift modeling, despite its sound theoretical foundation, has faced challenges to show in practice a consistent advantage over classical predictive modeling. In this section, we present the body research that considers this issue.

Ascarza (2018), Devriendt, Berrevoets, and Verbeke (2021), and Wijaya et al. (2021) are the first papers to compare uplift modeling with the predictive approach. Their experimental results suggest that the uplift approach is superior for preventing churn, although their findings are only empirical, and based on a small number of datasets.

Fernández-Loria and Provost (2022a) argue that the true objective is to find *persuadable* individuals (as defined in Table 3.1), a task named causal classification, and that uplift modeling is only one of the possible ways to tackle causal classification. They derived an analytical criterion expressing when a model outperforms another in terms of classification error, which depends upon the bias and the variance of both models.

Their criterion has two major conceptual differences from the profit measure developed by Verbeke et al. (see Definition 3.5). First, it is conditional on a specific realization $\boldsymbol{x} = x$, while the profit measure takes into account the whole population. This is essential when the goal is to draw general conclusions about the population, rather than about specific individuals. Second, Provost's criterion is based on the probability that a model differs from the Bayes-optimal classifier. However, in practice, an imperfect model might have a large classification error (with respect to the Bayes-optimal classifier) but no loss in terms of profit, if the selected individuals generate a profit close to that of the individuals selected by the Bayes-optimal classifier.

Fernández-Loria and Provost (2022a,b) argue that the predictive approach can outperform the uplift approach under four different conditions:

1. When the positive outcomes are very rare;
2. When the outcomes are difficult to predict;
3. When the treatment effect ($S_0 - S_1$) is small;
4. Or when the treatment is correlated with the outcome.

Condition 1 translates the intuition that if S_1 is always close to zero, then the uplift, which is $S_0 - S_1$, will be correlated with the outcome S_0. The analysis by Fernández-Loria and Provost is based on the *monotonicity* assumption, i.e., that there is no negative individual causal effect (no *do-not-disturb* customers). Also, their analysis is based on the area under the ROC curve, which might not always represent the objective of a campaign.

Finally, Alaa and Schaar (2018) investigate the maximum performance of an uplift model using observational data, and provide guidelines to achieve the maximum performances. In the case of experimental data, which is the setting on which this thesis focuses, their results reduce to the conventional loss of a supervised learning algorithm.

3.2 Counterfactual identification

Counterfactual statements (or counterfactuals for short) concern the potential of events in situations different from the factual state of the world. An example of counterfactual statement is "I got no effect since I made no action, but something would have happened had I acted". Counterfactuals are used in many fields, ranging from algorithmic recourse (Karimi, Schölkopf, and Valera, 2021) to online advertisement and customer relationship management (A. Li and Pearl, 2019).

As an example, consider a company that plans to use direct marketing actions to prevent customers from churning. As we discussed in Section 3.1 (see Table 3.1), the behavior of customers in reaction to the two possible actions (contact or not) can be described in terms of counterfactual statements (Devriendt, Berrevoets, and Verbeke, 2021):

- *Sure thing*: customer not churning regardless of the action.
- *Persuadable*: customer churning only if not contacted.
- *Do-not-disturb*: customer churning only if contacted.
- *Lost cause*: customer churning regardless of the action.

Formula	Name
$P(\boldsymbol{y}_0 = 0 \mid \boldsymbol{t} = 1, \boldsymbol{y} = 1)$	Probability of necessity (PN)
$P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{t} = 0, \boldsymbol{y} = 0)$	Probability of sufficiency (PS)
$P(\boldsymbol{y}_0 = 0 \mid \boldsymbol{y} = 1)$	Probability of disablement (PD)
$P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{y} = 0)$	Probability of enablement (PE)
$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1)$	Probability of necessity and sufficiency (PNS), or probability of being a *do-not-disturb* customer
$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0)$	Probability of being a *persuadable* customer
$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0)$	Probability of being a *sure thing* customer
$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1)$	Probability of being a *lost cause* customer

Table 3.2 Various counterfactual probabilities defined by Pearl (2009, Sec. 9.2.1) and Verhelst, Mercier, et al. (2023b).

Although not observable, these quantities are relevant for adequate decision-making, and identifying counterfactuals can help reduce the uncertainty about the possible customer behaviors. Companies can also establish a profile of these four categories based on usage patterns and demographic information, a process called *customer segmentation* (Cooil, Aksoy, and Keiningham, 2008), which can reveal valuable insights and guide future business strategies.

Uplift modeling, as presented in Section 3.1, is another well-known approach for estimating causal effects. Counterfactuals and uplift are closely related, yet formally distinct notions. The counterfactual distribution describes the probability of each possible combination of realized and hypothetical outcomes, while the uplift describes the change in outcome probability due to treatment. Although the counterfactual distribution is more informative, it is also more difficult to estimate than the uplift. A. Li and Pearl (2019) mention that the similarity between these two notions can lead to confusion, especially since they become identical under the assumption of monotonicity (the absence of negative causal effects, discussed in Section 3.2.2).

The following sections are organized as follows. In Section 3.2.1, we provide the mathematical formulation of the counterfactual probabilities considered in this thesis and we define the concept of identifiability. We describe state-of-the-art results that provide exact inference for counterfactual probabilities in Section 3.2.2, and bounds on counterfactual probabilities in Section 3.2.3.

3.2.1 Problem formulation

A counterfactual expression is any expression that involves different potential outcomes (see Definition 2.20). Consequently, there is a large number of possible counterfactual expressions, even considering only two binary variables. Some of them, listed in Table 3.2, have been studied in the literature due to their relevance in causal decision-making (Heckman, 1991; Tian and Pearl, 2000). For example, the probability of sufficiency (PS) formalizes questions "*This candidate, who had a PhD, was hired. What is the probability that they would have been hired if they did not have a PhD?*". As another example, the probability of being a *do-not-disturb* customer is the same as the probability of necessity and sufficiency (PNS) described by Pearl (2009): the counterfactual expression "$\boldsymbol{y}_0 = 0$ and $\boldsymbol{y}_1 = 1$" indicates that the customer churns if and only if they are contacted. In logic, we say that the treatment $\boldsymbol{t} = 1$ is a necessary and sufficient

condition for the outcome $\boldsymbol{y} = 1$. In this thesis, we are particularly interested in the last four probabilities listed in Table 3.2. In customer management, they correspond to the customer categories we presented in Section 3.1. Because of their importance in this thesis, we define the following notation:

$$\alpha = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0) \qquad \alpha(x) = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x) \tag{3.30}$$

$$\beta = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0) \qquad \beta(x) = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x) \tag{3.31}$$

$$\gamma = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1) \qquad \gamma(x) = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x) \tag{3.32}$$

$$\delta = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1) \qquad \delta(x) = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x) \tag{3.33}$$

All the probabilities in Table 3.2 are related; for example, Pearl (2009, Lemma 9.2.6) shows that

$$\gamma = P(\boldsymbol{y} = 1, \boldsymbol{t} = 1)\text{PN} + P(\boldsymbol{y} = 0, \boldsymbol{t} = 0)\text{PS},$$

and, similarly, (Tian and Pearl, 2000, Thm. 2) show that, under the condition of strong exogeneity (Tian and Pearl, 2000, Def. 13), we have

$$\text{PN} = \frac{\gamma}{1 - S_0} \quad \text{and} \quad \text{PS} = \frac{\gamma}{S_1}.$$

Since the realizations of counterfactual statements cannot be directly observed, and we typically do not have a full knowledge of the causal model, the research focuses on methods to estimate their probabilities based on data and various assumptions. This task is called *counterfactual identification*. The data can be observational data, experimental data, or a mix of both. Identification procedures indicate when and how the probability of counterfactuals can be computed exactly (Correa, Sanghack Lee, and Bareinboim, 2021). This is called the *fully identifiable* setting. In situations where the exact probability of counterfactuals cannot be computed, an alternative consists in bounding this quantity. This approach, called *partial counterfactual identification*, was first developed by Tian and Pearl (2000), and more recently by Mueller, A. Li, and Pearl (2021) and J. Zhang, Tian, and Bareinboim (2022).

3.2.2 Estimation in fully identifiable settings

Under specific conditions, some counterfactual probabilities can be computed exactly from data. See Appendix B for an example of counterfactual computation with a very a simple causal model. In this section, we discuss two important results from the literature pertaining to more realistic settings.

The assumption of *monotonicity* states that treatment $\boldsymbol{t}$ does not have a negative effect on the outcome $\boldsymbol{y}$. For example, we might assume that sending a marketing email to potential buyers of a product cannot reduce the probability that the customers buy the product. With a binary outcome and treatment, this corresponds to the probability of being a *do-not-disturb* customer equal to zero:

$$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x) = 0 \quad \text{for all } x \in \mathcal{X}. \tag{3.34}$$

Note that, in the case where we seek to maximize the probability of the outcome $\boldsymbol{y} = 1$ (for example, in online retail), the assumption of monotonicity would be $P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x) = 0$ for all $x \in \mathcal{X}$. From Eq. (3.34), the joint probability distribution of $\boldsymbol{y}_0, \boldsymbol{y}_1$ given $\boldsymbol{x} = x$ can be fully recovered from the conditional marginal distributions:

$$\alpha(x) = 1 - S_0(x) \qquad \beta(x) = S_0(x) - S_1(x)$$

$$\gamma(x) = 0 \qquad \delta(x) = S_1(x)$$

where $S_t(x) = P(\boldsymbol{y}_t = 1 \mid \boldsymbol{x} = x)$, as defined in Eq. (3.1). We can see that the uplift $U(x) = S_0(x) - S_1(x)$ is now equal to the probability of being a *persuadable* customer. Therefore, uplift modeling can be used to estimate counterfactual probabilities under the assumption of monotonicity.

Correa, Sanghack Lee, and Bareinboim (2021) derived a systematic way to determine whether and how counterfactual probabilities can be computed from a set of observational and experimental data. They assume that, of a causal model $M = (P, \boldsymbol{U}, \boldsymbol{V}, G, F)$, we only have knowledge of the graph G, and we have a collection of datasets $\boldsymbol{D}_1, \dots, \boldsymbol{D}_n$ sampled from $\boldsymbol{V}$ distributed under diverse interventions $\text{do}(\boldsymbol{W}_1 = w_1), \dots, \text{do}(\boldsymbol{W}_n = w_n)$, with $\boldsymbol{W}_i \subseteq \boldsymbol{V}$. This can include the null intervention $\boldsymbol{W}_i = \emptyset$, which corresponds to observational data. Given a counterfactual probability of interest, the general idea is to formulate a set of conditions that indicate whether the probability can be computed, based on the graph G, the potential outcomes in the counterfactual expression, and the available data $\boldsymbol{D}_1, \dots, \boldsymbol{D}_n$. If the probability can be estimated, another procedure indicates how the counterfactual probability can be transformed into another expression involving only probabilities that can be estimated from the available data. It is worth stressing that this procedure is *complete*, that is, if the algorithm fails to provide a solution, then the counterfactual cannot be identified under this set of assumptions. As a special case, they provide a separate algorithm for conditional counterfactual probabilities, that is, probabilities of the form $P(\boldsymbol{Y}_* = y_* \mid \boldsymbol{Z}_* = z_*)$, where $\boldsymbol{Y}_*$ and $\boldsymbol{Z}_*$ are arbitrary conjunctions of counterfactuals.

The result by Correa, Sanghack Lee, and Bareinboim (2021) applies to an arbitrary graph G, an arbitrary collection of observational and experimental data, and any counterfactual probability to be estimated. As such, this generalizes a wide range of results in the causal inference literature, notably on inferring causal effects from observational data (Correa and Bareinboim, 2020; Pearl and Robins, 1995; Tian and Pearl, 2002). However, this requires full knowledge of the causal graph G, which is not always the case in practice. Note the conceptual difference with the assumption of monotonicity, which does not require knowledge of the graph G, but assumes that the function determining the value of $\boldsymbol{y}$ is monotonic in $\boldsymbol{t}$. Although we can consider that Correa, Sanghack Lee, and Bareinboim (2021) solved the problem of full counterfactual identification from assumptions on G, there is a much wider set of possible assumptions on the set of functions F that could lead to new identification results.

3.2.3 Estimation in partially identifiable settings

Although counterfactual probabilities cannot always be fully identified, that is, they cannot be computed exactly, they can always be bounded. This task is called *partial identification*. By definition, any probability is bounded by zero and one, but with assumptions on G, F, or the available data, tighter bounds can be derived.

Bounds on the probability of counterfactuals have first been derived in Tian and Pearl (2000). They consider various assumptions to derive bounds on the probability of necessity (PN), the probability of sufficiency (PS), and the probability of necessity and sufficiency (PNS), denoted γ in Eq. (3.32). In particular, they showed that γ can be bounded as

$$\max\{0, S_1 - S_0\} \leq \gamma \leq \min\{1 - S_0, S_1\}. \tag{3.35}$$

The bounds are derived from the classical Fréchet bounds (Fréchet, 1935) stating that

for any events A and B,

$$\max\{0, P(A) + P(B) - 1\} \leq P(A, B) \leq \min\{P(A), P(B)\}. \tag{3.36}$$

By replacing A with $\boldsymbol{y}_0 = 0$ and B with $\boldsymbol{y}_1 = 1$, we find Eq. (3.35).

Mueller, A. Li, and Pearl (2021) derived tighter bounds on γ for a variety of causal diagrams, such as in the presence of a mediator variable (a variable on the causal path from the treatment to the outcome). By assuming that the covariates $\boldsymbol{x}$ have a discrete probability distribution, they also show in their Theorem 5 that, if $\boldsymbol{x}$ satisfies the back-door criterion (Pearl, 2009, Def. 3.3.1), we have

$$\sum_{x \in \mathcal{X}} P(\boldsymbol{x} = x) \max\{0, S_1(x) - S_0(x)\} \leq \gamma \leq \sum_{x \in \mathcal{X}} P(\boldsymbol{x} = x) \min\{1 - S_0(x), S_1(x)\}.$$

In Chapter 6, we will derive similar bounds, but without restrictions on the distribution of the covariates.

A. Li and Pearl (2019) bound the quantity $\gamma(x)$ (see Eq. 3.32), which they call the x-specific PNS, without assuming unconfoundedness, by

$$\max \begin{Bmatrix} 0 \\ S_1(x) - S_0(x) \\ P(\boldsymbol{y}{=}1|x) - S_0(x) \\ S_1(x) - P(\boldsymbol{y}{=}1|x) \end{Bmatrix} \leq \gamma(x) \leq \min \begin{Bmatrix} 1 - S_0(x) \\ S_1(x) \\ P(\boldsymbol{y}{=}1, \boldsymbol{t}{=}1|x) + P(\boldsymbol{y}{=}0, \boldsymbol{t}{=}0|x) \\ S_1(x) - S_0(x) + P(\boldsymbol{y}{=}1, \boldsymbol{t}{=}0|x) + P(\boldsymbol{y}{=}0, \boldsymbol{t}{=}1|x) \end{Bmatrix}.$$

These bounds reduce to the Fréchet bounds in Eq. (3.35) when $S_0(x) = P(\boldsymbol{y} = 1 \mid \boldsymbol{t} = t, \boldsymbol{x} = x)$, which is a consequence of the assumption of unconfoundedness. Their main focus is the estimation of the profits generated by individuals with specific characteristics $\boldsymbol{x} = x$, assuming arbitrary gains for the four counterfactual categories. For example, keeping a persuadable customer (a customer churns only when not targeted) might be more beneficial than keeping a sure thing customer, besides the cost of the targeted action. Assuming that the counterfactual outcomes generate a gain $\boldsymbol{B}$, noted

$$\boldsymbol{B} = \begin{cases} a & \text{if } \boldsymbol{y}_0 = 0 \text{ and } \boldsymbol{y}_1 = 0 \\ b & \text{if } \boldsymbol{y}_0 = 1 \text{ and } \boldsymbol{y}_1 = 0 \\ c & \text{if } \boldsymbol{y}_0 = 0 \text{ and } \boldsymbol{y}_1 = 1 \\ d & \text{if } \boldsymbol{y}_0 = 1 \text{ and } \boldsymbol{y}_1 = 1, \end{cases}$$

they provide bounds on the expected gain, which is defined as

$$\mathbb{E}[\boldsymbol{B} \mid \boldsymbol{x} = x] = a\alpha(x) + b\beta(x) + c\gamma(x) + d\delta(x).$$

Furthermore, they show that the bounds on $\boldsymbol{B}$ reduce to a single value under the assumption of monotonicity (discussed in the previous section), or *gain equality*, defined as the setting where $b + c = a + d$. A. Li and Pearl (2022) further refine the bounds on the campaign benefit by using additional covariates and assumptions on the causal graph G, such as the mediator setting discussed by Mueller, A. Li, and Pearl (2021).

Finally, J. Zhang, Tian, and Bareinboim (2022) express partial counterfactual identification as a polynomial programming problem, providing tight bounds for any causal graph G and any combination of experimental and observational data, assuming that we have full knowledge of G. This can be seen as the equivalent for partial counterfactual identification of the procedure for full counterfactual identification developed by Correa, Sanghack Lee, and Bareinboim (2021).

Most of these articles provide bounds on counterfactual probabilities based on structural assumptions on G, however, besides those discussing the monotonic setting, none make any assumption on the functional dependencies (the set of functions F in the causal model). In Chapter 6, we explore the synergies between uplift modeling and counterfactual identification, resulting in tighter bounds without knowledge of G. We also provide different point estimates of counterfactual probabilities based on different assumptions about the functional dependencies between $\boldsymbol{y}, \boldsymbol{t}$ and $\boldsymbol{x}$.

Part II

Contributions

If we knew what we were doing,
we wouldn't call it research.

Albert Einstein, apocryphal

4

Experimental comparison of predictive and uplift modeling

Some of the results presented in this chapter have been published in the following articles:

- Théo Verhelst, Jeevan Shrestha, et al. (2021). "Predicting reach to find persuadable customers: Improving uplift models for churn prevention". In: *Discovery science*. Ed. by Carlos Soares and Luis Torgo. Cham: Springer International Publishing, pp. 44–54. ISBN: 978-3-030-88942-5

- Théo Verhelst, Denis Mercier, et al. (2023a). "A churn prediction dataset from the telecom sector: a new benchmark for uplift modeling". In: *ECML PKDD 2023 Workshops - Workshop on Uplift Modeling and Causal Machine Learning for Operational Decision Making*

In this chapter, we compare the performance of the conventional machine learning approach (described in Section 2.3) and that of uplift modeling (described in Section 3.1) for the problem of customer churn mitigation. Emphasis is placed on empirical results derived from benchmarks and real-world experiments, prioritizing practical insights over theoretical arguments. We examine this question from a theoretical point of view in Chapter 5.

The added value of uplift modeling over churn prediction has rarely been assessed empirically. Wijaya et al. (2021) evaluated this question with a focus on employee turnover, while the studies by Ascarza (2018) and Devriendt, Berrevoets, and Verbeke (2021) focus on customer retention. These three studies advocate the use of uplift modeling, but have evaluated a very limited number of uplift models. While it is clear that uplift is less biased than churn prediction to estimate causal effects, the performance gain is debated and context-dependent (Fernández-Loria and Provost, 2022a). In settings such as customer retention, characterized by non-linearity, low class separability, and high dimensionality, the theoretical advantages of uplift might be insufficient to outweigh its drawbacks with respect to the usual strategy of churn prediction.

The uplift literature faces another pressing issue: the low number of publicly available datasets designed specifically for uplift modeling. A recent uplift benchmark conducted by Rößler and Schoder (2022) listed only four public uplift datasets: Criteo

(Diemert Eustache, Renaudin, and Massih-Reza, 2018), Hillstrom (Hillstrom, 2008), Starbucks[1] and Lenta[2]. Furthermore, despite the fact that customer churn is often cited as a common application for uplift modeling, none of these public datasets is concerned with churn. This is a critical issue that hinders the reproducibility of studies on churn prevention with uplift modeling.

In this chapter, we conduct several experiments to compare the performance of uplift modeling with the predictive approach. Specifically, we assess the performance of several uplift and churn models in a benchmark on two real-world churn datasets from Orange Belgium, as well as on two publicly available uplift datasets. Then, we compare the best strategies according to our empirical results in a series of real customer retention campaigns. This experiment represents a unique contribution to the literature, where new models are always evaluated on historical data. Finally, we develop and assess a series of strategies to take advantage of information about which customers were reached during previous churn campaigns. To address the lack of publicly available churn datasets for uplift modeling, this chapter makes publicly available one of the two datasets from Orange Belgium. This dataset offers researchers and practitioners a new resource to evaluate strategies aimed at reducing churn and increasing customer retention within the telecommunications industry.

The contributions of this chapter can be summarized as follows:

- The publication of the first public churn dataset with anonymized customer data from Orange Belgium, allowing the research community to evaluate new uplift strategies on challenging and realistic data (Section 4.1).
- A benchmark of various uplift models on two churn datasets and two other publicly available datasets (Section 4.2), indicating that the classical predictive approach is competitive, if not more effective than uplift modeling.
- A new measure of the variability of a ranking, which plays an important role in the performance of a model (Section 4.2.2).
- The comparison of uplift and predictive modeling in a series of real customer retention campaigns (Section 4.3), confirming our results on historical data. This is the first time in the uplift literature that the performances of predictive and uplift modeling are compared in a live setting.
- The development of several strategies to integrate reach information into uplift modeling (Section 4.4), which significantly improve the performance of uplift modeling.

The remainder of this chapter is organized as follows. In Section 4.1, we present the data used in these experiments and the procedure used to make one of the datasets publicly available online. The experiments, divided in three stages, are described in Sections 4.2 to 4.4. A benchmark of various uplift models is performed in Section 4.2. These results are confirmed in real customer retention campaigns in Section 4.3. In Section 4.4, we explore how to use reach information to improve uplift models. Each of these three sections is divided into a description of the experimental setup and a

[1] https://github.com/joshxinjie/Data_Scientist_Nanodegree/tree/master/starbucks_portfolio_exercise, last accessed 2023-12-12.

[2] https://www.uplift-modeling.com/en/latest/api/datasets/fetch_lenta.html, last accessed 2023-12-12.

presentation of the results. Finally, we discuss our results and present our conclusions in Section 4.5.

4.1 Churn datasets

The two churn datasets used in this benchmark come from a series of marketing campaigns carried out by Orange Belgium during 2019 and 2020. The Churn 1 dataset contains six campaigns during 2019 and 2020, while the Churn 2 dataset contains three campaigns during 2020. The different number of campaigns in the two datasets results from specific circumstances at the time of creation of the datasets in terms of access rights to the databases and other technical considerations. In both datasets, the data from the individual campaigns are aggregated. We chose to perform our analyses on the aggregated data rather than on each campaign individually to obtain a higher statistical confidence in our results with a greater number of samples. This might blur some specific dynamics and effects that occur in some campaigns (e.g., if a competitor launched a large marketing campaign at the same time). However, our objective is not to understand the dynamics and the performance of each campaign, rather it is to obtain insights on the performance of different modeling approaches as a function of the characteristics of the data. We refer the reader to our previous work (Verhelst, 2018; Verhelst, Caelen, et al., 2020) for a more detailed analysis of churn dynamics in specific campaigns.

As described in Section 1.3, during each campaign, the probability of churn of each customer is estimated using a predictive model and the most risky customers are selected. A subset of these high-risk customers is randomly assigned to the control group, while the remaining customers form the target group. The list of customers in the target group is then shared with a call center tasked with contacting each customer and presenting them with a marketing offer, or recommending a new tariff plan based on their individual history. Customer churn is determined in a two-month window following the campaign, and any subsequent churn is not attributed to this specific campaign. The data from this campaign and the churn outcome are then recorded in the historical database, and the same campaign process is repeated the next month.

We also assessed datasets from other campaigns, such as the *add card* campaign where customers are contacted to suggest them to buy a new SIM card, for example for another member of their family. Another campaign we considered focused on customers who have poor network coverage at home, which requires a dedicated strategy to mitigate churn. Preliminary results suggested that the uplift approach does not perform significantly better on these datasets than on the Churn 1 and Churn 2 datasets. Therefore, we did not include these preliminary results in this chapter to simplify the presentation. See Massimetti (2021) for an assessment of uplift modeling on the churn and add card datasets.

4.1.1 Description

The characteristics of the two churn datasets from Orange Beligum, as well as the Criteo dataset (Diemert Eustache, Renaudin, and Massih-Reza, 2018), and the Hillstrom dataset (Hillstrom, 2008), are summarized in Table 4.1. In the Criteo dataset, customers in the target group are exposed to an online ad, whereas the Hillstrom dataset represents an email marketing campaign. In both cases, the reaction of the customer is recorded in terms of visiting the advertiser's website and possibly buying a product.

Table 4.1 Description of the churn dataset and two other uplift datasets. The Hillstrom dataset is more balanced than the other datasets, and the two Churn datasets have many more features. Criteo is characterized by a large number of samples.

Name	Features	Samples	Control response rate (%)	Target response rate (%)	Treatment rate (%)
Churn 1	145	11,268	4.8	4.0	66.3
Churn 2	178	11,896	3.6	3.4	75.7
Hillstrom	15	42,693	10.6	15.1	66.7
Criteo	12	25,309,483	4.2	4.9	84.6

We report the number of features, the number of samples, the response rate in the control and target groups (S_0 and S_1), and the treatment rate, $P(\boldsymbol{t} = 1)$.

The Churn 1 and Churn 2 datasets consists of 11,268 and 11,896 records, a relatively small number compared to other publicly available uplift datasets. However, they have a larger number of features, 145 and 178. These features encompass a diverse range of customer attributes:

- subscription details (e.g., type of tariff plan, number of products): 45 features in Churn 1, 42 in Churn 2.
- Usage metadata (e.g., number of calls, data consumption): 30 features in Churn 1, 37 in Churn 2.
- Revenue (e.g., total revenue, revenue due to data usage): 33 in Churn 1, 34 in Churn 2.
- Hardware (e.g., phone type): 4 in both Churn 1 and Churn 2.
- Sociodemographics (e.g., age, province): 19 in Churn 1, 48 in Churn 2.
- Service quality (e.g., number of calls to customer service): 14 in both Churn 1 and Churn 2.

To illustrate these features, we report in Fig. 4.1 the distribution of the *out-of-bundle* amount in the Churn 2 dataset, which is the additional fee paid by the customer for services not included in the provision of the tariff plan. The distribution of this feature is a strong indicator of *bill shock*, that is, the reaction of a customer when faced with a high bill, as discussed in Section 1.3. The distribution of the out-of-bundle amount is shown on a logarithmic scale, separately for churners and non-churners. The probability density function depicted in this figure is estimated using a Gaussian kernel. Consequently, the resulting distribution exhibits characteristics resembling a mixture of Gaussians, even though the underlying probability density is not necessarily Gaussian. A disparity appears between the distributions of churners and non-churners, with churners frequently having a larger out-of-bundle amount. This observation is frequently used in the formulation of expert rules during churn retention campaigns. Specifically, when a customer is deemed likely to churn and demonstrates a substantial out-of-bundle, it serves as a trigger to propose a tariff plan better suited to their usage profile. We refer the reader to (Verhelst, 2018) or a more comprehensive description of the various customer features used by Orange Belgium.

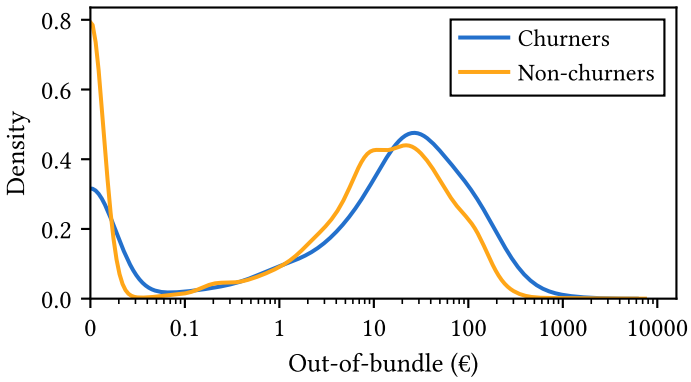


Figure 4.1 Probability density function of the *out-of-bundle* amount (the supplement paid by the customer for services not included in their standard allowances), on a logarithmic scale. We see that churners have a higher *out-of-bundle* than non-churners.

4.1.2 Data preparation

Several steps are performed to prepare the dataset for the experiments.

Categorical features Some features, such as the tariff plan or the province of residence, are categorical. Some of these features can have a large number of different values, but this can have a negative impact on the performance of some learning algorithms, such as random forests. To address this issue, we replace the less common values with a placeholder value "Other".

One-hot encoding The categorical features are then replaced by a binary vector containing a 1 at the position corresponding to the value of the feature. In principle, this should not be necessary for random forests (the machine learning model used in our experiments), since they can handle discrete values; however, we use the `sklearn` Python package for machine learning, which requires all features to be numerical.

Normalization Since some features have very different scales (e.g., the total amount paid by a customer and the number of contracts of a customer), the features are linearly normalized to have a minimum value of 0 and a maximum value of 1.

4.1.3 Mutual information between features and potential outcomes

One distinctive aspect of churn datasets is the inherent difficulty of accurately predicting the churn outcome. The complex dynamics of churn in the telecom sector make prediction a challenging task, requiring advanced modeling techniques to capture the underlying patterns and factors influencing customer behavior. Uplift modeling is even more difficult than predicting churn due to the small effect of retention campaigns. To quantify this aspect, we estimate the mutual information $I(\boldsymbol{x}; \boldsymbol{y}_t)$ (for $t = 0, 1$), which measures the difficulty in predicting the binary outcome $\boldsymbol{y}_t$ from the vector of features $\boldsymbol{x}$ (T. M. Cover and Thomas, 1991). It is computed using the identity in Eq. (2.21):

$$I(\boldsymbol{x}; \boldsymbol{y}_t) = H(\boldsymbol{y}_t) - H(\boldsymbol{y}_t \mid \boldsymbol{x}) = H(\boldsymbol{y}_t) - \int_{\mathcal{X}} H(\boldsymbol{y}_t \mid \boldsymbol{x} = x)\, \mathrm{d}x \tag{4.1}$$

Table 4.2 Estimates of the mutual information between the features and the outcomes.

	$H(\boldsymbol{y}_0)$	$H(\boldsymbol{y}_1)$	$\hat{I}(\boldsymbol{x};\boldsymbol{y}_0)$	$\hat{I}(\boldsymbol{x};\boldsymbol{y}_1)$	$\frac{\hat{I}(\boldsymbol{x};\boldsymbol{y}_0)}{H(\boldsymbol{y}_0)}$	$\frac{\hat{I}(\boldsymbol{x};\boldsymbol{y}_1)}{H(\boldsymbol{y}_1)}$
Churn 1	0.19	0.17	0.1871	0.1584	3.54%	6.18%
Churn 2	0.16	0.15	0.0008	0.0025	0.54%	1.71%
Hillstrom	0.34	0.43	0.0112	0.0123	3.32%	2.90%
Criteo	0.16	0.20	0.0429	0.0573	24.63%	29.32%

where the term $H(\boldsymbol{y}_t)$ is estimated from the prior distribution of $\boldsymbol{y}_t$ using Eq. (2.10). Noting the dataset as $D = \{x^{(i)}, y^{(i)}, t^{(i)}\}_{i=1}^{N}$, the integral above can be estimated by its plug-in estimator, as a sum over the samples in the dataset:

$$\int_{\mathcal{X}} H(\boldsymbol{y}_t \mid \boldsymbol{x} = x)\,\mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N} H\left(\boldsymbol{y}_t \mid \boldsymbol{x} = x^{(i)}\right). \tag{4.2}$$

Individual terms $H\left(\boldsymbol{y}_t \mid \boldsymbol{x} = x^{(i)}\right)$ are estimated using the probabilities predicted by a T-learner uplift model (see Section 3.1.2) as

$$H(\boldsymbol{y}_t \mid x) = P(\boldsymbol{y}_t = 0 \mid x)\log P(\boldsymbol{y}_t = 0 \mid x) + P(\boldsymbol{y}_t = 1 \mid x)\log P(\boldsymbol{y}_t = 1 \mid x). \tag{4.3}$$

Details of the experimental setup used to train the uplift model are presented in Section 4.2. The estimates of the mutual information are given in Table 4.2. In the last two columns, the mutual information estimate is divided by the entropy of the prior distribution, indicating the proportion of uncertainty of the outcome explained by the features. The Churn 1 and Churn 2 datasets have a low outcome probability similar to that of the Criteo dataset, while also having a very low mutual information, like the Hillstrom dataset.

4.1.4 Randomization

Since the churn datasets come from randomized campaigns, the treatment should be independent of the outcomes. To validate this independence, we performed the Classifier 2 Sample Test used in (Diemert Eustache, Renaudin, and Massih-Reza, 2018) on the Churn 2 dataset. We trained a random forest classifier (Breiman, 2001) to predict the treatment indicator and compared its Hamming loss with the loss distribution obtained under the null hypothesis (i.e., assuming the treatment is indeed randomized), which is sampled by training models to predict random splits. The treatment predictor has a loss of 23.82% (close to the proportion of control samples, 24.26%), which corresponds to a p-value of 0.26 under the null hypothesis. This result is shown in Fig. 4.2. This indicates that the treatment cannot be predicted based on available features, hence the randomization of treatment assignment can be considered appropriate and unbiased.

4.1.5 Online availability

Despite the fact that uplift modeling is often illustrated with the problem of customer churn, none of the public uplift datasets contain churn data. During this research project, we had the opportunity to publish the Churn 2 dataset online. This is the first public uplift dataset concerning customer churn, allowing other researchers to develop and assess uplift models and causal inference methods on realistic and challenging data.

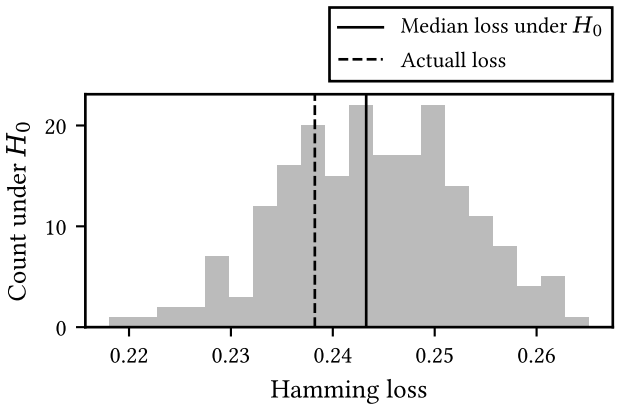


Figure 4.2 Prediction loss on the treatment indicator for the Churn 2 dataset. The histogram represents the distribution of the loss with a randomized treatment. If the treatment was not randomized, we would expect the actual loss to be far in the left tail of the loss distribution under H_0. The actual loss when predicting the treatment is 23.82%, which corresponds to a p-value of 0.26, indicating that the treatment is correctly randomized.

To ensure the privacy of customers and the confidentiality of the features used by Orange Belgium, the data is anonymized by using a Principal Component Analysis (PCA) projection of the numerical features. This allows effective for analyses and modeling while protecting sensitive information. Adopting this strategy has proven to be effective in preserving predictive accuracy while safeguarding privacy in the domain of fraud detection (Dal Pozzolo, Caelen, Johnson, et al., 2015). There are many other anonymization approaches (Majeed and Sungchang Lee, 2020), however, the PCA projection has the advantage of both ensuring the privacy of the customers and to guarantee the confidentiality of the features. All categorical features and their levels are anonymized by giving them generic names. The dataset is available on the OpenML website[3], or by running the following Python code, after installing the package `openml`:

```
from openml.datasets import get_dataset
dataset = get_dataset("churn-uplift-orange")
```

4.2 Benchmark of uplift models

In this section, we compare the performance of several uplift models against a conventional churn prediction model. This benchmark constitutes a stepping stone in this thesis to understand the performance of uplift modeling to prevent customer churn. Several studies have assessed the performance of different uplift models (Devriendt, Moldovan, and Verbeke, 2018; Kayaalp, 2017; Rößler and Schoder, 2022), however, to the best of our knowledge, only three studies have empirically compared predictive and uplift approaches. The study by Wijaya et al. (2021) is focused on employee turnover, while the studies by Ascarza (2018) and Devriendt, Berrevoets, and Verbeke (2021) focus on customer retention. These three studies advocate the use of uplift modeling, but have evaluated a very limited number of uplift models. We aim to bridge this gap in the

[3] https://www.openml.org/search?type=data&id=45580, last accessed 2023-12-12.

literature by comparing a larger number of uplift models with the predictive approach on various datasets. Note that we do not seek to achieve the highest performance with uplift modeling on the datasets at hand. Rather, we aim to compare the various approaches on an equal footing. Therefore, the hyperparameters are chosen to achieve realistic performances, but were not subject to extensive optimization.

4.2.1 Experimental setup

We benchmark the following uplift models, which have been described in Section 3.1.2:

- `Outcome-RF`: A random forest (Breiman, 2001) trained to predict the probability of churn $P(\boldsymbol{y} = 1 \mid x)$. This represents the classical predictive approach.
- `S-learner`: The S-learner uplift model using a random forest as base learner.
- `T-learner`: The T-learner uplift model using random forests as base learners.
- `X-learner`: The X-learner uplift model by Künzel et al. (2019) using random forests as base learners.
- `Z-target`: The modified target uplift model estimating the probability distribution of $\boldsymbol{z} = \boldsymbol{y}(1 - \boldsymbol{t}) + (1 - \boldsymbol{y})\boldsymbol{t}$, proposed by Jaskowski and Jaroszewicz (2012), using a random forest as base learner.
- `Uplift-RF`: The uplift random forest model proposed by Guelman, Guillén, and Pérez-Marín (2015).

The choice of the random forest as the base learner is based on previous work on churn prediction and uplift modeling (Massimetti, 2021; Verhelst, Caelen, et al., 2020) and related work in fraud detection Dal Pozzolo and Bontempi (2015), which shares many similarities, such as class imbalance and low separability. In these previous experiments, random forests were consistently the best performing models, sometimes equaled by boosting models. This is also consistent with benchmarks performed by data scientists at Orange Belgium. Given this previous experience, we chose random forests in all our experiments to have a consistent and realistic basis to compare the performance of different approaches. We did not perform extensive hyperparameter optimization; instead, we observed that the performance of random forests in churn prediction plateaued when using more than 100 trees. Limiting the depth of trees and setting a minimum number of samples per leaf was also found to be important to achieve good performances. Therefore, all models use 100 trees, a maximum depth of 20, and a minimum of 10 samples per leaf.

Given the high class imbalance of the datasets, we rely on the EasyEnsemble strategy (X.-Y. Liu, Wu, and Zhou, 2009) for class balancing. Ensemble techniques for class balancing are known to perform well for churn prediction (Zhu, Baesens, and Broucke, 2017). As described in Section 2.3.6, EasyEnsemble consists in training k base learners on the whole set of positive instances (churners) and an equally sized random set of negative instances. The predictions of all the base learners are averaged to obtain the final prediction. We set the number of base learners for EasyEnsemble at $k = 10$. We did not observe significant differences in performance between the different split criteria for `Uplift RF`. Therefore, we chose the Euclidean distance, given its empirical superiority according to Rzepakowski and Jaroszewicz (2012).

We used the k-fold cross-validation procedure with $k = 3$ folds. In each iteration of the k-fold cross-validation, $k - 1$ folds are used as a training set, and the remaining fold is used as a test set. We estimate the performance of the different models using the uplift curve (defined in Definition 3.3) evaluated on the held-out folds. To obtain a single quantitative measure of performance, we compute the area under the uplift curve (AUUC). The AUUC suffers from a large variance due to the low number of churners; therefore, we repeat the whole experiment 10 times to reduce the impact of random sampling. An alternative would be to increase the value of k (for example, to $k = 30$); however, this approach would result in very small validation sets, some of which may not even contain churners. This would further increase the variance of the resulting uplift curve. On the other hand, repeating the k-fold procedure maintains a relatively large size for the validation sets, while decreasing the variance of the AUUC.

The literature suggests that the variance of the estimator plays an important role in determining whether the uplift approach outperforms predictive modeling (Fernández-Loria and Provost, 2022a). To evaluate this aspect, we estimate the variance of the models using the procedure presented by Webb (2000). Given that we repeat the k-fold cross-validation procedure 10 times, we have 10 different predictions of the score for each sample, coming from model instances trained on different training sets. For each sample, we compute the variance of the 10 predictions, and report the average variance across all data samples for each model.

4.2.2 Ranking variance

The performance of a model in terms of the uplift curve is determined by the ranking induced by the scores predicted by the model. This ranking can vary even with a low estimator variance, if a similar score is assigned to all samples. For example, `S-learner` provides scores close to zero on datasets where the treatment has a low causal effect, because the model ignores the effect of the treatment on the outcome given the other more informative features.

There is no standard and well-established way to estimate the stability of a ranking. Gao et al. (2010) suggest an approach that requires repeating the training process as many times as there are samples in the training set, which is prohibitively long for medium to large datasets. Perini, Galvin, and Vercruyssen (2020) suggest a three steps process:

1. Train ℓ models from ℓ different subsets of the training set.
2. Compute, for each test sample, the variance of the rank of this sample induced by the ℓ models.
3. Average the rank variance of all test samples to obtain a unique measure.

We propose a new measure of ranking variance that is less computationally intensive than that proposed by Gao et al. (2010). Our measure is based on the notion of KL-divergence (presented in Definition 2.8), which we expect to be more amenable to theoretical analysis than the notion of variance of the rank of a sample used by Perini, Galvin, and Vercruyssen (2020). However, we have not theoretically analyzed the validity of our measure. Given a model $\mathscr{M}$ trained on a random dataset $\boldsymbol{D}_{\text{tr}}$ and a sample $x^{(i)}$ from a test set $D_{\text{te}} = \{(x^{(i)}, y^{(i)}, t^{(i)})\}_{i=1}^{N}$, we assume that the scores are independent and normally distributed:

$$\mathscr{M}(x^{(i)}, \boldsymbol{D}_{\text{tr}}) \sim \mathcal{N}(\mu_i, \sigma_i^2). \tag{4.4}$$

We quantify the robustness of the ranking between two samples $x^{(i)}$ and $x^{(j)}$ by using the Kullback–Leibler (KL) divergence between the two score distributions (see Definition 2.8). A higher KL divergence indicates that the score distributions are further apart, hence it is less likely that the relative position of these two samples in the ranking changes. In the case of normal distributions (Belov and Armstrong, 2011), the KL divergence reduces to

$$D(\mathscr{M}(x^{(i)}, \boldsymbol{D}_{\text{tr}}) \parallel \mathscr{M}(x^{(j)}, \boldsymbol{D}_{\text{tr}})) = \log \frac{\sigma_j}{\sigma_j}\sigma_i + \frac{\sigma_i^2 + (\mu_i - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2}. \tag{4.5}$$

The final measure is the inverse of the average KL divergence between all pairs of samples in the test set D_{te}:

$$\text{RankVar}(\mathscr{M}, D_{\text{te}}) = \left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} D(\mathscr{M}(x^{(i)}, \boldsymbol{D}_{\text{tr}}) \parallel \mathscr{M}(x^{(j)}, \boldsymbol{D}_{\text{tr}})) \right)^{-1}. \tag{4.6}$$

The average KL divergence (the quantity between parentheses) is higher when the ranking is more stable. Therefore, we take the inverse of this quantity to express the variability of the ranking.

4.2.3 Results

The results of the benchmark are presented in Tables 4.3 to 4.5. As shown in Table 4.3, our observations reveal that `Outcome-RF` is superior to other approaches in terms of the area under the uplift curve (AUUC) across most datasets, particularly excelling on the Churn 1 and Criteo datasets. The second position is claimed by `X-learner` on the Churn 1 dataset, and by `T-learner` on the Criteo dataset. On the Churn 2 dataset, most models, except `S-learner`, display similar performances. Note these results exhibit a high degree of uncertainty, as indicated by the large standard deviations.

The Hillstrom dataset displays results quite different from those of the other data sets. The uplift models far outperform `Outcome-RF`, with `S-learner` and `T-learner` achieving the highest AUUC. We also observe that all models have a higher AUUC on the Hillstrom dataset than on the other datasets. We attribute this difference to the fact that the outcome is balanced in the Hillstrom dataset, whereas it is unbalanced and more difficult to predict in the Churn 2 dataset, as reported in Table 4.2.

In order to assess the impact on performances of the PCA projection, we conducted an identical experiment on the original, non-anonymized Churn 2 dataset. It appears that the performance is lower on the anonymized dataset than on the original, most notably for `X-learner`. However, the other models do not suffer as much from the anonymization procedure, and the performance of `Z-target` is not impacted at all.

From Table 4.4 we see that `S-learner` has the lowest estimator variance, by several orders of magnitude, except on the Hillstrom dataset. As discussed in Section 4.2.1, this can be explained by the low importance given by `S-learner` to the treatment indicator compared to other features, leading it to predict almost the same value in the treatment ($\boldsymbol{t} = 1$) and control ($\boldsymbol{t} = 0$) scenarios, resulting in a predicted uplift close to zero. `Outcome-RF` has the lowest estimator variance after `S-learner`. This is a possible reason for its superiority to the uplift models, as already suggested in the literature (Fernández-Loria and Provost, 2022a,b). This is further demonstrated by the ranking variance, reported in Table 4.5. `Outcome-RF` has the lowest ranking variance across all datasets, except for the Churn 2 dataset, where it is second after `Z-target`.

Table 4.3 Mean and standard deviation of the AUUC in percentage points. The highest AUUC is underlined for each dataset.

Model	Churn 1	Churn 2	Churn 2 (original)	Hillstrom	Criteo
`Outcome-RF`	0.91 ± 0.45	0.26 ± 0.47	0.33 ± 0.37	2.20 ± 0.32	1.01 ± 0.25
`S-learner`	0.48 ± 0.37	0.12 ± 0.31	0.19 ± 0.27	2.74 ± 0.32	0.69 ± 0.21
`T-learner`	0.25 ± 0.18	0.25 ± 0.38	0.33 ± 0.39	2.72 ± 0.28	0.86 ± 0.18
`X-learner`	0.72 ± 0.36	0.24 ± 0.39	0.45 ± 0.37	2.52 ± 0.27	0.58 ± 0.16
`Z-target`	0.53 ± 0.41	0.26 ± 0.32	0.25 ± 0.39	1.85 ± 0.32	0.09 ± 0.05
`Uplift-RF`	0.43 ± 0.38	0.20 ± 0.37	0.18 ± 0.29	2.31 ± 0.28	0.89 ± 0.23

Table 4.4 Average variance of the estimators. The lowest variance is underlined for each dataset.

Model	Churn 1	Churn 2	Hillstrom	Criteo
`Outcome-RF`	2.93×10^{-3}	2.07×10^{-3}	3.49×10^{-3}	1.23×10^{-3}
`S-learner`	6.81×10^{-6}	5.77×10^{-7}	4.70×10^{-3}	9.40×10^{-5}
`T-learner`	4.13×10^{-3}	3.78×10^{-3}	7.58×10^{-3}	1.94×10^{-3}
`X-learner`	3.01×10^{-3}	7.12×10^{-3}	4.05×10^{-2}	9.31×10^{-3}
`Z-target`	1.04×10^{-2}	6.07×10^{-3}	1.70×10^{-2}	6.36×10^{-3}
`Uplift-RF`	3.04×10^{-3}	3.06×10^{-3}	6.21×10^{-4}	2.05×10^{-3}

Table 4.5 Variance of the ranking. The lowest ranking variance is underlined for each dataset.

Model	Churn 1	Churn 2	Hillstrom	Criteo
`Outcome-RF`	0.112	0.729	0.134	0.002
`S-learner`	1.434	1.588	0.460	0.232
`T-learner`	0.646	0.929	0.538	0.054
`X-learner`	0.148	1.396	0.814	0.002
`Z-target`	0.695	0.647	0.640	0.021
`Uplift-RF`	5.130	4.619	4.322	0.624

4.3 Customer retention campaigns

Most causal studies are limited to experiments on historical datasets or simulations. An important opportunity offered by this research project is the ability to validate causal models in a real setting. Through a collaboration with the data science department and direct marketing department (i.e., the service responsible for marketing actions that reach out directly to customers by email, phone, etc.) at Orange Belgium, we are able to establish the list of customers for several churn prevention campaigns, which follow the process described in Section 1.3. This allows us to assess the added value of uplift modeling by comparing the results of such a campaign designed with a causal model with the campaigns based on the classical predictive approach. During the course of this thesis, we had the opportunity to perform this experiment in four different campaigns. To the best of our knowledge, this represents a unique aand new contribution to the uplift and churn literature.

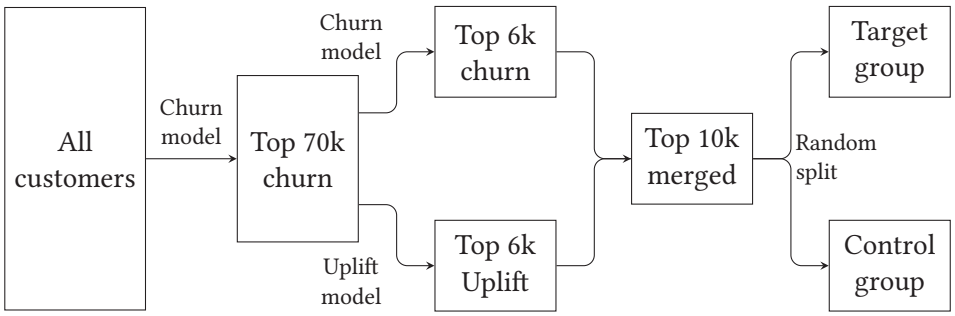


Figure 4.3 Methodology for comparing the performance of two models during a customer retention campaign. In this example, there is an overlap of 2000 between the customers selected by the churn and uplift models.

4.3.1 Experimental setup

Our objective is to evaluate whether an uplift model performs better than the classical churn model used by Orange Belgium. A simple approach would be to use only the predictions of an uplift model for new campaigns and compare the causal effect of these new campaigns with that of past campaigns which used the churn approach. However, wide-scale marketing actions by the company or competitors can affect customer behavior independently of the campaign. More generally, the churn rate and the effect of the campaign vary over time, making it difficult to attribute a performance difference solely to the uplift model. To address this issue, we established an experimental protocol in which both the churn model and the uplift model are used concurrently in the same campaign. The procedure is shown in Fig. 4.3, and involves the following steps:

1. The entire customer base is ranked by a conventional churn prediction model.
2. The top 70k customers are selected, based on this ranking.
3. An uplift model trained on historical data predicts the uplift of these 70k customers, which generates a second ranking.
4. Customers are selected by going down both ranking in parallel until the number required by the call center is reached, accounting for overlap between the rankings. An example of this process is given in Table 4.6.
5. The resulting group of customers is split randomly into a target group and a control group.

Note that an uplift model requires data from previous retention campaigns and, as such, unlike the churn model, it cannot be trained on the whole customer base. This implies that uplift models are trained only on high-risk customers that have been selected in past retention campaigns. Therefore, the output of an uplift model on low-risk customers is uncertain, and we first need to select only high-risk customers from the current customer base using the churn model. We could predict the uplift only on the top 10k customers (assuming the uplift model was trained on data coming from campaigns with 10k customers), but this would not give the opportunity to the uplift model to select new customers that the churn model would not have already placed in the top 10k. Using 70k customers is, in our opinion, a good trade-off between ensuring the validity

Table 4.6 Example of merging the rankings of the uplift and churn models on a population of 50 customers. We show the first 10 customers in both rankings. Blue cells indicate selected customers, red cells indicate customers already selected by the other model earlier in the ranking, and orange cells indicate customers selected simultaneously by both models. By going down to rank 7 in both rankings, 9 customers are selected , which means that there is an overlap of 5 customers between the rankings.

Uplift model		Churn model	
Customer ID	Score	Customer ID	Score
27	0.74	5	0.83
24	0.67	37	0.81
5	0.65	27	0.77
40	0.65	40	0.76
44	0.44	24	0.76
48	0.42	21	0.66
6	0.41	44	0.64
35	0.35	6	0.61
23	0.33	48	0.61
36	0.31	38	0.56
⋮		⋮	

of predicted uplift scores and allowing the uplift model to discover new persuadable customers.

By keeping track of which model selected the customers in the resulting control and target groups (a customer might be selected by both models), we can evaluate and compare the performances of both models simultaneously. More precisely, we compute the average treatment effect of the campaign on the customers selected by either the churn model or the uplift model. Let $\boldsymbol{w}_C$ be a binary random variable indicating whether the customer was selected by the churn model after Step 4 of the process described above, and let $\boldsymbol{w}_U$ be the corresponding variable for the uplift model. Let $D = \{(y^{(i)}, t^{(i)}, w_C^{(i)}, w_U^{(i)})\}_{i=1}^N$ be the list of selected customer after Step 4, where N is the number of customers in the campaign, $y^{(i)}$ is the churn indicator, and $t^{(i)}$ is the treatment indicator ($t^{(i)} = 1$ for the treatment group and $t^{(i)} = 0$ for the control group). First, we estimate the churn rate in the control and target groups for both models as

$$\hat{S}_{0C} = \frac{\sum_{i=1}^N (1 - t^{(i)}) w_C^{(i)} y^{(i)}}{\sum_{i=1}^N (1 - t^{(i)}) w_C^{(i)}} \qquad \hat{S}_{1C} = \frac{\sum_{i=1}^N t^{(i)} w_C^{(i)} y^{(i)}}{\sum_{i=1}^N t^{(i)} w_C^{(i)}} \tag{4.7}$$

$$\hat{S}_{0U} = \frac{\sum_{i=1}^N (1 - t^{(i)}) w_U^{(i)} y^{(i)}}{\sum_{i=1}^N (1 - t^{(i)}) w_U^{(i)}} \qquad \hat{S}_{1U} = \frac{\sum_{i=1}^N t^{(i)} w_U^{(i)} y^{(i)}}{\sum_{i=1}^N t^{(i)} w_U^{(i)}}. \tag{4.8}$$

Then, the average treatment effect for the churn and uplift models is estimated as

$$\hat{U}_C = \hat{S}_{0C} - \hat{S}_{1C} \quad \text{and} \quad \hat{U}_U = \hat{S}_{0U} - \hat{S}_{1U}. \tag{4.9}$$

We computed the 90% confidence interval on the average treatment effect following the procedure proposed by X. Li and P. Ding (2016, Sec. 2.1).

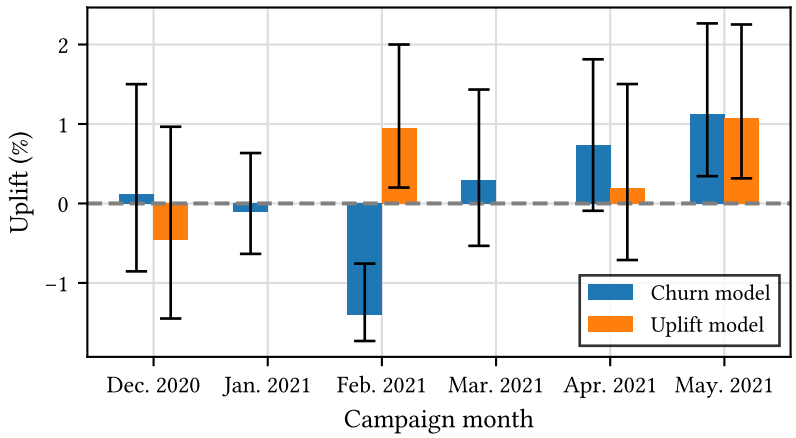


Figure 4.4 Performance of both the uplift and churn approaches over the 6 different campaigns. The whiskers represent a 90% confidence interval on the magnitude of the effect. The uplift model was not used for the campaigns of January and March 2021. The higher the bar, the better the performance of the corresponding model. We see that, due to the large uncertainty in the results, it is difficult to be sure that any model is better than the other, with the exception of February 2021.

We had the opportunity to carry out this experiment in December 2020, and in February, April and May 2021. For the December 2020 and the February 2021 experiments, the uplift model is a X-learner, while we used the R-lower approach described in Section 4.4.1. At the time of conducting these experiments, we had not yet obtained all the results presented in Section 4.2 and Section 4.4. Consequently, the selection of the uplift model was based on the information that was accessible to us during that period.

4.3.2 Results

The results of the campaigns are reported in Figure 4.4. For most campaigns, the 0% uplift line is in the 90% confidence interval, suggesting that the hypothesis that the campaign has no effect on churn cannot be rejected. We can only observe a positive uplift during the May 2021 campaign. In terms of comparing the uplift and predictive approaches, the uplift model shows slightly worse performance than the churn model in three of the four months where both models were used. We see that in February 2021, the uplift model seems to have selected more persuadable customers than the churn model.

These results do not allow us to conclude that the uplift model brings a consistent advantage over the churn model. The uncertainty observed in our results can be attributed to the relatively small number of customers contacted during each campaign, coupled with the low churn rate. This situation leads to a limited number of churners in each campaign, consequently resulting in a high uncertainty in the uplift measurements.

4.4 Using reach information to improve uplift estimation

In this section, we suggest using information about the reaction of the customers to the campaign to improve uplift estimation. In the marketing domain, the *reach* denotes

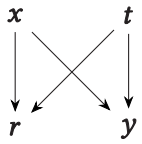


Figure 4.5 Causal graph of customer behavior, where the reach indicator $\boldsymbol{r}$ has the same parents as the churn indicator $\boldsymbol{y}$, and can be used to improve uplift prediction.

the proportion of the population exposed to the campaign, more specifically for advertisement campaigns (Farris et al., 2010). Here, we instead define *reach* as the reaction of the customer to the attempted call, that is, whether or not the customer picked up the phone and had a conversation with the phone operator. More specifically, some customers will not pick up the phone, will hang up immediately, or, more generally, will not respond positively to the call. This information, recorded by the call center, is a strong marker of customer receptivity. This is related to the notion of click-through rate (Winer, 2001) or response rate (B. J. Hansotia and Bradley Rukstales, 2002) in online and email advertising. Although response models have been developed to improve direct marketing (Bose and Chen, 2009; Guido et al., 2011; B. J. Hansotia and Bradley Rukstales, 2002), the current literature on uplift modeling ignores this information during the learning process. Expert knowledge from Orange Belgium indicates that customers who do not pick up the phone or hang up immediately should be avoided because targeting them can increase their propensity to churn.

It is important to note that the reach is only known after the campaign. Thus, it cannot be simply added as input to the model as an additional feature. We have to devise a dedicated approach to incorporate it into the learning process. In this sense, reach serves as an inductive bias for the uplift model. This section investigates whether an uplift model properly adapted to account for this new source of information provides an improvement over the state of the art.

4.4.1 Strategies for integrating reach

We denote with $\boldsymbol{r} = 1$ *reached* customers, i.e., customers who picked up the phone and had a dialogue with the phone operator. Otherwise, the customer is considered *unreached* ($\boldsymbol{r} = 0$). The causal process is represented in Fig. 4.5, where the features $\boldsymbol{x}$ represent the individual characteristics of the customers, and $\boldsymbol{t}$ is the call indicator.[4] Although there is no direct causal link between $\boldsymbol{r}$ and $\boldsymbol{y}$, there is a strong statistical dependency between them, and consequently between reach and uplift. We present four ways to integrate the information about reach to improve uplift estimation. The resulting equations are summarized in Table 4.7.

Reach probability as a feature

The first approach, `R-feature`, consists in building a predictive model of reach from historical data, and integrating the probability of reaching the customer, noted $\hat{r}$, among the input features of the uplift model. We note the reach prediction model as

$$\hat{r}(x) \approx P(\boldsymbol{r} = 1 \mid \boldsymbol{x} = x). \tag{4.10}$$

[4] $\boldsymbol{t}$ is a cause of $\boldsymbol{r}$ because a customer is necessarily unreached ($\boldsymbol{r} = 0$) when no call attempt is made ($\boldsymbol{t} = 0$).

The R-feature approach estimates the uplift as

$$\begin{aligned} \texttt{R-feature}(x) &= S_0(x, \hat{r}(x)) - S_1(x, \hat{r}(x)) && (4.11)\\ &= P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x, \hat{r}(\boldsymbol{x}) = \hat{r}(x)) - P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x, \hat{r}(\boldsymbol{x}) = \hat{r}(x)). && (4.12) \end{aligned}$$

Decomposition of probability

The second approach, R-decomp, is based on the decomposition of the probability of $S_1(x)$ with respect to the reach indicator:

$$\begin{aligned} S_1(x) &= P(\boldsymbol{y}_1 = 1 \mid x) && (4.13)\\ &= P(\boldsymbol{r}_1 = 0 \mid x)P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0) + P(\boldsymbol{r}_1 = 1 \mid x)P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1) && (4.14)\\ &= P(\boldsymbol{r}_1 = 0 \mid x)P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0) + (1 - P(\boldsymbol{r}_1 = 0 \mid x))P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1) && (4.15)\\ &= P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1) + P(\boldsymbol{r}_1 = 0 \mid x)\big[P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1)\big] && (4.16) \end{aligned}$$

Using this decomposition, R-decomp estimates the uplift as

$$\begin{aligned} \texttt{R-decomp}(x) &= S_0(x) - S_1(x) && (4.17)\\ &= P(\boldsymbol{y}_0 = 1 \mid x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1)\\ &\quad - P(\boldsymbol{r}_1 = 0 \mid x)\big[P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1)\big]. && (4.18) \end{aligned}$$

The formula contains 5 terms but can be estimated with two uplift models and a simple classifier. The first two terms, $P(\boldsymbol{y}_0 = 1 \mid x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1)$, can be estimated with a uplift model by restricting the target group to reached customers. The third term, $P(\boldsymbol{r}_1 =\mid x)$, can be estimated using a predictive model of reach. The last two terms between brackets, $P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0)$, can also be estimated by an uplift model, but using the reach indicator $\boldsymbol{r}$ instead of $\boldsymbol{t}$ as the treatment indicator for the model.

Bounds on uplift

As discussed by Radclifte and Simpson (2008), direct interventions, especially intrusive ones such as phone calls, can act as a trigger that leads to customer churn that could otherwise have been prevented, or at least postponed. The experience of direct marketing experts at Orange Belgium confirms this hypothesis and suggests that this behavior is typically associated with unreached customers. Therefore, not reaching a customer has a doubly detrimental effect: the resources of the call center are wasted, and the customer is more likely to churn than if no call had been made. This domain knowledge may be translated into the inequality

$$\underbrace{P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0)}_{\text{Probability of churn when not reached}} \geq \underbrace{P(\boldsymbol{y}_0 = 1 \mid x)}_{\text{Probability of churn with no action}}. \qquad (4.19)$$

To obtain a new estimator that takes advantage of this inequality, we start by marginalizing $S_1(x)$ over the distribution of $\boldsymbol{r}$:

$$\begin{aligned} S_1(x) &= P(\boldsymbol{y}_1 = 1, \boldsymbol{r} = 0 \mid x) + P(\boldsymbol{y}_1 = 1, \boldsymbol{r} = 1 \mid x) && (4.20)\\ &= P(\boldsymbol{r}_1 = 0 \mid x)P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0) + P(\boldsymbol{r}_1 = 1 \mid x)P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1). && (4.21) \end{aligned}$$

Table 4.7 Summary of the approaches used to integrate reach in uplift modeling.

Approach	Equation
`R-feature`	$S_0(x, \hat{r}(x)) - S_1(x, \hat{r}(x))$
`R-decomp`	$S_0(x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1) + P(\boldsymbol{r}_1 = 0 \mid x)(P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0))$
`R-upper`	$P(\boldsymbol{r}_1 = 1 \mid x)(S_0(x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1))$
`R-lower`	$P(\boldsymbol{r}_1 = 0 \mid x)(S_0(x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0))$

Substituting $P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0)$ by $P(\boldsymbol{y}_0 = 1 \mid x)$, we obtain

$$S_1(x) \geq P(\boldsymbol{r}_1 = 0 \mid x)S_0(x) + P(\boldsymbol{r}_1 = 1 \mid x)P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1).$$

This is a lower bound on $S_1(x)$; since $U(x) = S_0(x) - S_1(x)$, we have also an upper bound on $U(x)$:

$$U(x) \leq S_0(x) - S_0(x)P(\boldsymbol{r}_1 = 0 \mid x) - P(\boldsymbol{r}_1 = 1 \mid x)P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1). \tag{4.22}$$

We can rearrange this expression as

$$\begin{aligned} U(x) &\leq S_0(x)P(\boldsymbol{r}_1 = 1 \mid x) - P(\boldsymbol{r}_1 = 1 \mid x)P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1) && (4.23) \\ &= P(\boldsymbol{r}_1 = 1 \mid x)(S_0(x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1)). && (4.24) \end{aligned}$$

The third approach, `R-upper`, consists in using the upper bound as an estimation of the uplift:

$$\texttt{R-upper}(x) = P(\boldsymbol{r}_1 = 1 \mid x)(S_0(x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 1)). \tag{4.25}$$

Using `R-upper` as an uplift estimator assumes that the less-than relationship in Eq. (4.19) is an equality, which, intuitively, indicates that unreached customers will ignore the call attempt and behave as if no call attempt was made, viz., the probability of churn when called and unreached is equal to $S_0(x)$. Estimating Eq. (4.25) requires two models: a simple predictive model of the reach indicator (using only the target group) and an uplift model where the target group has been restricted to reached customers.

A symmetrical reasoning may lead to the hypothesis that a reached customer is less likely to churn than if not contacted:

$$P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x}, \boldsymbol{r}_1 = 1) \leq P(\boldsymbol{y}_0 = 1 \mid x). \tag{4.26}$$

We can derive an upper bound from this assumption and Eq. (4.14):

$$U(x) \geq P(\boldsymbol{r}_1 = 0 \mid x)(P(\boldsymbol{y}_0 = 1 \mid x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0)). \tag{4.27}$$

The fourth approach, `R-lower`, consists in using this lower bound as an uplift estimator:

$$\texttt{R-lower}(x) = P(\boldsymbol{r}_1 = 0 \mid x)(P(\boldsymbol{y}_0 = 1 \mid x) - P(\boldsymbol{y}_1 = 1 \mid x, \boldsymbol{r}_1 = 0)) \tag{4.28}$$

4.4.2 Experimental setup

In this benchmark, we evaluate the approaches described in Section 4.4 on the Churn 2 dataset. We use the X-learner algorithm as the base uplift model required for the implementation of the reach approaches, and the random forest algorithm for base predictive models. We compare the four reach approaches against three baselines:

- `X-learner`: A conventional X-learner uplift model (Künzel et al., 2019) with no information on reach.
- `Outcome-RF`: A random forest (Breiman, 2001) trained to predict the probability of churn $P(\boldsymbol{y} = 1 \mid x)$. This represents the classical predictive approach.
- `Reach-RF`: A random forest trained to predict the probability of reaching the customer $P(\boldsymbol{r} = 1 \mid x)$.

The first two baselines are the models that perform best in the benchmark of Section 4.2. The third baseline `R-target` is introduced to check whether the reach alone can be used to find persuadable customers. As in Section 4.2, the class imbalance between churners and non-churners is addressed with the EasyEnsemble strategy (X.-Y. Liu, Wu, and Zhou, 2009). To obtain a measure of the variability of the performance, we create 50 independent random splits of the dataset into training and test sets, in proportion 80% / 20%. Each of these splits is used to train each model, and we report the area under the uplift curve (AUUC) on the test set, averaged over the 50 runs. We also estimate the estimator variance and the ranking variance, as described in Section 4.2.1.

We also evaluated several variations of the models presented in Table 4.7. But, since they did show remarkable performances, we did not include them in the results. These variations are: i) the average of `R-lower` and `R-upper`, ii) the average of `X-learner` and `R-target`, and iii) the product of `X-learner` and `R-target`.

4.4.3 Results

Table 4.8 reports the mean and standard deviation of the area under the uplift curve (AUUC) over 50 runs for each model. The AUUC is also reported as a boxplot in Fig. 4.6. We see that `R-feature` is the best performing model in terms of area under the uplift curve. Among the other reach approaches, `R-decomp` and `R-lower` perform similarly, while `R-upper` does not outperform the baselines. The two baselines `X-learner` and `Outcome-RF` have similar performances and, as expected, `Reach-RF` performs quite poorly.

`R-upper` has the lowest estimator variance, which might be due to the fact that it consists of a product of the predictions of an uplift model and the predictions of a predictive model of reach. Since both the uplift and the probability of reach are by definition lower than or equal to one, their product will be smaller than any of the two quantities separately. This is also the case for `R-decomp` and `R-lower`, which also have a low estimator variance. However, it is surprising that `X-learner` has a lower variance than `Outcome-RF`, which is the opposite of the results in the previous experiment, reported in Table 4.4.

As expected, the two predictive models `Outcome-RF` and `Reach-RF` have a lower ranking variance than the uplift models. Note that `R-feature` has a higher ranking variance and estimator variance than `X-learner`, even though the two models differ only in that `R-feature` uses the probability of reach as an additional feature.

Table 4.8 AUUC (with standard deviation) and estimator variance on the Churn 2 dataset over 50 runs. The best values (highest AUUC and lowest variances) are underlined.

Model	AUUC (%)	Estimator variance	Ranking variance
`R-feature`	$\underline{0.857 \pm 0.547}$	8.10×10^{-4}	0.150
`R-decomp`	0.584 ± 0.549	4.67×10^{-4}	0.122
`R-upper`	0.427 ± 0.507	$\underline{1.06 \times 10^{-4}}$	0.219
`R-lower`	0.674 ± 0.575	2.07×10^{-4}	0.355
`X-learner`	0.541 ± 0.509	4.36×10^{-4}	0.107
`Outcome-RF`	0.604 ± 0.621	7.06×10^{-4}	$\underline{0.023}$
`Reach-RF`	0.247 ± 0.397	8.46×10^{-4}	0.031

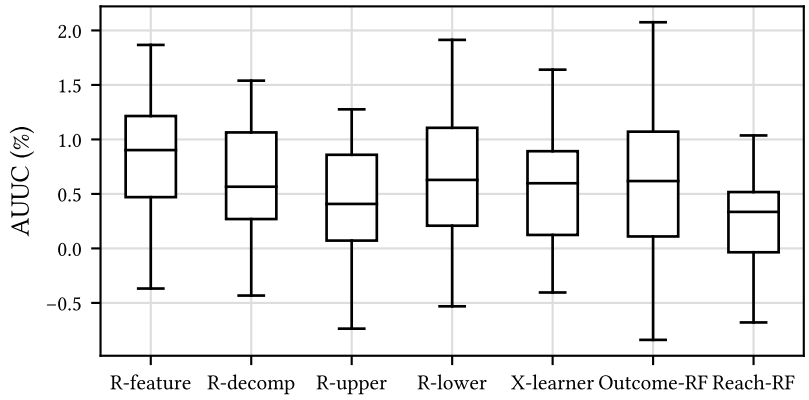


Figure 4.6 Boxplots representing the distribution of the AUUC on the Churn 2 dataset over 50 runs. The middle line represents the median, the extents of the box represents the first and third quartiles (hence, the boxes contain half of the points), and the whiskers extend to the minimum and maximum values.

In general, the best performing model in terms of AUUC is `R-feature`, as it clearly outperforms all other reach models and the three baselines. Furthermore, it has a reasonable ranking variance with respect to the other uplift models. This indicates that the reach information can be successfully exploited to improve the quality of uplift predictions.

4.5 Conclusion

In this chapter, we conducted three experimental comparisons of uplift modeling and the conventional predictive approach.

In the first experiment, we compared the performance of several state-of-the-art uplift models in terms of area under the uplift curve. We observed that the predictive approach is competitive or performs better than the uplift models, except in the Churn 2 and Hillstrom datasets. The Hillstrom dataset is characterized by a less severe class imbalance and more informative features. We hypothesize that the superiority of the predictive approach is due to its lower variance, and, to assess this hypothesis, we used a measure of the variance of the ranking generated by the models. The predictive

approach has the lowest ranking variance, except on the Churn 2 dataset where it is slightly higher than that of the modified outcome strategy. These findings indicate a strong association between the performance of a model and the stability of its ranking. Especially in scenarios characterized by outcome imbalance and uninformative features, it appears that the bias reduction achieved through the use of uplift modeling may not be sufficient to outweigh the performance loss due to its higher variance.

In the second experiment, we conducted several customer retention campaigns using simultaneously an uplift model and a churn prediction model, allowing us to compare their performance in a real setting. This experiment was repeated over four different months, the first two months using an `X-learner` uplift model and the last two months using the `R-upper` approach described in Section 4.4.1. We observed that in three of the four campaigns, the churn model and the uplift model had very similar performances, with an average treatment effect close to zero. The uplift showed a significant advantage over the churn model only during the February 2021 campaign. Due to time constraints, we did not have the opportunity to carry out this experiment with the most promising uplift model, `R-feature`.

In the third experiment, we showed the potential of reach information to improve the estimation of uplift. Reach information is not readily accessible prior to a campaign, and, as a result, specific strategies must be employed to leverage it effectively. The strategy showing the greatest potential, `R-feature`, consists in adding the predicted probability of reaching the customer as an additional feature to an uplift model. This strategy outperformed both the predictive approach and the `X-learner` uplift model, which showed the best performance in the benchmark of Section 4.2.

The `R-feature` strategy has the advantage of being relevant to a wider range of use cases than churn prevention. It is common for telecom companies to run different campaigns using the voice call channel, such as *up-sell* (to propose a better product to the customer), or *cross-sell* (to present additional products). A reach model can be used in these contexts as well, while using the same training data. This is a significant advantage, both in terms of computation time and volume of data. However, this approach requires access to records detailing the response of customers to calls, which may not be readily available, especially for companies without prior experience in direct marketing. Additionally, given the relatively novel nature of uplift modeling and the limited availability of publicly accessible uplift datasets online, none of these datasets incorporate information related to reach. Consequently, it is difficult to assess novel approaches that exploit reach information beyond the scope of a collaborative arrangement with a private company.

The findings from these experiments collectively indicate that uplift modeling does not consistently outperform conventional predictive modeling, with the exception of datasets like the Hillstrom dataset characterized by balanced outcomes and informative features. In the following chapter, we investigate this finding from a theoretical perspective by examining various characteristics of the data that may explain the discrepancy in performance between the two approaches.

5

Theoretical analysis of uplift modeling

Results presented in this chapter have been published in the following article: Théo Verhelst, Wouter Verbeke, et al. (2023). "Uplift vs. Predictive Modeling: a Theoretical Analysis". In: *Submitted to Journal of Machine Learning Research.*

In the previous chapter, we assessed the added value of causal-oriented strategies with respect to the purely predictive approach through a series of experiments. We believe that it is important to assess whether the expected benefit of uplift strategies (deriving from a bias reduction in the estimation of causal effect) is still noticeable in settings such as churn prediction where the data distribution is characterized by a large number of dimensions, non-linearity, class imbalance, and low-class separability. Such an empirical comparison has few precedents in the literature (Ascarza, 2018; Devriendt, Berrevoets, and Verbeke, 2021; Wijaya et al., 2021). There are also very few articles addressing this question from a theoretical perspective. Fernández-Loria and Provost (2022a,b) develop quantitative measures and qualitative arguments that indicate when the predictive approach should be preferred. We extend these papers by comprehensively treating the question, starting from theoretical foundations and studying the influence of different characteristics of the setting (distribution of the outcome, variance of the estimators, etc.) on the performance of the uplift and predictive approaches.

A critical aspect of comparing the two approaches is the need for a meaningful and sensible measure of model performance. In this chapter, we extend the work of Verbeke, Olaya, Berrevoets, et al. (2021) by developing a new formulation of the profit generated by a campaign in which individuals targeted by interventions are selected by a machine learning model. By incorporating the concept of profit, we go beyond traditional evaluation metrics and we consider the economic impact of decision-making strategies. Our measure of profit generalizes Verbeke's by accommodating varying costs and benefits across individuals. This flexibility is beneficial, for example in churn prediction, where prioritizing higher-value customers is crucial. By selecting an appropriate measure of performance, we ensure a fair and accurate comparison between uplift and predictive models, enabling decision-makers to make informed choices based on the true effectiveness and suitability of each approach.

This chapter seeks to establish firm theoretical foundations for uplift modeling and to answer the question "When does uplift modeling outperform predictive modeling?". While we focus on the example of customer churn prediction, our findings have broad applicability across domains including marketing, telecommunications, healthcare, and finance. Our main conclusions are as follows: the estimator variance (see Definition 2.23) plays a critical role in determining the performance of a model, and in most cases, the predictive approach outperforms the uplift approach when the variance of the uplift estimator exceeds a certain threshold. We also show the important impact of three other aspects: cost sensitivity, the mutual information between the features and the potential outcomes, and the distribution of the potential outcomes. While the importance of cost sensitivity and the distribution of the potential outcomes has been discussed in the literature, respectively, by Verbeke, Olaya, Berrevoets, et al. (2021) and Fernández-Loria and Provost (2022a), to the best of our knowledge, the impact of mutual information has not been assessed before. We show that it has an important impact on the performance, independently of these other aspects (estimator variance, cost sensitivity, and distribution of potential outcomes).

Note, however, that we do not address the question of how to adapt uplift modeling to account for cost sensitivity or the other aspects mentioned above. Our contributions pertain to model *evaluation*, rather than model *optimization*. As such, it is left for future work to assess the effectiveness of cost sensitive models in terms of the metrics developed in this chapter. On that topic, Gubela and Lessmann (2021) propose a value-driven ranking method for targeted marketing campaigns.

The main contributions of this chapter are:

- A new formulation of the measure of profit, bringing the focus on individual cost sensitivity and on the stochastic nature of the machine learning model used to rank individuals (Section 5.2.2). We show its equivalence with the profit measure developed by Verbeke, Olaya, Berrevoets, et al. (2021) in Section 5.2.3.

- A proof that the uplift curve (an evaluation curve often used in the uplift literature, see Definition 3.3) is an estimator of the measure of profit, highlighting the strict conditions necessary for the validity of the uplift curve (Section 5.2.4).

- An empirical estimator of the measure of profit, which is a cost sensitive generalization of the uplift curve (Section 5.2.5).

- A demonstration through theoretical analyses and simulations of the conditions under which the predictive approach still outperforms uplift modeling, and, notably, the important role of the mutual information between the input features and the potential outcomes, which has not yet been discussed in the literature (Section 5.3).

The rest of this chapter is organized as follows. Section 5.1 introduces the notations and notions used throughout this chapter. Our contributions are presented in Sections 5.2 and 5.3: we present the new formulation of the measure of profit in Section 5.2 and we assess when the predictive approach outperforms the uplift approach in Section 5.3. We discuss our results and the limitations of our approach in Section 5.4. Concluding remarks and recommendations for practitioners are given in Section 5.5. Proofs of the theorems are provided in Appendices C and D.

5.1 Background

In this section, we introduce the notations and present the key concepts used throughout this chapter.

5.1.1 Notation

We use the notation presented in Section 3.1.1, where $\boldsymbol{x}$ is a random vector of features (or covariates), $\boldsymbol{y}$ is the binary outcome, and $\boldsymbol{t}$ is the assignment of treatment. We assume the treatment assignment to be randomized (see Definition 3.1). S_0 and S_1 are, respectively, the probabilities of a positive outcome in the control and target groups (Eqs. 3.4 and 3.5), $U = S_0 - S_1$ is the average uplift (Eq. 3.6), and $S_0(x), S_1(x), U(x)$ are the same quantities conditioned on a particular realization $\boldsymbol{x} = x$ (Eqs. 3.1 to 3.3), which are learned by a machine learning model. Note that, for example in the literature pertaining to retail or online advertisement, the uplift is defined as $U = S_1 - S_0$ (and similarly $U(x) = S_1(x) - S_0(x)$). This choice depends on whether the probability of the positive outcome ($\boldsymbol{y} = 1$) should be minimized (e.g., in churn prevention) or maximized (e.g., in sales). The uplift is then defined so that a positive uplift corresponds to a beneficial outcome. Since we apply our results mainly to churn prevention, we use the convention $U = S_0 - S_1$.

Let $\mathscr{M}$ be a model that is used to rank individuals such that only the individuals with the highest scores should be targeted. The model $\mathscr{M}$ is trained on a dataset $D_{\text{tr}} = \{(x^{(i)}, y^{(i)}, t^{(i)})\}_{i=1}^{N}$ of N iid[1] realizations of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$. We assume that D_{tr} is the result of a random process and we note it as a random variable as $\boldsymbol{D}_{\text{tr}}$. We consider $\mathscr{M}(x, D_{\text{tr}})$ as a learning algorithm taking a dataset D_{tr} and a vector of features x as input, and returning a score, for example an estimation of $U(x)$. We denote the fact that $\mathscr{M}(x, D_{\text{tr}})$ is trained to estimate some function $g(x)$ as $\mathscr{M}(x, D_{\text{tr}}) \approx g(x)$.

A threshold τ is used to determine which individuals should be targeted: the model $\mathscr{M}$ prescribes targeting all individuals with a score $\mathscr{M}(x, D_{\text{tr}}) \geq \tau$ and not targeting the remaining individuals. Since different models can provide scores with different distributions, the threshold τ depends on the model being used. Therefore, in order to compare the performance of different models in a consistent way, we let $\rho \in (0, 1)$ be the proportion of individuals who should be targeted. We call ρ the *prescription rate*. The corresponding threshold τ can be determined as the smallest value that satisfies $P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) > \tau) \geq \rho$. The threshold τ is also a function of D_{tr}, since different training sets induce different score distributions, hence different thresholds. The dependency of τ in ρ is implicit in our notation, but we explicitly note $\tau = \tau(D_{\text{tr}})$, as the dependency in the training set is important in some equations of this chapter. Therefore, we formally define $\tau(D_{\text{tr}})$ as

$$\tau(D_{\text{tr}}) = \inf\{\tau' \,:\, P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau') \geq \rho\}. \tag{5.1}$$

See Fig. 5.1 for an illustration of this definition based on the cumulative distribution function of the scores. Also, note that $\mathscr{M}(x, D_{\text{tr}})$ is a deterministic function of x (for a fixed D_{tr}), therefore, in Eq. (5.1), $\mathscr{M}(\boldsymbol{x}, D_{\text{tr}})$ is a random variable due to the random nature of the population represented by the features $\boldsymbol{x}$. The probability $P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) > \tau(D_{\text{tr}}))$ represents the probability that the score given to an individual selected at random in the population is greater than the threshold $\tau(D_{\text{tr}})$.

[1] The independence assumption might be violated in applications such as churn with, for example, a word-of-mouth effect generating a second order of treatment.

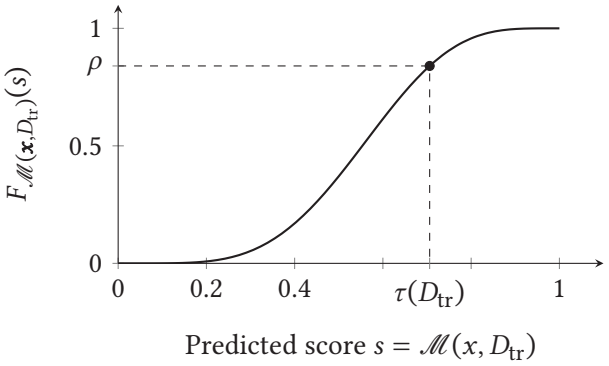


Figure 5.1 Graphical representation of the definition of $\tau(D_{\text{tr}})$ as a function of ρ from the cumulative distribution function of $\mathscr{M}(\boldsymbol{x}, D_{\text{tr}})$.

5.1.2 Uplift and predictive approaches

We designate by *predictive approach* the process of ranking individuals using a machine learning model $\mathscr{M}_p$ that estimates the conditional probability $S_0(x) = P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x)$:

$$\mathscr{M}_p(x, D_{\text{tr}}) \approx S_0(x). \tag{5.2}$$

And we designate by *uplift approach* the process of ranking individuals using a machine learning model $\mathscr{M}_u$ that estimates the uplift $U(x) = S_0(x) - S_1(x)$:

$$\mathscr{M}_u(x, D_{\text{tr}}) \approx U(x). \tag{5.3}$$

Note that the definition of these approaches can vary in the literature. For example, Fernández-Loria and Provost (2022a) focus on online advertisement, in which the outcome should be maximized. As such, they define three approaches: the *treatment difference* (TD) approach, which ranks individuals by $S_1(x) - S_0(x)$ (the opposite of the uplift approach as defined in this chapter), the *outcome most* (OM) approach, which ranks individuals by $S_1(x)$, and the *outcome least* approach (OL), which ranks individuals by $1 - S_0(x)$. The TD and OM approaches are equivalent to, respectively, our uplift and predictive approaches, up to a change of label for the values of $\boldsymbol{t}$.

5.2 Measure of profit

This section presents our contributions to the evaluation of the performance of a model in the context of causal decision making. We start by introducing the concept of causal profit for individuals, which measures the incremental profit gained by targeting specific individuals with interventions. Next, we extend this definition to the campaign level, where the cumulative profit is assessed by considering the overall impact of targeting a group of individuals. To establish the connection between the causal profit and existing measures in the literature, we first prove its equivalence with Verbeke's original definition of profit (Verbeke, Olaya, Berrevoets, et al., 2021). Then, we prove that the uplift curve is an estimator of the causal profit, under a specific assumption about the values of costs and benefits generated by individuals. Finally, we propose an empirical estimator of the profit measure that leverages the data and a ranking model to estimate the potential profit of targeting specific individuals. This new empirical performance metric is a cost sensitive generalization of the uplift curve.

5.2.1 Individual profit

Our measure of profit for a campaign, presented in the next section, relies on the notion of *individual causal profit*. This notion is called *revenue uplift* by Gubela and Lessmann (2021) and Gubela, Lessmann, and Jaroszewicz (2020). It is a generalization of the individual uplift $U(x)$ to cost sensitive settings. Our contribution with respect to the definition by Gubela, Lessmann, and Jaroszewicz lies in its definition in terms of a cost-benefit matrix, similarly to the cost-benefit matrix CB defined by Verbeke, Olaya, Berrevoets, et al. (2021), presented in Section 3.1.3. We extend the definition of the cost-benefit matrix to allow it to vary between individuals. This generalization of CB to the individual level is useful in settings such as churn prediction, where different customers have different values (e.g., some customers might have a more expensive tariff plan than others); hence, in order to maximize profits, retention efforts should be focused on high-value customers.

Definition 5.1 (Cost-benefit matrix)**.** The *cost-benefit matrix* $\mathrm{CB}(x)$ expresses the sum of the costs and benefits for the two possible actions ($\boldsymbol{t} = 0$ or $\boldsymbol{t} = 1$) and the two possible outcomes ($\boldsymbol{y} = 0$ or $\boldsymbol{y} = 1$) for an individual with features $\boldsymbol{x} = x$. It is noted

$$\mathrm{CB}(x) = \begin{matrix} \\ \boldsymbol{y} = 0 \\ \boldsymbol{y} = 1 \end{matrix} \begin{matrix} \boldsymbol{t} = 0 & \boldsymbol{t} = 1 \\ \end{matrix} \begin{bmatrix} \mathrm{CB}_{00}(x) & \mathrm{CB}_{01}(x) \\ \mathrm{CB}_{10}(x) & \mathrm{CB}_{11}(x) \end{bmatrix} \tag{5.4}$$

Note that although the actual value generated by an individual is inherently random (e.g., the data consumption of a customer for the next month is unknown), $\mathrm{CB}(x)$ expresses the expected cost-benefits for all individuals with features $\boldsymbol{x} = x$. From this matrix and the probability distribution of the outcome $\boldsymbol{y}$, one can define the expected profit that a given action generates:

Definition 5.2 (Individual profit)**.** When the action $\boldsymbol{t} = t$ is carried out for an individual $\boldsymbol{x} = x$, we define the *individual profit* of that action as

$$\begin{aligned} \pi_t(x) &= \mathrm{CB}_{0t}(x) P(\boldsymbol{y}_t = 0 \mid x) + \mathrm{CB}_{1t}(x) P(\boldsymbol{y}_t = 1 \mid x) && (5.5) \\ &= \mathrm{CB}_{0t}(x)(1 - S_t(x)) + \mathrm{CB}_{1t}(x) S_t(x). && (5.6) \end{aligned}$$

Intuitively, the equation above expresses the individual profit as the cost-benefit value when the customer does not churn multiplied by the probability that they indeed do not churn, plus the cost-benefit value when the customer churns multiplied by the probability that they churn.

As discussed by Verbeke, Olaya, Berrevoets, et al. (2021), the performance of a model should be measured relative to a baseline scenario, rather than in absolute terms. That is because, even when no action is carried out, some outcome will always occur and, therefore, the success of an action should be compared with the outcome resulting from the absence of action. Formally, we name this difference the *individual causal profit*:

Definition 5.3 (Individual causal profit)**.** The *individual causal profit* for an individual with features $\boldsymbol{x} = x$ is defined as

$$\pi(x) = \pi_1(x) - \pi_0(x). \tag{5.7}$$

We can readily develop this expression to obtain

$$\pi(x) = \mathrm{CB}_{01}(x)(1 - S_1(x)) + \mathrm{CB}_{11}(x)S_1(x) - \mathrm{CB}_{00}(x)(1 - S_0(x)) - \mathrm{CB}_{10}(x)S_0(x) \quad (5.8)$$

We can see that the causal profit is a function of $S_0(x)$, $S_1(x)$ and $\mathrm{CB}(x)$. This quantity was defined similarly by Verbeke, Olaya, Berrevoets, et al. (2021, Eq. 42). We now illustrate these definitions in the context of churn prevention with marketing campaigns.

Example 5.4. Let us suppose that the costs and benefits of a churn mitigation campaign are defined as follows:

- Calling a customer has a fixed operating cost C;
- We gain value $V_0(x)$ when the customer does not churn, a value called the *customer lifetime value* (Gupta et al., 2006), and we gain $V_1(x)$ when they churn (typically €0);
- We offer a marketing incentive of cost $I(x)$ to the customer when we call them and if they do not churn.

From this, we can compute the cost-benefit matrix $\mathrm{CB}(x)$ as

$$\mathrm{CB}(x) = \begin{matrix} \\ \boldsymbol{y}=0 \\ \boldsymbol{y}=1 \end{matrix} \begin{matrix} \boldsymbol{t}=0 & \boldsymbol{t}=1 \end{matrix} \begin{bmatrix} V_0(x) & V_0(x) - C - I(x) \\ V_1(x) & V_1(x) - C \end{bmatrix}$$

and the causal profit $\pi(x)$ is

$$\begin{aligned} \pi(x) &= -C - I(x) - (V_0(x) - I(x) - V_1(x))S_1(x) + (V_0(x) - V_1(x))S_0(x) \\ &= (V_0(x) - V_1(x))U(x) - C - I(x)P(\boldsymbol{y}_1 = 0 \mid x). \end{aligned}$$

From this last equation, we can see that if the magnitude of the uplift and the expected lifetime value of the customer are greater than the expected cost of the retention action, then we can expect a positive causal profit by calling the customer. To show this, let us suppose that the lifetime value of a given customeris $V_0(x) = €120$, $V_1(x) = €0$, the call has an operating cost $C = €1$, the marketing incentive has a cost $I(x) = €20$, and this customer has the probabilities of churn $S_0(x) = 0.15$ and $S_1(x) = 0.05$, hence an uplift of $U(x) = 0.1$. The cost-benefit matrix is

$$\mathrm{CB}(x) = \begin{matrix} \\ \boldsymbol{y}=0 \\ \boldsymbol{y}=1 \end{matrix} \begin{matrix} \boldsymbol{t}=0 & \boldsymbol{t}=1 \end{matrix} \begin{bmatrix} 120 & 99 \\ 0 & -1 \end{bmatrix}$$

and the causal profit is $\pi(x) = 120 \times 0.1 - 1 - 20 \times 0.95 = -8$. Given the low uplift of this customer and the high cost of the incentive, this customer should therefore not be targeted by this marketing campaign.

5.2.2 Campaign profit

The causal profit $\pi(x)$, as per Definition 5.3, is defined for an individual $\boldsymbol{x} = x$, but in business applications, a campaign is carried out on a large number of different individuals. As presented in Section 3.1.3, Verbeke, Olaya, Guerry, et al. (2022) define a cost sensitive measure of the causal profit of a campaign based on the causal confusion

matrix and the cost-benefit matrix. In this section, we provide another definition of the causal profit of a campaign, which we then show to be equivalent to the definition given by Verbeke et al. Our contribution lies in the fact that our measure emphasizes the individual causal profit, the influence of the individual cost and benefits, and the stochastic nature of the machine learning estimator. We also give a formal proof that the uplift curve, which is widely used in the uplift literature (see Section 3.1.3), is equivalent to the measure of profit. In particular, this equivalence highlights the strict assumption necessary for the validity of the uplift curve as an evaluation measure.

As defined in Section 5.1.1, let us assume we have a prescription rate ρ, that is, the action $\text{do}(\boldsymbol{t} = 1)$ is carried out on the proportion ρ of individuals with the highest scores, and the action $\text{do}(\boldsymbol{t} = 0)$ is carried out on the other individuals. If we have a population of N individuals, then $\lceil N\rho \rceil$ individuals will be targeted. Then, given a predictive model $\mathscr{M}(\boldsymbol{x}, D_{\text{tr}})$, we can find the threshold $\tau(D_{\text{tr}})$ on the scores that would separate the proportion ρ of individuals with the highest scores from the rest. As for the individual profit defined in Section 5.2.1, it is important to notice that even if a campaign is not carried out, some profit will be generated anyway. Therefore, the performance of an uplift model should be evaluated in terms of profit with respect to a baseline scenario where no action is taken. Here, we first define the profit induced by carrying out the campaign (Definition 5.5) and the profit of the baseline scenario where the campaign is not carried out (Definition 5.6). Finally, the causal profit is defined to be the difference between these two quantities (Definition 5.7).

Definition 5.5 (Campaign action profit)**.** The *action profit* of a campaign with a prescription rate ρ and a model $\mathscr{M}$ trained on a dataset D_{tr} is defined as

$$\Pi_1(\rho, D_{\text{tr}}) = \rho \mathbb{E}_{\boldsymbol{x}}[\pi_1(\boldsymbol{x}) \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau(D_{\text{tr}})] + (1-\rho)\mathbb{E}_{\boldsymbol{x}}[\pi_0(\boldsymbol{x}) \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau(D_{\text{tr}})] \tag{5.9}$$

where $\tau(D_{\text{tr}})$ is defined as in Eq. (5.1).

Intuitively, this quantity is the sum of the expected profit of the group of individuals being targeted (i.e., those with a score $\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau(D_{\text{tr}})$), and the expected profit of the group of individuals not being targeted (i.e., those with a score $\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau(D_{\text{tr}})$). It is important to note that the quantity in Definition 5.5, as well as in the following definitions, is independent of the size of the population. As such, Eq. (5.9) represents the expected profit generated, on average, per individual. As a very simple example, if we have a campaign on a population where targeted individuals generate on average a profit of €20, and non-targeted individuals generate on average €10, a prescription rate of $\rho = 0.5$ would result in an action profit $\Pi_1(\rho, D_{\text{tr}}) = 0.5 \times €10 + 0.5 \times €20 = €15$.

Now, we define the baseline profit of the campaign as follows.

Definition 5.6 (Campaign baseline profit)**.** The *baseline profit* of running no campaign is defined as

$$\Pi_0 = \mathbb{E}_{\boldsymbol{x}}[\pi_0(\boldsymbol{x})]. \tag{5.10}$$

The causal profit of the campaign is logically defined as the difference between these two quantities.

Definition 5.7 (Campaign causal profit)**.** The *causal profit* of a campaign with a prescription rate ρ and a model $\mathscr{M}$ trained on a dataset D_{tr} is defined as

$$\Pi(\rho, D_{\text{tr}}) = \Pi_1(\rho, D_{\text{tr}}) - \Pi_0. \tag{5.11}$$

Now, recall that the training set D_{tr} is fixed in the previous definitions, which does not take into account the stochastic nature of the sampling process of $\boldsymbol{D}_{\text{tr}}$. We would like to consider the training set as a random variable, which leads the model $\mathscr{M}$ to provide different scores depending on the realization of the training set. This generalization is necessary if one wants to formalize the bias and variance of the estimator $\mathscr{M}$, as presented in Definitions 2.22 and 2.23, Section 2.3.6. To the best of our knowledge, none of the evaluation measures for uplift modeling in the literature takes into account the random nature of $\boldsymbol{D}_{\text{tr}}$. This generalization is essential to the comparison of the uplift and predictive approaches in Section 5.3. Concretely, we define the expected causal profit as the expected value of the causal profit over the distribution of $\boldsymbol{D}_{\text{tr}}$.

Definition 5.8 (Campaign expected causal profit)**.** The *expected causal profit* of a campaign with a prescription rate ρ and a model $\mathscr{M}$ trained on a random dataset $\boldsymbol{D}_{\text{tr}}$ is defined as

$$\overline{\Pi}(\rho) = \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[\Pi(\rho, \boldsymbol{D}_{\text{tr}})]. \tag{5.12}$$

These various definitions can be unwrapped and expressed in terms of $\pi(x)$ (see Definition 5.3) and the score $\mathscr{M}(x, D_{\text{tr}})$ over the distribution of $\boldsymbol{x}$. In particular, the causal profit can be computed from the causal profit of only the targeted individuals. This is formalized in the following theorem.

Theorem 5.1. *Let the model $\mathscr{M}(\boldsymbol{x}, D_{\text{tr}})$ be a continuous random variable for all realizations D_{tr} of $\boldsymbol{D}_{\text{tr}}$. The causal profit can be expressed as*

$$\Pi(\rho, D_{\text{tr}}) = \int_{\mathcal{X}} \pi(x)\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) \geq \tau(D_{\text{tr}})]\,\mathrm{d}x \tag{5.13}$$

$$= \mathbb{E}_{\boldsymbol{x}}[\pi(\boldsymbol{x})\mathbb{I}[\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau(D_{\text{tr}})]] \tag{5.14}$$

where $\tau(D_{\text{tr}})$ is defined as in Eq. (5.1), and the expected causal profit can be expressed as[2]

$$\overline{\Pi}(\rho) = \int_{\mathcal{X}} \pi(x)P(\mathscr{M}(x, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}}))\,\mathrm{d}x \tag{5.15}$$

$$= \mathbb{E}_{\boldsymbol{x}}[\pi(\boldsymbol{x})P(\mathscr{M}(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}}))]. \tag{5.16}$$

Proof. Let $f_{\boldsymbol{x}}(\cdot)$ be the probability density function of $\boldsymbol{x}$. Using the definition of the conditional expected value (Definition 2.2), we can develop $\Pi_1(\rho, D_{\text{tr}})$ in Eq. (5.9) as

$$\begin{aligned}\Pi_1(\rho, D_{\text{tr}}) =& \frac{\rho}{P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau(D_{\text{tr}}))} \int f_{\boldsymbol{x}}(x)\pi_1(x)\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) \geq \tau(D_{\text{tr}})]\,\mathrm{d}x \\ &+ \frac{1-\rho}{P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau(D_{\text{tr}}))} \int f_{\boldsymbol{x}}(x)\pi_0(x)\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) < \tau(D_{\text{tr}})]\,\mathrm{d}x. \end{aligned} \tag{5.17}$$

$\mathscr{M}(\boldsymbol{x}, D_{\text{tr}})$ is a continuous random variable, therefore its cumulative distribution function is nondecreasing. Therefore, the value $\tau(D_{\text{tr}})$ that satisfies its definition $\tau(D_{\text{tr}}) = \inf\{\tau' : P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau') \geq \rho\}$ in fact satisfies exactly $\rho = P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau(D_{\text{tr}}))$.

[2] Note that in Eqs. (5.15) and (5.16), the probability $P(\mathscr{M}(x, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}}))$ is taken over the distribution of training sets $\boldsymbol{D}_{\text{tr}}$.

Table 5.1 Expected profits of the population of six individuals from Example 5.9.

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$
$\pi_0(x)$	−0.1	0.1	0.15	0.1	0.2	0
$\pi_1(x)$	0.2	0.05	−0.05	0.1	−0.1	0.1
$\pi(x)$	0.3	−0.05	−0.2	0	−0.3	0.1

Hence, Eq. (5.17) simplifies to

$$\begin{aligned}
&\Pi_1(\rho, D_{\text{tr}}) \\
&= \int f_{\boldsymbol{x}}(x)\pi_1(x)\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) \geq \tau(D_{\text{tr}})]\,\mathrm{d}x + \int f_{\boldsymbol{x}}(x)\pi_0(x)\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) < \tau(D_{\text{tr}})]\,\mathrm{d}x \\
&= \int f_{\boldsymbol{x}}(x)(\pi_1(x)\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) \geq \tau(D_{\text{tr}})] + \pi_0(x)(1 - \mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) \geq \tau(D_{\text{tr}})]))\,\mathrm{d}x \\
&= \int f_{\boldsymbol{x}}(x)(\pi_0(x) + \pi(x)\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) \geq \tau(D_{\text{tr}})])\,\mathrm{d}x \\
&= \Pi_0 + \int_{\mathcal{X}} \pi(\boldsymbol{x})\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) \geq \tau(D_{\text{tr}})]\,\mathrm{d}x.
\end{aligned}$$

Therefore, the causal profit of the campaign can be expressed as

$$\Pi(\rho, D_{\text{tr}}) = \Pi_1(\rho, D_{\text{tr}}) - \Pi_0 = \int_{\mathcal{X}} \pi(\boldsymbol{x})\mathbb{I}[\mathscr{M}(x, D_{\text{tr}}) \geq \tau(D_{\text{tr}})]\,\mathrm{d}x.$$

The expected causal profit can be developed as

$$\begin{aligned}
\overline{\Pi}(\rho) &= \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[\Pi(\rho, \boldsymbol{D}_{\text{tr}})] \\
&= \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}\left[\int_{\mathcal{X}} \pi(x)\mathbb{I}[\mathscr{M}(x, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}})]\,\mathrm{d}x\right] \\
&= \int_{\mathcal{X}} \pi(x)\mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[\mathbb{I}[\mathscr{M}(x, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}})]]\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \pi(x)P(\mathscr{M}(x, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}}))\,\mathrm{d}x.
\end{aligned}$$

□

Let's illustrate these definitions and Theorem 5.1 with a numerical example.

Example 5.9. Let us suppose that we have a population of six individuals with feature values $x^{(1)}, \ldots, x^{(6)}$. The expected profits are given in Table 5.1, and the resulting ranking from a model $\mathscr{M}$ is depicted in Fig. 5.2. Note that the ranking provided by $\mathscr{M}$ is based here on the expected profit, but in general this not always the case.

In this example, we use a prescription rate of $\rho = 0.5$. As such, we obtain a threshold $\tau(D_{\text{tr}})$ that prescribes the treatment $\boldsymbol{t} = 1$ to the three individuals with the highest score $\mathscr{M}(x, D_{\text{tr}})$ and prescribes $\boldsymbol{t} = 0$ to the three remaining individuals. We see from Fig. 5.2 that the treated individuals are those with features $x^{(2)}$, $x^{(6)}$ and $x^{(1)}$. The action profit

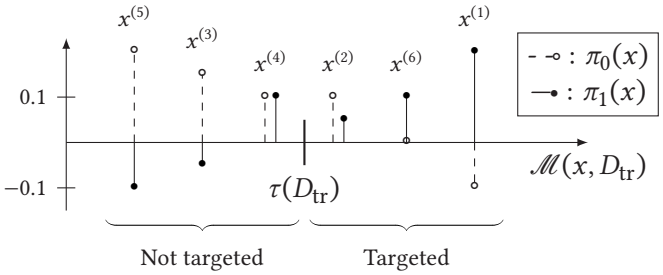


Figure 5.2 Ranking of the population of 6 individuals from Example 5.9. The profit of each individual (vertical axis) is plotted with respect to its score (horizontal axis). Individual action profits and individual baseline profits are represented, respectively, by full and hollow dots.

of the campaign (Definition 5.5) is

$$\begin{aligned}\Pi_1(\rho, D_{\text{tr}}) =& \rho\frac{1}{3}\left(\pi_1(x^{(2)}) + \pi_1(x^{(6)}) + \pi_1(x^{(1)})\right) \\ &+ (1-\rho)\frac{1}{3}\left(\pi_0(x^{(5)}) + \pi_0(x^{(3)}) + \pi_0(x^{(4)})\right) \\ =& \frac{1}{6}(0.05 + 0.1 + 0.2 + 0.2 + 0.15 + 0.1) = 0.133\ldots\end{aligned}$$

Performing this campaign produces an average profit of $0.133\ldots$ per individual. But this should be contrasted with the baseline profit (Definition 5.6):

$$\Pi_0 = \frac{1}{6}\sum_{i=1}^{6}\pi_0(x^{(i)}) = \frac{0.45}{6} = 0.075.$$

This represents a significant profit, even though no customers were contacted. The causal profit, in this example, is $\Pi(\rho, D_{\text{tr}}) = \Pi_1(\rho, D_{\text{tr}}) - \Pi_0 = 0.05833\ldots$, indicating that it is still beneficial to perform the campaign. From Theorem 5.1, we know that this quantity depends only on the causal profit of the targeted individuals, with indices 2, 6 and 1 in this example. We can verify this, by applying Eq. (5.14):

$$\Pi(\rho, D_{\text{tr}}) = \frac{1}{6}\left(\pi(x^{(2)}) + \pi(x^{(6)}) + \pi(x^{(1)})\right) = \frac{1}{6}(-0.05 + 0.1 + 0.3) = 0.05833\ldots$$

In this example, the prescription rate ρ can be adjusted to further increase the causal profit: with $\rho = 1/3$, only $x^{(1)}$ and $x^{(6)}$ would be targeted and this would result in a causal profit of $0.066\ldots$

5.2.3 Equivalence with the profit from Verbeke et al.

The profit measure developed in the previous section is mostly equivalent to the profit measure $\text{CP}(\tau, D_{\text{tr}})$ developed by Verbeke, Olaya, Guerry, et al. (2022), defined in Definition 3.5. The main technical difference is that our measure $\Pi(\rho, D_{\text{tr}})$ naturally accepts individual variations of the cost-benefit matrix $\text{CB}(x)$, that is, an instance-dependent cost-benefit matrix. Another advantage of our profit measure comes from Theorem 5.1, which shows that the profit measure can be expressed as $\overline{\Pi}(\rho) = \mathbb{E}_{\boldsymbol{x}}[\pi(\boldsymbol{x})P(\mathscr{M}(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}}))]$, highlighting the fact that the profit is determined by two terms: the individual profit $\pi(\boldsymbol{x})$, which is intrinsic to the population, and the (stochastic) estimator

$\mathscr{M}(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}})$. We discuss this aspect in detail in Section 5.3.1. We now show the equivalence of $\Pi(\rho, D_{\text{tr}})$ with Verbeke's measure.

Theorem 5.2. *Let* $\text{CB}(x)$ *be identical for all* x*, that is,* $\text{CB}(x) = \text{CB}$*. For any model* $\mathscr{M}$ *training on a set* D_{tr}*, and for any threshold* τ*, we have* $\text{CP}(\tau, D_{\text{tr}}) = \Pi(\rho, D_{\text{tr}})$ *with* $\rho = P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau)$*.*

Proof. We use the notation described in Section 3.1.3, where $F_{yt}^{D_{\text{tr}}}$ is the cumulative distribution function of the score from a model $\mathscr{M}$ trained on a data set D_{tr}, conditional on a particular realization of the potential outcome $\boldsymbol{y}_t = y$ (Eq. 3.22). Also, $\text{CF}(\tau, D_{\text{tr}})$ is the causal confusion matrix (Eqs. 3.23 and 3.25), $\text{E}(\tau, D_{\text{tr}})$ is the causal effect matrix (Eq. 3.24), and $\oplus$ denotes the sum of the components of the componentwise product of two matrices (Eq. 3.27). First, from Eq. (3.28), we develop

$$\text{CP}(\tau, D_{\text{tr}}) = \text{E}(\tau, D_{\text{tr}}) \oplus \text{CB} = \text{CF}(\tau, D_{\text{tr}}) \oplus \text{CB} - \text{CF}(\infty, D_{\text{tr}}) \oplus \text{CB}\,.$$

Let's expand the two terms of this difference separately. The first term is

$$\begin{aligned}\text{CF}(\tau, D_{\text{tr}}) \oplus \text{CB} =& (1 - S_0) F_{00}^{D_{\text{tr}}}(\tau)\, \text{CB}_{00} + S_0 F_{10}^{D_{\text{tr}}}(\tau)\, \text{CB}_{10} \\ &+ (1 - S_1)(1 - F_{01}^{D_{\text{tr}}}(\tau))\, \text{CB}_{01} + S_1 (1 - F_{11}^{D_{\text{tr}}}(\tau))\, \text{CB}_{11}\,.\end{aligned}$$

Let's focus on

$$\begin{aligned}(1 - S_0) F_{00}^{D_{\text{tr}}}(\tau) &= (1 - S_0) P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau \mid \boldsymbol{y}_0 = 0) \\ &= P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau, \boldsymbol{y}_0 = 0) \\ &= P(\boldsymbol{y}_0 = 0 \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau) P(\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau) \\ &= P(\boldsymbol{y}_0 = 0 \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau)(1 - \rho) \\ &= \mathbb{E}[(1 - S_0(\boldsymbol{x})) \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau](1 - \rho).\end{aligned}$$

Similarly, we can show that

$$\begin{aligned}S_0 F_{10}^{D_{\text{tr}}}(\tau) &= \mathbb{E}[S_0(\boldsymbol{x}) \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau](1 - \rho) \\ (1 - S_1)(1 - F_{01}^{D_{\text{tr}}}(\tau)) &= \mathbb{E}[(1 - S_1(\boldsymbol{x})) \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau]\rho \\ S_1 (1 - F_{11}^{D_{\text{tr}}}(\tau)) &= \mathbb{E}[S_1(\boldsymbol{x}) \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau)]\rho.\end{aligned}$$

Hence $\text{CF}(\tau, D_{\text{tr}}) \oplus \text{CB}$ can be expressed as

$$\begin{aligned}\text{CF}(\tau, D_{\text{tr}}) \oplus \text{CB} =& \mathbb{E}[(1 - S_0(\boldsymbol{x}))\, \text{CB}_{00} + S_0(\boldsymbol{x})\, \text{CB}_{10} \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau](1 - \rho) \\ &+ \mathbb{E}[(1 - S_1(\boldsymbol{x}))\, \text{CB}_{01} + S_1(\boldsymbol{x})\, \text{CB}_{11} \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau]\rho \\ =& \mathbb{E}[\pi_0(\boldsymbol{x}) \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) < \tau](1 - \rho) + \mathbb{E}[\pi_1(\boldsymbol{x}) \mid \mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) \geq \tau]\rho \\ =& \Pi_1(\rho, D_{\text{tr}}).\end{aligned}$$

We can use the linearity of the expected value operator to show

$$\begin{aligned}\text{CF}(\infty, D_{\text{tr}}) \oplus \text{CB} &= (1 - S_0)\, \text{CB}_{00} + S_0\, \text{CB}_{10} = (1 - \mathbb{E}[S_0(\boldsymbol{x})])\, \text{CB}_{00} + \mathbb{E}[S_0(\boldsymbol{x})]\, \text{CB}_{10} \\ &= \mathbb{E}[(1 - S_0(\boldsymbol{x}))\, \text{CB}_{00} + S_0(\boldsymbol{x})\, \text{CB}_{00}] = \mathbb{E}[\pi_0(\boldsymbol{x})] = \Pi_0.\end{aligned}$$

Wrapping up, we have

$$\text{CP}(\tau, D_{\text{tr}}) = \text{CF}(\tau, D_{\text{tr}}) \oplus \text{CB} - \text{CF}(\infty, D_{\text{tr}}) \oplus \text{CB} = \Pi_1(\rho, D_{\text{tr}}) - \Pi_0 = \Pi(\rho, D_{\text{tr}})$$

□

This theorem shows that two seemingly different definitions of the profit of a campaign lead to the same mathematical quantity. We hope that this firmly establishes the measure of profit as the most general measure of performance for uplift models. Given its generality, in particular stemming from the use of a cost-benefit matrix, the profit measure subsumes most other performance metrics developed in the literature (Fernández-Loria and Provost, 2022b; Gubela, Lessmann, and Jaroszewicz, 2020; Haupt and Lessmann, 2022), which are reviewed in Section 3.1.3.

5.2.4 Relationship with the uplift curve

The uplift curve, as defined in Definition 3.3, is widely used in the literature to evaluate the performance of uplift models (Gutierrez and Gérardy, 2016). While it has been used for more than 20 years (V. S. Y. Lo, 2002), its definition has always been accepted without a formal proof of its validity. Proving this validity requires a definition of the underlying objective of the campaign, and a demonstration of the correspondence between this objective and the uplift curve. Verbeke, Olaya, Berrevoets, et al. (2021) propose the causal profit as the definition of the underlying objective of the campaign, and we provide another definition of causal profit in Definition 5.8. In this section, we demonstrate that the uplift curve converges to the measure of profit as the number of samples in the test set increases. Importantly, this convergence is true only under a strong assumption about the cost-benefit matrix, which we call the *unitary value assumption*.

Definition 5.10 (Unitary value assumption)**.** The *unitary value assumption* posits that the cost-benefit matrix does not depend on $\boldsymbol{x}$ and is

$$\mathrm{CB}(x) = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \tag{5.18}$$

Intuitively, the unitary value assumption represents the case where all individuals have an equal value, only the value of the outcome $\boldsymbol{y}$ should be taken into account, and the treatment has no cost. These assumptions are rather restrictive, especially the last one. Definition 5.10 has the following corollary.

Result 5.3. *Under the unitary value assumption, the causal profit $\pi(x)$ is equal to the uplift $U(x)$.*

Proof. From Eq. (5.8), by replacing the values of $\mathrm{CB}(x)$, we obtain

$$\pi(x) = 1 - S_1(x) - (1 - S_0(x)) = S_0(x) - S_1(x) = U(x).$$

□

We are now ready to give the main theorem of this section.

Theorem 5.4. *Let D_{tr} be a training set of iid realizations of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$, and let $\boldsymbol{D}_{\mathrm{te}}$ be a test set of N tuples $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{t}^{(i)})$ iid to $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$, where $\boldsymbol{t}$ is randomized (see Definition 3.2). Let $\mathscr{M}$ be a model such that $\mathscr{M}(\boldsymbol{x}, D_{\mathrm{tr}})$ is a continuous random variable. Let $\rho \in (0, 1)$ be the prescription rate, and $k = \lceil N\rho \rceil$. Under the unitary value assumption, the value of the uplift curve at index k, noted $\mathrm{Uplift}(k, D_{\mathrm{tr}}, \boldsymbol{D}_{\mathrm{te}})$, converges to the causal profit of a campaign at the corresponding prescription rate ρ. This is expressed formally as*

$$\lim_{N \to \infty} \frac{1}{N} \mathrm{Uplift}(k, D_{\mathrm{tr}}, \boldsymbol{D}_{\mathrm{te}}) = \Pi(\rho, D_{\mathrm{tr}}) \quad \textit{in probability.} \tag{5.19}$$

Due to its length, the proof is given in Appendix C. This theorem establishes a theoretical foundation for the uplift curve, which is widely used in the uplift literature. It also shows that the unitary value assumption is necessary for the validity of the uplift curve, an assumption which has not been explicitly formulated before. We argue that this assumption, and in particular the implication that the treatment has no cost, might not hold for a large number of practical applications. For example, in churn prediction, different customers represent different values for the company. Marketing efforts should be focused on customers sensitive to the action, and who represent a large value. The profit measure takes these two aspects into account, whereas the uplift curve disregards the value of the customer. In Section 3.1.4, we will illustrate through simulated examples that the effectiveness of both uplift and predictive approaches is significantly influenced by the value of the cost-benefit matrix. This underscores the importance of assessing whether the unitary value assumption holds in practical scenarios.

5.2.5 Empirical profit curve

In this section, we propose an empirical estimator of the measure of profit as a generalization of the uplift curve to an arbitrary cost-benefit matrix. We use a notation similar to that in the definition of the uplift curve (Definition 3.3).

Definition 5.11 (Empirical profit curve)**.** Let $\boldsymbol{D}_{\text{te}} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{t}^{(i)})\}_{i=1}^{N}$ be a dataset of N tuples of random variables distributed identically to $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$, such that the treatment t is randomized. Let $\mathscr{M}$ be a model trained on a dataset D_{tr}, and let $\boldsymbol{D}_{\text{te}}$ be sorted in decreasing order according to $\mathscr{M}$: for any $i < j$, we have $\mathscr{M}(\boldsymbol{x}^{(i)}, D_{\text{tr}}) \geq \mathscr{M}(\boldsymbol{x}^{(j)}, D_{\text{tr}})$. The *empirical profit curve* is defined, for $k \in \{1, \dots, N\}$, as

$$\widetilde{\Pi}(k, D_{\text{tr}}, \boldsymbol{D}_{\text{te}}) = \left(\frac{\tilde{\boldsymbol{r}}_0(k)}{\boldsymbol{n}_0(k)} - \frac{\tilde{\boldsymbol{r}}_1(k)}{\boldsymbol{n}_1(k)} \right) k \tag{5.20}$$

with $n_t(k)$ defined as in Definition 3.3, and

$$\tilde{\boldsymbol{r}}_t(k) = \sum_{i=1}^{k} \left(\text{CB}_{0t}(x^{(i)})(1 - \boldsymbol{y}^{(i)}) + \text{CB}_{1t}(x^{(i)}) \boldsymbol{y}^{(i)} \right) \mathbb{I}[\boldsymbol{t}^{(i)} = t]. \tag{5.21}$$

It is easy to see that this curve is equivalent to the uplift curve under the unitary value assumption. Indeed, in this case, we have $\text{CB}_{00}^{(i)} = 1$ and $\text{CB}_{10}^{(i)} = 0$, thus $\tilde{\boldsymbol{r}}_0(k) = \boldsymbol{r}_0(k)$. Also, $\text{CB}_{01}^{(i)} = 1$ and $\text{CB}_{11}^{(i)} = 0$, thus $\tilde{\boldsymbol{r}}_1(k) = \boldsymbol{r}_1(k)$. Therefore, Eq. (5.20) reduces to Eq. (3.18). Theorem 5.4 shows the convergence of the uplift curve to the causal profit under the unitary value assumption; however, we do not have a proof of whether the empirical profit curve converges to the causal profit without the unitary value assumption.

An important advantage of this empirical estimator is that only the values of the cost-benefit matrix relating to the observed outcome need to be defined: for a given individual with $\boldsymbol{y}^{(i)} = y$ and $\boldsymbol{t}^{(i)} = t$, only the value $\text{CB}_{yt}^{(i)}$ needs to be known. This is especially interesting when the profit in counterfactual scenarios cannot be computed with certainty. For example, if we contacted ($\boldsymbol{t} = 1$) a customer $\boldsymbol{x} = x$ who did not churn ($\boldsymbol{y} = 0$), and we evaluate their associated benefit minus the cost of the marketing

action as, say, 42, then we can fill one cell of the cost-benefit matrix as follows:

$$\mathrm{CB}(x) = \begin{array}{c} \begin{array}{cc} t=0 & t=1 \end{array} \\ \left[\begin{array}{cc} ? & 42 \\ ? & ? \end{array}\right] \begin{array}{l} y=0 \\ y=1 \end{array} \end{array}$$

The interrogation marks denote unknown values. Even this partial information is sufficient to estimate the empirical profit curve, since only the term $\mathrm{CB}_{01}(x)$ will be used in Eq. (5.21) for this customer.

5.3 Uplift vs predictive approach

In this section, we discuss the different aspects that influence the profit measure, and we apply this discussion to the comparison between the uplift and predictive approaches. In Section 5.3.1, we discuss the profit measure in general terms, by giving some intuition on its formula. Then in Sections 5.3.2 and 5.3.3, we compare the predictive approaches in two different simulations of increasing complexity. The simpler simulation of Section 5.3.2 uses a normal distribution for the features and the noise terms, while the more complex simulation of Section 5.3.3 is based on a Dirichlet distribution, which provides more flexibility and allows us to draw more general conclusions.

5.3.1 Parameters influencing the profit measure

The primary contribution of our profit measure lies in its ability to clearly illustrate the impact of different components, thereby providing better insights into the performance of a model. Recall from Theorem 5.1 that the expected causal profit generated by a model $\mathscr{M}$ is expressed as

$$\overline{\Pi}(\rho) = \mathbb{E}_{\boldsymbol{x}}[\pi(\boldsymbol{x}) P(\mathscr{M}(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}}) \geq \tau(\boldsymbol{D}_{\mathrm{tr}}))]. \tag{5.22}$$

We discuss in turn each of the two components in the expected value operator: the individual profit $\pi(\boldsymbol{x})$ and the probability distribution of the scores $P(\mathscr{M}(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}}) \geq \tau(\boldsymbol{D}_{\mathrm{tr}}))$.

Individual profit $\pi(\boldsymbol{x})$

$\pi(\boldsymbol{x})$ represents the causal profit of an individual with features $\boldsymbol{x}$. This is specific to the customer population of the campaign and does not depend on $\mathscr{M}$. Going back to the definition of $\pi(\boldsymbol{x})$ (Definition 5.3), we have

$$\pi(\boldsymbol{x}) = \mathrm{CB}_{01}(\boldsymbol{x})(1 - S_1(\boldsymbol{x})) + \mathrm{CB}_{11}(\boldsymbol{x}) S_1(\boldsymbol{x}) - \mathrm{CB}_{00}(\boldsymbol{x})(1 - S_0(\boldsymbol{x})) - \mathrm{CB}_{10}(\boldsymbol{x}) S_0(\boldsymbol{x}).$$

It is clear that the cost-benefit matrix and the distribution of the conditional probabilities $S_0(\boldsymbol{x})$ and $S_1(\boldsymbol{x})$ have an influence on the distribution of $\pi(\boldsymbol{x})$. In particular, when the cost-benefit matrix $\mathrm{CB}(\boldsymbol{x})$ is the same for all values of $\boldsymbol{x}$, only the conditional probabilities $S_0(\boldsymbol{x})$ and $S_1(\boldsymbol{x})$ matter. We will use the mutual information, defined in Definition 2.9, to make the following discussion clearer. It is defined as, for $t = 0, 1$,

$$\begin{aligned} I(\boldsymbol{x}, \boldsymbol{y}_t) &= H(\boldsymbol{y}_t) - H(\boldsymbol{y}_t \mid \boldsymbol{x}) && (5.23) \\ &= -P(\boldsymbol{y}_t = 0) \log P(\boldsymbol{y}_t = 0) - P(\boldsymbol{y}_t = 1) \log P(\boldsymbol{y}_t = 1) \\ &\quad + \int_{\mathscr{X}} P(\boldsymbol{y}_t = 0 \mid x) \log P(\boldsymbol{y}_t = 0 \mid x) + P(\boldsymbol{y}_t = 1 \mid x) \log P(\boldsymbol{y}_t = 1 \mid x) \, \mathrm{d}x, && (5.24) \end{aligned}$$

which is a function of the prior probabilities $S_t = P(\boldsymbol{y}_t = 1)$ and the posterior probabilities $S_t(x) = P(\boldsymbol{y}_t = 1 \mid \boldsymbol{x} = x)$. Recall that both $H(\boldsymbol{y}_t)$ and $H(\boldsymbol{y}_t \mid \boldsymbol{x})$ are positive. Let us consider the distribution of $\boldsymbol{y}_t$ to be fixed and, hence, $H(\boldsymbol{y}_t)$ to be constant, leading us to consider three different cases:

- The mutual information $I(\boldsymbol{x}; \boldsymbol{y}_t)$ is maximum, corresponding to $H(\boldsymbol{y}_t \mid \boldsymbol{x}) = 0$. In this case, the scores $S_t(x)$ are either 0 or 1, and we can perfectly distinguish between the four counterfactual categories of individuals: persuadable, do-not-disturb, sure thing, and lost cause. The problem of optimal targeting is solved, because we can select only the persuadable individuals.

- The mutual information is zero, corresponding to the case where $H(\boldsymbol{y}_t \mid \boldsymbol{x}) = H(\boldsymbol{y}_t)$. In this case, the scores $S_t(\boldsymbol{x})$ are equal to S_t for all values of $\boldsymbol{x}$. Since we also assumed $\text{CB}(\boldsymbol{x})$ to be the same for all $\boldsymbol{x}$, the causal profit $\pi(\boldsymbol{x})$ is the same for all $\boldsymbol{x}$. All individuals have the same individual profit, and there is no point in trying to rank them. Any model $\mathscr{M}$ would generate the same benefit as a random model.

- The mutual information is between these two extremes. This corresponds to realistic scenarios. The causal profit is influenced by $\text{CB}(\boldsymbol{x})$ and the scores $S_t(\boldsymbol{x})$, but also by the prediction model $\mathscr{M}$, as discussed below.

In more general settings, where $\text{CB}(\boldsymbol{x})$ is not the same for all values of $\boldsymbol{x}$, it is more difficult to draw any conclusion on the distribution of $\pi(\boldsymbol{x})$ without any other assumption.

Probability distribution of the score

The second term in Eq. (5.22), $P(\mathscr{M}(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}}))$, represents the probability that a given individual with features $\boldsymbol{x}$ has a score higher than the threshold $\tau(\boldsymbol{D}_{\text{tr}})$, with respect to the distribution of training sets $\boldsymbol{D}_{\text{tr}}$. This quantifies the stochastic nature of the learning process: the predicted score varies depending on $\boldsymbol{D}_{\text{tr}}$, and thus the probability of this score being higher than a threshold varies depending on $\boldsymbol{D}_{\text{tr}}$ as well. To give some intuition, let us suppose that we have two individuals with features $\boldsymbol{x} = x_1$ and $\boldsymbol{x} = x_2$, such that $\pi(x_1) < \pi(x_2)$. A model aiming to rank the most profitable individuals first should rank x_2 before x_1. However, in practice, we only have the sampling distributions $\mathscr{M}(x_1, \boldsymbol{D}_{\text{tr}})$ and $\mathscr{M}(x_2, \boldsymbol{D}_{\text{tr}})$. These estimators can be characterized in terms of their bias and variances (see Definitions 2.22 and 2.23). On the one hand, a higher variance always increases the probability of misclassification, that is, the probability that $\mathscr{M}(x_1, \boldsymbol{D}_{\text{tr}}) > \mathscr{M}(x_2, \boldsymbol{D}_{\text{tr}})$. On the other hand, the bias might have a detrimental or positive effect, depending on its sign. If the bias on x_2 is much larger than that on x_1, then the profit estimates, although biased, will increase the probability of correct classification. Fernández-Loria and Provost (2022a) derived an analytical criterion to determine when a model has a higher risk of misclassification than another depending on their biases and variances.

Interaction of the two terms

The causal profit is determined not only by the two terms in Eq. (5.22) independently, but also by their interaction. To illustrate this, we highlight two important interactions:

- Let $\mathscr{M}_1$ and $\mathscr{M}_2$ be two different models that give a high score to different subpopulations, i.e., $P(\mathscr{M}_1(x, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}}))$ is high only for x in a set $\mathscr{X}_1$ and $P(\mathscr{M}_2(x, \boldsymbol{D}_{\text{tr}}) \geq \tau(\boldsymbol{D}_{\text{tr}}))$ is high only for x in a different set $\mathscr{X}_2$. Yet, if the expected profits $\pi(x)$ on $\mathscr{X}_1$ and $\mathscr{X}_2$ are similar, then the causal profit will be similar as well.

- Suppose that we have a vector of features $\boldsymbol{x}$ that is not very informative. In this case, the posterior probabilities $S_t(x)$ will be close to the prior probability S_t. This affects the distribution of $\pi(x)$, so that most individuals have a causal profit $\pi(x)$ close to $\mathbb{E}_{\boldsymbol{x}}[\pi(\boldsymbol{x})]$ (if $\text{CB}(x)$ does not vary too much across x as well). Since $\pi(x)$, which is estimated by $\mathscr{M}$, does not vary much between different values of $\boldsymbol{x}$, a slight estimation error by $\mathscr{M}$ can easily lead to a misclassification (i.e., ranking in the wrong order different individuals).

These two scenarios show that the causal profit is impacted by the model estimation error and the distribution of $\pi(x)$ in a non-trivial way. In the following sections, we will assess the impact of four components of the problem: the distribution of scores $S_0(x)$ and $S_1(x)$, the cost-benefit matrix $\text{CB}(x)$, the mutual information $I(\boldsymbol{x}; \boldsymbol{y}_t)$, and variance of the estimator $\mathscr{M}(x, \boldsymbol{D}_{\text{tr}})$. In terms of estimator bias, we will focus on the bias inherent in the uplift and predictive approaches. The uplift approach estimates $U(\boldsymbol{x})$, which is a potentially biased estimator of $\pi(x)$ (when the unitary value assumption does not hold, see Result 5.3), and the predictive approach estimates $S_0(x)$, which is definitely biased, since the impact of $S_1(x)$ is not taken into account.

5.3.2 Simulation study with normally-distributed features

In this section, we illustrate the results of the previous sections using a simple data-generating process. Although the results of this simulation might not generalize well to real-life situations, they nevertheless provide qualitative intuitions about the impact of various parameters such as the mutual information on the performance of the model. This also shows that the uplift approach is not always the best option even in very simple settings.

First, we give the mathematical definition of the data-generating process. Let $\boldsymbol{x}$ be a vector of n features $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]$, which are all independent random variables with a standard normal distribution $\mathcal{N}(0, 1)$. The binary potential outcomes $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are determined using a linear combination of $\boldsymbol{x}$ with coefficient vectors λ_0 and $\lambda_1 \in \mathbb{R}^n$ and thresholds $\eta_0, \eta_1 \in \mathbb{R}$. More precisely, the outcome $\boldsymbol{y}_t$, for $t = 0, 1$, is defined as $\boldsymbol{y}_t = \mathbb{I}[\lambda_t^T \boldsymbol{x} + \boldsymbol{\varepsilon} \geq \eta_t]$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1)$. Higher values of η_t lead to a lower probability that $\boldsymbol{y}_t = 1$. Furthermore, higher values in λ_0 and λ_1 induce a lower impact of the noise $\boldsymbol{\varepsilon}$ on the value of $\boldsymbol{y}_t$, hence the features $\boldsymbol{x}$ are more informative about the outcome $\boldsymbol{y}$ (the mutual information $I(\boldsymbol{x}; \boldsymbol{y}_t)$ is higher). Finally, the treatment indicator $\boldsymbol{t}$ is sampled from a Bernoulli distribution with parameter $p \in (0, 1)$, and the observed outcome $\boldsymbol{y}$ is defined accordingly as $\boldsymbol{y} = \boldsymbol{y}_0(1 - \boldsymbol{t}) + \boldsymbol{y}_1 \boldsymbol{t}$.

With this data-generating process, a training dataset D_{tr} of size N_{tr} and a test dataset D_{te} of size N_{te} are generated. The training dataset is used to train an uplift model and a predictive model, and their predictions on the test set are then compared in terms of the profit measure. In this simulation setting, we have access to the exact conditional probabilities $S_0(x)$ and $S_1(x)$ and to the cost-benefit matrix and, therefore, we know the exact value of the individual profit $\pi(x)$. The conditional probabilities $S_0(x)$ and $S_1(x)$

can easily be retrieved from the distribution of $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ as follows:

$$\begin{aligned} S_t(x) &= P(\boldsymbol{y}_t = 1 \mid \boldsymbol{x} = x) = P(\lambda_t^T x + \boldsymbol{\varepsilon} \geq \eta_t) \\ &= P(\boldsymbol{\varepsilon} \geq \eta_t - \lambda_t^T x) = 1 - \Phi(\eta_t - \lambda_t^T x) \\ &= \Phi(\lambda_t^T x - \eta_t) \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. To obtain a quantitative measure of performance without relying on the choice of the prescription rate ρ, we compute the area under the empirical profit curve:

$$\text{AUPC} = \frac{1}{N_{\text{te}}} \sum_{k=1}^{N_{\text{te}}} \tilde{\Pi}(k, D_{\text{tr}}, D_{\text{te}}).$$

In this simulation, we also want to assess the impact of the mutual information $I(\boldsymbol{x}; \boldsymbol{y}_0)$ and $I(\boldsymbol{x}; \boldsymbol{y}_1)$ on the performance of the uplift and predictive approaches. The mutual information is computed as $I(\boldsymbol{x}; \boldsymbol{y}_t) = H(\boldsymbol{y}_t) - H(\boldsymbol{y}_t \mid \boldsymbol{x})$. These two terms are themselves computed as

$$H(\boldsymbol{y}_t) = -S_t \log S_t - (1 - S_t) \log(1 - S_t) \tag{5.25}$$

$$H(\boldsymbol{y}_t \mid \boldsymbol{x}) = -\int_{\mathcal{X}} S_t(\boldsymbol{x}) \log S_t(\boldsymbol{x}) + (1 - S_t(\boldsymbol{x})) \log(1 - S_t(\boldsymbol{x}))\, \mathrm{d}x. \tag{5.26}$$

The term $S_t = P(\boldsymbol{y}_t = 1)$ in Eq. (5.25) is computed as $S_t = \mathbb{E}_{\boldsymbol{x}}[S_t(\boldsymbol{x})]$. Expected values over the distribution of $\boldsymbol{x}$ (for computing S_t or $H(\boldsymbol{y}_t \mid \boldsymbol{x})$ in Eq. 5.26) are computed by averaging on the test data set D_{te}.

The parameters of the experiment are set as follows. We use $n = 10$ features. The treatment rate is $p = 0.04$ to induce a greater variance for the uplift approach, by providing the estimator of $S_1(x)$ with fewer samples. We have $N_{\text{tr}} = 1000$ training samples (a low number of samples to induce a high estimator variance) and $N_{\text{te}} = 10000$ test samples (to reduce the variance of the empirical profit curve). The values in the vectors λ_0 and λ_1 are chosen randomly according to, respectively, $\mathcal{N}(1.2c, c^2)$ and $\mathcal{N}(c, c^2)$, where c is a scale parameter varying from 10^{-2} to 10 in different runs of the experiment. The thresholds are chosen manually to $\eta_0 = 1.12$ and $\eta_1 = 0.87$ to generate a distribution of potential outcomes close to $S_0 = 0.4$ and $S_1 = 0.4$. The predictive approach is the logistic regression model of the Scikit-learn Python package (Pedregosa et al., 2011) trained on the samples with $\boldsymbol{t} = 0$. The uplift approach is a T-learner, implemented as the difference between two logistic regression models (also from the Scikit-learn package) trained, respectively, on the control samples ($\boldsymbol{t} = 0$) and the target samples ($\boldsymbol{t} = 1$), with a regularization parameter $C = 10$. We assume a unitary cost-benefit matrix to evaluate the empirical profit curve. The experiment is repeated 100 times, sampling new values for λ_0 and λ_1 at each iteration.

Fig. 5.3 shows the performance of the uplift and predictive approaches as a function of the mutual information $I(\boldsymbol{x}; \boldsymbol{y}_0)$. Note that we report the mutual information $I(\boldsymbol{x}; \boldsymbol{y}_0)$ as a ratio between zero and its maximum value $H(\boldsymbol{y}_0)$. Due to the low number of treated samples in the training set ($p = 0.04$ and $N_{\text{tr}} = 1000$), the uplift approach suffers from a higher variance than the predictive approach, the latter using only the more numerous control samples. This leads to the predictive approach performing better in terms of area under the profit curve (AUPC) in the low information regime. However, as the features become more informative, the uplift approach starts to outperform the predictive approach.

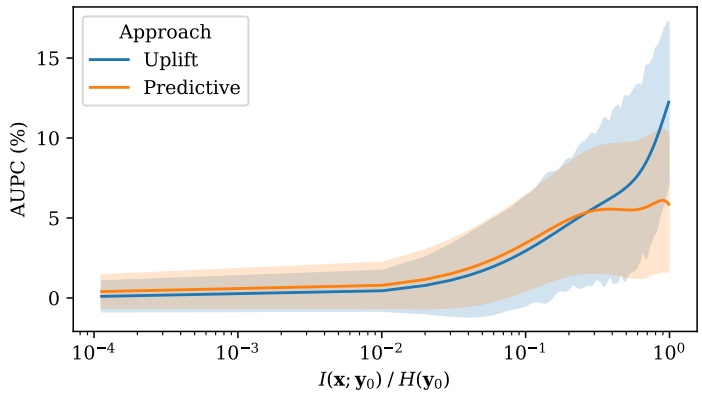


Figure 5.3 Performance of the uplift and predictive approach as a function of the proportion of mutual information between the features $\boldsymbol{x}$ and $\boldsymbol{y}_0$ in the simulation with normally distributed features. The lines represent the mean AUPC of all experiments with similar mutual information, which is determined by weighting the different experiments using a Gaussian kernel moving along the horizontal axis. The bands represent the 95% confidence interval assuming a normal distribution. While the uplift approach performs better when all the information is available, the low information regime is dominated by the predictive approach.

This simulated experiment shows that even in very simple settings, with normally distributed features and linear models, the uplift approach does not always provide the best performance. A low number of treated samples and relatively uninformative features can lead the predictive approach to outperform the uplift approach.

5.3.3 Simulation study with a Dirichlet distribution

In the simulation of Section 5.3.2, the distribution of potential outcomes, the mutual information $I(\boldsymbol{x}; \boldsymbol{y}_t)$, and the variance of the uplift and predictive models had to be calculated after the fact, because they were influenced by the simulation parameters in a complex way. Moreover, the choice of the distribution of the covariates $\boldsymbol{x}$, the functional dependency between $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{t}$, and the choice of the uplift models have a significant influence on the results of our analysis. In this section, we develop a more complex simulation that allows more flexibility and whose characteristics are easier to compute than the simple data generation process of Section 5.3.2. In particular, the distribution of potential outcomes, the mutual information, and the variance of the estimators can be specified directly and independently of each other.

The simulation of this section is based on the insight that the process of training an uplift model and predicting the uplift of samples from a test set results in a distribution of scores, and that only these scores affect the result of our analysis. The exact distribution of $\boldsymbol{x}$ or its dependency with $\boldsymbol{y}$ and $\boldsymbol{t}$ has no impact beyond the predicted scores. Therefore, if we can generate scores according to a distribution similar to those predicted by an uplift model, then modeling the distribution of $\boldsymbol{x}$ in our simulation is unnecessary. More precisely, we generate conditional probabilities $S_0(\boldsymbol{x})$ and $S_1(\boldsymbol{x})$ without explicitly modeling $\boldsymbol{x}$. Afterward, we can emulate the scores given by the uplift and predictive approaches $\mathscr{M}_u(x, D_{\text{tr}})$ and $\mathscr{M}_p(x, D_{\text{tr}})$ as noisy estimates of $U(x)$ and $S_0(x)$ (which are easily computed from the joint distribution of $\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x}$). Since we

sample the distribution $\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x}$ and the estimated scores but we do not sample $\boldsymbol{x}$ directly, we will denote individual samples with superscript i rather than as functions of x.

We use the notation for the joint distribution of the potential outcomes defined in Section 3.2.1.

$$\alpha = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0) \qquad \beta = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0) \tag{5.27}$$

$$\gamma = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1) \qquad \delta = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1). \tag{5.28}$$

We also note $\mu = [\alpha, \beta, \gamma, \delta]$. From this notation, one can easily show that

$$S_0 = \beta + \delta \quad \text{and} \quad S_1 = \gamma + \delta. \tag{5.29}$$

The sampling process is as follows:

1. First, we generate N independent samples $\{(\boldsymbol{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)}, \boldsymbol{\gamma}^{(i)}, \boldsymbol{\delta}^{(i)})\}_{i=1}^{N}$ according to a Dirichlet distribution with parameter vector $m = [a, b, c, d]$:

$$(\boldsymbol{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)}, \boldsymbol{\gamma}^{(i)}, \boldsymbol{\delta}^{(i)}) \sim \mathrm{Dir}(a, b, c, d). \tag{5.30}$$

   They model the joint probabilities of the potential outcomes as in Eqs. (5.27) to (5.28), but at the individual level for each individual i. The Dirichlet distribution is a natural candidate to sample numbers in a probability simplex (i.e., such that $\boldsymbol{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)}, \boldsymbol{\gamma}^{(i)}$ and $\boldsymbol{\delta}^{(i)}$ are all positive and sum up to 1), since it is the conjugate prior of the multinomial distribution (Lin, 2016). This distribution has a number of properties that make it particularly suited to our setting, which we demonstrate in Appendix D.

2. Then, we derive the value of the conditional probabilities $\boldsymbol{S}_0^{(i)}$ and $\boldsymbol{S}_1^{(i)}$ with the identities $\boldsymbol{S}_0^{(i)} = \boldsymbol{\beta}^{(i)} + \boldsymbol{\delta}^{(i)}$ and $\boldsymbol{S}_1^{(i)} = \boldsymbol{\gamma}^{(i)} + \boldsymbol{\delta}^{(i)}$ from Eq. (5.29). This results in $\boldsymbol{S}_0^{(i)}$ and $\boldsymbol{S}_1^{(i)}$ having marginal beta distributions, which is convenient since the beta distribution is the conjugate prior of the Bernoulli distribution, and $\boldsymbol{y}_t^{(i)}$ follows a Bernoulli distribution $\mathrm{Bern}(\boldsymbol{S}_t^{(i)})$. This procedure is based on the bivariate beta distribution of Olkin and Trikalinos (2015). We note the two steps

$$(\boldsymbol{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)}, \boldsymbol{\gamma}^{(i)}, \boldsymbol{\delta}^{(i)}) \sim \mathrm{Dir}(a, b, c, d) \tag{5.31}$$

$$\boldsymbol{S}_0^{(i)} = \boldsymbol{\beta}^{(i)} + \boldsymbol{\delta}^{(i)} \tag{5.32}$$

$$\boldsymbol{S}_1^{(i)} = \boldsymbol{\gamma}^{(i)} + \boldsymbol{\delta}^{(i)}. \tag{5.33}$$

   as

$$(\boldsymbol{S}_0^{(i)}, \boldsymbol{S}_1^{(i)}) \sim \mathrm{BB}(a, b, c, d). \tag{5.34}$$

3. The score of the predictive approach is emulated as a binomial distribution with parameters $\boldsymbol{S}_0^{(i)}$ and n_p normalized to have a maximum value of 1:

$$\mathscr{M}_p^{(i)} \sim \frac{1}{n_p} \mathrm{Bin}(\boldsymbol{S}_0^{(i)}, n_p). \tag{5.35}$$

   This emulates the behavior of tree-based models that estimate the conditional probability of the outcome by computing the ratio of positive outcomes among the samples close to the input sample in the feature space. Larger values of n_p induce a lower estimator variance.

4. Similarly, the score of the uplift approach is emulated as the difference between two normalized binomial distributions, with parameters $S_0^{(i)}, n_u$ and $S_1^{(i)}, n_u$:

$$\mathscr{M}_u^{(i)} \sim \frac{1}{n_u}\text{Bin}(S_0^{(i)}, n_u) - \frac{1}{n_u}\text{Bin}(S_1^{(i)}, n_u). \tag{5.36}$$

5. Finally, the binary outcomes $\boldsymbol{y}_0^{(i)}, \boldsymbol{y}_1^{(i)}$ are sampled according to a categorical distribution of probability vector $\boldsymbol{\mu}^{(i)} = [\boldsymbol{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)}, \boldsymbol{\gamma}^{(i)}, \boldsymbol{\delta}^{(i)}]$:

$$P(\boldsymbol{y}_0^{(i)} = 0, \boldsymbol{y}_1^{(i)} = 0 \mid \boldsymbol{\mu}^{(i)} = \mu^{(i)}) = \alpha^{(i)}, \tag{5.37}$$
$$P(\boldsymbol{y}_0^{(i)} = 1, \boldsymbol{y}_1^{(i)} = 0 \mid \boldsymbol{\mu}^{(i)} = \mu^{(i)}) = \beta^{(i)}, \tag{5.38}$$
$$P(\boldsymbol{y}_0^{(i)} = 0, \boldsymbol{y}_1^{(i)} = 1 \mid \boldsymbol{\mu}^{(i)} = \mu^{(i)}) = \gamma^{(i)}, \tag{5.39}$$
$$P(\boldsymbol{y}_0^{(i)} = 1, \boldsymbol{y}_1^{(i)} = 1 \mid \boldsymbol{\mu}^{(i)} = \mu^{(i)}) = \delta^{(i)}. \tag{5.40}$$

This sampling process is particularly convenient because most of its parameters are directly related to the quantities we are interested in:

- The parameters a, b, c, d are proportional to the probabilities of the distribution of the potential outcomes, $\alpha, \beta, \gamma, \delta$ (see Eqs. (5.27) to (5.28)). For example, using the moments of the Dirichlet distribution described in Section 2.1.5, we have

$$\begin{aligned}\beta = P(\boldsymbol{y}_0^{(i)} = 1, \boldsymbol{y}_1^{(i)} = 0) &= \int_{\mathcal{S}} P(\boldsymbol{y}_0^{(i)} = 1, \boldsymbol{y}_1^{(i)} = 0 \mid \boldsymbol{\mu}^{(i)} = \mu^{(i)}) f_{\boldsymbol{\mu}^{(i)}}(\mu^{(i)})\, \mathrm{d}\mu^{(i)} \\ &= \int_{\mathcal{S}} \beta^{(i)} f_{\boldsymbol{\mu}^{(i)}}(\mu^{(i)})\, \mathrm{d}\mu^{(i)} = \mathbb{E}[\boldsymbol{\beta}^{(i)}] = \frac{b}{M}\end{aligned}$$

  where $M = a + b + c + d$, $\mathcal{S}$ is the unit 4-simplex, and $f_{\boldsymbol{\mu}^{(i)}}(\cdot)$ is the probability density function of $\boldsymbol{\mu}^{(i)}$.

- The mutual information $H(\boldsymbol{y}_0^{(i)}, \boldsymbol{y}_1^{(i)}; \boldsymbol{\mu}^{(i)})$ decreases as the sum of the parameters $M = a + b + c + d$ increases, as shown in Fig. 5.4. The analytical relationship between M and the mutual information is given in Appendix D. This mutual information is equivalent in our simulation setting to the mutual information between the features and the outcomes $H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x})$ discussed in Section 5.3.1, which has an important impact on the measure of profit.

- The variance of the estimators $\mathscr{M}_u$ and $\mathscr{M}_p$ can be easily adjusted independently of all other parameters by choosing the appropriate value of n_u and n_p. In fact, we have

$$\text{Var}(\mathscr{M}_u^{(i)}) = \frac{1}{n_u}\left(S_0^{(i)}(1 - S_0^{(i)}) + S_1^{(i)}(1 - S_1^{(i)})\right), \tag{5.41}$$
$$\text{Var}(\mathscr{M}_p^{(i)}) = \frac{1}{n_p}\left(S_0^{(i)}(1 - S_0^{(i)})\right). \tag{5.42}$$

  Note that in practice, the variance of the uplift and predictive approaches might be similar since both are typically trained on the same dataset.

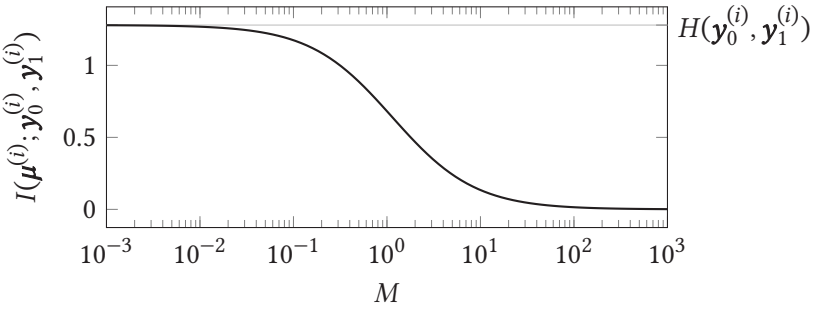


Figure 5.4 Mutual information as a function of $M = a+b+c+d$, for a fixed distribution of potential outcomes $[\alpha, \beta, \gamma, \delta] = [0.4, 0.3, 0.2, 0.1]$.

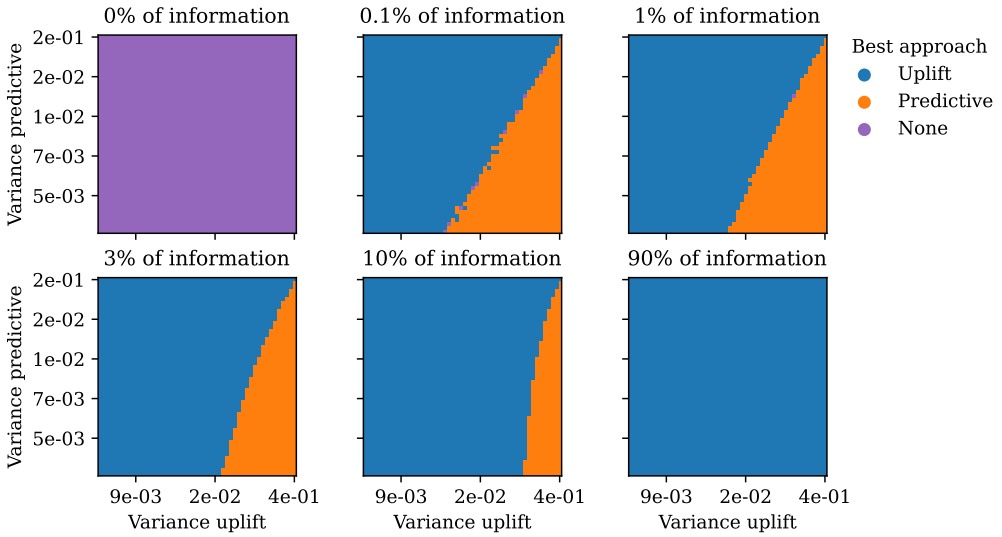


Figure 5.5 Best approach as a function of the estimator variance for different levels of information between the features and the outcome. Here, $\alpha = 0.6, \beta = 0.2, \gamma = 0.1$ and $\delta = 0.1$, and we use the unitary cost-benefit matrix.

The sampling process is repeated for different values of the parameters n_u and n_p (ranging both independently from 1 to 50), which influence the variance of the uplift and predictive approaches. Figures 5.5 to 5.7 show which of the uplift and predictive approaches performs the best for each value of n_u and n_p, and for different values of other parameters.

Figure 5.5 varies the mutual information between the features and the outcome. In the first panel, where no information is available, both approaches perform similarly, since they are equivalent to random selection. However, we see in the remaining panels that a lower mutual information increases the proportion of cases where the predictive approach performs better.

Figure 5.6 varies the cost-benefit matrix used to compute the profit measure. The third panel represents the unitary value assumption. The remaining panels highlight various scenarios that favor either the predictive approach or the uplift approach, emphasizing the significant effect of the cost-benefit matrix on their relative performance. We hypothesize that the large difference between, for example, the first and second panels is due to the cost of treatment, that is, the difference between CB_{00} and CB_{10}. It

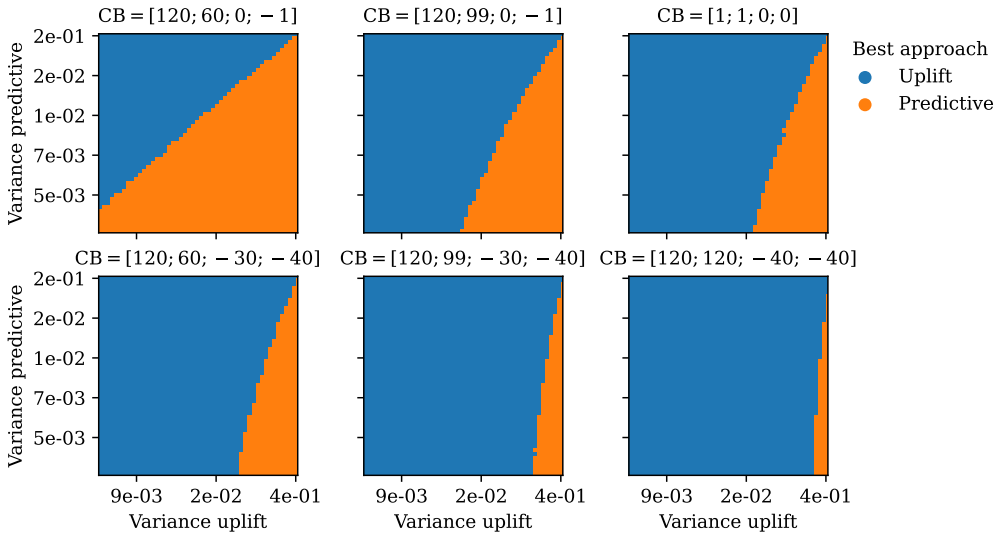


Figure 5.6 Best approach as a function of the estimator variance for different values of CB(x). Here, we have 1% of mutual information between the features and the outcomes. The probabilities of potential outcomes are $\alpha = 0.6, \beta = 0.2, \gamma = 0.1$ and $\delta = 0.1$.

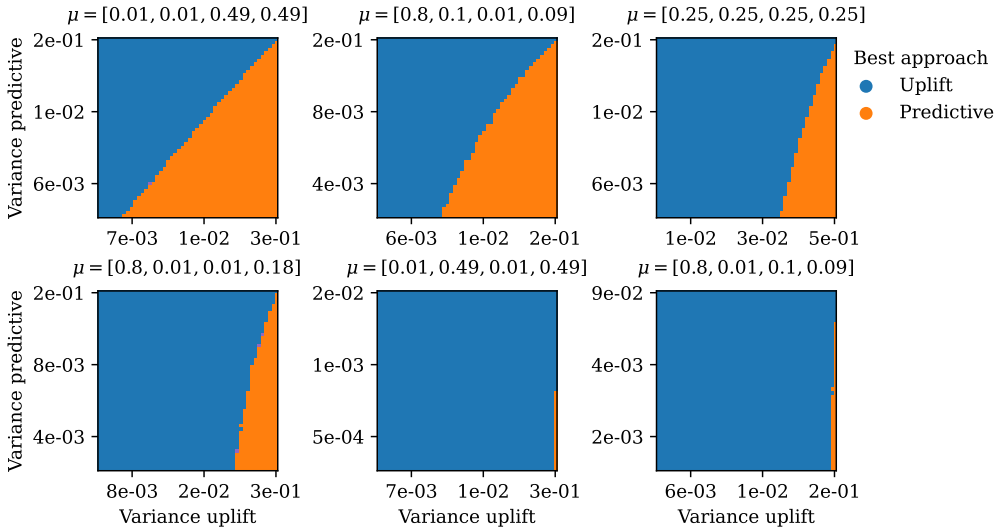


Figure 5.7 Best approach as a function of the estimator variance for different values of μ. Here, we have 1% of mutual information between the features and the outcomes.

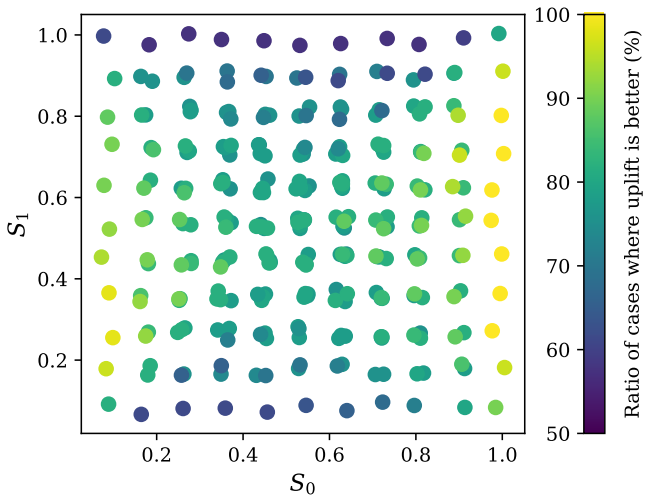


Figure 5.8 Ratio of runs where the uplift approach is better, where each dot represents several experiments with different estimator variances but a fixed distribution of potential outcomes. The dots are arranged on the space of values of S_0 and S_1 (recall from Eq. (5.29) that $S_0 = \beta + \delta$ and $S_1 = \gamma + \delta$). Jitter is added to help distinguish overlapping dots.

is equal to 60 in the first panel and 21 in the second.

Figure 5.7 varies the joint distribution of the potential outcomes between the different panels. We can observe that it has an important impact on the performance of the uplift and predictive approaches. In the first panel, where $\gamma = 0.49$ and $\delta = 0.49$, which indicates a negative causal effect and a high probability S_1, the predictive approach performs better than the uplift approach when its variance is lower. On the other hand in the fifth panel, where $\beta = 0.49$ and $\delta = 0.49$, the situation is the opposite: the causal effect is high, hence large benefits can be generated by selecting individuals with the uplift approach in almost every case. The other panels indicate intermediate situations between these two extremes.

To have a more comprehensive understanding of the impact of the distribution of potential outcomes, we repeat the experiment for different values of μ chosen uniformly. Then, for each value of μ, we repeat the experiment by varying the variance of both approaches. Figure 5.8 depicts the ratio of cases where the uplift approach outperforms the predictive approach for each value of μ. The data points are organized according to S_0 and S_1 (recall that $S_0 = \beta + \delta$ and $S_1 = \gamma + \delta$ from Eq. 5.29). We observe that, for a given marginal distribution of potential outcomes (characterized by S_0 and S_1), the joint distribution of potential outcomes (characterized by μ) has a relatively minor impact. In particular, when S_1 is close to zero or one while S_0 is not, the uplift approach consistently outperforms the predictive approach. On the contrary, when S_0 is close to zero or one while S_1 is not, the predictive approach becomes the favorable choice in approximately 50% of the cases.

5.4 Discussion and limitations

The results presented in this study shed light on the performance of the uplift and predictive approaches under various parameter settings. Our findings illustrate the crucial

role of factors such as estimator variance, mutual information, cost-benefit matrix, and the distribution of potential outcomes in determining the performance dynamics between these approaches.

One of the key findings of this study is the crucial role that the variance of the uplift and predictive approaches plays in determining their relative performance. The results consistently show that, in almost all settings, one approach will outperform the other if its variance is significantly lower. This highlights the importance of carefully considering the higher variance of the uplift approach compared to the predictive approach.

We summarize our findings as follows. The uplift approach should be preferred when

- The outcome is easy to predict from the features, that is, there is a high mutual information between the outcome and the features (see Fig. 5.5);
- The probability of the outcome in the control group (S_0) is close to zero or one, while the probability of the outcome in the target group (S_1) is not (see Fig. 5.8);
- The uplift approach has a lower or the same variance as the predictive approach (see Figs. 5.5 to 5.7).

On the other hand, the predictive approach is more effective when its variance is low enough with respect to the uplift approach and one of the following conditions is satisfied:

- When the mutual information is low (see Fig. 5.5);
- The probability of the outcome in the target group (S_1) is close to zero or one, while the probability of the outcome in the control group (S_0) is not (see Fig. 5.8);
- The treatment has a significant cost (see Fig. 5.6).

The condition of S_0 being close to zero or one was already noted in articles comparing the uplift and predictive approaches (Fernández-Loria and Provost, 2022a,b), but this simulation provides a more systematic illustration. In particular, we observe a symmetric condition where the uplift approach is better when the roles of S_0 and S_1 are swapped (see Fig. 5.8).

Our analysis is limited by several factors. First, we assumed that the cost-benefit matrix does not vary between individuals, which could have an important impact on our results. Second, the uplift and predictive approaches have no bias with respect to the quantity they aim to estimate. Machine learning estimators might be biased, which could further impact their performance. Finally, the choice of modeling the joint distribution of the potential outcomes from a Dirichlet distribution, as well as simulating the estimators from binomial distributions, is obviously a limiting factor. We consider this limitation to be not critical because the space of distributions that can be simulated this way is large enough to represent real-world distributions.

5.5 Conclusion

In this chapter, we investigated the effectiveness of uplift modeling compared to the classical predictive approach. To perform this comparison from a sound theoretical

basis, we proposed a new formulation of the measure of profit. It emphasizes individual cost sensitivity and the stochastic nature of the underlying machine learning model. We showed the equivalence of the measure of profit to a preexisting definition in the literature, and the convergence of the uplift curve to the measure of profit. We highlighted the strict conditions necessary for the validity of the uplift curve for performance evaluation.

The variance of the estimator plays a crucial role in the performance of the uplift and predictive approaches: in almost every case, the predictive approach is preferred if the variance of the uplift approach is high enough. This result is critical because the uplift approach typically exhibits higher variance than the predictive approach. The higher variance arises from the fact that the uplift approach estimates the difference between two probabilities, introducing additional uncertainty into the estimation process. We also showed that the mutual information between input features and the outcome, as well as the distribution of the potential outcomes, plays an important role in determining when the predictive approach outperforms uplift modeling. Lastly, a proper definition of the cost-benefit matrix is essential, as the performance of both approaches can vary widely depending on it.

Overall, this chapter provides firm theoretical foundations for uplift modeling and answers the question of when uplift modeling outperforms predictive modeling. Our findings have important implications for practitioners in various domains, such as marketing, telecommunications, healthcare, and finance, who rely on machine learning techniques for decision-making. In particular, we suggest practitioners to carefully estimate the various parameters highlighted above to evaluate whether the uplift approach is likely to bring benefits.

6

Counterfactual identification

Some of the results presented in this chapter have been published or are expected to be published in the following articles:

- Théo Verhelst, Denis Mercier, et al. (Mar. 2023b). "Partial counterfactual identification and uplift modeling: theoretical results and real-world assessment". en. In: *Machine Learning*. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-023-06317-w. URL: https://link.springer.com/10.1007/s10994-023-06317-w (visited on 05/03/2023)
- Théo Verhelst and Gianluca Bontempi (2024). "Identifying counterfactual probabilities using bivariate distributions and uplift modeling". In: *to be submitted*
- Théo Verhelst, Mercier Denis, et al. (2024). "Customer segmentation from counterfactual probabilities: new insights for the telecom sector". In: *to be submitted*

Counterfactual statements (or counterfactuals for short) concern the potential of events in situations different from the actual state of the world, such as "Would this customer have stayed if we did not call them?". We introduced the concept of counterfactuals and its place in Pearl's causal hierarchy in Section 2.2.1. In the setting of customer churn, we are particularly interested in classifying customers into four categories, as presented in Sections 3.1.1 and 3.2.1: *persuadable*, *sure thing*, *lost cause* and *do-not-disturb* customers. The probability of counterfactual statements cannot be estimated purely from data without any assumption; however, it is possible to bound these probabilities. This task is called *partial counterfactual identification*, which was first addressed by Tian and Pearl (2000) and more recently by Mueller, A. Li, and Pearl (2021) and J. Zhang, Tian, and Bareinboim (2022). We describe these works in Section 3.2.3.

These existing works on partial counterfactual identification make structural assumptions on the causal model to derive bounds whose estimate requires a combination of experimental and observational data. In this chapter, we propose original bounds and point estimators on the probability of counterfactuals based on the uplift terms. The originality of our approach stems from the fact that they depend on terms (like uplift) for which a wide range of estimators already exist in the literature. This is of particular interest in big data applications, where structural assumptions are hard to validate, but a large number of observations about individual descriptors (covariates) and past behavior are available.

The main contributions of this chapter are as follows:

- A set of original bounds on the probability of counterfactuals, expressed in terms of uplift terms (Section 6.2).
- A point estimator of the counterfactual probabilities based on the conditional independence assumption (Section 6.3).
- A Bayesian posterior distribution on the counterfactual probabilities based on a bivariate beta distribution, and three variations of this approach that make less restrictive assumptions at the cost of higher computational complexity (Section 6.4).[1]
- An assessment with two different simulations of the proposed counterfactual estimators (Section 6.5).[1]
- An assessment of the proposed counterfactual estimators with a dataset of customer churn campaigns from Orange Belgium and a discussion of the potential benefits suggested by the results (Section 6.6).
- A characterization of the different types of customer using our counterfactual estimators and other customer descriptive features, giving new insights on the reaction of customers to retention efforts (Section 6.6.3).[1]

The rest of this chapter is organized as follows. In Section 6.1, we give a formal description of the problem addressed in this chapter. In Sections 6.2 and 6.3, we derive bounds and point estimates on the probability of counterfactuals. We describe the estimators based on a bivariate beta distribution in Section 6.4. We analyze the behavior of these estimators under various conditions with simulated examples in Section 6.5. We apply our estimator to a real-world dataset from Orange Belgium and perform various analyses from the estimated distribution of counterfactuals in Section 6.6. We discuss our results in Section 6.7 and give our conclusions in Section 6.8.

6.1 Problem setting

The mathematical notation used throughout this chapter is defined in Sections 3.1.1 and 3.2.1, which we briefly summarize. The random variable $\boldsymbol{y}$ denotes the binary outcome of interest, $\boldsymbol{t}$ represents the binary treatment, and $\boldsymbol{x}$ is a random vector of descriptive features. The domains of these variables are, respectively, $\mathcal{Y} = \{0, 1\}$, $\mathcal{T} = \{0, 1\}$, and $\mathcal{X} \subseteq \mathbb{R}^n$. The feature vector $\boldsymbol{x}$ is continuous and characterized by a probability density function $f_{\boldsymbol{x}}$.

Suppose that we observe $\boldsymbol{y} = 1$ after having assigned the treatment $\boldsymbol{t} = 0$ to a given individual, leading to the realization of the potential outcome $\boldsymbol{y}_0 = 1$. Although we do not know the value of the counterfactual outcome $\boldsymbol{y}_1$, we can reason about it. If $\boldsymbol{y}_1 = 0$, the treatment would have a causal impact on the outcome, since the outcome $\boldsymbol{y}$ changes by intervening on $\boldsymbol{t}$. Otherwise, if $\boldsymbol{y}_1 = 1$, the treatment would not have a causal influence on the outcome of this individual. More generally, the joint values of $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ define four different counterfactual expressions, summarized in Table 6.1. Our

[1] These contributions are part of two unpublished articles mentioned at the beginning of this chapter, which we intend to submit for publication early 2024.

Table 6.1 The four categories of customers for churn prevention in terms of counterfactual outcomes.

	$\boldsymbol{y}_0 = 0$	$\boldsymbol{y}_0 = 1$
$\boldsymbol{y}_1 = 0$	Sure thing	Persuadable
$\boldsymbol{y}_1 = 1$	Do-not-disturb	Lost cause

objective in this chapter is the estimation of the probability of these four counterfactual expressions, noted

$$\alpha = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0) = P(\text{sure thing}) \tag{6.1}$$
$$\beta = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0) = P(\text{persuadable}) \tag{6.2}$$
$$\gamma = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1) = P(\text{do-not-disturb}) \tag{6.3}$$
$$\delta = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1) = P(\text{lost cause}). \tag{6.4}$$

From which we can derive

$$S_0 = P(\boldsymbol{y}_0 = 1) = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0) + P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1) = \beta + \delta \tag{6.5}$$
$$S_1 = P(\boldsymbol{y}_1 = 1) = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1) + P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1) = \gamma + \delta. \tag{6.6}$$

The probabilities in Eqs. (6.1) to (6.4) represent the probability of *any* individual, picked at random in the population, to correspond to one of the four categories. We name these probabilities *population-level* counterfactuals. Given the size of the population, this also indicates the expected number of individuals in each category. In practice, we are also interested in estimating the probability that a *specific* individual belongs to each category. This individual is described by the realization of the random vector $\boldsymbol{x} = x$, and the corresponding counterfactual probabilities are noted

$$\alpha(x) = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x) = P(\text{surething} \mid \boldsymbol{x} = x) \tag{6.7}$$
$$\beta(x) = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x) = P(\text{persuadable} \mid \boldsymbol{x} = x) \tag{6.8}$$
$$\gamma(x) = P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x) = P(\text{do} - \text{not} - \text{disturb} \mid \boldsymbol{x} = x) \tag{6.9}$$
$$\delta(x) = P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x) = P(\text{lostcause} \mid \boldsymbol{x} = x). \tag{6.10}$$

These probabilities are called *individual-level* counterfactuals.

Contrary to most of the literature on counterfactual identification presented in Sections 3.2.2 and 3.2.3, we do not make any assumption on the causal graph G of the underlying causal model. Instead, we only assume to have access to estimators of the probabilities $S_t = P(\boldsymbol{y}_t = 1)$ and $S_t(x) = P(\boldsymbol{y}_t = 1 \mid \boldsymbol{x} = x)$ for $t = 0, 1$. The probability S_t can easily be estimated from a dataset $D = \left\{(x^{(i)}, y^{(i)}, t^{(i)})\right\}_{i=1}^{N}$ of N iid realizations of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$ under the assumption of unconfoundedness, defined in Definition 3.2. Unconfoundedness implies that the conditional and the interventional distributions are identical, that is,

$$S_t = P(\boldsymbol{y}_t = 1) = P(\boldsymbol{y} = 1 \mid \boldsymbol{t} = t), \tag{6.11}$$

and thus we can use the maximum likelihood estimator

$$S_t \approx \frac{\sum_{i=1}^{N} \mathbb{I}[y^{(i)} = 1 \text{ and } t^{(i)} = t]}{\sum_{i=1}^{N} \mathbb{I}[t^{(i)} = t]}. \tag{6.12}$$

Similarly, under unconfoundedness, the interventional probability $S_t(x)$ is equal to the observational probability,

$$S_t(x) = P(\boldsymbol{y}_t = 1 \mid \boldsymbol{x} = x) = P(\boldsymbol{y} = 1 \mid \boldsymbol{t} = t, \boldsymbol{x} = x), \tag{6.13}$$

and an uplift model can be used to estimate $P(\boldsymbol{y} = 1 \mid \boldsymbol{t} = t, \boldsymbol{x} = x)$ for any t and x. Randomization of treatment guarantees unconfoundedness, but, in the absence of randomized data, an adjustment set (i.e., satisfying the *back-door criterion* as described by Pearl (2009, Def. 3.3.1)) and a suitable learning algorithm can enable an unbiased estimation of $S_0(x)$ and $S_1(x)$. This is possible, for example, using the X-learner strategy (see Section 3.1.2) with propensity scores (Künzel et al., 2019) and, more recently, with double machine learning estimators (Jung, Tian, and Bareinboim, 2021).

6.2 Bounds on the probability of counterfactuals

As presented in Section 3.2.3, bounds on the probability of counterfactuals have first been derived in Tian and Pearl (2000), where the authors focus on $P(\boldsymbol{y}_0 = 0 \mid \boldsymbol{t} = 1, \boldsymbol{y} = 1)$, $P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{t} = 0, \boldsymbol{y} = 0)$, and $P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1)$ (the latter is denoted γ in Eq. 6.3) under various assumptions. These bounds derive from the classical Fréchet bounds (Fréchet, 1935) stating that for any pair of events A and B,

$$\max\{0, P(A) + P(B) - 1\} \leq P(A, B) \leq \min\{P(A), P(B)\}. \tag{6.14}$$

For instance, by replacing A with $\boldsymbol{y}_0 = 0$ and B with $\boldsymbol{y}_1 = 1$, it is easy to derive the bounds in Eq. (3.35) of Section 3.2.3. Tighter bounds on counterfactual probabilities are derived in (Mueller, A. Li, and Pearl, 2021; J. Zhang, Tian, and Bareinboim, 2022) by making structural assumptions on the causal directed acyclic graph (DAG).

In this chapter, we focus on a setting where (i) no structural assumptions may be made (besides unconfoundedness) and (ii) an estimation of the uplift is possible on the basis of historical data. For this reason, we derive a set of original bounds that depend on the conditional probabilities terms $S_0(x) = P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x)$ and $S_1(x) = P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x)$.

Our derivation consists in first generalizing the bounds on γ by Tian and Pearl (2000) to all four counterfactual probabilities, by substituting A with $\boldsymbol{y}_0 = 0$ or $\boldsymbol{y}_0 = 1$, and B with $\boldsymbol{y}_1 = 0$ or $\boldsymbol{y}_1 = 1$ in Eq. (6.14):

$$\max\{0, P(\boldsymbol{y}_0 = 0) - P(\boldsymbol{y}_1 = 1)\} \leq \alpha \leq \min\{P(\boldsymbol{y}_0 = 0), P(\boldsymbol{y}_1 = 0)\} \tag{6.15}$$
$$\max\{0, P(\boldsymbol{y}_0 = 1) - P(\boldsymbol{y}_1 = 1)\} \leq \beta \leq \min\{P(\boldsymbol{y}_0 = 1), P(\boldsymbol{y}_1 = 0)\} \tag{6.16}$$
$$\max\{0, P(\boldsymbol{y}_1 = 1) - P(\boldsymbol{y}_0 = 1)\} \leq \gamma \leq \min\{P(\boldsymbol{y}_0 = 0), P(\boldsymbol{y}_1 = 1)\} \tag{6.17}$$
$$\max\{0, P(\boldsymbol{y}_0 = 1) - P(\boldsymbol{y}_1 = 0)\} \leq \delta \leq \min\{P(\boldsymbol{y}_0 = 1), P(\boldsymbol{y}_1 = 1)\}. \tag{6.18}$$

Then, we use the conditional scores $S_0(x)$ and $S_1(x)$ to refine the bounds on $\alpha, \ldots, \delta$ by leveraging Jensen's inequality (Durrett, 2019, Thm. 1.5.1). Jensen's inequality, in its probabilistic form, states that for a convex function f and a scalar random variable $\boldsymbol{v}$, we have

$$f(\mathbb{E}[\boldsymbol{v}]) \leq \mathbb{E}[f(\boldsymbol{v})]. \tag{6.19}$$

We detail here the derivation of our new lower bound on β, but the same reasoning can easily be extended to the other bounds as well.

$$\begin{aligned}
\max\{0, P(\boldsymbol{y}_0 = 1) - P(\boldsymbol{y}_1 = 1)\} &= \max\{0, S_0 - S_1\} && (6.20)\\
&= \max\{0, \mathbb{E}[S_0(\boldsymbol{x}) - S_1(\boldsymbol{x})]\} && (6.21)\\
&\leq \mathbb{E}[\max\{0, S_0(\boldsymbol{x}) - S_1(\boldsymbol{x})\}] && (6.22)\\
&\leq \mathbb{E}[\beta(\boldsymbol{x})] = \beta. && (6.23)
\end{aligned}$$

The quantity in Eq. (6.21) is the Fréchet bound on β, which by Jensen's inequality is lower than Eq. (6.22). Applying Fréchet's lower bound again, this time inside the expected value operator, we find that the quantity in Eq. (6.22) is between β and its conventional Fréchet bound. Therefore, we derived a tighter lower bound than the Fréchet lower bound. By applying the same reasoning on all Fréchet bounds in Eqs. (6.15) to (6.18), we propose to bound $\alpha, \dots, \delta$ as follows:

$$\begin{aligned}
\mathbb{E}[\max\{0, 1 - S_0(\boldsymbol{x}) - S_1(\boldsymbol{x})\}] \leq \alpha \leq \mathbb{E}[\min\{1 - S_0(\boldsymbol{x}), 1 - S_1(\boldsymbol{x})\}] && (6.24)\\
\mathbb{E}[\max\{0, S_0(\boldsymbol{x}) - S_1(\boldsymbol{x})\}] \leq \beta \leq \mathbb{E}[\min\{S_0(\boldsymbol{x}), 1 - S_1(\boldsymbol{x})\}] && (6.25)\\
\mathbb{E}[\max\{0, S_1(\boldsymbol{x}) - S_0(\boldsymbol{x})\}] \leq \gamma \leq \mathbb{E}[\min\{1 - S_0(\boldsymbol{x}), S_1(\boldsymbol{x})\}] && (6.26)\\
\mathbb{E}[\max\{0, S_0(\boldsymbol{x}) + S_1(\boldsymbol{x}) - 1\}] \leq \delta \leq \mathbb{E}[\min\{S_0(\boldsymbol{x}), S_1(\boldsymbol{x})\}]. && (6.27)
\end{aligned}$$

Hereafter we will refer to those bounds as the *uplift bounds* (UB) since they are defined in terms of the uplift terms.

6.2.1 Bounds span

To assess whether these bounds improve upon the state-of-the-art Fréchet bounds, we consider their respective spans (i.e., the difference between the upper and lower bounds). It can be shown that the span of uplift bounds, noted Span_{UB}, is the same for all four counterfactual probabilities. Here, we base our derivation on Eq. (6.25).

$$\begin{aligned}
\text{Span}_{\text{UB}} &= \mathbb{E}[\min\{S_0(\boldsymbol{x}), 1 - S_1(\boldsymbol{x})\}] - \mathbb{E}[\max\{0, S_0(\boldsymbol{x}) - S_1(\boldsymbol{x})\}] && (6.28)\\
&= \mathbb{E}[\min\{S_0(\boldsymbol{x}), 1 - S_1(\boldsymbol{x})\} - \max\{0, S_0(\boldsymbol{x}) - S_1(\boldsymbol{x})\}] && (6.29)\\
&= \mathbb{E}[\min\{S_0(\boldsymbol{x}), 1 - S_1(\boldsymbol{x})\} + \min\{0, S_1(\boldsymbol{x}) - S_0(\boldsymbol{x})\}] && (6.30)\\
&= \mathbb{E}[\min\{S_0(\boldsymbol{x}), S_1(\boldsymbol{x}), 1 - S_0(\boldsymbol{x}), 1 - S_1(\boldsymbol{x})\}] && (6.31)
\end{aligned}$$

where in Eq. (6.30) we used the equality $-\max\{a, b\} = \min\{-a, -b\}$, and in Eq. (6.31) the equality $\min\{a, b\} + \min\{c, d\} = \min\{a + c, a + d, b + c, b + d\}$. Compare this with the span of the Fréchet bounds, denoted by Span_{Fr}, which is

$$\text{Span}_{\text{Fr}} = \min\{S_0, S_1, 1 - S_0, 1 - S_1\}$$

for all four counterfactual probabilities. Note that Span_{Fr} depends only on the marginal terms S_0 and S_1 (i.e., the average probability of the outcome in the control and target groups), while Span_{UB} is a function of the descriptive features $\boldsymbol{x}$. This means that in the case of informative features (that is, when the conditional entropy of $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ given $\boldsymbol{x}$ is smaller than the marginal entropy), the uplift bounds are tighter than the Fréchet bounds. In the case of perfect knowledge (i.e., when $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are deterministic functions of $\boldsymbol{x}$), $S_0(x)$ and $S_1(x)$ are either 0 or 1, the span of the uplift bounds collapses to zero, and the counterfactual distribution is fully determined. In the case

of non-informative features (i.e., when the conditional entropy of $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ given $\boldsymbol{x}$ is equal to the marginal entropy) the uplift bounds reduce to the Fréchet bounds. These considerations are formalized in the following theorem.

Theorem 6.1. *When the conditional entropy $H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x})$ is zero, the uplift bounds on the probability $P(\boldsymbol{y}_0 = y_0, \boldsymbol{y}_1 = y_1)$ collapse to the exact value of that probability. Conversely, when the conditional entropy $H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x})$ is equal to the entropy $H(\boldsymbol{y}_0, \boldsymbol{y}_1)$, the uplift bounds reduce to the Fréchet bounds.*

Proof. First, let us prove that the span of the uplift bounds is zero when $H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x}) = 0$. From Definition 2.7, we have

$$H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x}) = -\int_{\mathcal{X}} \sum_{(y_0,y_1)\in\mathcal{Y}^2} P(y_0, y_1 \mid x) \log P(y_0, y_1 \mid x) f_{\boldsymbol{x}}(x)\, \mathrm{d}x$$

$$\begin{aligned} H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x}) = -\int_{\mathcal{X}} &(\alpha(x)\log(\alpha(x)) + \beta(x)\log(\beta(x)) \\ &+ \gamma(x)\log(\gamma(x)) + \delta(x)\log(\delta(x)))f_{\boldsymbol{x}}(x)\, \mathrm{d}x. \end{aligned}$$

It is minimized (in fact, equal to zero) when one of $\alpha(x), \dots, \delta(x)$ is equal to one and the three other ones are equal to zero for all $x \in \mathcal{X}$. Also, the span of the uplift bounds is

$$\begin{aligned} \mathrm{Span}_{\mathrm{UB}} &= \mathbb{E}[\min\{S_0(\boldsymbol{x}), S_1(\boldsymbol{x}), 1 - S_0(\boldsymbol{x}), 1 - S_1(\boldsymbol{x})\}] \\ &= \int_{\mathcal{X}} \min\{\beta(x) + \delta(x), \gamma(x) + \delta(x), \alpha(x) + \gamma(x), \alpha(x) + \beta(x)\} f_{\boldsymbol{x}}(x)\, \mathrm{d}x. \end{aligned}$$

When one of $\alpha(x), \dots, \delta(x)$ is equal to one and the three other values are equal to zero for all $x \in \mathcal{X}$, this expression collapses to zero, since two of the four terms in the minimum will be equal to zero. In this case, the bounds collapse to the true value of the counterfactual probability. This proves the first part of the theorem.

For the second part of the theorem, assume that $H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{x}) = H(\boldsymbol{y}_0, \boldsymbol{y}_1)$. In terms of statistical independence, this is expressed as $(\boldsymbol{y}_0, \boldsymbol{y}_1) \perp \boldsymbol{x}$. By the definition of statistical independence, we know that $P(y_0 \mid x) = P(y_0)$ and $P(y_1 \mid x) = P(y_1)$ for all values $y_0, y_1 \in \mathcal{Y}$ and $x \in \mathcal{X}$. Therefore, the uplift bounds on $P(y_0, y_1)$ simplify to

$$\mathbb{E}_{\boldsymbol{x}}[\max\{0, P(y_0) + P(y_1) - 1\}] \leq P(y_0, y_1) \leq \mathbb{E}_{\boldsymbol{x}}[\min\{P(y_0), P(y_1)\}].$$

The expected value is on the distribution of $\boldsymbol{x}$, but since the terms in the expected value do not depend on $\boldsymbol{x}$, the bounds reduce to

$$\max\{0, P(y_0) + P(y_1) - 1\} \leq P(y_0, y_1) \leq \min\{P(y_0), P(y_1)\},$$

which are the Fréchet bounds. □

6.2.2 Plug-in estimator

The main motivation underlying the derivation of the uplift bounds is that in real-world settings characterized by large historical datasets (such as churn modeling), it is possible to derive sample-based estimates of the terms that bound counterfactual probabilities. In particular, we advocate the adoption of a plug-in estimator from an

uplift model $\hat{S}_0(x, D_{\text{tr}}), \hat{S}_1(x, D_{\text{tr}})$ trained on a dataset D_{tr}, evaluated on a test dataset $D_{\text{te}} = \{x^{(1)}, \dots, x^{(N)}\}$. In this case, a sample-based version of the lower bound on β is

$$\widehat{\text{LB}}_\beta = \frac{1}{N} \sum_{i=1}^{N} \max\left\{0, \hat{S}_0(x^{(i)}, D_{\text{tr}}) - \hat{S}_1(x^{(i)}, D_{\text{tr}})\right\} \tag{6.32}$$

And similarly for the other bounds on $\alpha, \dots, \delta$, noted $\widehat{\text{LB}}_\alpha, \widehat{\text{UB}}_\alpha$, etc.

6.3 Point estimate of counterfactual probabilities

In the previous section, we proposed an original approach to bound counterfactual probabilities. However, it is sometimes desirable to compute a point estimate of those probabilities, even if this requires stronger assumptions. Here, we present a point estimator of the probabilities $\alpha, \dots, \delta$ based on the conditional independence between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ given $\boldsymbol{x}$. The introduction of specific assumptions is required since counterfactual probabilities cannot be estimated from observational or experimental data, as one of the two outcomes will necessarily be unobserved. This can be understood as a fundamental limitation postulated by the causal hierarchy, as presented in Section 2.2.1: counterfactual probabilities belong to the third layer, while information that can be learned directly from data, be it observational or experimental, belongs to the first or second layer. To bridge the gap to the third layer, we must exploit knowledge or assumptions pertaining to the third layer. Let us assume the conditional independence between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ given $\boldsymbol{x} = x$. We discuss the meaning and validity of this assumption below. Formally, this assumption is expressed as $\boldsymbol{y}_0 \perp \boldsymbol{y}_1 \mid \boldsymbol{x} = x$, and allows us to develop the probability $\alpha(x)$ in Eq. (6.7) as

$$\alpha(x) = P(\boldsymbol{y}_0 = 0 \mid \boldsymbol{x} = x)P(\boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x). \tag{6.33}$$

Since we assume to have access to estimators of the scores $S_t(x) = P(\boldsymbol{y}_t = 1 \mid \boldsymbol{x} = x)$, estimating $\alpha(x)$ or any of the other counterfactuals is easy under this assumption of conditional independence. To assess the impact of this assumption, we define the difference between $P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x)$ and its approximation $P(\boldsymbol{y}_0 = 0 \mid \boldsymbol{x} = x)P(\boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x)$ as $\phi(x)$. The same quantity appears in the other conditional probabilities as follows:

$$\alpha(x) = P(\boldsymbol{y}_0 = 0 \mid \boldsymbol{x} = x)P(\boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x) + \phi(x) \tag{6.34}$$
$$\beta(x) = P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x)P(\boldsymbol{y}_1 = 0 \mid \boldsymbol{x} = x) - \phi(x) \tag{6.35}$$
$$\gamma(x) = P(\boldsymbol{y}_0 = 0 \mid \boldsymbol{x} = x)P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x) - \phi(x) \tag{6.36}$$
$$\delta(x) = P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x)P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x) + \phi(x). \tag{6.37}$$

The quantity $\phi(x)$ can be interpreted as a measure of the dependency between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ given $\boldsymbol{x} = x$: when $\phi(x)$ is equal to zero, then $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are independent (given $\boldsymbol{x} = x$), when it is positive, the two potential outcomes are positively correlated, and when it is negative, they are negatively correlated. In that sense, it is similar to classical binary dependency measures, like the odd ratio, Yule's Q coefficient, or the difference coefficient (Edwards, 1957). The assumption of conditional independence $\boldsymbol{y}_0 \perp \boldsymbol{y}_1 \mid \boldsymbol{x} = x$ is thus equivalent to assuming that $\phi(x) = 0$.

Now that we have formulated an assumption to estimate $\alpha(x), \dots, \delta(x)$ from data, we now move on to estimate the probabilities $\alpha, \dots, \delta$. We develop α to obtain

$$\begin{aligned}
\alpha &= P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1) && (6.38)\\
&= \int_{\mathcal{X}} f_{\boldsymbol{x}}(x) P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1 \mid \boldsymbol{x}) \, \mathrm{d}x && (6.39)\\
&= \mathbb{E}[\alpha(\boldsymbol{x})] && (6.40)\\
&= \mathbb{E}[P(\boldsymbol{y}_0 = 0 \mid \boldsymbol{x}) P(\boldsymbol{y}_1 = 0 \mid \boldsymbol{x}) + \phi(\boldsymbol{x})] && (6.41)\\
&= \mathbb{E}[(1 - S_0(\boldsymbol{x}))(1 - S_1(\boldsymbol{x}))] + \phi && (6.42)
\end{aligned}$$

where we use the notation $\phi = \mathbb{E}[\phi(\boldsymbol{x})]$. We will see in Theorem 6.2 that

$$\phi = \alpha\delta - \beta\gamma - \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})). \tag{6.43}$$

This means that ϕ depends both on the distribution of counterfactuals (α, β, γ and δ) and the dependency between the scores $S_0(x)$ and $S_1(x)$. If we assume $\phi = 0$, or even $\boldsymbol{y}_0 \perp \boldsymbol{y}_1 \mid \boldsymbol{x} = x$ for all $x \in \mathcal{X}$ (which is a stronger assumption and implies $\phi = 0$), then the counterfactual probabilities can be estimated as

$$\begin{aligned}
\alpha &= \mathbb{E}[(1 - S_0(\boldsymbol{x}))(1 - S_1(\boldsymbol{x}))] && (6.44)\\
\beta &= \mathbb{E}[S_0(\boldsymbol{x})(1 - S_1(\boldsymbol{x}))] && (6.45)\\
\gamma &= \mathbb{E}[(1 - S_0(\boldsymbol{x}))S_1(\boldsymbol{x})] && (6.46)\\
\delta &= \mathbb{E}[S_0(\boldsymbol{x})S_1(\boldsymbol{x})]. && (6.47)
\end{aligned}$$

The question of the dependency between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ has already been discussed in the causal inference literature (G. W. Imbens and Donald B Rubin, 2015, Sec. 8.6). When there is a lack of evidence in favor of or against the dependency between the potential outcomes, a cautious approach would be to minimize risk by considering the worst-case scenario, for example by assuming the highest possible level of dependency between the potential outcomes. Alternatively, one could make no a priori preference between a positive and negative association between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ (i.e., $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ taking similar or opposite values), thus assuming no association. Since there is no a priori good answer in absence of some preexisting knowledge, we can reason about the dependency between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ as follows:

- A positive correlation[2] between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ means that they are often equal, indicating that the treatment has little effect on the outcome. When the correlation is maximum, the upper bounds on α and δ in Eqs. (6.24) and (6.27) are met.

- A negative correlation between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ indicates that the treatment has either a strongly positive or negative impact on the outcome. When the correlation is maximally negative, the upper bounds on β and γ in Eqs. (6.25) and (6.26) are met.

- The absence of dependency indicates an even mix of the two previous cases. This corresponds to the point estimator presented in this section.

[2]The correlation between $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ refers to the tendency of $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ to take identical or complentary values.

6.3.1 Point estimate and uplift estimation

Given estimators $\hat{S}_0(x, D_{\text{tr}}), \hat{S}_1(x, D_{\text{tr}})$ of the uplift terms trained on a training dataset D_{tr}, and given an evaluation dataset $D_{\text{te}} = \{x^{(i)}\}_{i=1,\dots,N}$, we propose to estimate $\alpha, \dots, \delta$ as follows:

$$\hat{\alpha} = \frac{1}{N} \sum_i (1 - \hat{S}_0(x^{(i)}, D_{\text{tr}})(1 - \hat{S}_1(x^{(i)}, D_{\text{tr}})) \tag{6.48}$$

$$\hat{\beta} = \frac{1}{N} \sum_i \hat{S}_0(x^{(i)}, D_{\text{tr}})(1 - \hat{S}_1(x^{(i)}, D_{\text{tr}})) \tag{6.49}$$

$$\hat{\gamma} = \frac{1}{N} \sum_i (1 - \hat{S}_0(x^{(i)}, D_{\text{tr}}))\hat{S}_1(x^{(i)}, D_{\text{tr}}) \tag{6.50}$$

$$\hat{\delta} = \frac{1}{N} \sum_i \hat{S}_0(x^{(i)}, D_{\text{tr}})\hat{S}_1(x^{(i)}, D_{\text{tr}}). \tag{6.51}$$

The bias of these estimators is expressed in Theorem 6.2.

Theorem 6.2. *Given that $\hat{S}_0(x, \boldsymbol{D}_{\text{tr}})$ and $\hat{S}_1(x, \boldsymbol{D}_{\text{tr}})$, trained on a training dataset sampled from a distribution $\boldsymbol{D}_{\text{tr}}$, are unconfounded and unbiased estimators of $S_0(x)$ and $S_1(x)$, in the large sample limit the bias of $\hat{\alpha}, \dots, \hat{\delta}$ estimated on a test dataset $D_{\text{te}} = \{x^{(i)}\}_{i=1}^N$ iid to $\boldsymbol{D}_{\text{tr}}$ is*

$$\begin{aligned}
\text{Bias}[\hat{\beta}] = \text{Bias}[\hat{\gamma}] = -\text{Bias}[\hat{\alpha}] = -\text{Bias}[\hat{\delta}] & \qquad (6.52)\\
= \alpha\delta - \beta\gamma - \text{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{x}}[\text{cov}_{\boldsymbol{D}_{\text{tr}}}(\hat{S}_0(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}}), \hat{S}_1(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}})] & \qquad (6.53)\\
= \phi - \mathbb{E}_{\boldsymbol{x}}[\text{cov}_{\boldsymbol{D}_{\text{tr}}}(\hat{S}_0(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}}), \hat{S}_1(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}})]. & \qquad (6.54)
\end{aligned}$$

Proof. We will derive the bias of $\hat{\beta}$, and the bias of the three other estimators can be derived in a similar way. The expected value of $\hat{\beta}$ over the distribution of training sets $\boldsymbol{D}_{\text{tr}}$ is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[\hat{\beta}] &= \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}\left[\frac{1}{N}\sum_{i=1}^N \hat{S}_0(x^{(i)}, \boldsymbol{D}_{\text{tr}})(1 - \hat{S}_1(x^{(i)}, \boldsymbol{D}_{\text{tr}}))\right]\\
&= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}\left[\hat{S}_0(x^{(i)}, \boldsymbol{D}_{\text{tr}})(1 - \hat{S}_1(x^{(i)}, \boldsymbol{D}_{\text{tr}}))\right]\\
&= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[\hat{S}_0(x^{(i)}, \boldsymbol{D}_{\text{tr}})]\mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[1 - \hat{S}_1(x^{(i)}, \boldsymbol{D}_{\text{tr}})]\\
&\qquad + \text{cov}_{\boldsymbol{D}_{\text{tr}}}(\hat{S}_0(x^{(i)}, \boldsymbol{D}_{\text{tr}}), 1 - \hat{S}_1(x^{(i)}, \boldsymbol{D}_{\text{tr}}))\\
&= \frac{1}{N}\sum_{i=1}^N S_0(x^{(i)})(1 - S_1(x^{(i)})) - \text{cov}_{\boldsymbol{D}_{\text{tr}}}(\hat{S}_0(x^{(i)}, \boldsymbol{D}_{\text{tr}}), \hat{S}_1(x^{(i)}, \boldsymbol{D}_{\text{tr}})).
\end{aligned}$$

In the large sample limit ($N \to +\infty$), we can assume that this sum converges to

$$\lim_{N\to\infty} \mathbb{E}_{\boldsymbol{D}_{\text{tr}}}[\hat{\beta}] = \mathbb{E}_{\boldsymbol{x}}[S_0(\boldsymbol{x})(1 - S_1(\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{x}}[\text{cov}_{\boldsymbol{D}_{\text{tr}}}(\hat{S}_0(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}}), \hat{S}_1(\boldsymbol{x}, \boldsymbol{D}_{\text{tr}})].$$

Algorithm 1 Estimating the counterfactual probability β

Input: dataset $D = \left\{(x^{(i)}, y^{(i)}, t^{(i)})\right\}_{i=1,\dots,N}$
Output: Point estimate $\hat{\beta}$, and bounds $\widehat{\mathrm{LB}}_\beta$ and $\widehat{\mathrm{UB}}_\beta$ such that $\widehat{\mathrm{LB}}_\beta \leq \beta \leq \widehat{\mathrm{UB}}_\beta$
Split D into training set D_{tr} and test set D_{te}
Train uplift model on D_{tr} to obtain estimators $\hat{S}_0(x, D_{\mathrm{tr}}), \hat{S}_1(x, D_{\mathrm{tr}})$
$\hat{\beta} = \frac{1}{|D_{\mathrm{te}}|} \sum_i \hat{S}_0(x^{(i)}, D_{\mathrm{tr}})(1 - \hat{S}_1(x^{(i)}, D_{\mathrm{tr}}))$ on D_{te}
$\widehat{\mathrm{LB}}_\beta = \frac{1}{|D_{\mathrm{te}}|} \sum_i \max\{0, \hat{S}_0(x^{(i)}, D_{\mathrm{tr}}) - \hat{S}_1(x^{(i)}, D_{\mathrm{tr}})\}$ on D_{te}
$\widehat{\mathrm{UB}}_\beta = \frac{1}{|D_{\mathrm{te}}|} \sum_i \min\{\hat{S}_0(x^{(i)}, D_{\mathrm{tr}}), 1 - \hat{S}_1(x^{(i)}, D_{\mathrm{tr}})\}$ on D_{te}

The first term can be expanded as

$$\begin{aligned}
\mathbb{E}[S_0(\boldsymbol{x})(1 - S_1(\boldsymbol{x}))] &= \mathbb{E}[S_0(\boldsymbol{x})]\mathbb{E}[1 - S_1(\boldsymbol{x})] + \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), 1 - S_1(\boldsymbol{x})) \\
&= S_0(1 - S_1) - \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})) \\
&= (\beta + \delta)(\beta + \alpha) - \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})) \\
&= \beta(\beta + \delta + \alpha) + \alpha\delta - \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})) \\
&= \beta(1 - \gamma) + \alpha\delta - \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})) \\
&= \alpha\delta - \beta\gamma + \beta - \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})).
\end{aligned}$$

And thus

$$\mathbb{E}_{\boldsymbol{D}_{\mathrm{tr}}}[\hat{\beta}] = \alpha\delta - \beta\gamma + \beta - \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{x}}[\mathrm{cov}_{\boldsymbol{D}_{\mathrm{tr}}}(\hat{S}_0(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}}), \hat{S}_1(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}})].$$

Finally, the bias of $\hat{\beta}$ is

$$\begin{aligned}
\mathrm{Bias}[\hat{\beta}] &= \mathbb{E}_{\boldsymbol{D}_{\mathrm{tr}}}[\hat{\beta}] - \beta \\
&= \alpha\delta - \beta\gamma - \mathrm{cov}_{\boldsymbol{x}}(S_0(\boldsymbol{x}), S_1(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{x}}[\mathrm{cov}_{\boldsymbol{D}_{\mathrm{tr}}}(\hat{S}_0(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}}), \hat{S}_1(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}})],
\end{aligned}$$

which proves Eq. (6.53). Equation (6.54) is derived from

$$\mathbb{E}[S_0(\boldsymbol{x})(1 - S_1(\boldsymbol{x}))] = \mathbb{E}[\beta(\boldsymbol{x}) + \phi(\boldsymbol{x})] = \beta + \phi$$

And then

$$\begin{aligned}
\mathrm{Bias}[\hat{\beta}] &= \mathbb{E}_{\boldsymbol{D}_{\mathrm{tr}}}[\hat{\beta}] - \beta \\
&= \mathbb{E}_{\boldsymbol{x}}[S_0(\boldsymbol{x})(1 - S_1(\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{x}}[\mathrm{cov}_{\boldsymbol{D}_{\mathrm{tr}}}(\hat{S}_0(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}}), \hat{S}_1(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}})] - \beta \\
&= \phi - \mathbb{E}_{\boldsymbol{x}}[\mathrm{cov}_{\boldsymbol{D}_{\mathrm{tr}}}(\hat{S}_0(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}}), \hat{S}_1(\boldsymbol{x}, \boldsymbol{D}_{\mathrm{tr}})].
\end{aligned}$$

□

While the three first terms in Eq. (6.53) are inherent to the customer population, the last term depends also on the estimators $\hat{S}_0(x, \boldsymbol{D}_{\mathrm{tr}})$ and $\hat{S}_1(x, \boldsymbol{D}_{\mathrm{tr}})$, and the data distribution $\boldsymbol{D}_{\mathrm{tr}}$. Without assumptions about these processes, the last term cannot be further reduced.

The proposed procedure to compute $\hat{\beta}$, as well as the two uplift bounds on β presented in Section 6.2, is described in Algorithm 1, where we assume that we have unbiased estimators of the scores $S_0(x)$ and $S_1(x)$.

6.4 Posterior distribution of counterfactuals with a bivariate beta distribution

In the experimental section of Chapter 5, we used the bivariate beta distribution developed by Olkin and Trikalinos (2015) to simulate uplift modeling without having to generate a dataset and train an uplift model directly. The bivariate beta distribution can generate samples that represent the output of an uplift model trained under various conditions, such as varying degrees of class imbalance and mutual information between the emulated features and the potential outcomes. Having control over these characteristics of the simulation is crucial to understand the empirical results of Chapter 4, where uplift modeling seems to perform well on the Hillstrom dataset (where the potential outcomes are balanced and the features are informative) but not on the churn datasets (where the potential outcomes are unbalanced and the features are not very informative). A very convenient property of the bivariate beta distribution used in Section 5.3.3 is that it first generates the probability of counterfactuals for each individual before summing them to obtain the uplift scores. This property allowed us to examine the distribution of counterfactuals resulting in a given set of uplift scores in the simulations.

In this section, we use the same bivariate beta distribution, but instead of directly specifying the parameters of the distribution, we fit the parameters to an existing dataset. This procedure provides an estimator for the probability of counterfactuals for this dataset, since the counterfactual distribution is modeled by the bivariate beta distribution under the hood. More precisely, we obtain a point estimate for the population-level counterfactuals $\alpha, \dots, \delta$, and a posterior distribution over the possible values for the individual-level counterfactuals $\alpha(x), \dots, \delta(x)$ for any realization x of the features $\boldsymbol{x}$.

As mentioned in Section 2.2.1, quantities belonging the third layer of the causal hierarchy (i.e., counterfactual probabilities) cannot be uniquely determined from information or assumptions pertaining to the lower layers (observational and experimental data). Fitting a bivariate beta distribution to the uplift scores represents an inductive bias over the distribution of counterfactuals and, as such, is an assumption on the data-generating process governing the potential outcomes. This is the critical element that allows us to infer an exact value for counterfactual probabilities based on experimental data, much like the conditional independence assumption used in Section 6.3.

6.4.1 Summary of the approach

The bivariate beta distribution, noted $(S_0, S_1) \sim \text{BB}(m)$ for for a vector of positive parameters $m = [a, b, c, d]$, is a bivariate distribution with beta marginals. We derive several properties of this distribution in Appendix D. Sampling from this distribution is done in two steps. First, we sample a four-valued random vector $\boldsymbol{\mu} = [\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}]$ (that we also note $[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4]$) from a Dirichlet distribution $\text{Dir}(m)$, which is noted

$$\boldsymbol{\mu} \sim \text{Dir}(m). \tag{6.55}$$

Then, the scores S_0, S_1 are defined as

$$S_0 = \boldsymbol{\beta} + \boldsymbol{\delta} \quad \text{and} \quad S_1 = \boldsymbol{\gamma} + \boldsymbol{\delta}. \tag{6.56}$$

Using this procedure, we can show from the properties of the Dirichlet distribution (Lin, 2016) that $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ have marginal beta distributions, noted

$$\boldsymbol{S}_0 \sim \text{Beta}(b+d, a+c) \quad \text{and} \quad \boldsymbol{S}_1 \sim \text{Beta}(c+d, a+b). \tag{6.57}$$

Finally, the binary potential outcomes $\boldsymbol{y}_0, \boldsymbol{y}_1$ are defined as following a categorical distribution $(\boldsymbol{y}_0, \boldsymbol{y}_1) \sim \text{Cat}(\boldsymbol{\mu})$, which is a shorthand notation for

$$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0 \mid \boldsymbol{\mu} = \mu) = \alpha \tag{6.58}$$
$$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0 \mid \boldsymbol{\mu} = \mu) = \beta \tag{6.59}$$
$$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1 \mid \boldsymbol{\mu} = \mu) = \gamma \tag{6.60}$$
$$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1 \mid \boldsymbol{\mu} = \mu) = \delta. \tag{6.61}$$

In this section, the random variables $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{S}_0$ and $\boldsymbol{S}_1$ denote probabilities at the individual level, similarly to $\alpha(x), \beta(x), \gamma(x), \delta(x), S_0(x)$, and $S_1(x)$. Thus, $\boldsymbol{\alpha}$ *does not* represent the population-level probability $P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1)$. Instead, we have $P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0) = \mathbb{E}[\boldsymbol{\alpha}]$. In Section 5.3.3, we used the superscript $\boldsymbol{\alpha}^{(i)}$ to make this fact evident. In this section, we do not use superscript notation to avoid confusion with samples in a training set $D = \left\{(x^{(i)}, y^{(i)}, t^{(i)})\right\}_{i=1}^{N}$, at the risk of being slightly more confusing.

Our approach consists of two phases. First, the learning phase consists of

1. Using a dataset $D = \left\{(x^{(i)}, y^{(i)}, t^{(i)})\right\}_{i=1}^{N}$ to train an uplift model, providing a set of predicted scores $F = \left\{(\hat{S}_0(x^{(i)}), \hat{S}_1(x^{(i)}))\right\}_{i=1}^{N}$ where $\hat{S}_t(x^{(i)})$ is an estimator of $P(\boldsymbol{y}_t = 1 \mid \boldsymbol{x} = x^{(i)})$ for $t = 0, 1$.
2. Fitting a bivariate beta distribution $\text{BB}(a, c, b, d)$ on this set of scores.

Then, we infer the counterfactual probabilities in the inference phase. Using the moments of the Dirichlet distribution, the population-level counterfactuals are readily available as

$$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0) = \mathbb{E}[\boldsymbol{\alpha}] = \frac{a}{M} \qquad P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0) = \mathbb{E}[\boldsymbol{\beta}] = \frac{b}{M} \tag{6.62}$$
$$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1) = \mathbb{E}[\boldsymbol{\gamma}] = \frac{c}{M} \qquad P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1) = \mathbb{E}[\boldsymbol{\delta}] = \frac{d}{M} \tag{6.63}$$

where $M = a+b+c+d$. Individual-level counterfactual probabilities for a new sample x' are estimated by computing the posterior probability distribution of $\alpha(x'), \ldots, \delta(x')$ that is compatible with the scores $\hat{S}_0(x')$ and $\hat{S}_1(x')$. We propose to estimate $\alpha(x'), \ldots, \delta(x')$ with the expected value of this posterior distribution.

6.4.2 Learning phase

In this section, we describe the mathematical details of the learning phase. We use several properties of the bivariate beta distribution that are derived in Appendix D.

1. Train an uplift model on D to estimate the probability scores

$$\hat{S}_0(x^{(i)}) \approx P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x^{(i)}) \tag{6.64}$$
$$\hat{S}_1(x^{(i)}) \approx P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x} = x^{(i)}) \tag{6.65}$$

We note the set of predicted scores as $F = \left\{(\hat{S}_0(x^{(i)}), \hat{S}_0(x^{(i)}))\right\}_{i=1}^{N}$.

2. Fit a bivariate beta distribution $\text{BB}(a, b, c, d)$ on F using the method of moments. More precisely, we use an optimization procedure to find the value of m that minimizes the square difference between the raw sample moments from F and the raw distribution moments of $\text{BB}(m)$. We use the Trust Region Reflective algorithm (Branch, Coleman, and Yuying Li, 1999) implemented by the Python package SciPy.[3] This function requires four elements as input:

   a) The raw sample moments, for $r, s \in \mathbb{N}$:

   $$\widehat{R}_{rs} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{S}_0(x^{(i)})\right)^r \left(\hat{S}_1(x^{(i)})\right)^s \tag{6.66}$$

   In practice, we use the first 7 raw moments, that is, $(r, s) \in \{(0, 1), (1, 0), (2, 0), (1, 1), (0, 2), (3, 0), (0, 3)\}$. Preliminary results suggest that the choice of the moments has a minimal impact on the results as long as we use at least the first four moments.

   b) The loss, defined as the square difference between the sample moments $\widehat{R}_{rs}$ and the distribution moments $R^m_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1)$ for the current value of m:

   $$L(m) = \frac{1}{2} \sum_{r,s} (R^m_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1) - \widehat{R}_{rs})^2. \tag{6.67}$$

   The analytical formula for $R^m_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1)$ is given in Result D.2.

   c) The gradient vector of the loss, which is computed from the Jacobian matrix of the moments with respect to the components of m. The analytical formula is given in Result D.3.

   d) An initial solution $a_{(0)}, b_{(0)}, c_{(0)}, d_{(0)}$. We found the following initial solution using the first four moments $(r, s) \in \{(1, 0), (0, 1), (2, 0), (0, 2)\}$. The derivation is given in Result D.4.

   $$M_{(0)} = \frac{1}{2} \left(\frac{\widehat{R}_{10} - \widehat{R}_{20}}{\widehat{R}_{20} - \widehat{R}^2_{10}} + \frac{\widehat{R}_{01} - \widehat{R}_{02}}{\widehat{R}_{02} - \widehat{R}^2_{01}} \right) \tag{6.68}$$

   $$a_{(0)} = M_{(0)}(1 - \widehat{R}_{10})(1 - \widehat{R}_{01}) \tag{6.69}$$

   $$b_{(0)} = M_{(0)}\widehat{R}_{10}(1 - \widehat{R}_{01}) \tag{6.70}$$

   $$c_{(0)} = M_{(0)}(1 - \widehat{R}_{10})\widehat{R}_{01} \tag{6.71}$$

   $$d_{(0)} = M_{(0)}\widehat{R}_{10}\widehat{R}_{01}. \tag{6.72}$$

6.4.3 Inference phase

Once the distribution has been learned and we have fitted the parameters $m = [a, b, c, d]$, we can proceed to the inference phase. This consists in estimating, for any set of uplift scores $\hat{S}_0(x^{(i)}), \hat{S}_1(x^{(i)})$, which, in this section, we note S_0 and S_1 for simplicity (these should not be confused with Eqs. 3.4 and 3.5), the posterior probability distribution $\boldsymbol{\alpha}, \ldots, \boldsymbol{\delta} \mid \boldsymbol{S}_0 = S_0, \boldsymbol{S}_1 = S_1$. In fact, we directly compute the expected value of this posterior distribution. The following two new results facilitate this computation.

[3] `scipy.optimzie.least_squares`

Result 6.3. *The expected value of $\boldsymbol{\mu}_j$ for $j = 1, \dots, 4$ (i.e., the jth component of $\boldsymbol{\mu}$, corresponding to one of $\boldsymbol{\alpha}, \dots, \boldsymbol{\delta}$) given that we observe the realization S_0, S_1 of $\boldsymbol{S}_0, \boldsymbol{S}_1 \sim \mathrm{BB}(m)$ can be expressed as*

$$\mathbb{E}[\boldsymbol{\mu}_j \mid S_0, S_1] = \frac{m_j}{M} \frac{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m'; S_0, S_1)}{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m; S_0, S_1)}. \tag{6.73}$$

where $f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m; \cdot)$ is the probability density function of $\boldsymbol{S}_0, \boldsymbol{S}_1$ with parameter vector m, and m' is the vector m with the jth component incremented by one, that is, $m' = [m_1, \dots, m_j + 1, \dots, m_4]$, and $M = m_1 + m_2 + m_3 + m_4$.

Proof. By the definition of the conditional expected value (Definition 2.2), we have

$$\mathbb{E}[\boldsymbol{\mu}_j \mid S_0, S_1] = \frac{\int_0^1 \mu_j f_{\boldsymbol{\mu}_j,\boldsymbol{S}_0,\boldsymbol{S}_1}(m; \mu_j, S_0, S_1)\, \mathrm{d}\mu_j}{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m; S_0, S_1)} \tag{6.74}$$

where $f_{\boldsymbol{\mu}_j,\boldsymbol{S}_0,\boldsymbol{S}_1}(m; \mu_j, S_0, S_1)$ is the joint pdf of $\boldsymbol{\mu}_j, \boldsymbol{S}_0, \boldsymbol{S}_1$. The pdf $f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m; S_0, S_1)$ is defined as (Olkin and Trikalinos, 2015)

$$f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(S_0, S_1) = \int_{\Lambda(S_0,S_1)} f_{\boldsymbol{\mu}}(m; \mu)\, \mathrm{d}\mu \tag{6.75}$$

where $f_{\boldsymbol{\mu}}(m, \cdot)$ is the pdf the random vector $\boldsymbol{\mu}$ following a Dirichlet distribution (see Eq. 2.39), and where the integration domain is

$$\Lambda(S_0, S_1) = \{\mu = [\alpha, \beta, \gamma, \delta] \in \Lambda \mid \beta + \delta = S_0, \gamma + \delta = S_1\} \tag{6.76}$$

with Λ the four-dimensional unit simplex (see Eq. 2.38). The joint pdf $f_{\boldsymbol{\mu}_j,\boldsymbol{S}_0,\boldsymbol{S}_1}(m; \cdot, \cdot)$ in the numerator of Eq. (6.74) can be marginalized by integrating over the other components of $\boldsymbol{\mu}$. For example, if $\boldsymbol{\mu}_j$ is $\boldsymbol{\beta}$ (that is, $j = 2$), we have

$$f_{\boldsymbol{\beta},\boldsymbol{S}_0,\boldsymbol{S}_1}(m; \beta, S_0, S_1) = \int_0^{1-\beta} \int_0^{1-\alpha-\beta} \int_0^{1-\alpha-\beta-\gamma} f_{\boldsymbol{\mu},\boldsymbol{S}_0,\boldsymbol{S}_1}(m; [\alpha, \beta, \gamma, \delta], S_0, S_1)\, \mathrm{d}\delta\, \mathrm{d}\gamma\, \mathrm{d}\alpha.$$

Since $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ are defined as $\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta}$ and $\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta}$, we have

$$f_{\boldsymbol{S}_0,\boldsymbol{S}_1|\mu}(m; S_0, S_1) = \begin{cases} 1 & \text{if } \beta + \delta = S_0 \text{ and } \gamma + \delta = S_1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} f_{\boldsymbol{\mu},\boldsymbol{S}_0,\boldsymbol{S}_1}(m; \mu, S_0, S_1) &= f_{\boldsymbol{S}_0,\boldsymbol{S}_1|\mu}(m; S_0, S_1) f_{\boldsymbol{\mu}}(m; \mu) \\ &= \begin{cases} f_{\boldsymbol{\mu}}(m; \mu) & \text{if } \beta + \delta = S_0 \text{ and } \gamma + \delta = S_1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The set of values of $\boldsymbol{\mu}$ that respect these two constraints ($S_0 = \beta + \delta$ and $S_1 = \gamma + \delta$) is noted $\Lambda(S_0, S_1)$. This allows us to develop the numerator of Eq. (6.74) as

$$\int_0^1 \mu_j f_{\boldsymbol{\mu}_j,\boldsymbol{S}_0,\boldsymbol{S}_1}(m; \mu_j, S_0, S_1)\, \mathrm{d}\mu_j = \int_{\Lambda(S_0,S_1)} \mu_j f_{\boldsymbol{\mu}}(m; \mu)\, \mathrm{d}\mu,$$

leading to

$$\mathbb{E}[\boldsymbol{\mu}_j \mid S_0, S_1] = \frac{\int_{\Lambda(S_0,S_1)} \mu_j f_{\boldsymbol{\mu}}(m;\mu)\,\mathrm{d}\mu}{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m;S_0,S_1)}. \tag{6.77}$$

We now transform the numerator of Eq. (6.77) into an expression similar to the pdf $f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m,\cdot)$. As in the statement of the result, let m' be the vector m with the jth component incremented by one, that is, $m' = [m_1,\dots,m_j+1,\dots,m_4]$. The numerator of Eq. (6.77) can be developed as

$$\begin{aligned}
\int_{\Lambda(S_0,S_1)} \mu_j f_{\boldsymbol{\mu}}(m;\mu)\,\mathrm{d}\mu &= \frac{1}{\mathrm{B}(m)} \int_{\Lambda(S_0,S_1)} \mu_j \prod_{k=1}^{4} \mu_k^{m_k-1}\,\mathrm{d}\mu && \text{(by Eq. 2.39)}\\
&= \frac{\mathrm{B}(m')}{\mathrm{B}(m)\mathrm{B}(m')} \int_{\Lambda(S_0,S_1)} \mu_j^{m_j} \prod_{k\neq j} \mu_k^{m_k-1}\,\mathrm{d}\mu\\
&= \frac{\mathrm{B}(m')}{\mathrm{B}(m)} f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m';S_0,S_1) && \text{(by Eq. 6.75)}\\
&= \frac{m_j}{M} f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m';S_0,S_1)
\end{aligned}$$

where the last equality follows from the definition of the beta function in terms of the gamma function (see Eq. 2.31), using the property $\Gamma(x+1) = x\Gamma(x)$:

$$\frac{\mathrm{B}(m')}{\mathrm{B}(m)} = \frac{\Gamma(m_j+1)\prod_{i\neq j}\Gamma(m_i)}{\Gamma(M+1)} \frac{\Gamma(M)}{\prod_{i=1}^{4}\Gamma(m_i)} = \frac{\Gamma(m_j)m_j\Gamma(M)}{\Gamma(M)M\Gamma(m_j)} = \frac{m_j}{M}.$$

In summary, we can express Eq. (6.77) as

$$\mathbb{E}[\boldsymbol{\mu}_j \mid S_0, S_1] = \frac{m_j}{M} \frac{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m';S_0,S_1)}{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m;S_0,S_1)}.$$

□

According to this result, if we can compute the value of $f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m;S_0,S_1)$ for any vector m, we can estimate the expected value of the counterfactual distribution. The following result expresses $f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m;S_0,S_1)$ into an expression that can be easily computed with any numerical integration routine, such as the function `scipy.integrate.quad` in the the Python package SciPy.

Result 6.4. *The probability density function of* $(\boldsymbol{S}_0,\boldsymbol{S}_1) \sim \mathrm{BB}(m)$ *can be expressed as*

$$f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m;S_0,S_1) = \frac{1}{\mathrm{B}(m)} \int_{\max\{0,S_0+S_1-1\}}^{\min\{S_0,S_1\}} (S_0+S_1-\delta)^{a-1}(S_0-\delta)^{b-1}(S_1-\delta)^{c-1}\delta^{d-1}\,\mathrm{d}\delta. \tag{6.78}$$

Proof. As shown in Eq. (6.75), the pdf of $\boldsymbol{S}_0,\boldsymbol{S}_1$ is defined as an integral over a set $\Lambda(S_0,S_1)$ of four-dimensional vectors μ. Since the components of μ must sum up to one, the set has only three degrees of freedom; for example, given the value of β, γ and δ, the value of α is determined as $1-\beta-\gamma-\delta$. The two additional constraints $\beta+\delta = S_0$ and $\gamma+\delta = S_1$ remove two other degrees of freedom; if we know S_0, S_1 and, for example, δ, we can determine the value of α, β and γ. Hence, the set $\Lambda(S_0,S_1)$ is in fact a one-dimensional line in the four-dimensional unit simplex. In the following, we will fix the

value of δ, which has to lie between the Fréchet bounds (see Eq. 6.18), otherwise one of α, β or γ would have to be negative to satisfy the constraints $\beta + \delta = S_0$ and $\gamma + \delta = S_1$. The Fréchet bounds on δ are

$$\max\{0, S_0 + S_1 - 1\} \leq \delta \leq \min\{S_0, S_1\}. \tag{6.79}$$

Given S_0, S_1 and δ, the other probabilities are

$$\alpha = S_0 + S_1 - \delta, \tag{6.80}$$
$$\beta = S_0 - \delta, \tag{6.81}$$
$$\gamma = S_1 - \delta. \tag{6.82}$$

Using this result, the set $\Lambda(S_0, S_1)$ can be defined by varying a single parameter:

$$\Lambda(S_0, S_1) = \big\{[S_0 + S_1 - \delta, S_0 - \delta, S_1 - \delta, \delta] \mid \max\{0, S_0 + S_1 - 1\} \leq \delta \leq \min\{S_0, S_1\}\big\}. \tag{6.83}$$

This allows us to express Eq. (6.75) into an integral over a real interval:

$$f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(m; S_0, S_1) = \frac{1}{\mathrm{B}(m)} \int_{\max\{0, S_0+S_1-1\}}^{\min\{S_0,S_1\}} (S_0+S_1-\delta)^{a-1}(S_0-\delta)^{b-1}(S_1-\delta)^{c-1}\delta^{d-1}\,\mathrm{d}\delta. \tag{6.84}$$

□

6.4.4 Generalized Dirichlet distribution

As discussed above, the bivariate beta distribution BB(m) enforces an inductive bias on the distribution fitted to the scores $S_0(\boldsymbol{x})$ and $S_1(\boldsymbol{x})$. In particular, this distribution ensures that the scores have marginal beta distributions Beta(a_0, b_0) and Beta(a_1, b_1) such that $a_0 + b_0 = a_1 + b_1$. Furthermore, the bivariate beta distribution puts constraints on the possible covariance structures between the counterfactuals $\boldsymbol{\alpha}, \dots, \boldsymbol{\delta}$, since the distribution has only four parameters, while the covariance matrix of $\boldsymbol{\alpha}, \dots, \boldsymbol{\delta}$ contains 10 possiblity different values. In fact, Ongaro and Migliorati (2013) showed that the covariance between counterfactuals is restricted to be proportional to the product between their marginal expected values. Although these two constraints can be useful inductive biases, it can be interesting to see if alleviating them can improve the estimation of counterfactuals.

To reduce the inductive bias imposed by the bivariate beta distribution, we adapt the procedure described in Section 6.4.1 with the *generalized Dirichlet distribution* (Connor and Mosimann, 1969). This distribution, as its name suggests, is a generalization of the Dirichlet distribution. It has the same domain, but more parameters. When considering four-dimensional vectors, it is defined with six positive real parameters, which we note $a_1, a_2, a_3, b_1, b_2, b_3$. The generalized Dirichlet reduces to the Dirichlet Dir(a, b, c, d) when the following equalities hold:

$$a_1 = a \qquad a_2 = b \qquad a_3 = c \tag{6.85}$$
$$b_1 = a_2 + b_2 \qquad b_2 = a_3 + b_3 \qquad b_3 = d \tag{6.86}$$

We note a random vector $\boldsymbol{\mu} = [\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}]$ following the generalized Dirichlet distribution as $\boldsymbol{\mu} \sim \mathrm{GD}(a_1, \dots, b_3)$. Its probability density function is

$$f_{\boldsymbol{\mu}}(\mu) = \frac{1}{\prod_{j=1}^{3} \mathrm{B}(a_j, b_j)} \alpha^{a_1-1}\beta^{a_2-1}\gamma^{a_3-1}\delta^{b_3-1}(\beta+\gamma+\delta)^{b_1-a_2-b_2}(\gamma+\delta)^{b_2-a_3-b_3}. \tag{6.87}$$

All the equations and results used in Sections 6.4.2 and 6.4.3 can be adapted to the generalized Dirichlet distribution, but we defer the derivations to Appendix D.2. The resulting bivariate distribution is named the *generalized bivariate beta distribution* (GBB for short).[4] This distribution is noted $(\boldsymbol{S}_0, \boldsymbol{S}_1) \sim \text{GBB}(a_1, \dots, b_3)$.

6.4.5 Noisy predictions

Due to the finite size of the training set, among other factors, the scores $\hat{S}_0(x), \hat{S}_0(x)$ predicted by the uplift model in the first step of the learning phase described in Section 6.4.2 are subject to estimation errors. However, in the inference phase, we implicitly assume that we have the exact scores $S_0(x)$ and $S_1(x)$ to infer the counterfactual probabilities. It could be desirable to take into account the uncertainty of the predicted scores.

For that, we develop another extension of the bivariate beta distribution from Olkin and Trikalinos (2015) where some noise is added to the scores $\boldsymbol{S}_0, \boldsymbol{S}_1$. In the simulation of Section 5.3.3, we considered noisy scores following a normalized binomial distribution $\frac{1}{v}\text{B}(S_t, v)$ (with a positive integer v) to emulate the fact that scores are learned from a set of realizations of the binary outcome $\boldsymbol{y}$. However, the possible values that can be generated are limited to the set $\{1/v, 2/v, \dots, 1\}$. In this section, the noisy estimates follow instead a beta distribution. More precisely, the the noisy estimates $\hat{\boldsymbol{S}}_t$ of the scores $\boldsymbol{S}_t$ are sampled as

$$\hat{\boldsymbol{S}}_0 \sim \text{Beta}(\lambda_0 S_0, \lambda_0(1 - S_0)) \tag{6.88}$$

$$\hat{\boldsymbol{S}}_1 \sim \text{Beta}(\lambda_1 S_1, \lambda_1(1 - S_1)) \tag{6.89}$$

where $\lambda_0, \lambda_1 > 0$ are real scale parameters. A higher value of λ_t leads to a lower estimator variance. Also, the expected value of $\hat{\boldsymbol{S}}_0$ given the observation $\boldsymbol{S}_0 = S_0$ is S_0, emulating an unbiased estimator. In fact, we have, for $t = 0, 1$,

$$\mathbb{E}[\hat{\boldsymbol{S}}_t \mid \boldsymbol{S}_t = S_t] = \frac{\lambda_t S_t}{\lambda_t} = S_t, \tag{6.90}$$

$$\text{Var}(\hat{\boldsymbol{S}}_t \mid \boldsymbol{S}_t = S_t) = \frac{\lambda_t S_t \lambda_t (1 - S_t)}{\lambda_t^2(\lambda_t + 1)} = \frac{S_t(1 - S_t)}{\lambda_t + 1}. \tag{6.91}$$

This confirms that higher values of λ_t induce a lower variance of the noise component.

The resulting bivariate distribution, which we name the noisy bivariate beta distribution, noted $\text{NBB}(m, \lambda_0, \lambda_1)$, is sampled as follows.

$$\boldsymbol{\mu} \sim \text{Dir}(m) \tag{6.92}$$

$$\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta} \tag{6.93}$$

$$\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta} \tag{6.94}$$

$$\hat{\boldsymbol{S}}_t \sim \text{Beta}(\lambda_t \boldsymbol{S}_t, \lambda_t(1 - \boldsymbol{S}_t)) \quad \text{for } t = 0, 1. \tag{6.95}$$

We can adapt the procedure of Sections 6.4.2 and 6.4.3 to fit the value of m from the data and predict the probability of counterfactuals. We defer the derivation of these adaptations to Appendix D.3. An important aspect to consider is also to fit the value of the parameters λ_0 and λ_1, which are proportional to the variance of the estimators of the uplift terms. We propose to repeat n times a k-fold cross-validation scheme to generate a set of n different predictions $\left\{\hat{S}_t^{(1)}(x^{(i)}), \dots, \hat{S}_t^{(n)}(x^{(i)})\right\}$ for each sample $x^{(i)}$,

[4] This is a slight abuse of terminology, since the marginals do not necessarily have beta distributions.

and then fit a beta distribution on this set of predictions. The scale parameter $\lambda_t^{(i)}$ for the sample $x^{(i)}$ is found using the standard method of moments,

$$\lambda_t^{(i)} = \frac{R_1^{(i)} - R_2^{(i)}}{R_2^{(i)} - \left(R_1^{(i)}\right)^2} \tag{6.96}$$

where we use

$$R_1^{(i)} = \frac{1}{n}\sum_{j=1}^{n} \hat{S}_t^{(j)}(x^{(i)}) \quad \text{and} \quad R_2^{(i)} = \frac{1}{n}\sum_{j=1}^{n} (\hat{S}_t^{(j)}(x^{(i)}))^2. \tag{6.97}$$

We then assume that all samples have, on average, the same noise scale parameter $\lambda_t^{(i)}$, hence we compute the overall scale parameter λ_t with the average

$$\lambda_t = \frac{1}{N}\sum_{i=1}^{N} \lambda_t^{(i)}. \tag{6.98}$$

The parameters λ_0, λ_1 can be further optimized, together with m, using the method of moments described in Section 6.4.3. However, we found that optimizing λ_0, λ_1 does not bring about any significant improvement over fixing their values with the heuristic described above.

6.4.6 Combining the two previous approaches

We can combine the variations described in Sections 6.4.4 and 6.4.5 by using the following sampling procedure:

$$\boldsymbol{\mu} \sim \mathrm{GD}(a_1, \dots, b_3) \tag{6.99}$$
$$\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta} \tag{6.100}$$
$$\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta} \tag{6.101}$$
$$\hat{\boldsymbol{S}}_t \sim \mathrm{Beta}(\lambda_t \boldsymbol{S}_t, \lambda_t(1 - \boldsymbol{S}_t)) \quad \text{for } t = 0, 1. \tag{6.102}$$

We note this distribution as $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NGBB}(a_1, \dots, b_3, \lambda_0, \lambda_1)$. The properties of this distribution required for the learning phase and the inference phase are derived in Appendix D.4.

6.5 Assessment with simulations

In this section, we assess the bounds and estimators presented in Sections 6.2 to 6.4 with simulated data. This allows to compare the estimators with the ground truth, which is not feasible with real data. As in Sections 5.3.2 and 5.3.3, we use two different simulations: a data-generating process based on a Dirichlet distribution, and another one based on a Gaussian distribution. With both simulations, we generate a large number of samples and evaluate the various models developed in the previous sections. This experiment is repeated a number of times with different parameters to obtain average statistics of the performance of the different models.

6.5.1 Dirichlet simulation

The Dirichlet simulation is similar to the simulation used in Section 5.3.3 and is based on the same probability distribution as that used in the model of Section 6.4. The main difference from the simulation described in Section 5.3.3 is that the noisy estimators are sampled from a beta distribution rather than a normalized binomial distribution. We discussed the advantages of the beta distribution for emulating noisy estimators in Section 6.4.5. The data-generating process is as follows.

$$\boldsymbol{\mu} \sim \text{Dir}(a, b, c, d) \tag{6.103}$$

$$(\boldsymbol{y}_0, \boldsymbol{y}_1) \sim \text{Cat}(\boldsymbol{\mu}) \tag{6.104}$$

$$\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta} \tag{6.105}$$

$$\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta} \tag{6.106}$$

$$\hat{\boldsymbol{S}}_t \sim \text{Beta}(\lambda_t \boldsymbol{S}_t, \lambda_t(1 - \boldsymbol{S}_t)) \quad \text{for } t = 0, 1. \tag{6.107}$$

The Dirichlet simulation is defined by seven parameters: N (then number of samples), a, b, c, d, λ_0 and λ_1. Various properties of this distribution are derived in Appendix D. Here, we state the properties relevant to understanding the results presented in the following sections.

- The parameter N represents the size of the dataset on which the bounds and estimators are evaluated. Lower values of N lead to a larger variance in the final results.
- Parameters a, b, c, d are proportional to the distribution of counterfactuals $P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1), \dots, P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1)$. For example, using the moments of the Dirichlet distribution, we have

  $$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0) = \mathbb{E}[\boldsymbol{\beta}] = \frac{b}{M} \tag{6.108}$$

  where $M = a + b + c + d$.
- The value of M influences the distribution of $\boldsymbol{\alpha}, \dots, \boldsymbol{\delta}$. High values of M lead to samples of $\boldsymbol{\alpha}, \dots, \boldsymbol{\delta}$ to be more concentrated around their expected values (which can be computed from Eq. 6.108), while low values of M lead to samples where one of $\boldsymbol{\alpha}, \dots, \boldsymbol{\delta}$ is close to one while the three other values are close to zero. This has an impact on the scores $\boldsymbol{S}_0, \boldsymbol{S}_1$ as well: they are close to their expected values when M is large, and close to either zero or one when M is low. In loose terms, the quantity M represents the amount of information that the emulated covariates $\boldsymbol{x}$ brings about the outcomes $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$: when the features are not informative, the scores $S_0(x), S_1(x)$ are close to their prior probabilities $P(\boldsymbol{y}_0 = 1)$ and $P(\boldsymbol{y}_1 = 1)$, while when the features are highly informative, the scores are close to either zero or one. The exact relationship between the value of M and the entropy of the potential outcomes is formalized in Result D.1.
- Parameters λ_0 and λ_1 influence the variance of the simulated uplift model. Higher values of λ_t (for $t = 0, 1$) induce a lower variance, as demonstrated in Eq. (6.91).

In this experiment, we generate $N = 5000$ samples. The Dirichlet parameters a, b, c, d are determined as $[a, b, c, d] = M[\alpha, \beta, \gamma, \delta]$ where M is sampled between 0.01 and 10, and the vector $[\alpha, \beta, \gamma, \delta]$ is sampled uniformly over the simplex. The noise parameters

λ_0 and λ_1 are sampled uniformly in $[500, 1500]$. The whole experiment (sampling the data and evaluating the different models) is repeated 30 times.

Theorem 6.2 indicates that the bias of the point estimators $\hat{\alpha}, \dots, \hat{\delta}$ based on the conditional independence assumption of Section 6.3 (transposed to the notation of this section) is

$$\mathbb{E}[\boldsymbol{\phi}] - \mathbb{E}_{\boldsymbol{\alpha},\dots,\boldsymbol{\delta}}[\text{cov}(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1)] \tag{6.109}$$

where $\boldsymbol{\phi} = \boldsymbol{\alpha}\boldsymbol{\delta} - \boldsymbol{\beta}\boldsymbol{\gamma}$. The second term in Eq. (6.109) is null because we sample $\hat{\boldsymbol{S}}_0$ and $\hat{\boldsymbol{S}}_1$ independently in Eq. (6.107), but we can show using the product moments of the Dirichlet distribution (Lin, 2016) that the first term $\mathbb{E}[\boldsymbol{\phi}]$ is

$$\mathbb{E}[\boldsymbol{\phi}] = \mathbb{E}[\boldsymbol{\alpha}\boldsymbol{\delta} - \boldsymbol{\beta}\boldsymbol{\gamma}] = \frac{ad - bc}{M(M+1)}. \tag{6.110}$$

In this experiment, the parameters a, b, c, d are sampled uniformly; therefore, the expression in Eq. (6.110) will generally be different from zero, and the bias of the point estimators $\hat{\alpha}, \dots, \hat{\delta}$ will also be different from zero. This is desirable to assess how violations of the hypothesis underlying our estimators affect their performance.

6.5.2 Gaussian simulation

In the case of the Gaussian simulation, the procedure is the same as that used in Section 5.3.2. The data-generating process is as follows.

$$\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \sim \mathcal{N}(0, I_n) \tag{6.111}$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1) \tag{6.112}$$

$$\boldsymbol{y}_t = \mathbb{I}[\lambda_t^T \boldsymbol{x} + \boldsymbol{\varepsilon} \geq \eta_t] \quad \text{for } t = 0, 1 \tag{6.113}$$

$$S_t(x) = \Phi(\lambda_t^T x - \eta_t) \tag{6.114}$$

where I_n is the identity matrix of size $n \times n$, $\lambda_t \in \mathbb{R}^n$ is a vector of the weights of the features $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ used to determine the value of $\boldsymbol{y}_t$, together with the threshold η_t. In total, the Gaussian distribution is defined by six parameters: $N, n, \lambda_0, \lambda_1, \eta_0$ and η_1.

- Similarly to the Dirichlet simulation, the parameter N is the number of samples used to evaluate the estimators.

- The parameter n is the number of features of the emulated uplift model. Since all features $\boldsymbol{x}_i$ and the noise $\boldsymbol{\varepsilon}$ have the same distribution, a larger number of features gives less importance to the noise $\boldsymbol{\varepsilon}$ in Eq. (6.113). Therefore, this emulates the fact that the feature vector contains more information about the potential outcomes.

- $\lambda_t = [\lambda_{t,1}, \dots, \lambda_{t,n}]$ is a vector of parameters corresponding to the weight of each feature in the value of $\boldsymbol{y}_t$. Larger weights have a similar impact as a larger number of features n, in that the importance of the noise $\boldsymbol{\varepsilon}$ is reduced, emulating more informative features.

- η_t is the threshold to determine the value of $\boldsymbol{y}_t$, which in turn influences the counterfactuals probabilities $\boldsymbol{\alpha}, \dots, \boldsymbol{\delta}$.

Table 6.2 Mean and standard deviation of the squared error of the different models when estimating population-level counterfactuals $\alpha, \ldots, \delta$. Note the different exponents for the first two models. The models based on bivariate beta distribution consistently perform best.

Model	Dirichlet simulation	Gaussian simulation
Independence	$(1.79 \pm 5.62) \times 10^{-3}$	$(1.25 \pm 0.32) \times 10^{-2}$
Midpoint	$(1.16 \pm 3.63) \times 10^{-3}$	$(1.91 \pm 0.55) \times 10^{-2}$
BB	$(0.87 \pm 1.76) \times 10^{-5}$	$(0.36 \pm 1.24) \times 10^{-5}$
NBB	$(0.88 \pm 1.75) \times 10^{-5}$	$(0.35 \pm 1.24) \times 10^{-5}$
GBB	$(0.86 \pm 1.72) \times 10^{-5}$	$(0.78 \pm 3.17) \times 10^{-6}$
NGBB	$(0.88 \pm 1.75) \times 10^{-5}$	$(1.14 \pm 3.32) \times 10^{-6}$

Table 6.3 Mean and standard deviation of the squared error of the different models when estimating $\alpha(x), \ldots, \delta(x)$. Results for NBB and NGBB are not shown due to prohibitively long computation times on the Dirichlet simulation.

Model	Dirichlet simulation	Gaussian simulation
Independence	$(0.33 \pm 1.12) \times 10^{-2}$	$(1.46 \pm 0.56) \times 10^{-2}$
Midpoint	$(0.23 \pm 0.85) \times 10^{-2}$	$(2.23 \pm 0.71) \times 10^{-2}$
BB	$(0.18 \pm 3.11) \times 10^{-2}$	$(0.67 \pm 1.56) \times 10^{-4}$
NBB	–	$(0.72 \pm 1.45) \times 10^{-4}$
GBB	$(0.14 \pm 2.78) \times 10^{-2}$	$(0.69 \pm 1.91) \times 10^{-4}$
NGBB	–	$(0.71 \pm 1.38) \times 10^{-4}$

6.5.3 Results

The performance of the different models for the estimation of population-level counterfactuals $\alpha, \ldots, \delta$ are shown in Table 6.2. We report the squared error between the true value and the estimated value, averaged over the four counterfactuals and all 30 runs. To have a baseline for comparison, we also report the performance of the midpoint of the uplift bounds used as a point estimator. We see that the uplift bounds midpoint has an error slightly lower than the independence estimator. All four methods based on fitting a bivariate beta distribution perform much better than the independence and the midpoint estimators, with an error lower by two orders of magnitude. The generalized beta approaches, GBB and NGBB, provide a slight increase in performance, which is especially noticeable on the Gaussian simulation.

The results for the estimation of individual-level counterfactuals $\alpha(x), \ldots, \delta(x)$ are shown in Table 6.3 and Fig. 6.1. Although fitting NBB and NGBB to the data takes approximately as long as BB and GBB, computing counterfactuals involves two additional levels of integration, which significantly increases computation time. Therefore, we are not able to report their results for individual-level counterfactual estimation on the Dirichlet simulation. We see in Table 6.3 that the independence estimator performs the worst, while GBB shows the best performance, although not by a large margin. As in the case of population-level counterfactuals, the generalized beta distribution seems to provide a slight improvement in performance for estimating individual-level counterfactuals.

The average span of the Fréchet and uplift bounds on the population-level counter-

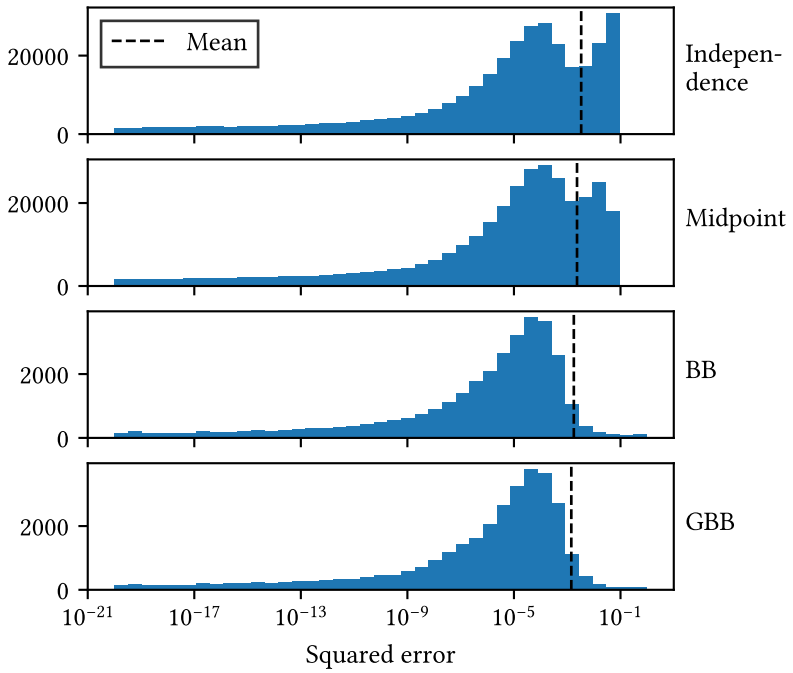


Figure 6.1 Distribution of the squared error of the different models when estimating $\alpha(x), \dots, \delta(x)$ on the Dirichlet simulation. Results for NBB and NGBB are not shown due to prohibitively long computation times.

Table 6.4 Mean and standard deviation of the span of the Fréchet and Uplift bounds.

Model	Dirichlet simulation (%)	Gaussian simulation (%)
Fréchet bounds	11.8 ± 12.6	31.7 ± 2.3
Uplift bounds	3.4 ± 5.9	27.4 ± 4.3

factuals $\alpha, \dots, \delta$ is shown in Table 6.4 and Fig. 6.2. The uplift bounds are significantly tighter than the Fréchet bounds with the Dirichlet simulation, with a span reduced by a factor of approximately 3.5 on average. The improvement is not as dramatic with the Gaussian simulation, but it is still significant.

We can summarize the results presented in this section as follows:

- The four models based on fitting the bivariate beta distribution show very good performances for the estimation of population-level counterfactuals.
- The models based on the generalized Dirichlet distribution (GBB and NGBB, described in Section 6.4.4) provide an improvement in performances compared to BB and NBB, at the cost of an increased number of parameters
- Models incorporating the noise of the uplift scores in the model (NBB and NGBB, as described in Section 6.4.5), do not seem to provide a significant advantage in terms of performances, while suffering from a large increase in computation time for individual-level counterfactuals.
- The BB and GBB approaches are the most effective for estimating individual-level counterfactuals.

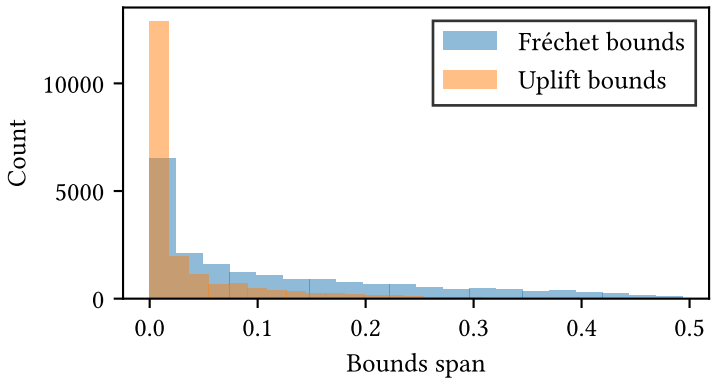


Figure 6.2 Distribution of the bounds span for the Fréchet and uplift bounds on the Dirichlet simulation.

- The uplift bounds significantly improve over the state-of-the-art Fréchet bounds.

6.5.4 Sensitivity analysis

In this section, we assess the influence of the training data on the quality of the estimation in terms of the number of samples, the variance of the estimator, or the amount of information between features and outcomes. We plot the span of the uplift bounds and the error of the point estimator while varying one of the parameters M, N or λ_0, λ_1 and keeping the other parameters fixed. The values of the fixed parameters are selected to clearly show the influence of the varying parameter. In particular, we set $[\alpha, \beta, \gamma, \delta] = [0.947, 0.020, 0.017, 0.017]$ based on the results of Section 6.6, which represents the distribution of counterfactuals in a typical scenario of customer churn in telecom. The main conclusions of this sensitivity analysis are:

- The span of the uplift bounds decreases as M decreases (Fig. 6.3), which is itself directly linked to the mutual information between $\boldsymbol{y}_0, \boldsymbol{y}_1$ and $\boldsymbol{x}$. This is an empirical illustration of Theorem 6.1. The relationship between M and the mutual information is formalized in Result D.1. We see that as M approaches zero (which emulates very informative features), the bounds span converges towards zero as well.

- The variance of the point estimator decreases as the number of samples N increases (Fig. 6.4) or the estimator variance $\mathrm{Var}(\hat{\boldsymbol{S}}_t)$ decreases (Fig. 6.5). In fact, the error converges towards the bias derived in Theorem 6.2. This demonstrates the convergence of our estimator in the large sample scenario.

- The span of the uplift bounds increases as the estimator variance $\mathrm{Var}(\hat{\boldsymbol{S}}_t)$ (for $t = 0, 1$) decreases (Fig. 6.6). This is because a model with high variance often predicts scores lower or higher than the expected score. Since the bounds span is $\mathbb{E}[\min\{\boldsymbol{S}_0, \boldsymbol{S}_1, 1 - \boldsymbol{S}_0, 1 - \boldsymbol{S}_1\}]$ (see Eq. 6.31), this artificially reduces the bounds span.

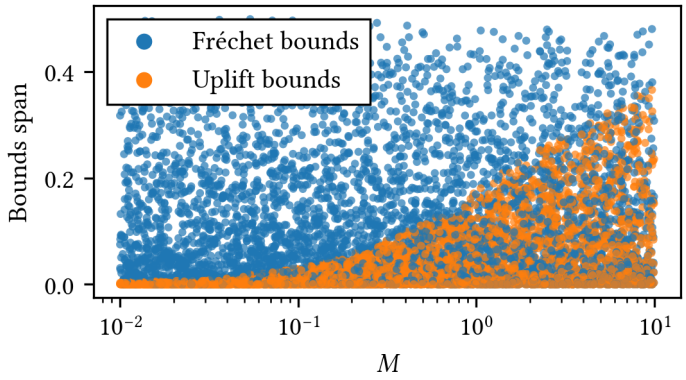


Figure 6.3 The bounds span as a function of the parameter M, which is directly linked to the conditional entropy $H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu})$, on the Dirichlet simulation. We fixed the counterfactual probabilities $[\alpha, \beta, \gamma, \delta] = [0.947, 0.020, 0.017, 0.017]$, $\lambda_0 = \lambda_1 = 200$ and $N = 5000$.

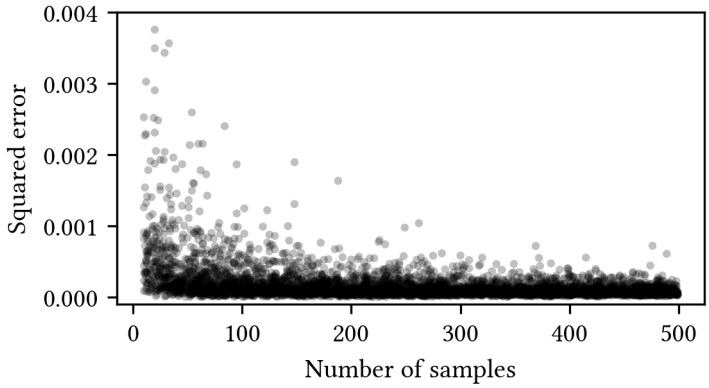


Figure 6.4 The error of the point estimator as a function of the number of samples in the evaluation dataset on the Dirichlet simulation. We fixed $[\alpha, \beta, \gamma, \delta] = [0.947, 0.020, 0.017, 0.017]$, $\lambda_0 = \lambda_1 = 200$ and $M = 1$.

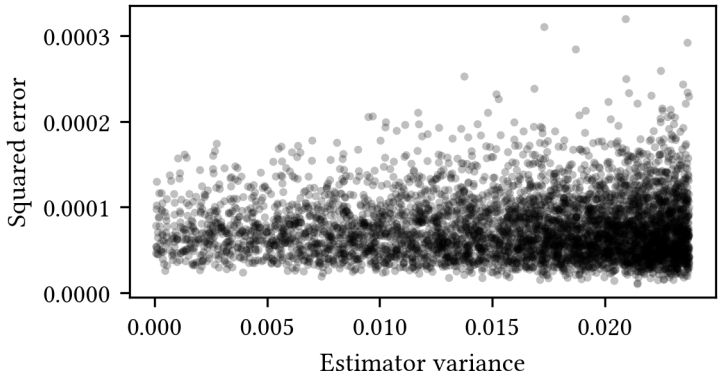


Figure 6.5 The error of the point estimator as a function of estimator variance $\mathrm{Var}(\hat{S}_t)$. We fixed $[\alpha, \beta, \gamma, \delta] = [0.947, 0.020, 0.017, 0.017]$, and $N = 2000$ and $M = 1$. As the variance decreases, the estimator bias converges towards to its theoretical value in the large sample scenario.

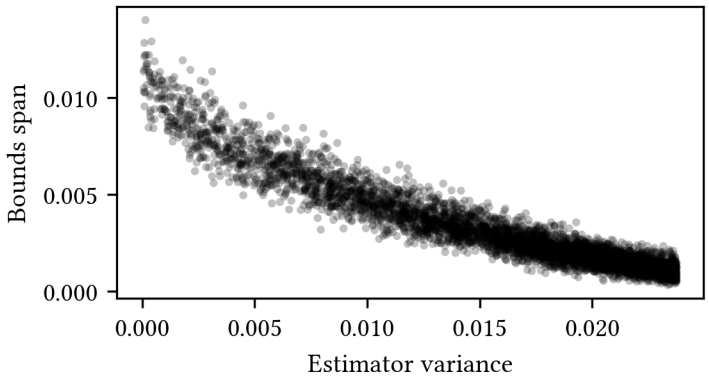


Figure 6.6 The uplift bounds span as a function of the estimator variance $\mathrm{Var}(\hat{S}_t)$. We fixed $[\alpha, \beta, \gamma, \delta] = [0.947, 0.020, 0.017, 0.017]$, and $N = 2000$ and $M = 1$. A lower estimator variance is shown here to be associated with larger bounds. In fact, as the variance goes to zero (left side of the plot), the bounds span converges towards its theoretical value. A model with a high variance predicts more often low values, which artificially reduces the bounds span.

6.6 Evaluation with real data

In this section, we apply the estimators and theoretical results presented in this chapter to a real-world dataset provided by our industrial partner Orange Belgium, the Churn 2 dataset, presented in Section 4.1. First, in Section 6.6.1, we detail our methodology and report the estimated distribution of counterfactuals over the whole campaign and separately for each month in Section 6.6.2. Then, we analyze the distribution of some customer features according to their inferred counterfactual category (*persuadable*, *do-not-disturb*, etc.) in Section 6.6.3. Finally, we perform a profit analysis highlighting the potential gains suggested by the estimated number of persuadable customers in Section 6.6.4.

6.6.1 Methodology

We train an T-learner uplift model (see Section 3.1.2) on the Churn 2 dataset, augmented with the `R-feature` methodology presented in Section 4.4.1. In addition to having good performances, this model provides separate estimators for $S_0(x)$ and $S_1(x)$, which is essential for computing the uplift bounds and the point estimators. Similarly to the benchmark in Section 4.2, we use a random forest as the base learner with 100 trees, a maximum depth of 20 and a minimum of 10 samples per leaf. Given the high imbalance of the datasets, we also rely on the EasyEnsemble strategy (X.-Y. Liu, Wu, and Zhou, 2009) for class balancing, described in Section 2.3.6.

When using a resampling strategy such as EasyEnsemble to obtain a balanced dataset, the prior probability of churn is modified (Batista, Prati, and Monard, 2004), and the scores predicted by the model are biased. We correct this bias with the calibration formula derived by Dal Pozzolo, Caelen, Johnson, et al. (2015). Let $\boldsymbol{s}$ be a binary random variable equal to zero for samples discarded by the resampling strategy, and equal to one for the selected samples. For example, with EasyEnsemble, all positive samples are selected ($\boldsymbol{s} = 1$ if $\boldsymbol{y} = 1$), and a fraction of the negative samples are selected uniformly at random, that is, $P(\boldsymbol{s} = 1 \mid \boldsymbol{y} = 0)$ is equal to the proportion of under-

sampling for negative outcomes. Let $\tilde{S}(x) = P(\boldsymbol{y} = 1 \mid \boldsymbol{x} = x, \boldsymbol{s} = 1)$ be the posterior probability of a sample x predicted by a model after resampling. The objective of score calibration is to find the real posterior probability $P(\boldsymbol{y} = 1 \mid \boldsymbol{x} = x)$. The calibration formula is

$$P(\boldsymbol{y} = 1 \mid \boldsymbol{x} = x) = \frac{\mu\tilde{S}(x)}{\mu\tilde{S}(x) - \tilde{S}(x) + 1} \tag{6.115}$$

where $\mu = P(\boldsymbol{s} = 1 \mid \boldsymbol{y} = 0)$ is the probability of selecting a negative sample during resampling. In practice, we compute it using the identity

$$\mu = \frac{P(\boldsymbol{y} = 0 \mid \boldsymbol{s} = 1)P(\boldsymbol{y} = 1)}{P(\boldsymbol{y} = 1 \mid \boldsymbol{s} = 1)P(\boldsymbol{y} = 0)}. \tag{6.116}$$

In our case, the scores $\tilde{S}_0(x)$ and $\tilde{S}_1(x)$ predicted by the uplift model are calibrated separately using samples from the control and target groups.

To avoid overfitting a given train-test split and to reduce the impact of sampling error, we use a k-fold cross-validation scheme with $k = 3$, and this scheme is repeated 10 times. This results in a total of 10 predictions by 10 different models for each data sample. The different predictions for each sample are averaged before computing the bounds and point estimators.

6.6.2 Estimated counterfactual probabilities

The estimated distribution of counterfactuals is reported in Fig. 6.7 and Table 6.5. In Fig. 6.7, we report the point estimates and the bounds on all four counterfactuals. As in Section 6.5.3, we use the midpoint of the uplift bounds as a baseline point estimate. The point estimates of GBB and NGBB are not shown in Fig. 6.7 because they are on the same position as the point estimate of NBB. The bounds are reported as diagonal lines to highlight the interdependence between the counterfactuals; if, for example, α takes its maximum value according to the uplift bounds, then β takes its minimum value, and similarly for γ and δ. Therefore, the set of possible counterfactual probabilities according to some bounds forms two diagonal lines in Fig. 6.7.

We observe that the uplift bounds are tighter than the Fréchet bounds, although not by a large margin. Also, point estimates based on bivariate beta distributions have similar values and they all predict smaller values for β and γ than the independence estimator and the uplift bounds midpoint. This indicates that they predict a lower number of persuadable or do-not-disturb customers, which corresponds to a smaller causal effect of the campaign, either positive or negative.

We show the probability density function (pdf) of the distributions fitted by BB and GBB in Fig. 6.8. We can see that the pdf of GBB is less symmetrical than that of BB, highlighting its greater flexibility which allows it to fit more complex patterns in the data. We computed the log-likelihood of the dataset from both models, obtaining 74185 for BB and 73745 for GBB, which seems to contradict the intuition that GBB fits better the data. However, we repeated the learning process several times using randomized initial parameters, resulting in a more thorough exploration of the parameter space. While BB systematically converged to the same parameter values, GBB converged to different parameter values, resulting in a better log-likelihood than BB in approximately 75% of the cases. The best log-likelihood that we obtained with this process is 75472.

The proportion of persuadable customers is estimated by BB as $\hat{\beta} = 1.7\%$, with a lower uplift bound of 0.70% and an upper uplift bound of 3.64%. This amounts to 203 customers, bounded between 84 and 433. This indicates that a maximum of 433 customers

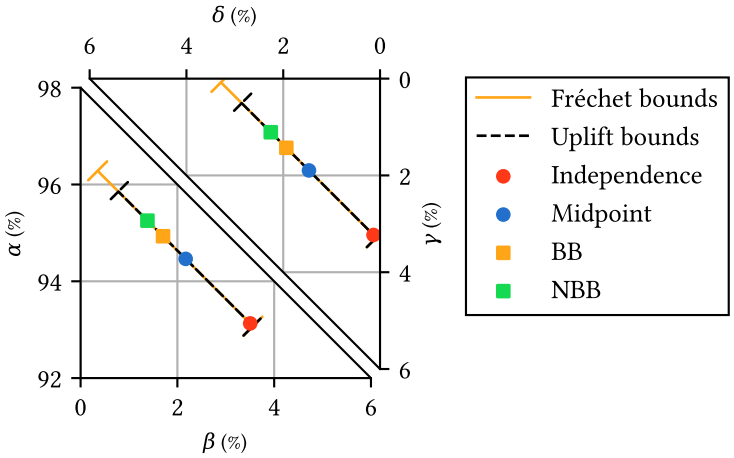


Figure 6.7 Point estimate and bounds on $\alpha, \ldots, \delta$ on the Churn 2 dataset. We display the four axes in two separate plots, which can be arranged in a single square thanks to the geometry of the simplex. Pay attention to the fact that the coordinate systems are different in the two triangles. In particular, note the different vertical axis for α. The point estimates of GBB and NGBB are not shown as they are almost equal to that of NBB.

Table 6.5 Point estimates and bounds on $\alpha, \ldots, \delta$ on the Churn 2 dataset.

	α (%)	β (%)	γ (%)	δ (%)
Fréchet bounds	93.00 – 96.36	0.27 – 3.64	0.00 – 3.37	0.00 – 3.37
Uplift bounds	93.00 – 95.93	0.70 – 3.64	0.43 – 3.37	0.00 – 2.94
Independence	93.13	3.50	3.23	0.13
Midpoint	94.46	2.17	1.90	1.47
BB	94.93	1.70	1.43	1.94
NBB	95.25	1.38	1.11	2.26
GBB	95.23	1.41	1.13	2.23
NGBB	95.23	1.40	1.13	2.24

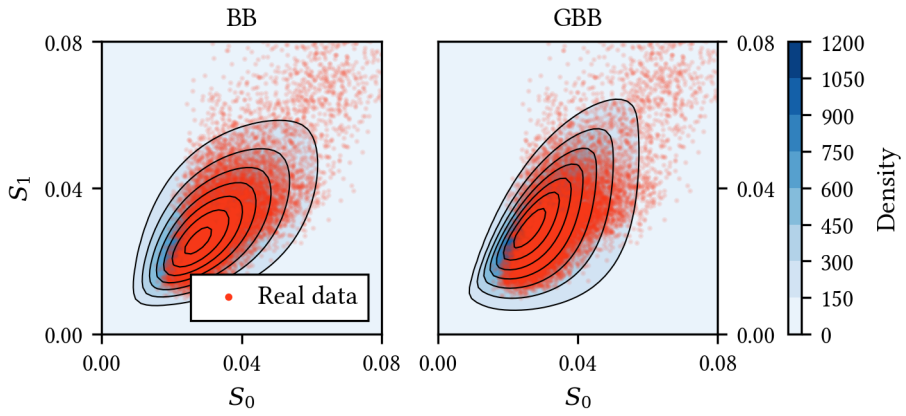


Figure 6.8 Probability density distribution fitted by BB and GBB on the Churn 2 dataset. Notice the asymmetry in the pdf of GBB, which more closely fits the data.

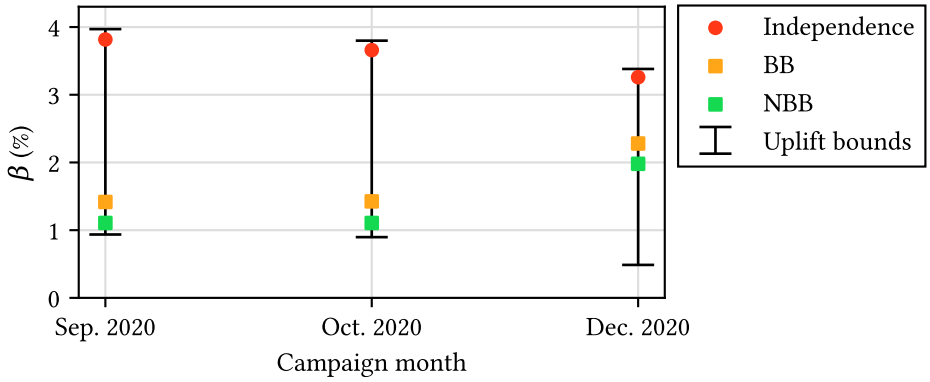


Figure 6.9 Point estimates and uplift bounds on β, for each month of the campaign. The estimations by GBB and NGBB are not shown for clarity as they overlap that of BB and NBB.

should have been called during the 3 campaigns, while in practice 9010 customers have been called. We applied the same methodology separately for each month instead of on the whole dataset at once, and the results are reported in Fig. 6.9. We observe a decrease in the overall efficacy of the campaign (this is confirmed by measuring the campaign uplift over the three months, which is not shown for confidentiality reasons), although the bivariate beta estimators seem to detect an increasing number of persuadable customers. This indicates that campaigners targeted more and more customers without being able to find persuadable customers, leading to a decrease in uplift. As in the previous results, the independence estimator is close to the upper uplift bound. This is because both $\hat{S}_0(x)$ and $\hat{S}_1(x)$ tend to be close to zero, and $\hat{\beta}(x)$ is estimated as $\hat{S}_0(x)(1-\hat{S}_1(x))$ in Eq. (6.49). Therefore, $\hat{\beta}(x)$ is typically close to $\hat{S}_0(x)$, and the upper bound $\min\left\{\hat{S}_0(x), 1-\hat{S}_1(x)\right\}$ from Eq. (6.25) is often equal to $\hat{S}_0(x)$.

6.6.3 Customer profiles with counterfactual estimation

Counterfactual probabilities provide a description of customer behavior in terms of their potential responses to retention efforts, but can also be used to establish a business profile of the different types of customers. In this section, we assign to each existing customer one of the four labels (*sure thing*, *lost cause*, *do-not-disturb*, or *persuadable*), then we analyze the distribution of customer characteristics within these groups. This approach allows for a better understanding of the customer base, the identification of patterns and trends, and enables businesses to optimize marketing efforts and personalize communication strategies.

We apply the counterfactual estimator based on the bivariate beta distribution (BB in Section 6.4) on the Churn 2 dataset to estimate both the population-level counterfactuals $\alpha, \beta, \gamma, \delta$ and the individual-level counterfactuals $\alpha(x^{(i)})$, $\beta(x^{(i)})$, $\gamma(x^{(i)})$, $\delta(x^{(i)})$ for $i = 1, \dots, N$, where N is the number of customers and $x^{(i)}$ represents the features of the client at index i in the dataset. Since we expect to have αN *sure thing* customers, we assign the label *sure thing* to the αN customers with the highest probability $\alpha(x^{(i)})$. The same process is repeated for the *persuadable*, *do-not-disturb* and *lost cause* customers. Note that this strategy does not result in a partition of the N customers, since it is possible for a customer to have two or three labels simultaneously. Our objective here is

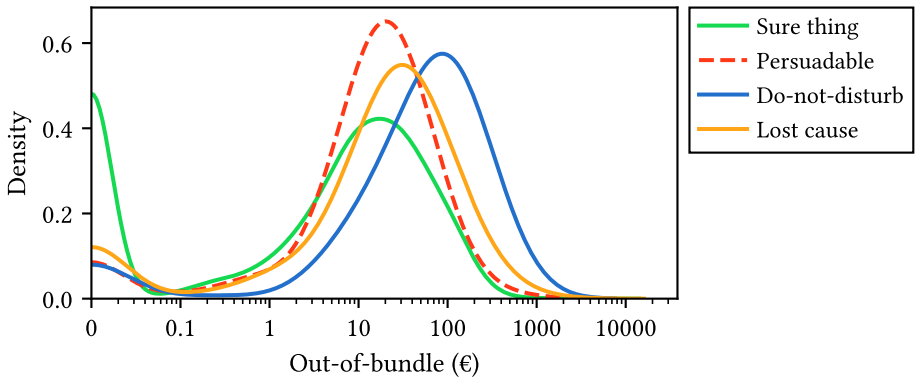


Figure 6.10 Probability density function of the *out-of-bundle* amount (the supplement paid by the customer for services not included in their standard allowances). We see that *persuadable* customers (dotted red line) have a higher *out-of-bundle* than *sure thing* customers, but lower than the two other classes, which are more likely to churn.

to establish a customer profile of each category rather than to find an exact partition, therefore, we do not consider this issue to be critical.

We report the distribution of three different features for each customer category in Figs. 6.10 to 6.12. We report an estimation of the probability density function using a Gaussian kernel, which gives a clearer presentation than histograms. We hide the scale of the feature distribution for confidentiality reasons.

Fig. 6.10 reports the distribution of the *out-of-bundle* amount, that is, the additional fee charged to customers for services not covered in their standard allowances. This out-of-bundle amount is sometimes unexpected for customers and is understood by business experts to be an important driver of customer churn, a phenomenon called *bill shock*, presented in Section 1.3. We see, as expected, that *sure thing* customers, who do not churn regardless of the campaign call, have the lowest out-of-bundle. The pattern of the distribution for the three other categories is interesting; the average out-of-bundle amount is increasing across the *persuadable*, *lost cause* and *do-not-disturb* customers. This indicates that if the extra fee is reasonable, a customer can be convinced to stay, but beyond a certain threshold, retention efforts have the opposite effect. The fact that *do-not-disturb* are the highest spenders can possibly be understood with the graph in Fig. 6.11. We see that *do-not-disturb* customers are markedly younger than the rest of the customer base. This can indicate that young customers tend to be not receptive to marketing calls, and to have a higher out-of-bundle amount.

In Fig. 6.12, we report the tenure, which refers to the length of time that a customer has been associated with the company. We see that new customers are much more likely to be *do-not-disturb*, while long-term customers are most likely *sure thing*. As in the previous graphs, *persuadable* customer lies between the two extremes.

This analysis shows the potential of counterfactual probabilities to unveil important business insights which cannot be inferred using only the probability of churn or the uplift. For example, in Fig. 6.10, *sure thing* and *lost cause* customers display a distinctive behavior, while a model based on uplift alone would not differentiate them because both categories have an uplift close to zero. Similarly, in this same figure, a classical churn model that ranks individuals according to their probability of churn $S_0(x) = P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{x} = x)$ would put the *sure thing* and *do-not-disturb* customers in the same class because

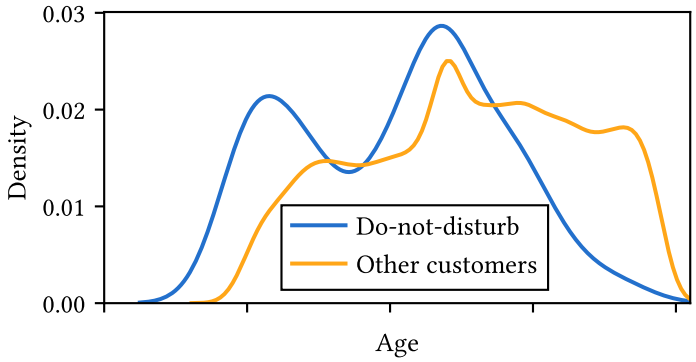


Figure 6.11 Probability density function of the age for do-not-disturb customers. We observe that younger customers are more likely to react negatively to the campaign call.

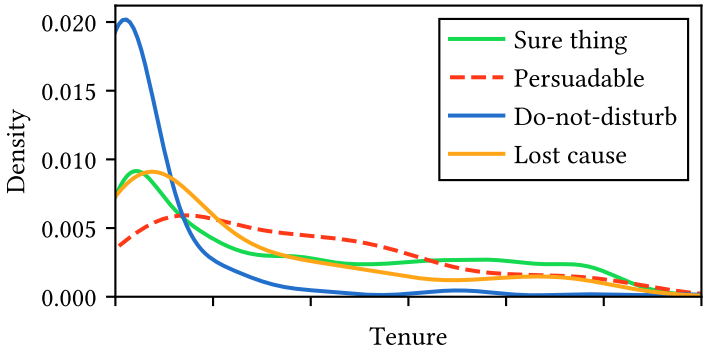


Figure 6.12 Probability density function of the tenure, that is, the duration the customer has been with the company. We observe that new customers are more likely to be negatively impacted by the campaign call (higher probability of being a *do-not-disturb*), and that *sure thing* customers have a higher tenure.

they both have a low score $S_0(x)$, although they have very different characteristics.

6.6.4 Profit analysis

In this section, we perform a simplistic profit analysis based on the estimation of counterfactual probabilities reported in Section 6.6.2. Let us suppose that each call has a cost $C = 0.25€$, and that the average customer lifetime value is $V = 120€$ (a customer pays on average 20€ per month and stays 6 months). The benefit due to the campaign as it actually happened can be computed as

$$\text{Profit} = NUV - NC \tag{6.117}$$

Where N is the number of contacted customers and $U = S_0 - S_1$ is the campaign uplift (approximately 0.3% in our case). The term NUV is the benefit generated by converting customers. The benefit of calling do-not-disturb customers cancels out the benefit of calling persuadable customers, since $U = \beta - \gamma$.[5] The term NC in Eq. (6.117) is the

[5]This can be shown by decomposing $U = P(\boldsymbol{y}_0 = 1) - P(\boldsymbol{y}_1 = 1)$ in terms of β, γ and δ.

cost of calling N customers. To use the notation of Section 5.2.1, this corresponds to a cost-benefit matrix

$$\text{CB} = \begin{bmatrix} 0 & -C \\ V & V-C \end{bmatrix} = \begin{matrix} \\ \boldsymbol{y}=0 \\ \boldsymbol{y}=1 \end{matrix} \begin{matrix} \boldsymbol{t}=0 & \boldsymbol{t}=1 \\ \end{matrix} \begin{bmatrix} 0 & -0.25 \\ 120 & 119.75 \end{bmatrix}$$

By evaluating Eq. (6.117) on the Churn 2 dataset, we find that the campaign incurred a profit of €724. However, if we suppose that we were able to call only the 271 persuadable customers, the potential profit would be

$$\text{Potential profit} = N\beta(V-C) \tag{6.118}$$

where $N\beta$ is the number of contacted customers in this ideal campaign, and $V-C$ is the profit generated by convincing them to stay, minus the cost of the call. For the Churn 2 dataset, this results in a profit of up to 32479€. Note that this is a simplistic way to evaluate the profit generated by a campaign. For more detailed formulas of the profit of a campaign, we refer the reader to Section 5.2.2.

6.7 Discussion

In this section, we discuss some of the potential shortcomings of our estimators and experimental results.

The improvement of the uplift bounds with respect to the Fréchet bounds is directly related to the quantity of information between the features and the outcome (see Theorem 6.1). The small improvement observed in practical applications, as shown in Fig. 6.7, indicates that the uplift terms, and in turn counterfactual probabilities, are difficult to accurately estimate in real-world settings such as customer churn prediction. A possible solution would be to add more informative features or design a more powerful uplift model. The uplift bounds can also be further refined when observational data is available (i.e., data where the treatment assignment is not randomized), as demonstrated in (Mueller and Pearl, 2022). The results of this chapter provide nonetheless valuable insights for Orange Belgium on the potential value of past retention campaigns and on the distribution of the different customer categories.

The four models based on the bivariate beta distribution, presented in Section 6.4, perform much better in terms of population-level counterfactual probabilities than the model based on the conditional independence assumption of Section 6.3 or the midpoint of the uplift bounds. These four models have varying degrees of complexity, which allows practitioners to choose the most appropriate model in the context of application. The two models which incorporate the variance of the uplift scores estimator, NBB and NGBB, suffer from a much larger computation time when computing individual counterfactuals. This can limit its applicability in practical scenarios involving large datasets.

All of our estimators are influenced by the choice of the underlying uplift model. The uplift model should be unbiased, and the quality of the estimator depends on the quality of the uplift model. Since the two uplift terms $S_0(x)$ and $S_1(x)$ are used independently in our estimators, we are also limited to uplift estimators that can provide an estimation of these two terms separately.

The results of this chapter do not indicate which customers should be targeted to maximize the profit from the retention campaign. This is the objective of uplift modeling, as discussed in Chapter 5, in which we examined whether uplift modeling is

always the best approach for causal decision-making. We showed that uplift models are suboptimal under some circumstances and that proxy targets such as the probability of the outcome are sometimes more effective for accurate causal decision-making. A. Li and Pearl (2019) consider the case where each of the four categories of customers (*persuadable*, *sure thing*, *lost cause* and *do-not-disturb*, see Table 6.1) have arbitrary associated costs. In this case, counterfactual identification is essential for accurate decision-making. Although we have not explored this area of research, we nonetheless believe that counterfactual identification brings about very valuable insights, even when counterfactual probabilities are not used directly for decision-making. This is shown in Section 6.6.3, where we established a basic profile of the different types of customers depending on their inferred counterfactual outcomes.

6.8 Conclusion

In this chapter, we have derived and empirically assessed new bounds and point estimators on the probability of counterfactuals for binary outcomes under the assumption of unconfoundedness based on the scores predicted by an uplift model.

The proposed uplift bounds improve upon the classical Fréchet bounds by leveraging the scores estimated by an uplift model. We have demonstrated theoretically that the bounds improve as the quality of the uplift estimation increases. Simulated examples indicate that the uplift bounds typically provide a significant improvement over the Fréchet bounds, without requiring any new assumption. This differs from most of the literature on partial counterfactual identification (reviewed in Section 3.2.3), where new results are typically based on specific assumptions on the causal graph.

We have derived a point estimator assuming the conditional independence between the potential outcomes $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ given $\boldsymbol{x}$. The bias of this estimator is theoretically quantified in Theorem 6.2, and simulated examples demonstrate that the estimator is still close to the true value even when the conditional independence assumption is not satisfied.

We have derived four different point estimators by fitting a bivariate beta distribution on the uplift scores, and using the internal representation of the distribution as an estimator of counterfactual probabilities. In particular, population-level counterfactuals are derived from the moments of the fitted distribution, and individual-level counterfactuals are estimated as an expected value conditioned on the observed uplift scores. The four different point estimators are variations of the same basic approach, either incorporating the variance of the uplift terms in the model, or providing a more flexible distribution by increasing the number of parameters. These four models have similar performances in our experiments and largely outperform the other approaches. We recommend that practitioners use BB (Section 6.4.1) or GBB (Section 6.4.4), given their good performance and reasonable computation times compared to the other two variations, NBB and NGBB.

Finally, an evaluation of our counterfactual estimators on customer data from Orange Belgium reveals their potential for discovering interesting patterns of customer behavior. In particular, we found that the behavior of *persuadable* customers lies between that of *sure thing* customers and *do-not-disturb* customers, in terms of tenure and out-of-bundle amount. A simplistic profit analysis also revealed the very large increase in benefit induced by a campaign that would target only persuadable customers. While our approach to customer segmentation based on counterfactual estimation is

more complex and theoretically sophisticated compared to traditional methods relying solely on customer features, this complexity is indispensable for uncovering highly practical insights into customer behavior linked with counterfactual categories.

Part III

Conclusion

The more you know, the less you don't know
Unless what you know is not true

Icelandic proverb
@greipjokes on Instragram

7

Conclusions and future work

Customer churn remains an important concern for large corporations, particularly within the telecommunications sector. Customer retention campaigns are typically conducted to maintain customer loyalty. However, the effectiveness of these efforts is often hindered by the difficulty of precisely identifying and targeting customers based on their historical profiles. The complexity of consumer behavior requires a more nuanced and sophisticated approach.

Uplift modeling, an approach within the field of causal analysis, emerges as an important solution to this issue. Unlike traditional predictive approaches that focus on identifying customers likely to churn, uplift modeling takes into account the causal aspect of customer behavior. It not only predicts the likelihood that a customer will churn but also whether the customer will be positively influenced by targeted retention strategies. The growing adoption of uplift modeling in both industry and academia suggests its potential to improve customer retention strategies. As businesses increasingly recognize the limitations of conventional predictive models, there is a growing interest in methodologies that go beyond mere prediction to understand the causal mechanisms at play.

Despite its increasing prominence, the added value of uplift modeling compared to the traditional predictive approach has rarely been quantified in the existing literature. The assessment of its added value is essential for companies to optimize their customer retention efforts, and for academia to obtain a more nuanced view on modern causal modeling techniques. Evaluating the performance of uplift modeling performance in real-world scenarios, benchmarking it against traditional predictive models, and quantifying its impact on customer retention campaigns are critical steps towards unlocking its full potential and understanding its place in the broader landscape of data-driven decision-making.

7.1 Summary of the contributions

Our research started with an empirical comparison of predictive and uplift modeling. In Section 4.2 we performed a first benchmark of the two approaches on four different datasets, two from our industrial partner Orange Belgium and two publicly available uplift datasets. Although other uplift modeling benchmarks have been published in the literature, very few of them include the predictive approach as a baseline. It appeared

evident from our results that the predictive approach performs as well, if not better, than all other uplift models in three of the four datasets. To confirm this result, in Section 4.3 we compared the performance of both approaches in a series of real customer retention campaigns with a carefully designed experimental protocol. We observed that for all but one month, the uplift model did not outperform the predictive model. Finally, in Section 4.4, we propose a series of new strategies to improve the performance of uplift modeling by integrating information about whether a customer is likely to be reached during the phone campaign (that is, whether they pick up the phone and discuss with the agent). The most efficient strategy improves the performance of an uplift model beyond that of predictive modeling and is applicable to other settings as well. However, its scope is limited to applications where the reach information is available.

In Chapter 5 we provided a theoretical examination of uplift modeling with a dual focus. Firstly, in Section 5.2, we determined the optimal performance measure that is most aligned with the objective of organizations dealing with customer churn. We developed from the first principles a profit measure that can be adjusted to any scenario and encodes the potential losses and benefits associated with each outcome for each customer. We showed the equivalence of this measure with another profit measure recently proposed in the literature and with the well-known uplift curve. In particular, the equivalence with the uplift curve is conditioned on a particular assumption that we call the *unitary cost assumption*. Expressing the assumption underlying our methods is an essential aspect of the scientific method, and, as such, this formulation is an important contribution to the uplift literature. Secondly, we conducted a comparison of the strengths and weaknesses of the predictive and uplift approaches in terms of the profit measure in Section 5.3. We used simulated examples to compare the performance of both approaches in various settings. We showed the important role of the estimator variance and the mutual information between the input features and the potential outcomes. The distribution of the potential outcomes and the cost-benefit matrix also play a major role in this problem. This generalizes other results in the literature that compare uplift and predictive modeling.

In Chapter 6, we investigated the inference of the probability of counterfactual expressions. Counterfactual expressions are of particular interest in churn mitigation, as they represent the four categories of customers that can be delineated according to their reaction to retention efforts (*persuadable*, *do-not-disturb*, etc.). Estimating counterfactuals allows one to understand more precisely the performance of past campaigns and to establish a business profile of customers who react positively, negatively, or neutrally to the campaign. Counterfactual probabilities cannot be inferred exactly from data, but we proposed bounds and point estimators that improve upon the state of the art. The uplift bounds, based on the predictions of an uplift model, perform better as the customer features are more informative about the outcome. We proposed different point estimators of counterfactual probabilities, either by assuming the conditional independence between the potential outcomes, or by fitting a bivariate beta distribution on the uplift terms. Several variations of the latter approach are provided to allow for more flexibility at the expense of computation time. In Section 6.5, simulated examples showed the improvement of the uplift bounds over the state of the art, and the superiority of the point estimators based on the bivariate beta distribution. We applied these methods to data from our industrial partner in Section 6.6. We analyzed the typical behavior patterns of *persuadable*, *lost cause*, *sure thing* and *do-not-disturb* customers, which revealed insights unavailable using predictive or uplift modeling alone. In particular, we observed that the behavior of *persuadable* customers is between two extremes:

sure thing customers with low service usage, low bill, and long tenure, and the *do-not-disturb* customers, who are comparatively younger, churn more often, and have higher bills.

7.2 Recommendations for practitioners

We can summarize our findings in a series of recommendations for data scientists who plan to use uplift modeling on a problem with binary outcome and binary treatment. Before even choosing the features or the uplift model, we can give the following rules:

- If the potential outcome $\boldsymbol{y}_1$ is balanced but $\boldsymbol{y}_0$ is not, the uplift approach is more effective.

- On the contrary, if the potential outcome $\boldsymbol{y}_0$ is balanced but $\boldsymbol{y}_1$ is not, the predictive approach is more effective.

If neither of the two conditions above applies (such as when both outcomes are unbalanced, which is often the case for customer churn), the quality of the descriptive features must be assessed. We can summarize our findings with respect to the descriptive features as follows:

- If we have access to an indicator of which customers were reached in past campaigns or, more generally, an indicator of receptivity or compliance to the intervention, the estimated probability distribution of this indicator should be added as a feature.

- If the features are still not informative about the potential outcomes, the predictive approach is likely to be more effective.

When the features are reasonably informative, we recommend to perform an empirical comparison of both approaches with the available data. At this stage, it is important to take into account the costs and benefits associated with each outcome and each individual, as it can greatly influence the results. The profit measure can be used as a metric to compare different models while taking into account individual costs and benefits. The variance of the models is the last important factor that will influence the outcome of this comparison.

Lastly, deeper insights can be obtained on the problem at hand by estimating the distribution of counterfactuals. We recommend using the model based on the generalized bivariate beta distribution (GBB), as it is more flexible than the other proposed approaches for the same computational cost. If the uplift model suffers from a high variance, it may be beneficial to take this variability into account with the model named NGBB, although it is more computationally intensive. With the estimated distribution of counterfactuals, the practitioner can examine the distribution of other features within each counterfactual category and analyze the performance of the campaign in terms of positive and negative effects. The implementation of these methods can be challenging because of the complexity of the learning and inference process described in Section 6.4. This complexity is due to the fact that counterfactual events cannot be observed directly, requiring more sophisticated methods of inference.

7.3 Added value for the company

In this section, we discuss how the outcome of this research project can be valorized by Orange Belgium.

- Until now, Orange has not explored the use of uplift modeling. While the primary focus of this thesis has been to address customer churn, Orange Belgium conducts several other campaigns in various domains, such as *up-sell* (proposing a more expensive tariff plan), or *cross-sell* (proposing additional products). The potential applicability of uplift modeling extends beyond churn management, which presents an opportunity for Orange Belgium to enhance the effectiveness of these other campaigns. Our theoretical analysis of uplift modeling in Chapter 5, summarized as a set of guidelines in Section 7.2, provides a valuable framework for Orange Belgium to discern situations in which uplift modeling can prove beneficial. By incorporating uplift modeling into their repertoire, Orange Belgium can optimize resource allocation, target the right audience, and ultimately improve the overall impact of diverse campaigns.
- Due to its simple and generic nature, the *reach as a feature* methodology, developed in Section 4.4, can potentially improve the performance of any other model trained from customer data at Orange Belgium. This is probably the contribution of this thesis that has the widest scope of application for our industrial partner. This methodology can be further developed by considering the probability of reach through different channels, such as email, phone call, SMS, etc. Furthermore, a given customer does not need to have been contacted through each of these channels in order to implement the methodology, thanks to the generalization abilities of the underlying machine learning model.
- Estimating counterfactual probabilities provides a more complete description of the causal impact of retention efforts on customer churn. This estimation can be coupled with an analysis of customer features, which can reveal important insights for Orange Belgium's business operations. For example, we can establish a profile of the typical *persuadable* customer, and subsequent campaigns can be adjusted to take this new information into account.

7.4 Open issues and future work

This thesis represents an important contribution to the field of causal analysis applied to customer management. We have provided a detailed theoretical analysis of uplift modeling, both in terms of potential benefits (with the profit measure developed in Section 5.2) and compared to the classical predictive approach. To obtain general results, our analysis was based on a model-agnostic definition of the uplift approach and the predictive approach. This also implies that, in practice, the performance of both approaches can vary slightly depending on the specific model used. More specific results can be obtained by performing a detailed analysis of the different implementations of uplift modeling described in Section 3.1.2.

Our approach to counterfactual probability estimation is quite different from the rest of the literature on causal inference. We naturally built our inference strategies from an uplift model to obtain individual-level predictions, while in the causal inference literature, estimators are often based only on structural assumptions on the causal

graph (Mueller, A. Li, and Pearl, 2021; J. Zhang, Tian, and Bareinboim, 2022). Our approach based on uplift modeling opens the way for many other approaches to counterfactual inference. In the latter stages of research, we discovered the potential application of more advanced techniques:

- *Copulas* provide a way to encode the joint distribution of two random variables (Geenens, 2020). This can provide new possibilities to encode the dependency between the potential outcomes, while our estimators in Chapter 6 assume that the outcomes are conditionally independent or follow a given bivariate beta distribution.
- The *expectation-maximization* (EM) algorithm is an iterative procedure to find the distribution that best fits the data (M. Ding, 2022). Our approaches infer the distribution of counterfactuals in a single step from the data. Adapting the EM algorithm to the task of counterfactual inference can potentially refine and improve our results.
- *Stein discrepancy* is an approach to measure the discrepancy between data and a distribution (Barp et al., 2019). Stein discrepancy estimators can be used to tune the parameters of a distribution with potentially better results than the optimization procedure we used in Section 6.4. This field of research leans more heavily on advanced probability theory; therefore, whether this approach is applicable to the task of counterfactual inference is yet to be determined.

Lastly, we should stress that the results presented in this thesis pertain to a much broader scope than customer churn. Any task involving a binary action, a binary outcome, and descriptive features can benefit from our results. We hope that our contributions will extend beyond the domain of business analytics, serving as a small building block to address some of today's challenges, such as climate change, social equality, medicine, and other important tasks that can benefit from the new capabilities of causal inference and machine learning.

Part IV

Appendix

A

Introduction to probability theory

In this appendix, we give an introduction to the basic concepts of probability theory that are used throughout this thesis. This presentation is based on the handbook for the course *Statistical foundations of machine learning* by Bontempi (2017). Probability theory was synthesized in its current form by Kolmogoroff (1933). A formal definition requires notions of measure theory; however, a significant number of the technical details of such a formal definition are not relevant to this work, and therefore, we will strategically omit them. See (Durrett, 2019) for a more complete introduction, or (Barbe et al., 2007) in French.

A.1 Modeling uncertainty

The concept of *random experiment* is at the core of probability theory. It is a process that is possibly repeatable and whose output is uncertain; that is, the output of this process can be any one of multiple alternatives. The classic example of a repeatable random experiment is to throw a dice. An example of a non-repeatable random experiment would be the life expectancy of the King of Belgium. The result of both processes cannot be determined with certainty in the current state of our knowledge.

The set of possible outcomes for a particular random experiment is called the *sample space*, denoted Ω. In the case of the dice experiment, we would have $\Omega = \{1, 2, 3, 4, 5, 6\}$. In the example of life expectancy, Ω would be the set of positive real numbers $\mathbb{R}^+$, which could reasonably be reduced to positive real numbers between the King's current age and 130. This last example shows that the choice of the sample space is the result of a modeling process, and therefore different sample spaces could be suited to a given setting.

We denote a particular outcome, also called a *realization*, as ω. A single run of the random experiment is called a *trial*. Generally, we are not interested in whether an individual realization occurs. In the life expectancy example, it does not make sense to ask whether the life expectancy would be exactly 80 years, 0 day, 0 minute, 0 second, etc. We are more interested in modeling the probability that the outcome falls in a subset of the sample space, for example between 80 and 85 years.

To formalize this idea, we call any subset E of Ω an *event*. The set of events that we are interested in is called the event space, noted $\mathcal{F}$. Therefore, $\mathcal{F}$ is a set of subsets of Ω. Furthermore, we require $\mathcal{F}$ to fulfill the definition of a σ-algebra, that is, it must satisfy

the following three properties. These rules are essential to ensure the mathematical consistency of probabilities.

- The sample space Ω is in $\mathscr{F}$,
- $\mathscr{F}$ is closed under complements: for all events $E \in \mathscr{F}$, its complement $\Omega \setminus E$ is also in $\mathscr{F}$,
- $\mathscr{F}$ is closed under countable unions: for a sequence of events $E_1, E_2, \ldots$, their union $\bigcup_{i=1}^{\infty} E_i$ is also in $\mathscr{F}$.

Finally, the last element necessary to formalize the notion of uncertainty is to quantify the probability of occurrence of events in the event space $\mathscr{F}$. This is achieved by defining a *probability measure*, noted P, as a function that returns a number in $[0, 1]$ for each event $E \in \mathscr{F}$. More precisely, this function must satisfy the three axioms of probability.

1. The probability of an event $E \in \mathscr{F}$ is positive: $P(E) \geq 0$,
2. The probability of the sample space is one: $P(\Omega) = 1$ (i.e., the probability that *any* outcome occurs is one),
3. The probability measure is additive under countable unions: for all sequences $E_1, E_2, \ldots$ of mutually exclusive events (i.e., such that $E_i \cap E_j = \emptyset$ for all pairs of distinct nonnegative integers i, j), we have

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

This list of requirements should not be understood as truth corresponding to the real nature of uncertainty, but rather as definitions that provide mathematical convenience and that seem to correspond to our human intuition of probability.

When considered together, the three mathematical objects defined above, the sample space Ω, the event space $\mathscr{F}$ and the probability measure P, are called a *probability space*, noted $(\Omega, \mathscr{F}, P)$.

A.2 Random variables

It is often impractical to write a probabilistic expression in terms of elements of the event space $\mathscr{F}$, especially when Ω is continuous and multidimensional, or even more so when it contains more complex objects such as functions. The concept of *random variable* allows one to concisely and intuitively express complex probabilistic expressions. In the following, we will use the example of a random experiment where we throw two random dice, as a way to give a concrete expression of these quite abstract definitions. In this example, the space of outcomes Ω is the set $\{(1, 1), (1, 2), \ldots, (6, 6)\}$.

Formally, a random variable is defined as a function $\boldsymbol{x}$ (note the bold font) from the outcome space Ω to a set $\mathscr{X}$. This set is called the *domain* of $\boldsymbol{x}$. We note a *realization* of $\boldsymbol{x}$, that is, a member of $\mathscr{X}$, as x (note the regular font). In the two dice example, a random variable representing the sum of the two dice is defined as $\boldsymbol{x}((d_1, d_2)) = d_1 + d_2$, and its domain is $\{2, \ldots, 12\}$.

Furthermore, for each subset I of $\mathcal{X}$, we define the inverse function $\boldsymbol{x}^{-1}$ as

$$\boldsymbol{x}^{-1}(I) = \{\omega \in \Omega \mid \boldsymbol{x}(\omega) \in I\}. \tag{A.1}$$

Intuitively, $\boldsymbol{x}^{-1}(I)$ is the set of outcomes in Ω that lead the random variable $\boldsymbol{x}$ to take a value in I. In our example, $\boldsymbol{x}^{-1}$ is defined as $\boldsymbol{x}^{-1}(s) = \{(d_1, d_2) \in \Omega \mid d_1 + d_2 = s\}$.

This definition can be combined with the probability measure P to compute the probability of the random variable $\boldsymbol{x}$ to be realized in a given subset I of $\mathcal{X}$. The function that goes from I to a probability is called *push-forward measure*. In this work, the distinction between this push-forward measure and the original probability measure P is not important; hence we will note it P as well. From this, we can write

$$P(\boldsymbol{x} \in I) = P(\boldsymbol{x}^{-1}(I)) = P(\{\omega \in \Omega \mid \boldsymbol{x}(\omega) \in I\}). \tag{A.2}$$

In our example, the probability that the sum of the dice is greater than 10 is written as $P(\boldsymbol{x} \in \{11, 12\}) = P(\{(4, 6), (5, 5), (6, 4), (5, 6), (6, 5)\})$. Technically, the domain $\mathcal{X}$ must be accompanied by its own σ-algebra, I must be a measurable set, and $\boldsymbol{x}$ must be a *measurable function*, but these properties are not important in this work; therefore, we avoid defining them for simplicity.

We can also intuitively understand the notion of a random variable as a formalization of experimental measurement: the outcome space Ω represents the state of the universe (or, at least, the part of the universe relevant to the experiment), and the random variable $\boldsymbol{x}$ represents a single numerical quantity that represents an aspect of interest of the universe. For example, in a closed room filled with some gas, Ω would represent the state of each molecule in the room (their position and momentum), and $\boldsymbol{x}$ would be the average temperature of the room at thermal equilibrium. The random variable $\boldsymbol{x}$ represents the value of a temperature measurement, as a complex function of the state of each molecule; however, it is not necessary to fully describe the set Ω and the functional definition of $\boldsymbol{x}$ to reason about the equilibrium temperature.

A.3 Discrete and absolutely continuous random variables

When the domain of a random variable is countable (either with a finite or infinite number of elements), then we say that it is a *discrete* random variable. Examples include the dice random experiment, or the number of heads before a tail comes up when repeatedly flipping a coin. The countable nature of the domain allows us to assign a probability to each possible realization of the random variable.

Definition A.1 (Probability mass function)**.** The *probability mass function* of a discrete random variable $\boldsymbol{x}$, noted $P_{\boldsymbol{x}}$, is a function defined, for all $x \in \mathcal{X}$, as

$$P_{\boldsymbol{x}}(x) = P(\boldsymbol{x} \in \{x\}).$$

We also use the notation $P(\boldsymbol{x} = x) = P_{\boldsymbol{x}}(x)$. When it is clear from the context, we also write $P(\boldsymbol{x} = x) = P(x)$.

A random variable $\boldsymbol{x}$ is said to be *absolutely continuous* if its domain is uncountable, such as a subset of the real line. We may sometimes use the term *continuous* instead of *absolutely continuous*. More formally, $\boldsymbol{x}$ is absolutely continuous if it has a *probability density function*, defined as follows:

$P(x, y)$	$\boldsymbol{x} = 0$	$\boldsymbol{x} = 1$
$\boldsymbol{y} = 0$	0.1	0.2
$\boldsymbol{y} = 1$	0.3	0.4

Table A.1 Joint probability table of two binary variables.

Definition A.2 (Probability density function)**.** Given a random variable $\boldsymbol{x}$ with domain $\mathcal{X} \subseteq \mathbb{R}$, the *probability density function* (abbreviated *pdf*) $f_{\boldsymbol{x}}$ of $\boldsymbol{x}$ is a function from $\mathcal{X}$ to $\mathbb{R}^+$ such that, for all intervals $[a, b] \subseteq \mathcal{X}$, we have

$$P(a \leq \boldsymbol{x} \leq b) = \int_a^b f_{\boldsymbol{x}}(x)\,\mathrm{d}x. \tag{A.3}$$

Then, we can define the *cumulative distribution function* of $\boldsymbol{x}$, noted $F_{\boldsymbol{x}}$, as the probability of $\boldsymbol{x}$ being less than a given value:

Definition A.3 (Cumulative distribution function)**.** The *cumulative distribution function* $F_{\boldsymbol{x}}$ of an absolutely continuous random variable $\boldsymbol{x}$ is defined as

$$F_{\boldsymbol{x}}(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = P(\boldsymbol{x} \leq x). \tag{A.4}$$

Note that all the random variables we consider in this thesis are either discrete or absolutely continuous, but this categorization is not exhaustive. For example, the domain of the Cantor distribution (Hewitt and Stromberg, 2013) is the entire interval $[0, 1]$, however, it has no pdf satisfying Definition A.3.

A.4 Multiple random variables

In most applications, for example in machine learning, we want to model multiple random variables at a time. In this section, we define the notions of joint probability, conditional probability, and independence of two variables. We limit ourselves to two variables for the sake of clarity, but these definitions can be easily extended to any number of variables.

A.4.1 Joint probability

The probability that two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ take their values into two sets A and B is defined as the probability of the outcomes compatible simultaneously with A and B.

Definition A.4 (Joint probability)**.** The *joint probability distribution* of two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined, for all $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$, as

$$P(\boldsymbol{x} \in A, \boldsymbol{y} \in B) = P(\boldsymbol{x}^{-1}(A) \cap \boldsymbol{y}^{-1}(B)). \tag{A.5}$$

In the case of discrete variables, this probability distribution is often represented as a probability table; see, for example, Table A.1. In this example, the probability that $\boldsymbol{x} = 0$ and $\boldsymbol{y} = 1$ is $P(\boldsymbol{x} = 0, \boldsymbol{y} = 1) = 0.3$.

In the case of two absolutely continuous random variables, the joint pdf is defined in terms of their joint cumulative distribution function. This requires that their joint cumulative distribution function is well-defined and is twice differentiable.

Definition A.5 (Joint cumulative distribution function)**.** The *joint cumulative distribution function* of two absolutely continuous random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, as

$$F_{\boldsymbol{x},\boldsymbol{y}}(x, y) = P(\boldsymbol{x} \leq x, \boldsymbol{y} \leq y) = P(\boldsymbol{x}^{-1}(]-\infty, x]) \cap \boldsymbol{y}^{-1}(]-\infty, y])). \tag{A.6}$$

Definition A.6 (Joint probability density function)**.** The *joint probability density function* of two absolutely continuous random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, as

$$f_{\boldsymbol{x},\boldsymbol{y}}(x, y) = \frac{\partial^2 F_{\boldsymbol{x},\boldsymbol{y}}(x, y)}{\partial x \partial y}. \tag{A.7}$$

A.4.2 Conditional probability

An important aspect of scientific inquiry is the ability to update our model given new evidence. For example, suppose that we wish to predict whether it will rain tomorrow. Without any other information, the outcome will be quite uncertain. Then, someone tells us that we are in Belgium in the middle of February, and that it has been raining for the past three days. It is now reasonable to assign a much higher probability of rain for tomorrow given this new information. This operation is formalized in probability theory as conditional probabilities.

Definition A.7 (Conditional probability)**.** The *conditional probability distribution* of a random variable $\boldsymbol{y}$ given that we observe $\boldsymbol{x} \in A$, such that $P(\boldsymbol{x} \in A) > 0$, is defined, for all $B \subseteq \mathcal{Y}$, as

$$P(\boldsymbol{y} \in B \mid \boldsymbol{x} \in A) = \frac{P(\boldsymbol{x} \in A, \boldsymbol{y} \in B)}{P(\boldsymbol{x} \in A)}. \tag{A.8}$$

Note that this definition does not cover the case where we condition on an event of probability zero, such as the realization $\boldsymbol{x} = x$ for an absolutely continuous random variable. For example, we could wish to compute the probability distribution of the age of an individual given that we have a measure of their size. Using measure theory, it is possible to give a general definition of conditional probability that takes into account this case. However, to avoid going into technical details, we provide here a definition of conditional probability for two absolutely continuous random variables that is a corollary of this more general definition.

Definition A.8 (Conditional probability density function)**.** Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two absolutely continuous random variables with pdf $f_{\boldsymbol{x}}$ and $f_{\boldsymbol{y}}$ and joint pdf $f_{\boldsymbol{x},\boldsymbol{y}}$. The *conditional probability density function* of $\boldsymbol{y}$ given the observation $\boldsymbol{x} = x$, noted $f_{\boldsymbol{y}|x}$ is defined where $f_{\boldsymbol{x}}(x) > 0$ as

$$f_{\boldsymbol{y}|x}(y) = \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x, y)}{f_{\boldsymbol{x}}(x)}. \tag{A.9}$$

In contexts involving more than one random variable, the distribution of an individual random variable is called *marginal distribution*, in contrast to its joint or conditional distributions with another random variable.

A.4.3 Independence

The notion of independence formalizes the notion that two events are unrelated, that is, observing the occurrence of one does not provide any information about the occurrence of the other. For example, in the game of roulette, observing the outcome of a game does not provide any information on the outcome of the next game. There exists a definition of the independence between two events; however, we will directly focus on the independence between two random variables.

Definition A.9 (Independence)**.** Two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are *independent* if, for all $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$, we have

$$P(\boldsymbol{x} \in A, \boldsymbol{y} \in B) = P(\boldsymbol{x} \in A)P(\boldsymbol{y} \in B). \tag{A.10}$$

This is noted $\boldsymbol{x} \perp \boldsymbol{y}$. In particular, two absolutely continuous random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ with pdf $f_{\boldsymbol{x}}$ and $f_{\boldsymbol{y}}$ and joint pdf $f_{\boldsymbol{x},\boldsymbol{y}}$ are independent if, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have

$$f_{\boldsymbol{x},\boldsymbol{y}}(x, y) = f_{\boldsymbol{x}}(x) f_{\boldsymbol{y}}(y). \tag{A.11}$$

Combining Definitions A.7 and A.9 has the following rather intuitive consequence: the distribution of a random variable does not change after observing another random variable from which it is independent. This is formally expressed as

$$P(\boldsymbol{y} \in B \mid \boldsymbol{x} \in A) = P(\boldsymbol{y} \in B) \quad \text{if } \boldsymbol{x} \perp \boldsymbol{y}. \tag{A.12}$$

For absolutely continuous random variables, we also have

$$f_{\boldsymbol{y}|x}(y) = f_{\boldsymbol{y}}(y) \quad \text{if } \boldsymbol{x} \perp \boldsymbol{y}. \tag{A.13}$$

The notion of independence can be generalized to the notion of conditional independence. Conditional independence is an essential tool used in many causal inference algorithms, such as the PC algorithm (Spirtes, Glymour, and Review, 1991).

Definition A.10 (Conditional independence)**.** Two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are *conditionally independent* given another random variable $\boldsymbol{z}$ if, for all $A \subseteq \mathcal{X}$, $B \subseteq \mathcal{Y}$, and $C \subseteq \mathcal{Z}$, we have

$$P(\boldsymbol{x} \in A, \boldsymbol{y} \in B \mid \boldsymbol{z} \in C) = P(\boldsymbol{x} \in A \mid \boldsymbol{z} \in C)P(\boldsymbol{y} \in B \mid \boldsymbol{z} \in C) \tag{A.14}$$

whenever $P(\boldsymbol{z} \in C) > 0$. This is noted $\boldsymbol{x} \perp \boldsymbol{y} \mid \boldsymbol{z}$. In particular, two absolutely continuous random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are *conditionally independent* given another random variable $\boldsymbol{z}$ if, for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z}$, we have

$$f_{\boldsymbol{x},\boldsymbol{y}|z}(x, y) = f_{\boldsymbol{x}|z}(x) f_{\boldsymbol{y}|z}(y) \tag{A.15}$$

whenever $f_{\boldsymbol{z}}(z) > 0$.

B

Computing counterfactual probabilities from a causal model

In this appendix, we give an example of computation of counterfactual probabilities when we have full knowledge of the causal model. This is an extreme case of the *fully identifiable* setting described in Section 3.2.2. Let $M_1, \dots, M_m$ be a sequence of models derived by interventions from $M = (P, \boldsymbol{U}, \boldsymbol{V}, G, F)$, as in Definition 2.19, possibly including M. We note the components of each interventional model as $M_i = (P, \boldsymbol{U}, \boldsymbol{V}, G_i, F_i)$ (the unobserved variables $\boldsymbol{U}$, the observed variables $\boldsymbol{V}$, and their probability measure P are not modified by interventions, although the distribution of $\boldsymbol{V}$ is). Recall that the observed variables are noted $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_d)$. Let $\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(m)}$ be a sequence of random variables such that $\boldsymbol{y}^{(i)} \in \boldsymbol{V}$ for all $i \in \{1, \dots, m\}$. These random variables may be identical. We note their domains $\mathscr{Y}_1, \dots, \mathscr{Y}_m$. Let us compute the probability

$$P(\boldsymbol{y}^{(1)}_{M_1} \in A_1, \dots, \boldsymbol{y}^{(m)}_{M_m} \in A_m)$$

where $A_i \subseteq \mathscr{Y}_i$ for all $i = 1, \dots, m$. This can be done in a two-step process:

1. For each $i = 1, \dots, m$, find the subset Λ_i of $\mathscr{U}$ that satisfies $\boldsymbol{y}^{(i)}_{M_i} \in A_i$ in M_i:

$$\Lambda_i = \{U \in \mathscr{U} \mid \boldsymbol{y}^{(i)}_{M_i}(U) \in A_i\}$$

   This is done by solving for U a system of equations based on the functions $f_1, \dots, f_d$ in F_i:

$$\begin{aligned} f_{\boldsymbol{y}^{(i)}}(\mathrm{PA}_{G_i}(\boldsymbol{y}^{(i)}), U) &\in A_i \\ f_1(\mathrm{PA}_{G_i}(\boldsymbol{v}_1), U) &= v_1 \\ &\vdots \\ f_d(\mathrm{PA}_{G_i}(\boldsymbol{v}_d), U) &= v_d \end{aligned}$$

   The system above contains d equations: one for the random variable $\boldsymbol{y}^{(i)}$ present in the counterfactual expression $\boldsymbol{y}^{(i)}_{M_i} \in A_i$, and $d - 1$ for the remaining variables in $\boldsymbol{V}$.

2. Given the sets $\Lambda_1, \dots, \Lambda_m$, compute the probability

$$P(\boldsymbol{y}_{M_1}^{(1)} \in A_1, \dots, \boldsymbol{y}_{M_m}^{(m)} \in A_m) = P(U \in \Lambda_1 \cap \cdots \cap \Lambda_m).$$

As an example, let $\boldsymbol{x}$ be a discrete random variable with domain $\{1/4, 3/4\}$ such that both outcomes have an equal probability of $1/2$. Let $\boldsymbol{y}$ be a Bernoulli random variable with parameter $p = \boldsymbol{x}$. Let us compute the probability

$$P(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 0, \boldsymbol{y}_{\boldsymbol{x}=3/4} = 1).$$

To compute such a counterfactual probability, we need to define the system as a causal model following Definition 2.18, that is, with latent variables $\boldsymbol{u}_{\boldsymbol{x}}$ and $\boldsymbol{u}_{\boldsymbol{y}}$, and write $\boldsymbol{x}$ and $\boldsymbol{y}$ as deterministic functions of their parents. Here is a possible way to define this causal model:

$$\begin{aligned}
\boldsymbol{u}_{\boldsymbol{x}} &\sim \text{Bern}(0.5) \\
\boldsymbol{u}_{\boldsymbol{y}} &\sim U(0,1) \\
\boldsymbol{x} &= \frac{1}{4}(1 - \boldsymbol{u}_{\boldsymbol{x}}) + \frac{3}{4}\boldsymbol{u}_{\boldsymbol{x}} \\
\boldsymbol{y} &= \mathbb{I}[\boldsymbol{u}_{\boldsymbol{y}} < \boldsymbol{x}].
\end{aligned}$$

Here, $\boldsymbol{u}_{\boldsymbol{x}}$ and $\boldsymbol{u}_{\boldsymbol{y}}$ are independent and their joint domain is $\mathcal{U} = \{0,1\} \times [0,1]$. The first step is to determine Λ_1 and Λ_2, defined as

$$\begin{aligned}
\Lambda_1 &= \{(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) \in \mathcal{U} \mid \boldsymbol{y}_{\boldsymbol{x}=1/4}(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) = 0\} \\
\Lambda_2 &= \{(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) \in \mathcal{U} \mid \boldsymbol{y}_{\boldsymbol{x}=3/4}(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) = 1\}.
\end{aligned}$$

Let us focus on Λ_1. If $\boldsymbol{x}$ is set to $1/4$, then $\boldsymbol{y}$ is equal to 0 whenever $\boldsymbol{u}_{\boldsymbol{y}}$ is greater than or equal to $1/4$. Also, since the value of $\boldsymbol{x}$ is fixed, the value of $\boldsymbol{u}_{\boldsymbol{x}}$ has no influence on the system. This is formally written

$$\Lambda_1 = \{0,1\} \times \left[\tfrac{1}{4}, 1\right].$$

Similarly, when $\boldsymbol{x}$ is set to $3/4$, $\boldsymbol{y}$ is equal to 1 whenever $\boldsymbol{u}_{\boldsymbol{y}}$ is strictly less than $3/4$, implying

$$\Lambda_2 = \{0,1\} \times \left[0, \tfrac{3}{4}\right).$$

The probability $P(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 0, \boldsymbol{y}_{\boldsymbol{x}=3/4} = 1)$ is computed as

$$\begin{aligned}
P(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 0, \boldsymbol{y}_{\boldsymbol{x}=3/4} = 1) &= P((\boldsymbol{u}_{\boldsymbol{x}}, \boldsymbol{u}_{\boldsymbol{y}}) \in \Lambda_1 \cap \Lambda_2) \\
&= P\left((\boldsymbol{u}_{\boldsymbol{x}}, \boldsymbol{u}_{\boldsymbol{y}}) \in \{0,1\} \times \left[\tfrac{1}{4}, \tfrac{3}{4}\right)\right).
\end{aligned}$$

By independence of $\boldsymbol{u}_{\boldsymbol{x}}$ and $\boldsymbol{u}_{\boldsymbol{y}}$, this reduces to

$$P(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 0, \boldsymbol{y}_{\boldsymbol{x}=3/4} = 1) = P(\boldsymbol{u}_{\boldsymbol{x}} \in \{0,1\}) P\left(\tfrac{1}{4} \le \boldsymbol{u}_{\boldsymbol{y}} < \tfrac{3}{4}\right) = 1 \times \frac{1}{2}.$$

We can also compute other quantities, such as, given that we observed $\boldsymbol{y} = 1$ and $\boldsymbol{x} = 3/4$, the probability that $\boldsymbol{y}$ would still be 1 had $\boldsymbol{x}$ been $1/4$. This is formalized as

$$P\left(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 1 \mid \boldsymbol{y} = 1, \boldsymbol{x} = \tfrac{3}{4}\right) = \frac{P\left(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 1, \boldsymbol{y} = 1, \boldsymbol{x} = \frac{3}{4}\right)}{P\left(\boldsymbol{y} = 1, \boldsymbol{x} = \frac{3}{4}\right)}.$$

The denominator $P(\boldsymbol{y} = 1, \boldsymbol{x} = 3/4)$ does not involve potential outcomes and, therefore, can be computed from the definition of $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\begin{aligned} P\left(\boldsymbol{y} = 1, \boldsymbol{x} = \tfrac{3}{4}\right) &= P\left(\boldsymbol{y} = 1 \mid \boldsymbol{x} = \tfrac{3}{4}\right) P\left(\boldsymbol{x} = \tfrac{3}{4}\right) \\ &= P\left(\boldsymbol{u_y} < \tfrac{3}{4}\right) P(\boldsymbol{u_x} = 1) \\ &= \frac{3}{4} \times \frac{1}{2} = \frac{3}{8}. \end{aligned}$$

The numerator $P\left(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 1, \boldsymbol{y} = 1, \boldsymbol{x} = 3/4\right)$ is computed using the general procedure for counterfactuals, by defining

$$\begin{aligned} \Lambda_3 &= \{(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) \in \mathcal{U} \mid \boldsymbol{y}_{\boldsymbol{x}=1/4}(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) = 1\} \\ \Lambda_4 &= \{(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) \in \mathcal{U} \mid \boldsymbol{y}(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) = 1\} \\ \Lambda_5 &= \left\{(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) \in \mathcal{U} \mid \boldsymbol{x}(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}) = \tfrac{3}{4}\right\}. \end{aligned}$$

Using a similar reasoning as for Λ_1 and Λ_2, we find

$$\Lambda_3 = \{0, 1\} \times \left[0, \tfrac{1}{4}\right)$$

To determine Λ_4, we must take into account the two possible values for $\boldsymbol{u_x}$: if it is 0, then $\boldsymbol{u_y}$ must be between 0 and 1/4, whereas if it is 1, $\boldsymbol{u_y}$ must be between 0 and 3/4. This is expressed as

$$\Lambda_4 = \{0\} \times \left[0, \tfrac{1}{4}\right) \cup \{1\} \times \left[0, \tfrac{3}{4}\right).$$

For Λ_5, since we are only concerned with the value of $\boldsymbol{x}$, the variable $\boldsymbol{u_y}$ can take any value, leading to

$$\Lambda_5 = \{1\} \times [0, 1].$$

The intersection of Λ_3, Λ_4 and Λ_5 is

$$\Lambda = \Lambda_3 \cap \Lambda_4 \cap \Lambda_5 = \{1\} \times \left[0, \tfrac{1}{4}\right).$$

The counterfactual probability $P\left(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 1 \mid \boldsymbol{y} = 1, \boldsymbol{x} = 3/4\right)$ is computed as

$$P\left(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 1 \mid \boldsymbol{y} = 1, \boldsymbol{x} = \tfrac{3}{4}\right) = \frac{8}{3} P\left((\boldsymbol{u_x}, \boldsymbol{u_y}) \in \{1\} \times \left[0, \tfrac{1}{4}\right)\right).$$

By independence of $\boldsymbol{u_x}$ and $\boldsymbol{u_y}$, we have

$$\begin{aligned} P\left(\boldsymbol{y}_{\boldsymbol{x}=1/4} = 1 \mid \boldsymbol{y} = 1, \boldsymbol{x} = \tfrac{3}{4}\right) &= \frac{8}{3} P(\boldsymbol{u_x} = 1) P\left(0 \leq \boldsymbol{u_y} < \tfrac{1}{4}\right) \\ &= \frac{8}{3} \times \frac{1}{2} \times \frac{1}{4} = \frac{1}{3}. \end{aligned}$$

C

Convergence of the uplift curve to the profit measure

In this appendix, we prove the convergence of the uplift curve to the profit measure, formulated in Theorem 5.4:

Theorem 5.4. *Let D_{tr} be a training set of iid realizations of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$, and let $\boldsymbol{D}_{\text{te}}$ be a test set of N tuples $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{t}^{(i)})$ iid to $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$, where $\boldsymbol{t}$ is randomized (see Definition 3.2). Let $\mathscr{M}$ be a model such that $\mathscr{M}(\boldsymbol{x}, D_{\text{tr}})$ is a continuous random variable. Let $\rho \in (0, 1)$ be the prescription rate, and $k = \lceil N\rho \rceil$. Under the unitary value assumption, the value of the uplift curve at index k, noted $\text{Uplift}(k, D_{\text{tr}}, \boldsymbol{D}_{\text{te}})$, converges to the causal profit of a campaign at the corresponding prescription rate ρ. This is expressed formally as*

$$\lim_{N\to\infty} \frac{1}{N}\text{Uplift}(k, D_{\text{tr}}, \boldsymbol{D}_{\text{te}}) = \Pi(\rho, D_{\text{tr}}) \quad \textit{in probability.} \tag{5.1}$$

The overall structure of this appendix is as follows. First, we adapt the definition of the uplift curve in some technical aspects to facilitate the proof. Then, in Appendix C.1, we state Lemma C.1, which is the key result that enables the proof of Theorem 5.4. We then prove Theorem 5.4 assuming Lemma C.1 to be true. In Appendix C.2, we list a series of smaller technical results needed for the proof of Lemma C.1, and we then prove Lemma C.1. Finally, we prove these smaller technical results in Appendix C.3.

The training set D_{tr} is not important in the proof of Theorem 5.4, therefore, in this appendix, we note $\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) = \mathscr{M}(\boldsymbol{x})$ and $\tau(D_{\text{tr}}) = \tau$. Also, in this theorem, the test set is a random variable, noted $\boldsymbol{D}_{\text{te}}$, hence the uplift curve is also a random variable. However, in the original definition of the uplift curve (Definition 3.3), the test set is not random, and it is sorted according to the scores predicted by $\mathscr{M}$. We need a new definition that takes into account the random nature of the test set and does not require it to be sorted, such that it can represent iid samples of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$. For that, we define a binary random vector $\boldsymbol{B}(k)$ of length N that indicates which are the k samples from the test set with the highest scores according to $\mathscr{M}$. Since the fact that $\boldsymbol{B}(k)$ depends on k is not essential in our developments, we note $\boldsymbol{B} = \boldsymbol{B}(k)$.

Definition C.1 (Uplift curve)**.** Let $\boldsymbol{D}_{\text{te}} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{t}^{(i)})\}_{i=1}^{N}$ be a data set of N iid tuples of random variables, such that the treatment $t^{(i)}$ is randomized. Let $\mathscr{M}$ be a model, and let $k \in \{1, \dots, N\}$. Let $\boldsymbol{B} = \boldsymbol{B}(k)$ be a binary random vector indicating the k individuals

with the highest scores according to $\mathscr{M}$. Formally, $\boldsymbol{B}$ is defined as a binary random vector $[\boldsymbol{b}^{(i)}]_{i=1}^{N}$ of length N such that $\sum_{i=1}^{N} \boldsymbol{b}^{(i)} = k$ and $\mathscr{M}(\boldsymbol{x}^{(i)}) \geq \mathscr{M}(\boldsymbol{x}^{(j)})$ whenever $\boldsymbol{b}^{(i)} = 1$ and $\boldsymbol{b}^{(j)} = 0$. The *uplift curve* is defined as

$$\text{Uplift}(k, \boldsymbol{D}_{\text{te}}) = \left(\frac{\boldsymbol{r}_0(k)}{\boldsymbol{n}_0(k)} - \frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)} \right) k \tag{C.2}$$

where the following notation is used, for $t = 0, 1$:

$$\boldsymbol{r}_t(k) = \sum_{i=1}^{N} \boldsymbol{b}^{(i)} \mathbb{I}[\boldsymbol{y}^{(i)} \text{ and } \boldsymbol{t}^{(i)} = t] \qquad \boldsymbol{n}_t(k) = \sum_{i=1}^{N} \boldsymbol{b}^{(i)} \mathbb{I}[\boldsymbol{t}^{(i)} = t] \tag{C.3}$$

In the case $\boldsymbol{r}_t(k) = \boldsymbol{n}_t(k) = 0$, the quotient $\boldsymbol{r}_t(k)/\boldsymbol{n}_t(k)$ is defined as 0.

Given an unsorted realization of $\boldsymbol{D}_{\text{te}}$, we can always sort it by score and compute the uplift curve as in the original definition (Definition 3.3), and result will always be identical to the uplift curve computed from this realization of $\boldsymbol{D}_{\text{te}}$ with Definition C.1.

C.1 Convergence of the uplift curve

The key to the proof of Theorem 5.4 is the following lemma, which probably represents the intuition underlying the definition of the uplift curve. It links the probabilities that we intend to estimate when computing the uplift curve with their statistical estimates. More precisely, it indicates that the observed proportion of positive outcomes among the k individuals with the highest scores (the ratio $\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)$ in Definition C.1) converges to the theoretical probability of a positive outcome for individuals with a score higher than a threshold corresponding to k. To the best of our knowledge, this result is an original contribution to the literature.

Lemma C.1. *Let $\boldsymbol{D}_{\text{te}}$ be a test set of random variables iid to $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$, where $\boldsymbol{t}$ is randomized, and let $\mathscr{M}$ be a model such that $\mathscr{M}(\boldsymbol{x})$ is an absolutely continuous random variable. Let N be the size of $\boldsymbol{D}_{\text{te}}$, $\rho \in (0, 1)$ be the prescription rate, and $k = \lceil N\rho \rceil$. The ratio $\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)$, as defined in Definition C.1, converges to the probability $P(\boldsymbol{y} = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau, \boldsymbol{t} = 1)$, where $\tau = \inf\{\tau' : P(\mathscr{M}(\boldsymbol{x}) \geq \tau') \geq \rho)\}$. This is expressed formally as*

$$\lim_{N\to\infty} \frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)} = P(\boldsymbol{y} = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau, \boldsymbol{t} = 1) \quad \textit{in probability.} \tag{C.4}$$

The proof is given in the next section. The technical difficulty of the proof stems from the fact that the ratio $\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)$ is defined in terms of a random test set $\boldsymbol{D}_{\text{te}}$, while the probability it converges to depends only on $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{t}$. Lemma C.1 focuses on the target group ($\boldsymbol{t} = 1$), but it can be trivially adapted to the control group ($\boldsymbol{t} = 0$). The proof of Theorem 5.4 follows from this lemma rather simply:

Proof of Theorem 5.4 assuming Lemma C.1. Let us expand the definition of the uplift curve (Definition C.1):

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N} \text{Uplift}(k, D_{\text{tr}}, \boldsymbol{D}_{\text{te}}) &= \lim_{N\to\infty} \frac{k}{N} \left(\frac{\boldsymbol{r}_0(k)}{\boldsymbol{n}_0(k)} - \frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)} \right) \\ &= \left(\lim_{N\to\infty} \frac{\lceil N\rho \rceil}{N} \right) \left(\lim_{N\to\infty} \frac{\boldsymbol{r}_0(k)}{\boldsymbol{n}_0(k)} - \lim_{N\to\infty} \frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)} \right).\end{aligned}$$

It is easy to show that $\lim_{N\to\infty} \lceil N\rho \rceil / N = \rho$. Assuming Lemma C.1, which can be easily adapted to prove the convergence $\boldsymbol{r}_0(k)/\boldsymbol{n}_0(k)$, we have

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\text{Uplift}(k, D_{\text{tr}}, \boldsymbol{D}_{\text{te}}) \\
&= \rho(P(\boldsymbol{y} = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau, \boldsymbol{t} = 0) - P(\boldsymbol{y} = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau, \boldsymbol{t} = 1)) \\
&= \rho(P(\boldsymbol{y}_0 = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau) - P(\boldsymbol{y}_1 = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau)).
\end{aligned}$$

The last equality follows from the randomization of the treatment $\boldsymbol{t}$. Remember that τ is defined as the largest value that satisfies$P(\mathscr{M}(\boldsymbol{x}) \geq \tau) \geq \rho$. Since $\mathscr{M}(\boldsymbol{x})$ is an absolutely continuous random variable, its cumulative distribution function is absolutely continuous (with respect to the Lebesgue measure), therefore there will always exist a value τ such that $P(\mathscr{M}(\boldsymbol{x}) \geq \tau) = \rho$, leading to

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\text{Uplift}(k, D_{\text{tr}}, \boldsymbol{D}_{\text{te}}) \\
&= P(\boldsymbol{y}_0 = 1, \mathscr{M}(\boldsymbol{x}) \geq \tau) - P(\boldsymbol{y}_1 = 1, \mathscr{M}(\boldsymbol{x}) \geq \tau) \quad \text{in probability} \\
&= \int f_{\boldsymbol{x}}(x)(P(\boldsymbol{y}_0 = 1 \mid x) - P(\boldsymbol{y}_1 = 1 \mid x))\mathbb{I}[\mathscr{M}(x) \geq \tau]\,\mathrm{d}x \\
&= \int f_{\boldsymbol{x}}(x)U(x)\mathbb{I}[\mathscr{M}(x) \geq \tau]\,\mathrm{d}x \\
&= \mathbb{E}_{\boldsymbol{x}}[U(\boldsymbol{x})\mathbb{I}[\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) > \tau].
\end{aligned}$$

Under the unitary value assumption (Definition 5.10), and following Theorem 5.1, we have

$$\lim_{N\to\infty} \frac{1}{N}\text{Uplift}(k, D_{\text{tr}}, \boldsymbol{D}_{\text{te}}) = \mathbb{E}_{\boldsymbol{x}}[\pi(\boldsymbol{x})\mathbb{I}[\mathscr{M}(\boldsymbol{x}, D_{\text{tr}}) > \tau] = \Pi(\rho, D_{\text{tr}}) \quad \text{in probability.}$$

□

C.2 Proof of Lemma C.1

The proof of Lemma C.1 is based on a series of technical lemmas, which are themselves proven in the next section. Let us define domr additional notation to simplify our developments:

- Expressions involving multiple binary variables such as $P(\boldsymbol{t}^{(i)} = 1, \boldsymbol{b}^{(i)} = 1)$ are abbreviated as $P(\boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1)$.
- The product $\boldsymbol{y}^{(i)}\boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)}$ is noted $\boldsymbol{q}^{(i)}$.
- The probability $P(\boldsymbol{y} = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau, \boldsymbol{t} = 1)$, which is a function of τ, is simply noted $S(\tau) = S$.

The random variable $\boldsymbol{q}^{(i)}$ is equal to one only for individuals in the target group ($\boldsymbol{t}^{(i)} = 1$) with a positive outcome ($\boldsymbol{y}^{(i)} = 1$) and among the k individuals with the highest score ($\boldsymbol{b}^{(i)} = 1$). The following lemma indicates that the number of individuals assigned to the target group among those with the highest scores follows a binomial distribution, even after observing that the outcome $\boldsymbol{y}^{(i)}$, treatment $\boldsymbol{t}^{(i)}$ and selection indicator $\boldsymbol{b}^{(i)}$ of a given individual are all positive. This result will be used to derive a closed-form expression for some terms in the development of the uplift curve.

Lemma C.2. *Let $i \in \{1, \dots, N\}$. $\boldsymbol{n}_1(k) - 1 \mid \boldsymbol{q}^{(i)} = 1$ follows a binomial distribution* $\mathrm{Bin}(k-1, p)$.

This is generalized to the case of two observations as follows.

Lemma C.3. *Let $i, j \in \{1, \dots, N\}$, with $i \neq j$. $\boldsymbol{n}_1(k) - 2 \mid \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1$ follows a binomial distribution* $\mathrm{Bin}(k-2, p)$.

The following lemma states that, in the limit (i.e., as $N \to \infty$), the k individuals with the highest scores (i.e., such that $\boldsymbol{b}^{(i)} = 1$) are exactly those with a score higher than the threshold τ. This is not always the case for low enough N, since τ does not depend on N; therefore, the individual at rank k (when ranked by score) might have a score different from τ. Note that the index N has no special value in the statement of this lemma. This lemma could be formulated with any integer smaller or equal than N. We use N because it simplifies the notation during the proof of Lemma C.1, where this lemma is used.

Lemma C.4.

$$\lim_{N\to\infty} P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau \mid \boldsymbol{b}^{(N)} = 1) = \lim_{N\to\infty} P(\boldsymbol{b}^{(N)} = 1 \mid \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) = 1.$$

The following lemma is purely computational, and is required at the end of the proof of Lemma C.1.

Lemma C.5. *For any $0 < p < 1$, we have*

$$\lim_{k\to\infty} \sum_{n=1}^{k} \frac{1}{n}\binom{k}{n} p^n (1-p)^{k-n} = 0.$$

Armed with these results, we can now prove Lemma C.1.

Lemma C.1. *Let $\boldsymbol{D}_{\mathrm{te}}$ be a test set of random variables iid to $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$, where $\boldsymbol{t}$ is randomized, and let $\mathscr{M}$ be a model such that $\mathscr{M}(\boldsymbol{x})$ is an absolutely continuous random variable. Let N be the size of $\boldsymbol{D}_{\mathrm{te}}$, $\rho \in (0,1)$ be the prescription rate, and $k = \lceil N\rho \rceil$. The ratio $\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)$, as defined in Definition C.1, converges to the probability $P(\boldsymbol{y} = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau, \boldsymbol{t} = 1)$, where $\tau = \inf\{\tau' : P(\mathscr{M}(\boldsymbol{x}) \geq \tau') \geq \rho)\}$. This is expressed formally as*

$$\lim_{N\to\infty} \frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)} = P(\boldsymbol{y} = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau, \boldsymbol{t} = 1) \quad \textit{in probability.} \tag{C.5}$$

Proof. From Definition 2.10, the convergence in probability of $\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)$ to S (remember that we note $S = P(\boldsymbol{y} = 1 \mid \mathscr{M}(\boldsymbol{x}) \geq \tau, \boldsymbol{t} = 1)$) is formally expressed as, for any $\varepsilon > 0$,

$$\lim_{N\to\infty} P\left(\left|\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)} - S\right| \geq \varepsilon\right) = 0. \tag{C.6}$$

First, we will show the convergence of the expected value of $\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)$ to S:

$$\lim_{N\to\infty} \mathbb{E}\left[\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right] = S. \tag{C.7}$$

From Definition C.1, we have

$$\mathbb{E}\left[\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{N} \boldsymbol{q}^{(i)}}{\boldsymbol{n}_1(k)}\right] = \sum_{i=1}^{N} \mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}}{\boldsymbol{n}_1(k)}\right]. \tag{C.8}$$

By decomposing the expected value, we see that when $\boldsymbol{q}^{(i)}$ (which is either zero or one) is zero, this specific valuation will not contribute to the expected value. On the other hand, $\boldsymbol{n}_1(k)$ can take any integer value between 0 and k. Therefore, the expression $\boldsymbol{q}^{(i)}/\boldsymbol{n}_1(k)$ is either zero[1] or any of the values $1, 1/2, \dots, 1/k$. This leads to

$$\begin{aligned}\mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}}{\boldsymbol{n}_1(k)}\right] &= \sum_{n=1}^{k} \frac{1}{n} P\left(\frac{\boldsymbol{q}^{(i)}}{\boldsymbol{n}_1(k)} = \frac{1}{n}\right) = \sum_{n=1}^{k} \frac{1}{n} P(\boldsymbol{n}_1(k) = n, \boldsymbol{q}^{(i)} = 1) \\ &= \sum_{n=1}^{k} \frac{1}{n} P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)} = 1) P(\boldsymbol{q}^{(i)} = 1). \end{aligned} \tag{C.9}$$

Lemma C.2 shows that we can develop Eq. (C.9) as

$$\mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}}{\boldsymbol{n}_1(k)}\right] = P(\boldsymbol{q}^{(i)} = 1) \sum_{n=1}^{k} \frac{1}{n} P(\mathrm{Bin}(k-1, p) = n-1).$$

This sum can be transformed using the probability mass function of the binomial distribution:

$$\begin{aligned}\sum_{n=1}^{k} \frac{1}{n} P(\mathrm{Bin}(k-1,p) = n-1) &= \sum_{n=1}^{k} \frac{1}{n} \binom{k-1}{n-1} p^{n-1}(1-p)^{k-n} \\ &= \sum_{n=1}^{k} \frac{1}{n} \frac{(k-1)!}{(n-1)!(k-n)!} p^{n-1}(1-p)^{k-n} \\ &= \frac{1}{kp} \sum_{n=1}^{k} \binom{k}{n} p^{n}(1-p)^{k-n} = \frac{1}{kp} P(\mathrm{Bin}(k,p) \geq 1) \\ &= \frac{1}{kp}(1 - P(\mathrm{Bin}(k,p) = 0)) = \frac{1}{kp}\left(1 - (1-p)^k\right). \end{aligned}$$

Wrapping up, Eq. (C.8) can be developed as

$$\begin{aligned}\mathbb{E}\left[\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right] &= \sum_{i=1}^{N} \mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}}{\boldsymbol{n}_1(k)}\right] = \sum_{i=1}^{N} \frac{1}{kp} P(\boldsymbol{q}^{(i)} = 1)\left(1 - (1-p)^k\right) \\ &= \sum_{i=1}^{N} \frac{1}{kp} P(\boldsymbol{y}^{(i)}\boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1)\left(1 - (1-p)^k\right). \end{aligned}$$

Since $\boldsymbol{D}_{\text{te}}$ is iid, the choice of the index i in the probability $P(\boldsymbol{y}^{(i)}\boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1)$ does not matter. In what follows, we replace it by N. Note that $\boldsymbol{t}^{(N)} \perp \boldsymbol{b}^{(N)}$, and also $P(\boldsymbol{b}^{(N)} = 1) = k/N$ and $P(\boldsymbol{t}^{(N)} = 1) = P(\boldsymbol{t} = 1) = p$, from what we can deduce

$$\begin{aligned}P(\boldsymbol{y}^{(N)}\boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1) &= P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1) P(\boldsymbol{t}^{(N)} = 1) P(\boldsymbol{b}^{(N)} = 1) \\ &= \frac{kp}{N} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1), \end{aligned} \tag{C.10}$$

and then

$$\mathbb{E}\left[\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right] = P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1)\left(1 - (1-p)^k\right).$$

[1] Remember from Definition C.1 that the ratio $\boldsymbol{r}_0(k)/\boldsymbol{n}_0(k)$ is defined as 0 when $\boldsymbol{n}_0(k)$ is equal to zero, therefore, this case will not contribute to Eq. (C.8).

Since $k = \lceil N\tau \rceil$ and $p \in (0,1)$, it is clear that $\lim_{N\to\infty}(1-(1-p)^k) = 1$. To finish the proof of Eq. (C.7), it remains to show that

$$\lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1) = S = P(\boldsymbol{y} = 1 \mid \boldsymbol{t} = 1, \mathscr{M}(\boldsymbol{x}) \geq \tau).$$

From Lemma C.4, we can develop the limit of $P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1)$ as

$$\begin{aligned}
&\lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1) \\
&= \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1, \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau \mid \boldsymbol{b}^{(N)} = 1) \\
&\quad + \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)} b^{(N)} = 1, \mathscr{M}(\boldsymbol{x}^{(N)}) < \tau) P(\mathscr{M}(\boldsymbol{x}^{(N)}) < \tau \mid \boldsymbol{b}^{(N)} = 1) \\
&= \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1, \mathscr{M}(\boldsymbol{x}) \geq \tau).
\end{aligned}$$

Similarly, we can develop the limit of $P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)} = 1, \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau)$ as

$$\begin{aligned}
&\lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)} = 1, \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) \\
&= \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1, \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) P(\boldsymbol{b}^{(N)} = 1 \mid \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) \\
&\quad + \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)} = 1, \boldsymbol{b}^{(N)} = 0, \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) P(\boldsymbol{b}^{(N)} = 0 \mid \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) \\
&= \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1, \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau).
\end{aligned}$$

Using these two convergence results, we have

$$\begin{aligned}
\lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1) &= \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1, \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) \\
&= \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)} = 1, \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) \\
&= P(\boldsymbol{y} = 1 \mid \boldsymbol{t} = 1, \mathscr{M}(\boldsymbol{x}) \geq \tau)
\end{aligned}$$

where the last equality follows by the iid property of $\boldsymbol{D}_{\text{te}}$. This finishes the proof of Eq. (C.7):

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}\left[\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right] &= \lim_{N\to\infty} P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1) \\
&= P(\boldsymbol{y} = 1 \mid \boldsymbol{t} = 1, \mathscr{M}(\boldsymbol{x}) \geq \tau) = S.
\end{aligned}$$

Now, let us show that the variance of $\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)$ converges to 0:

$$\lim_{N\to\infty} \operatorname{Var}\left(\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right) = 0. \tag{C.11}$$

We will compute the variance as

$$\operatorname{Var}\left(\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right) = \mathbb{E}\left[\frac{\boldsymbol{r}_1(k)^2}{\boldsymbol{n}_1(k)^2}\right] - \mathbb{E}\left[\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right]^2.$$

Since we just have proven that $\mathbb{E}[\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)]$ converges to S, we focus on the other term, $\mathbb{E}[\boldsymbol{r}_1(k)^2/\boldsymbol{n}_1(k)^2]$:

$$\mathbb{E}\left[\frac{\boldsymbol{r}_1(k)^2}{\boldsymbol{n}_1(k)^2}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^N \boldsymbol{q}^{(i)} \sum_{j=1}^N \boldsymbol{q}^{(j)}}{\boldsymbol{n}_1(k)^2}\right] = \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}}{\boldsymbol{n}_1(k)^2}\right]. \tag{C.12}$$

Using a similar argument as for the expected value, note that $\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}$ is either equal to 0 or 1, and that $\boldsymbol{n}_1(k)$ can take any integer value between 0 and k. Hence the expression $\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}/\boldsymbol{n}_1(k)^2$ is either zero[2] or any of the values $1, 1/4, \ldots, 1/k^2$. This leads to

$$\begin{aligned}\mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}}{\boldsymbol{n}_1(k)^2}\right] &= \sum_{n=1}^k \frac{1}{n^2} P(\boldsymbol{n}_1(k) = n, \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1) \\ &= \sum_{n=1}^k \frac{1}{n^2} P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1) P(\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1). \end{aligned} \tag{C.13}$$

Using Lemma C.3, the sum in Eq. (C.13) can be developed as, for $i \neq j$,

$$\begin{aligned}\sum_{n=1}^k \frac{1}{n^2} P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1) &= \sum_{n=1}^k \frac{1}{n^2} P(\mathrm{Bin}(k-2, p) = n-2) \\ &= \sum_{n=1}^k \frac{n-1}{n^2(n-1)} \binom{k-2}{n-2} p^{n-2}(1-p)^{k-n} \\ &= \frac{1}{(k-1)kp^2} \sum_{n=1}^k \frac{n-1}{n} \binom{k}{n} p^n (1-p)^{k-n} \\ &= \frac{1}{(k-1)kp^2} \left(1 - (1-p)^k - \sum_{n=1}^k \frac{1}{n} \binom{k}{n} p^n (1-p)^{k-n}\right).\end{aligned}$$

Let us define the notation

$$A_k = \sum_{n=1}^k \frac{1}{n} \binom{k}{n} p^n (1-p)^{k-n}.$$

This allows to simplify Eq. (C.13), for $i \neq j$, as

$$\mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}}{\boldsymbol{n}_1(k)^2}\right] = P(\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1) \frac{1 - (1-p)^k - A_k}{(k-1)kp^2}.$$

Now, let us compute, for $i \neq j$, the probability $P(\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1)$. Remember that $\boldsymbol{t}^{(i)} \perp \boldsymbol{t}^{(j)}$ for $i \neq j$, and that $\boldsymbol{b}^{(i)} \perp \boldsymbol{t}^{(j)}$ for any i, j (possibly equal). We also use the fact that $\boldsymbol{y}^{(i)} \perp \boldsymbol{b}^{(j)}, \boldsymbol{t}^{(j)} \mid \boldsymbol{b}^{(i)}, \boldsymbol{t}^{(i)}$ for $i \neq j$.

$$\begin{aligned}P(\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1) &= P(\boldsymbol{y}^{(i)}\boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1 \mid \boldsymbol{y}^{(j)}\boldsymbol{t}^{(j)}\boldsymbol{b}^{(j)} = 1) P(\boldsymbol{y}^{(j)}\boldsymbol{t}^{(j)}\boldsymbol{b}^{(j)} = 1) \\ &= P(\boldsymbol{y}^{(i)} = 1 \mid \boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1) P(\boldsymbol{b}^{(i)} = 1 \mid \boldsymbol{b}^{(j)} = 1) P(\boldsymbol{t}^{(i)} = 1) \\ &\qquad P(\boldsymbol{y}^{(j)} = 1 \mid \boldsymbol{t}^{(j)}\boldsymbol{b}^{(j)} = 1) P(\boldsymbol{b}^{(j)} = 1) P(\boldsymbol{t}^{(j)} = 1) \\ &= \frac{(k-1)kp^2}{(N-1)N} P(\boldsymbol{y}^{(i)} = 1 \mid \boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1) P(\boldsymbol{y}^{(j)} = 1 \mid \boldsymbol{t}^{(j)}\boldsymbol{b}^{(j)} = 1).\end{aligned} \tag{C.14}$$

[2] Similarly to Eq. (C.7), the expression $\boldsymbol{q}^{(i)}/\boldsymbol{n}_1(k)$ is defined to be zero when $\boldsymbol{n}_1(k) = 0$ (in Definition C.1), therefore the ratio $\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}/\boldsymbol{n}_1(k)^2$ is well defined even in that case.

When $i = j$, we have

$$\mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}}{\boldsymbol{n}_1(k)^2}\right] = \sum_{n=1}^{k} \frac{1}{n^2} P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)} = 1) P(\boldsymbol{q}^{(i)} = 1).$$

From Lemma C.2, the sum can be developed as

$$\begin{aligned}\sum_{n=1}^{k} \frac{1}{n^2} P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)} = 1) &= \sum_{n=1}^{k} \frac{1}{n^2} P(\mathrm{Bin}(k-1, p) = n-1) \\ &= \sum_{n=1}^{k} \frac{1}{n^2} \binom{k-1}{n-1} p^{n-1}(1-p)^{k-n} \\ &= \frac{1}{kp} \sum_{n=1}^{k} \frac{1}{n} \binom{k}{n} p^{n}(1-p)^{k-n} = \frac{A_k}{kp}.\end{aligned}$$

Hence we have $\mathbb{E}\left[\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}/\boldsymbol{n}_1(k)^2\right] = P(\boldsymbol{q}^{(i)} = 1)A_k/(kp)$ for $i = j$. We can develop Eq. (C.12) as

$$\begin{aligned}\mathbb{E}\left[\frac{\boldsymbol{r}_1(k)^2}{\boldsymbol{n}_1(k)^2}\right] &= \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}}{\boldsymbol{n}_1(k)^2}\right] = \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \mathbb{E}\left[\frac{\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)}}{\boldsymbol{n}_1(k)^2}\right] + \sum_{i=1}^{N} \mathbb{E}\left[\frac{\boldsymbol{q}^{(i)2}}{\boldsymbol{n}_1(k)^2}\right] \\ &= \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \frac{1-(1-p)^k - A_k}{(k-1)kp^2} P(\boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1) + \sum_{i=1}^{N} \frac{A_k}{kp} P(\boldsymbol{q}^{(i)} = 1).\end{aligned}$$

Using Eqs. (C.10) and (C.14), we have

$$\begin{aligned}= \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \Bigg(&\frac{1-(1-p)^k - A_k}{(k-1)kp^2} \frac{(k-1)kp^2}{(N-1)N} P(\boldsymbol{y}^{(i)} = 1 \mid \boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1) \\ &P(\boldsymbol{y}^{(j)} = 1 \mid \boldsymbol{t}^{(j)}\boldsymbol{b}^{(j)} = 1)\Bigg) + \sum_{i=1}^{N} \frac{A_k}{kp}\frac{kp}{N} P(\boldsymbol{y}^{(i)} = 1 \mid \boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1).\end{aligned}$$

By the iid property of $\boldsymbol{D}_{\mathrm{te}}$, we replace every index i or j by N:

$$= (1-(1-p)^k - A_k) P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1)^2 + A_k P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1).$$

Let $S_N = P(\boldsymbol{y}^{(N)} = 1 \mid \boldsymbol{t}^{(N)}\boldsymbol{b}^{(N)} = 1)$. Coming back to Eq. (C.11), the variance is

$$\begin{aligned}\mathrm{Var}\left(\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right) &= \mathbb{E}\left[\frac{\boldsymbol{r}_1(k)^2}{\boldsymbol{n}_1(k)^2}\right] - \mathbb{E}\left[\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right]^2 \\ &= (1-(1-p)^k - A_k)S_N^2 + A_k S_N - (1-(1-p)^k)^2 S_N^2 \\ &= (1-(1-p)^k - 1 - (1-p)^{2k} + 2(1-p)^k - A_k)S_N^2 + A_k S_N \\ &= ((1-p)^k - (1-p)^{2k} - A_k)S_N^2 + A_k S_N.\end{aligned}$$

Its limit is

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right) = \lim_{N\to\infty} S_N(1-S_N)A_k.$$

By Lemma C.5, this converges to zero.

Be Chebychev's inequality, since $\mathbb{E}[\boldsymbol{r}_1(k)/\boldsymbol{n}_1(k)] = S$ (as shown in Eq. C.7), we know that for any fixed $\varepsilon > 0$,

$$P\left(\left|\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)} - S\right| \geq \varepsilon\right) \leq \operatorname{Var}\left(\frac{\boldsymbol{r}_1(k)}{\boldsymbol{n}_1(k)}\right)\frac{1}{\varepsilon^2}.$$

This converges to zero according to Eq. (C.11), which we proved above. □

C.3 Technical results

In this section, we prove the four lemmas used in the proof of Lemma C.1.

Lemma C.2. *Let $i \in \{1, \dots, N\}$. $\boldsymbol{n}_1(k) - 1 \mid \boldsymbol{q}^{(i)} = 1$ follows a binomial distribution* $\mathrm{Bin}(k-1, p)$.

Proof. Recall from Definition C.1 that $\boldsymbol{B}$ is a vector of binary random variables defined such that $\sum_i \boldsymbol{b}^{(i)} = k$ and $\mathscr{M}(\boldsymbol{x}^{(i)}) \geq \mathscr{M}(\boldsymbol{x}^{(j)})$ whenever $\boldsymbol{b}^{(i)} = 1$ and $\boldsymbol{b}^{(j)} = 0$ for any $i, j \in \{1, \dots, N\}$. By independence of the $\boldsymbol{x}^{(i)}$'s, $\boldsymbol{B}$ is uniformly distributed over the set of vectors with k values equal to one. That is, for any valuation $B = [b^{(i)}]_{i=1}^N$ of $\boldsymbol{B}$, one has

$$P(\boldsymbol{B} = B) = \begin{cases} 0 & \text{if } \sum_i b^{(i)} \neq k, \\ \binom{N}{k}^{-1} & \text{otherwise.} \end{cases} \tag{C.15}$$

By independence of the $\boldsymbol{x}^{(i)}$'s, the marginal distribution of each $\boldsymbol{b}^{(i)}$ is a Bernoulli trial of probability k/N. Now, let us fix the index $i \in \{1, \dots, N\}$, and let B be a realization of $\boldsymbol{B}$ such that $\sum_j b^{(j)} = k$ and $b^{(i)} = 1$. Then,

$$P(\boldsymbol{B} = B \mid \boldsymbol{b}^{(i)} = 1) = \frac{P(\boldsymbol{b}^{(i)} = 1 \mid \boldsymbol{B} = B)P(\boldsymbol{B} = B)}{P(\boldsymbol{b}^{(i)} = 1)} = \frac{N}{k}\binom{N}{k}^{-1} = \binom{N-1}{k-1}^{-1}. \tag{C.16}$$

And that probability is zero if B is a realization of $\boldsymbol{B}$ such that $\sum_j b^{(j)} = 1$ and $b^{(i)} = 0$. Now, note that $\boldsymbol{n}_1(k) \perp \boldsymbol{y}^{(i)} \mid \boldsymbol{t}^{(i)}, \boldsymbol{b}^{(i)}$, hence $P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)} = 1) = P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1)$. Also, remember that $\boldsymbol{n}_1(k) = \sum_i \boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)}$, hence the expression $\boldsymbol{n}_1(k) = n$ can be decomposed as

$$\begin{aligned} P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1) &= P\left(\boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1, \sum_{j=1, j\neq i}^{N} \boldsymbol{t}^{(j)}\boldsymbol{b}^{(j)} = n - 1 \mid \boldsymbol{t}^{(i)}\boldsymbol{b}^{(i)} = 1\right) \\ &= P\left(\sum_{j=1, j\neq i}^{N} \boldsymbol{t}^{(j)}\boldsymbol{b}^{(j)} = n - 1 \mid \boldsymbol{b}^{(i)} = 1\right) \end{aligned}$$

where the last equality follows from the independence $\boldsymbol{t}^{(i)} \perp \boldsymbol{t}^{(j)}$ for $i \neq j$ (since $\boldsymbol{t}$ is randomized), and the independence $\boldsymbol{t}^{(i)} \perp \boldsymbol{b}^{(j)}$ for any i, j (since the treatment does not influence the scores, and inversely, by randomization). Marginalizing on all possible

realizations B of $\boldsymbol{B}$ such that $|B| = \sum_j b^{(j)} = k$ and $b^{(i)} = 1$, we have

$$= \sum_{\substack{B \text{ s.t. } |B|=k, \\ b^{(i)}=1}} P(\boldsymbol{B} = B \mid \boldsymbol{b}^{(i)} = 1) P\left(\sum_{j=1, j\neq i}^{N} \boldsymbol{t}^{(j)} \boldsymbol{b}^{(j)} = n-1 \mid \boldsymbol{B} = B, \boldsymbol{b}^{(i)} = 1 \right)$$

$$= \sum_{\substack{B \text{ s.t. } |B|=k, \\ b^{(i)}=1}} \binom{N-1}{k-1}^{-1} P\left(\sum_{j=1, j\neq i}^{N} \boldsymbol{t}^{(j)} \boldsymbol{b}^{(j)} = n-1 \mid \boldsymbol{B} = B \right). \qquad \text{(by Eq. C.16)}$$

In the expression $\sum_{j=1,j\neq i}^{N} \boldsymbol{t}^{(j)} \boldsymbol{b}^{(j)}$, we know that $k-1$ of the $N-1$ terms $b^{(j)}$ are equal to one (because we condition on $\boldsymbol{B} = B$), therefore, for the sum to equal $n-1$, we need $n-1$ of the corresponding terms $t^{(j)}$ to equal one. Since they are all independent with a Bernoulli distribution Bern(p), the probability that $n-1$ of them are one is equal to the probability of a random variable with binomial distribution Bin($k-1, p$) to be equal to $n-1$. This leads to

$$P(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)} = 1) = \sum_{\substack{B \text{ s.t. } |B|=k, \\ b^{(i)}=1}} \binom{N-1}{k-1}^{-1} P(\text{Bin}(k-1, p) = n-1)$$
$$= P(\text{Bin}(k-1, p) = n-1).$$

The last equality follows from

$$\sum_{\substack{B \text{ s.t. } |B|=k, \\ b^{(i)}=1}} \binom{N-1}{k-1}^{-1} = 1.$$

□

Lemma C.3. *Let $i, j \in \{1, \dots, N\}$, with $i \neq j$. $\boldsymbol{n}_1(k) - 2 \mid \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1$ follows a binomial distribution* Bin($k-2, p$).

Proof. Let $B = [b^{(1)}, \dots, b^{(N)}]$ be a realization of the random vector $\boldsymbol{B} = [\boldsymbol{b}^{(1)}, \dots, \boldsymbol{b}^{(N)}]$ such that $b^{(i)} = b^{(j)} = 1$. Let us compute the probability

$$P(\boldsymbol{B} = B \mid \boldsymbol{b}^{(i)}\boldsymbol{b}^{(j)} = 1) = \frac{P(\boldsymbol{b}^{(i)}\boldsymbol{b}^{(j)} = 1 \mid \boldsymbol{B} = B) P(\boldsymbol{B} = B)}{P(\boldsymbol{b}^{(i)} = 1 \mid \boldsymbol{b}^{(j)} = 1) P(\boldsymbol{b}^{(j)} = 1)} \qquad \text{(C.17)}$$

$$= \binom{N}{k}^{-1} \frac{N-1}{k-1} \frac{N}{k} = \binom{N-2}{k-2}^{-1}. \qquad \text{(C.18)}$$

Using a similar reasoning as in the proof of Lemma C.2, we marginalize the probability

$P\left(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1\right)$ over the distribution of $\boldsymbol{B}$:

$$\begin{aligned}
&P\left(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1\right) \\
&\quad= \sum_{\substack{B \text{ s.t.} |B|=k \\ b^{(i)}b^{(j)}=1}} P\left(\boldsymbol{n}_1(k) = n \mid \boldsymbol{B} = B, \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1\right) P(\boldsymbol{B} = B \mid \boldsymbol{b}^{(i)}\boldsymbol{b}^{(j)} = 1) \quad \text{(by Eq. C.18)} \\
&\quad= \sum_{\substack{B \text{ s.t.} |B|=k \\ b^{(i)}b^{(j)}=1}} \binom{N-2}{k-2}^{-1} P\left(\sum_{\ell=1}^{N} \boldsymbol{t}^{(\ell)}\boldsymbol{b}^{(\ell)} = n \mid \boldsymbol{B} = B, \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1\right) \\
&\quad= \sum_{\substack{B \text{ s.t.} |B|=k \\ b^{(i)}b^{(j)}=1}} \binom{N-2}{k-2}^{-1} P\left(\sum_{\ell=1, \ell\neq i,j}^{N} \boldsymbol{t}^{(\ell)}\boldsymbol{b}^{(\ell)} = n-2 \mid \boldsymbol{B} = B\right). \qquad \text{(C.19)}
\end{aligned}$$

Again, the probability that $\sum_{\ell=1,\ell\neq i,j}^{N} \boldsymbol{t}^{(\ell)}\boldsymbol{b}^{(\ell)}$ is equal to $n-2$, given that $\boldsymbol{B} = B$, is the same as the probability that for the $k-2$ indices ℓ where $\boldsymbol{b}^{(\ell)} = 1$, $n-2$ of the associated Bernoulli-distributed terms $\boldsymbol{t}^{(\ell)}$ are equal to one. In equations, this is written

$$P\left(\sum_{\ell=1,\ell\neq i,j}^{N} \boldsymbol{t}^{(\ell)}\boldsymbol{b}^{(\ell)} = n-2 \mid \boldsymbol{B} = B\right) = P(\text{Bin}(k-2, p) = n-2)$$

for any B such that $|B| = k$ and $b^{(i)}b^{(j)} = 1$. By plugging this in Eq. (C.19), we obtain

$$\begin{aligned}
P\left(\boldsymbol{n}_1(k) = n \mid \boldsymbol{q}^{(i)}\boldsymbol{q}^{(j)} = 1\right) &= \sum_{\substack{B \text{ s.t.} |B|=k \\ b^{(i)}b^{(j)}=1}} \binom{N-2}{k-2}^{-1} P(\text{Bin}(k-2, p) = n-2) \\
&= P(\text{Bin}(k-2, p) = n-2).
\end{aligned}$$

The last equality follows from

$$\sum_{\substack{B \text{ s.t.} |B|=k \\ b^{(i)}b^{(j)}=1}} \binom{N-2}{k-2}^{-1} = 1.$$

□

Lemma C.4.

$$\lim_{N\to\infty} P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau \mid \boldsymbol{b}^{(N)} = 1) = \lim_{N\to\infty} P(\boldsymbol{b}^{(N)} = 1 \mid \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau) = 1.$$

Proof. As mentioned in the previous section, the index N has no special value here, and this lemma could be formulated with any integer smaller or equal than N. However, from a practical standpoint, any index other than 1 or N would require some special notation in order to be valid for any value of N (e.g., using index 3 would leave the case $N = 2$ undefined). We use N specifically because it simplifies the notation during the proof of Lemma C.1.

This proof requires a few additional definitions. Let $J = \{j \in \{1, \dots, N\} : \boldsymbol{b}^{(j)} = 0\}$ be the set of $N-k$ indices with the lowest scores according to $\mathscr{M}$. Since $\mathscr{M}(\boldsymbol{x})$ is absolutely

continuous, $P(\mathscr{M}(\boldsymbol{x}^{(i)}) = \mathscr{M}(\boldsymbol{x}^{(j)})) = 0$ for any two distinct indices i, j, therefore, the probability that $\boldsymbol{J}$ is not uniquely defined is zero. Let $\boldsymbol{\tau}_N = \max_{j \in \boldsymbol{J}}\{\mathscr{M}(\boldsymbol{x}^{(j)})\}$ be the threshold that separates the $N - k$ instances with the lowest scores from the k other instances. This is the empirical equivalent of τ with a dataset of size N. Since the dataset is a random variable, $\boldsymbol{\tau}_N$ is also a random variable. First, we will prove that $\boldsymbol{\tau}_N$ converges to τ almost surely. Recall that τ is defined as $\tau = \inf\{\tau' : P(\mathscr{M}(\boldsymbol{x}) \geq \tau') \geq \rho\}$ (see Eq. 5.1). Let $\boldsymbol{F}_N$ be the empirical cumulative distribution function of $\left\{\mathscr{M}(\boldsymbol{x}^{(i)})\right\}_{i=1}^{N}$, and let $F_{\mathscr{M}(\boldsymbol{x})}$ be the cumulative distribution function of $\mathscr{M}(\boldsymbol{x})$. Note that $\boldsymbol{F}_N$ is a random variable since it is defined in terms of the dataset $\boldsymbol{D}_{\text{te}}$. We have

$$\begin{aligned}
\lim_{N\to\infty} \boldsymbol{F}_N(\boldsymbol{\tau}_N) &= \lim_{N\to\infty} \frac{N-k+1}{N} && \text{(by def. of } \boldsymbol{F}_N\text{)} \\
&= 1 - \rho && \text{(by def. of } k\text{)} \\
&= P(\mathscr{M}(\boldsymbol{x}) < \tau) = F_{\mathscr{M}(\boldsymbol{x})}(\tau). && \text{(by def. of } \tau\text{)}
\end{aligned}$$

The Glivenko-Cantelli theorem states that $\sup_{t\in\mathbb{R}} |\boldsymbol{F}_N(t) - F_{\mathscr{M}(\boldsymbol{x})}(t)|$ converges almost surely to zero. By definition of the supremum, we know that, for any N,

$$|\boldsymbol{F}_N(\boldsymbol{\tau}_N) - F_{\mathscr{M}(\boldsymbol{x})}(\boldsymbol{\tau}_N)| \leq \sup_{t\in\mathbb{R}} |\boldsymbol{F}_N(t) - F_{\mathscr{M}(\boldsymbol{x})}(t)|.$$

Thus, $|\boldsymbol{F}_N(\boldsymbol{\tau}_N) - F_{\mathscr{M}(\boldsymbol{x})}(\boldsymbol{\tau}_N)|$ converges almost surely to zero, and $\boldsymbol{F}_N(\boldsymbol{\tau}_N) - F_{\mathscr{M}(\boldsymbol{x})}(\boldsymbol{\tau}_N)$ as well. We can now develop

$$\lim_{N\to\infty} \boldsymbol{F}_N(\boldsymbol{\tau}_N) - F_{\mathscr{M}(\boldsymbol{x})}(\boldsymbol{\tau}_N) = 0 \quad \text{almost surely,}$$

which means that

$$\lim_{N\to\infty} \boldsymbol{F}_N(\boldsymbol{\tau}_N) = \lim_{N\to\infty} F_{\mathscr{M}(\boldsymbol{x})}(\boldsymbol{\tau}_N) = F_{\mathscr{M}(\boldsymbol{x})}\left(\lim_{N\to\infty} \boldsymbol{\tau}_N\right)$$

by continuity of $F_{\mathscr{M}(\boldsymbol{x})}$. Wrapping up, we have $F_{\mathscr{M}(\boldsymbol{x})}(\lim_N \boldsymbol{\tau}_N) = F_{\mathscr{M}(\boldsymbol{x})}(\tau)$. If $F_{\mathscr{M}(\boldsymbol{x})}$ is strictly increasing at τ, then, by continuity, $\lim_N \boldsymbol{\tau}_N = \tau$, which is what we wanted to prove. Otherwise, there is a interval $[a, b]$ with $a < b$ such that $F_{\mathscr{M}(\boldsymbol{x})}(t) = F_{\mathscr{M}(\boldsymbol{x})}(\tau)$ for any $t \in [a, b]$. Since $F_{\mathscr{M}(\boldsymbol{x})}(\lim_N \boldsymbol{\tau}_N) = F_{\mathscr{M}(\boldsymbol{x})}(\tau)$, we have $\lim_N \boldsymbol{\tau}_N \in [a, b]$. Since $F_{\mathscr{M}(\boldsymbol{x})}(a) = F_{\mathscr{M}(\boldsymbol{x})}(b)$, we know that $P(a < \mathscr{M}(\boldsymbol{x}) < b) = 0$, hence $\lim_N \boldsymbol{\tau}_N$ is either a or b. In fact, $\boldsymbol{\tau}_N = a$ because $\boldsymbol{\tau}_N$ is defined as $\max_{j\in\boldsymbol{J}}\{\mathscr{M}(\boldsymbol{x}^{(j)})\}$, the maximum of the $N - k$ lowest scores. Also, by definition of τ, we have $\tau = \inf[a, b] = a$. Therefore, $\lim_N \boldsymbol{\tau}_N = \tau$.

Now, we are ready to prove that $\lim_{N\to\infty} P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq t \mid \boldsymbol{b}^{(N)} = 1) = 1$:

$$\begin{aligned}
P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau \mid \boldsymbol{b}^{(N)} = 1) &= P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau \mid \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \boldsymbol{\tau}_N) \\
&= 1 - P(\mathscr{M}(\boldsymbol{x}^{(N)}) < \tau \mid \mathscr{M}(\boldsymbol{x}^{(N)}) \geq \boldsymbol{\tau}_N) \\
&= 1 - \frac{P(\boldsymbol{\tau}_N \leq \mathscr{M}(\boldsymbol{x}^{(N)}) < \tau)}{P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \boldsymbol{\tau}_N)}.
\end{aligned}$$

Since $\boldsymbol{\tau}_N$ converges to τ, $P(\boldsymbol{\tau}_N < \mathscr{M}(\boldsymbol{x}^{(N)}) < \tau)$ converges to zero, and therefore $P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq t \mid \boldsymbol{b}^{(N)} = 1)$ converges to one. This proves the first part of the claim. We

can then use Bayes' theorem to develop

$$P(\boldsymbol{b}^{(N)} = 1 \mid \mathscr{M}(\boldsymbol{x}^{(N)}) \geq t) = P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau \mid \boldsymbol{b}^{(N)} = 1)\frac{P(\boldsymbol{b}^{(N)} = 1)}{P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau)}$$
$$= P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau \mid \boldsymbol{b}^{(N)} = 1)\frac{k/N}{\rho}$$

which converges to the same value as $P(\mathscr{M}(\boldsymbol{x}^{(N)}) \geq \tau \mid \boldsymbol{b}^{(N)} = 1)$, that is, 1. This proves the second part of Lemma C.4. □

Lemma C.5. *For any $0 < p < 1$, we have*

$$\lim_{k\to\infty} \sum_{n=1}^{k} \frac{1}{n}\binom{k}{n} p^n (1-p)^{k-n} = 0.$$

Proof.

$$\sum_{n=1}^{k} \frac{1}{n}\binom{k}{n} p^n (1-p)^{k-n} = \sum_{n=1}^{k} \frac{1}{n}\frac{n+1}{k+1}\binom{k+1}{n+1} p^n (1-p)^{k-n}$$
$$= \frac{1}{k+1}\sum_{n=1}^{k} \frac{n+1}{n}\binom{k+1}{n+1} p^n (1-p)^{k-n}.$$

Since $(n+1)/n \leq 2$, we have

$$\sum_{n=1}^{k} \frac{1}{n}\binom{k}{n} p^n (1-p)^{k-n} \leq \frac{2}{k+1}\sum_{n=1}^{k} \binom{k+1}{n+1} p^n (1-p)^{k-n}$$
$$= \frac{2}{p(k+1)}\sum_{n=2}^{k+1} \binom{k+1}{n} p^n (1-p)^{k+1-n}$$
$$\leq \frac{2}{p(k+1)}\sum_{n=0}^{k+1} \binom{k+1}{n} p^n (1-p)^{k+1-n}$$
$$= \frac{2}{p(k+1)}\sum_{n=0}^{k+1} P(\mathrm{Bin}(k+1, p) = n) = \frac{2}{p(k+1)}.$$

Therefore

$$\lim_{k\to\infty} \sum_{n=1}^{k} \frac{1}{n}\binom{k}{n} p^n (1-p)^{k-n} \leq \lim_{k\to\infty} \frac{2}{p(k+1)} = 0.$$

Since each term of the sum is positive, the limit is equal to zero. □

D

Properties of the bivariate beta distribution

In this appendix, we derive several properties of the bivariate beta distribution developed by Olkin and Trikalinos (2015), which is used in Sections 5.3.3, 6.4 and 6.5.1. Then, we derive some of these properties for the variations of the bivariate beta distribution proposed in Sections 6.4.4 to 6.4.6. We use the notation of Chapter 6, where $\boldsymbol{\mu}$ is a four-dimensional random vector with positive values summing up to one. In Chapter 5 we added a superscript (i) to highlight the fact that each individual is associated with a different realization of $\boldsymbol{\mu}$, and to avoid the confusion with population-level counterfactuals $\alpha, \dots, \delta$. As in Chapter 6, in this appendix we omit the superscripts to ease the already heavy notation, at the risk of being less clear. We note the components of $\boldsymbol{\mu}$ as $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$, or $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4$. Also, m is a 4-dimensional vector with positive values, whose components are noted a, b, c, d or m_1, m_2, m_3, m_4. Their sum is noted $M = a + b + c + d$.

D.1 Original bivariate beta distribution

In this section, we focus on the following data-generating process;

$$\boldsymbol{\mu} \sim \text{Dir}(m) \tag{D.1}$$
$$\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta} \tag{D.2}$$
$$\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta}. \tag{D.3}$$

We say that the pair of random variables $(\boldsymbol{S}_0, \boldsymbol{S}_1)$ follows a bivariate beta distribution $\text{BB}(m)$. Their joint probability density function, reproduced from Eqs. (6.75) and (6.78), is

$$f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(S_0, S_1) = \int_{\Lambda(S_0,S_1)} f_{\boldsymbol{\mu}}(\mu)\, \mathrm{d}\mu = \frac{1}{\text{B}(m)} \int_{\Lambda(S_0,S_1)} \prod_{j=1}^{4} \mu_j^{m_j - 1}\, \mathrm{d}\mu \tag{D.4}$$

$$= \frac{1}{\text{B}(m)} \int_{\max\{0, S_0+S_1-1\}}^{\min\{S_0,S_1\}} (1 - S_0 - S_1 + \delta)^{a-1} (S_0 - \delta)^{b-1} (S_1 - \delta)^{c-1} \delta^{d-1}\, \mathrm{d}\delta. \tag{D.5}$$

From this bivariate beta distribution, we sample the binary potential outcomes $\boldsymbol{y}_0, \boldsymbol{y}_1$ according to

$$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 0 \mid \boldsymbol{\mu} = \mu) = \alpha \tag{D.6}$$
$$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 0 \mid \boldsymbol{\mu} = \mu) = \beta \tag{D.7}$$
$$P(\boldsymbol{y}_0 = 0, \boldsymbol{y}_1 = 1 \mid \boldsymbol{\mu} = \mu) = \gamma \tag{D.8}$$
$$P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1 \mid \boldsymbol{\mu} = \mu) = \delta \tag{D.9}$$

or, more succinctly,

$$(\boldsymbol{y}_0, \boldsymbol{y}_1) \sim \mathrm{Cat}(\boldsymbol{\mu}). \tag{D.10}$$

From this, it is easy to show that $P(\boldsymbol{y}_0 = 1 \mid \boldsymbol{S}_0 = S_0) = S_0$ and $P(\boldsymbol{y}_1 = 1 \mid \boldsymbol{S}_1 = S_1) = S_1$.

First, we provide an analytical formula that relates the mutual information between $\boldsymbol{y}_0, \boldsymbol{y}_1$ and $\boldsymbol{\mu}$ to the sum of the distribution parameters. This is used in Section 5.3.3, and in Section 6.5.1 to illustrate Theorem 6.1.

Result D.1. *The mutual information between $\boldsymbol{y}_0, \boldsymbol{y}_1$ (either jointly or separately) and $\boldsymbol{\mu}$ is given by*

$$I(\boldsymbol{y}_0, \boldsymbol{y}_1; \boldsymbol{\mu}) = H(\boldsymbol{y}_0, \boldsymbol{y}_1) - \psi(M+1) + \sum_{j=1}^{4} \frac{m_j}{M} \psi(m_j + 1) \tag{D.11}$$

$$I(\boldsymbol{y}_0; \boldsymbol{\mu}) = H(\boldsymbol{y}_0) - \psi(M+1) + \frac{b+d}{M}\psi(b+d+1) + \frac{a+c}{M}\psi(a+c+1) \tag{D.12}$$

$$I(\boldsymbol{y}_1; \boldsymbol{\mu}) = H(\boldsymbol{y}_1) - \psi(M+1) + \frac{c+d}{M}\psi(c+d+1) + \frac{a+b}{M}\psi(a+b+1). \tag{D.13}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. Furthermore, as $M \to 0$, the mutual information (respectively, $I(\boldsymbol{y}_0, \boldsymbol{y}_1; \boldsymbol{\mu})$, $I(\boldsymbol{y}_0; \boldsymbol{\mu})$, and $I(\boldsymbol{y}_1; \boldsymbol{\mu})$) converges to the entropy (respectively, $H(\boldsymbol{y}_0, \boldsymbol{y}_1)$, $H(\boldsymbol{y}_0)$, and $H(\boldsymbol{y}_1)$), and, as $M \to \infty$, the mutual information converges to zero.

Proof. Using the identity $I(\boldsymbol{v}, \boldsymbol{w}) = H(\boldsymbol{v}) - H(\boldsymbol{v} \mid \boldsymbol{w})$ (see Eq. 2.21), we start by proving

$$H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu}) = \psi(M+1) - \sum_{j=1}^{4} \frac{m_j}{M} \psi(m_j + 1) \tag{D.14}$$

$$H(\boldsymbol{y}_0 \mid \boldsymbol{\mu}) = \psi(M+1) - \frac{b+d}{M}\psi(b+d+1) - \frac{a+c}{M}\psi(a+c+1) \tag{D.15}$$

$$H(\boldsymbol{y}_1 \mid \boldsymbol{\mu}) = \psi(M+1) - \frac{c+d}{M}\psi(c+d+1) - \frac{a+b}{M}\psi(a+b+1). \tag{D.16}$$

First, we show the derivation for $H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu})$. The conditional entropy of $\boldsymbol{y}_0, \boldsymbol{y}_1$ given a realization $\boldsymbol{\mu} = \mu$ is

$$\begin{aligned} H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu} = \mu) &= -\alpha \log \alpha - \beta \log \beta - \gamma \log \gamma - \delta \log \delta \\ &= -\sum_{j=1}^{4} \mu_j \log \mu_j. \end{aligned}$$

The conditional entropy $H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu})$ is the expected value of that quantity over the distribution of $\boldsymbol{\mu}$,

$$H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu}) = \int_{\Lambda} H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu} = \mu) f_{\boldsymbol{\mu}}(\mu) \, \mathrm{d}\mu,$$

where Λ is the unit 4-dimensional simplex $\Lambda = \{\mu \mid \mu_j \geq 0 \text{ and } \sum_j \mu_j = 1\}$, and $f_{\boldsymbol{\mu}}$ is the pdf of the Dirichlet distribution $\mathrm{Dir}(m)$. We can develop this as

$$\begin{aligned} H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu}) &= -\int_\Lambda f_{\boldsymbol{\mu}}(\mu) \sum_{j=1}^{4} \mu_j \log \mu_j \, \mathrm{d}\mu \\ &= -\frac{1}{\mathrm{B}(m)} \sum_{j=1}^{4} \int_\Lambda \mu_j \log \mu_j \prod_{k=1}^{4} \mu_k^{m_k - 1} \, \mathrm{d}\mu \\ &= -\frac{1}{\mathrm{B}(m)} \sum_{j=1}^{4} \mathscr{I}_j \end{aligned}$$

where we defined

$$\mathscr{I}_j = \int_\Lambda \mu_j \log \mu_j \prod_{k=1}^{4} \mu_k^{m_k - 1} \, \mathrm{d}\mu = \int_\Lambda \mu_j^{m_j} \log \mu_j \prod_{k \neq j} \mu_k^{m_k - 1} \, \mathrm{d}\mu.$$

We can separate the integral over the four-dimensional domain Λ to integrate first on μ_j and then the remaining dimensions, leading to

$$\mathscr{I}_j = \int_0^1 \mu_j^{m_j} \log \mu_j \int_{\Lambda(\mu_j)} \prod_{k \neq j} \mu_k^{m_k - 1} \, \mathrm{d}\mu_{-j} \, \mathrm{d}\mu_j$$

where we defined $\mu_{-j} = [\mu_k]_{k \neq j}$ as the vector of μ without μ_j, and $\Lambda(\mu_j)$ as the set of three-dimensional vectors summing up to $1 - \mu_j$:

$$\Lambda(\mu_j) = \left\{ \mu_{-j} \,:\, \mu_k > 0, \sum_{k \neq j} \mu_k = 1 - \mu_j \right\}.$$

We use the Leibniz integral rule and the fact that $\mathrm{d}a^x / \mathrm{d}x = a^x \log a$ to express $\mathscr{I}_j$ as

$$\begin{aligned} \mathscr{I}_j &= \int_0^1 \frac{\partial \mu_j^{m_j}}{\partial m_j} \int_{\Lambda(\mu_j)} \prod_{k \neq j} \mu_k^{m_k - 1} \, \mathrm{d}\mu_{-j} \, \mathrm{d}\mu_j \\ &= \frac{\partial}{\partial m_j} \int_0^1 \mu_j^{m_j} \int_{\Lambda(\mu_j)} \prod_{k \neq j} \mu_k^{m_k - 1} \, \mathrm{d}\mu_{-j} \, \mathrm{d}\mu_j \\ &= \frac{\partial}{\partial m_j} \int_\Lambda \mu_j^{m_j} \prod_{k \neq j} \mu_k^{m_k - 1} \, \mathrm{d}\mu. \end{aligned}$$

Note that this last expression is the partial derivative of the Beta function on a vector $m' = [m_1, \ldots, m_j + 1, \ldots, m_4]$. This partial derivative can also be expressed in terms of the digamma function:

$$\begin{aligned} \mathscr{I}_i &= \frac{\partial \mathrm{B}(m')}{\partial m_j} = \mathrm{B}(m')(\psi(m_j + 1) - \psi(M + 1)) \\ &= \frac{m_1}{M} \mathrm{B}(m)(\psi(m_j + 1) - \psi(M + 1)) \end{aligned}$$

where we used the identity

$$\mathrm{B}(m') = \frac{\Gamma(m_j + 1) \prod_{k \neq j} \Gamma(m_k)}{\Gamma(M + 1)} = \frac{m_j \Gamma(m_j) \prod_{k \neq j} \Gamma(m_k)}{M \Gamma(M)} = \frac{m_j}{M} \mathrm{B}(m).$$

Finally, the conditional entropy of $\boldsymbol{y}_0, \boldsymbol{y}_1$ is

$$H(\boldsymbol{y}_0, \boldsymbol{y}_1 \mid \boldsymbol{\mu}) = -\frac{1}{\mathrm{B}(m)} \sum_{j=1}^{4} \mathscr{I}_j = -\sum_{j=1}^{4} \frac{m_j}{M}(\psi(m_j + 1) - \psi(M + 1))$$
$$= \psi(M + 1) - \sum_{j=1}^{4} \frac{m_j}{M}\psi(m_j + 1).$$

Note that since the distribution of $\boldsymbol{y}_0$ only depends upon $\boldsymbol{S}_0$, we have that $H(\boldsymbol{y}_0 \mid \boldsymbol{\mu}) = H(\boldsymbol{y}_0 \mid \boldsymbol{S}_0)$. Using a similar reasoning as for joint entropy, the marginal entropy is developed as

$$H(\boldsymbol{y}_0 \mid \boldsymbol{\mu}) = H(\boldsymbol{y}_0 \mid \boldsymbol{S}_0) = -\int_0^1 (s_0 \log s_0 + (1 - s_0)\log(1 - s_0)) f_{\boldsymbol{S}_0}(s_0)\, \mathrm{d}s_0$$

where $f_{\boldsymbol{S}_0}$ is the pdf of $\boldsymbol{S}_0$. As shown by Olkin and Trikalinos (2015), $\boldsymbol{S}_0$ follows a beta distribution Beta$(b + d, a + c)$. Therefore, we can expand its pdf as

$$H(\boldsymbol{y}_0 \mid \boldsymbol{\mu}) = -\frac{1}{\mathrm{B}(b + d, a + c)} \int_0^1 (s_0 \log s_0 + (1 - s_0)\log(1 - s_0)) s_0^{b+d-1}(1 - s_0)^{a+c-1}\, \mathrm{d}s_0$$
$$= -\frac{1}{\mathrm{B}(b + d, a + c)} \int_0^1 (s_0^{b+d}(1 - s_0)^{a+c-1} \log s_0 + s_0^{b+d-1}(1 - s_0)^{a+c} \log(1 - s_0))\, \mathrm{d}s_0.$$

As above, we use the Leibniz integral rule and identity $\mathrm{d}a^x / \mathrm{d}x = a^x \log a$ to express the equation above in terms of the derivative of the Beta function and ultimately in terms of the digamma function:

$$H(\boldsymbol{y}_0 \mid \boldsymbol{\mu}) = -\frac{1}{\mathrm{B}(b + d, a + c)} \int_0^1 \left(\frac{\partial s_0^z}{\partial z}\bigg|_{z=b+d} (1 - s_0)^{a+c-1} + s_0^{b+d-1} \frac{\partial (1 - s_0)^z}{\partial z}\bigg|_{z=a+c} \right) \mathrm{d}s_0$$
$$= -\frac{1}{\mathrm{B}(b + d, a + c)} \left(\frac{\partial \mathrm{B}(b + d + 1, a + c)}{\partial (b + d)} + \frac{\partial \mathrm{B}(b + d, a + c + 1)}{\partial (a + c)} \right)$$
$$= -\left(\frac{b + d}{M}(\psi(b + d + 1) - \psi(M + 1)) + \frac{a + c}{M}(\psi(a + c + 1) - \psi(M + 1)) \right)$$
$$= \psi(M + 1) - \frac{b + d}{M}\psi(b + d + 1) - \frac{a + c}{M}\psi(a + c + 1).$$

The proof for $H(\boldsymbol{y}_1 \mid \boldsymbol{\mu})$ follows a similar development. □

Now, we provide an analytical formula for the raw moments of the bivariate beta distribution.

Result D.2. *The raw moments of* $(\boldsymbol{S}_0, \boldsymbol{S}_1) \sim \mathrm{BB}(m)$ *of order* $r, s > 0$*, noted* $R_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1)$*, are*

$$R_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1) = \mathbb{E}[\boldsymbol{S}_0^r \boldsymbol{S}_1^s] = \sum_{p=0}^{r} \sum_{q=0}^{s} \binom{r}{p} \binom{s}{q} \frac{\mathrm{B}(a, b + r - p, c + s - q, d + p + q)}{\mathrm{B}(a, b, c, d)}.$$

When r *and* s *are positive integers, this reduces to*

$$R_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1) = \mathbb{E}[\boldsymbol{S}_0^r \boldsymbol{S}_1^s] = \sum_{p=0}^{r} \sum_{q=0}^{s} \binom{r}{p} \binom{s}{q} \frac{b^{\overline{r-p}} c^{\overline{s-q}} d^{\overline{p+q}}}{M^{\overline{r+s}}}$$

where $x^{\overline{n}} = x(x + 1)\dots(x + n - 1)$ *is the rising factorial.*

Proof. Let us expand the definition of the raw moments from Definition 2.4:

$$\begin{aligned} R_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1) &= \mathbb{E}[\boldsymbol{S}_0^r \boldsymbol{S}_1^s] = \mathbb{E}[(\boldsymbol{\beta} + \boldsymbol{\delta})^r (\boldsymbol{\gamma} + \boldsymbol{\delta})^s] \\ &= \mathbb{E}\left[\left(\sum_{p=0}^{r} \binom{r}{p} \beta^{r-p} \delta^p\right)\left(\sum_{q=0}^{s} \binom{s}{q} \gamma^{s-q} \delta^q\right)\right] \\ &= \sum_{p=0}^{r}\sum_{q=0}^{s} \binom{r}{p}\binom{s}{q} \mathbb{E}[\beta^{r-p}\delta^p \gamma^{s-q}\delta^q] \\ &= \sum_{p=0}^{r}\sum_{q=0}^{s} \binom{r}{p}\binom{s}{q} \frac{\mathrm{B}(a, b+r-p, c+s-q, d+p+q)}{\mathrm{B}(a,b,c,d)}. \end{aligned}$$

This proves the first part of the result. We can expand the beta function to obtain

$$R_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1) = \sum_{p=0}^{r}\sum_{q=0}^{s} \binom{r}{p}\binom{s}{q} \frac{\Gamma(a)\Gamma(b+r-p)\Gamma(c+s-q)\Gamma(d+p+q)\Gamma(M)}{\Gamma(M+r+s)\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(d)}.$$

We can show that $\Gamma(x+n) = \Gamma(x)x^{\overline{n}}$ for a positive integer n. Then, when r and s are positive integers, the raw moments simplify to

$$R_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1) = \mathbb{E}[\boldsymbol{S}_0^r \boldsymbol{S}_1^s] = \sum_{p=0}^{r}\sum_{q=0}^{s} \binom{r}{p}\binom{s}{q} \frac{b^{\overline{r-p}} c^{\overline{s-q}} d^{\overline{p+q}}}{M^{\overline{r+s}}}.$$

□

Result D.3. *The partial derivative of the raw moments of* $(\boldsymbol{S}_0, \boldsymbol{S}_1) \sim \mathrm{BB}(m)$ *for integers* $r, s > 0$ *is*

$$\frac{\partial R_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1)}{\partial m_i} = \sum_{p=0}^{r}\sum_{q=0}^{s} \binom{r}{p}\binom{s}{q} \frac{b^{\overline{r-p}} c^{\overline{s-q}} d^{\overline{p+q}}}{M^{\overline{r+s}}} (m_i^* - M_{\overline{r+s}}) \tag{D.17}$$

where $x_{\overline{n}}$*, that we name the* harmonic difference*, is defined as*

$$x_{\overline{n}} = \sum_{i=0}^{n-1} \frac{1}{x+i}, \tag{D.18}$$

and

$$m_1^* = 0 \qquad m_2^* = b_{\overline{r-p}} \qquad m_3^* = c_{\overline{s-q}} \qquad m_4^* = d_{\overline{p+q}}.$$

Proof. The partial derivative of the raw moments can be readily expressed as

$$\begin{aligned} \frac{\partial R_{rs}(\boldsymbol{S}_0, \boldsymbol{S}_1)}{\partial m_i} &= \frac{\partial}{\partial m_i} \sum_{p=0}^{r}\sum_{q=0}^{s} \binom{r}{p}\binom{s}{q} \frac{b^{\overline{r-p}} c^{\overline{s-q}} d^{\overline{p+q}}}{M^{\overline{r+s}}} \\ &= \sum_{p=0}^{r}\sum_{q=0}^{s} \binom{r}{p}\binom{s}{q} \frac{\partial}{\partial m_i} \frac{b^{\overline{r-p}} c^{\overline{s-q}} d^{\overline{p+q}}}{M^{\overline{r+s}}}. \end{aligned} \tag{D.19}$$

In this expression, we have to compute the derivative of the rising factorial function. Using the product rule for differentiation, we have

$$\frac{\partial x^{\overline{n}}}{\partial x} = \frac{\partial}{\partial x} \prod_{i=0}^{n-1} (x+i) = \prod_{i=0}^{n-1} (x+i) \sum_{i=0}^{n-1} \frac{1}{x+i} \frac{\partial x+i}{\partial x} = x^{\overline{n}} x_{\overline{n}}.$$

Applying this result to Eq. (D.19), we can find

$$\begin{aligned}
\frac{\partial}{\partial a}\frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{M^{\overline{r+s}}} &= -\frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{(M^{\overline{r+s}})^2}M^{\overline{r+s}}M_{\overline{r+s}} = \frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{(M^{\overline{r+s}})}(-M_{\overline{r+s}}) \\
\frac{\partial}{\partial b}\frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{M^{\overline{r+s}}} &= \frac{b^{\overline{r-p}}b_{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}M^{\overline{r+s}} - b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}M^{\overline{r+s}}M_{\overline{r+s}}}{(M^{\overline{r+s}})^2} \\
&= \frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{M^{\overline{r+s}}}(b_{\overline{r-p}} - M_{\overline{r+s}}) \\
\frac{\partial}{\partial c}\frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{M^{\overline{r+s}}} &= \frac{b^{\overline{r-p}}c^{\overline{s-q}}c_{\overline{s-q}}d^{\overline{p+q}}M^{\overline{r+s}} - b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}M^{\overline{r+s}}M_{\overline{r+s}}}{(M^{\overline{r+s}})^2} \\
&= \frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{M^{\overline{r+s}}}(c_{\overline{s-q}} - M_{\overline{r+s}}) \\
\frac{\partial}{\partial b}\frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{M^{\overline{r+s}}} &= \frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}d_{\overline{p+q}}M^{\overline{r+s}} - b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}M^{\overline{r+s}}M_{\overline{r+s}}}{(M^{\overline{r+s}})^2} \\
&= \frac{b^{\overline{r-p}}c^{\overline{s-q}}d^{\overline{p+q}}}{M^{\overline{r+s}}}(d_{\overline{p+q}} - M_{\overline{r+s}}).
\end{aligned}$$

□

In the following result, we derive an approximate analytical solution for the value of m that (approximately) matches the first four sample moments.

Result D.4. *An initial approximate solution for the parameter vector $m = [a, b, c, d]$ to match the sample moments $\widehat{R}_{10}, \widehat{R}_{01}, \widehat{R}_{20}$ and $\widehat{R}_{02}$ is given by*

$$M = \frac{1}{2}\left(\frac{\widehat{R}_{10} - \widehat{R}_{20}}{\widehat{R}_{20} - \widehat{R}_{10}^2} + \frac{\widehat{R}_{01} - \widehat{R}_{02}}{\widehat{R}_{02} - \widehat{R}_{01}^2}\right) \tag{D.20}$$

$$a = M(1 - \widehat{R}_{10})(1 - \widehat{R}_{01}) \tag{D.21}$$

$$b = M\widehat{R}_{10}(1 - \widehat{R}_{01}) \tag{D.22}$$

$$c = M(1 - \widehat{R}_{10})\widehat{R}_{01} \tag{D.23}$$

$$d = M\widehat{R}_{10}\widehat{R}_{01}. \tag{D.24}$$

Proof. Equating the first four moments with the first four sample moments (given in Result D.2) gives

$$\widehat{R}_{10} = \frac{b+d}{M} \qquad \widehat{R}_{01} = \frac{c+d}{M}$$
$$\widehat{R}_{20} = \frac{(b+d)(b+d+1)}{M(M+1)} \qquad \widehat{R}_{02} = \frac{(c+d)(c+d+1)}{M(M+1)}.$$

Substituting the top twp equations into the bottom ones gives

$$\widehat{R}_{20} = \widehat{R}_{10}\frac{b+d+1}{M+1} \qquad \widehat{R}_{02} = \widehat{R}_{01}\frac{c+d+1}{M+1}.$$

We can derive two different constraints on M,

$$M_0 = \frac{\widehat{R}_{10} - \widehat{R}_{20}}{\widehat{R}_{20} - \widehat{R}_{10}^2} \quad \text{and} \quad M_1 = \frac{\widehat{R}_{01} - \widehat{R}_{02}}{\widehat{R}_{02} - \widehat{R}_{01}^2}.$$

This indicates that, in general, the system does not have a solution unless M_0 and M_1 are identical. In fact, if we denote the marginal beta distributions of $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ respectively as Beta(b_0+d_0, a_0+c_0) and Beta(c_1+d_1, a_1+b_1), we have a unique value for m when $a_0+b_0+c_0+d_0 = a_1+b_1+c_1+d_1$. The sum of the parameters of the beta distribution is sometimes called the *scale parameter*. To obtain a solution for m, we make two assumptions:

(i) The values of M_0 and M_1 are equal, which is equivalent to say that the sum of the parameters of the beta marginal distributions (i.e., the scale parameters) of $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ are equal.

(ii) The potential outcomes $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are independent, noted $\boldsymbol{y}_0 \perp \boldsymbol{y}_1$. This is a similar but stronger assumption than that used in Section 6.3.

Following the first assumption, we set

$$M = \frac{M_0 + M_1}{2}.$$

From the second assumption and the moments of the Dirichlet and bivariate beta distributions, we have

$$\begin{aligned} P(\boldsymbol{y}_0 = 1, \boldsymbol{y}_1 = 1) &= P(\boldsymbol{y}_0 = 1)P(\boldsymbol{y}_1 = 1) \\ \frac{d}{M} &= \left(\frac{b+d}{M}\right)\left(\frac{c+d}{M}\right) \\ d &= M\widehat{R}_{10}\widehat{R}_{01}, \end{aligned}$$

and similarly for a, b and c. We do not expect the two above assumptions to hold in the general case; instead, the solution derived from these assumptions is used as an initial guess for the optimization procedure in Section 6.4. The most important benefit of this initial solution is having a reasonable initial value for M, which helps to reduce the number of iterations given the large size of the space of possible values for m, viz. $\mathbb{R}^4_{>0}$. □

D.2 Generalized bivariate beta distribution

The generalized bivariate beta (GBB), used in Section 6.4.4, is similar to the bivariate beta described in Appendix D.1, with the exception that the random vector $\boldsymbol{\mu}$ is sampled from a generalized Dirichlet distribution, which is more flexible than the usual Dirichlet distribution (Connor and Mosimann, 1969). Let $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$ be three independent random variables with distributions Beta(a_1, b_1), Beta(a_2, b_2), and Beta(a_3, b_3). Then, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_4]$ is defined as

$$\boldsymbol{\mu}_i = \boldsymbol{z}_i \prod_{j=1}^{i-1}(1 - \boldsymbol{z}_j) \quad \text{for } i = 1, \dots, 4$$

with $\boldsymbol{z}_4 = 1$. We say that $\boldsymbol{\mu}$ follows a generalized Dirichlet distribution, noted $\boldsymbol{\mu} \sim$ GD$(a_1, a_2, a_3, b_1, b_2, b_3)$. More explicitly, we have

$$\boldsymbol{\alpha} = \boldsymbol{\mu}_1 = \boldsymbol{z}_1 \tag{D.25}$$

$$\boldsymbol{\beta} = \boldsymbol{\mu}_2 = \boldsymbol{z}_2(1 - \boldsymbol{z}_1) \tag{D.26}$$

$$\boldsymbol{\gamma} = \boldsymbol{\mu}_3 = \boldsymbol{z}_3(1 - \boldsymbol{z}_1)(1 - \boldsymbol{z}_2) \tag{D.27}$$

$$\boldsymbol{\delta} = \boldsymbol{\mu}_4 = (1 - \boldsymbol{z}_1)(1 - \boldsymbol{z}_2)(1 - \boldsymbol{z}_3). \tag{D.28}$$

The corresponding bivariate beta distribution, noted $(\boldsymbol{S}_0, \boldsymbol{S}_1) \sim \mathrm{GBB}(a_1, \dots, b_3)$, is sampled as follows.

$$\boldsymbol{\mu} \sim \mathrm{GD}(a_1, a_2, a_3, b_1, b_2, b_3) \tag{D.29}$$
$$\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta} \tag{D.30}$$
$$\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta}. \tag{D.31}$$

We now derive the properties of the generalized Dirichlet distribution and of the generalized bivariate beta distribution used in Section 6.4.4.

Result D.5. *The probability density function of* $\boldsymbol{\mu} \sim \mathrm{GD}(a_1, \dots, b_3)$ *is*

$$f_{\boldsymbol{\mu}}(\mu) = \frac{1}{\prod_{j=1}^{3} \mathrm{B}(a_j, b_j)} \alpha^{a_1-1} \beta^{a_2-1} \gamma^{a_3-1} \delta^{b_3-1} (\beta+\gamma+\delta)^{b_1-a_2-b_2} (\gamma+\delta)^{b_2-a_3-b_3}. \tag{D.32}$$

The probability density function of $(\boldsymbol{S}_0, \boldsymbol{S}_1) \sim \mathrm{GBB}(a_1, \dots, b_3)$ *is*

$$f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(S_0, S_1) = \int_{\Lambda(S_0,S_1)} f_{\boldsymbol{\mu}}(\mu)\, \mathrm{d}\mu \tag{D.33}$$

$$= \frac{1}{\prod_{j=1}^{3} \mathrm{B}(a_j, b_j)} \int_{\max\{0, S_0+S_1-1\}}^{\min\{S_0,S_1\}} (1 - S_0 - S_1 + \delta)^{a_1-1} (S_0 - \delta)^{a_2-1} (S_1 - \delta)^{a_3-1} \delta^{b_3-1} (S_0 + S_1 - \delta)^{b_1-a_2-b_2} S_1^{b_2-a_3-b_3}\, \mathrm{d}\delta. \tag{D.34}$$

This result is assumed without an explicit proof in (Connor and Mosimann, 1969). We believe that the following proof provides useful insights into the behavior of the generalized Dirichlet distribution.

Proof. By definition of the pdf, the probability that $\boldsymbol{\mu}$ belongs to some set S is

$$P(\boldsymbol{\mu} \in S) = \int_S f_{\boldsymbol{\mu}}(\mu)\, \mathrm{d}\mu.$$

Let ϕ be the function that maps the values of $\boldsymbol{\mu}$ to $\boldsymbol{z}$, for $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3]$. Using the formula for integration by substitution, we find

$$P(\boldsymbol{\mu} \in S) = P(\boldsymbol{z} \in \phi(S)) = \int_{\phi(S)} f_{\boldsymbol{z}}(z)\, \mathrm{d}z = \int_S f_{\boldsymbol{z}}(\phi(\mu)) |\det J_\phi(\mu)|\, \mathrm{d}\mu$$

where $|\det(J_\phi(\mu))|$ is the absolute value of the determinant of the Jacobian of ϕ. Since this is true for any set S, we can conclude that

$$f_{\boldsymbol{\mu}}(\mu) = f_{\boldsymbol{z}}(\phi(\mu)) |\det(J_\phi(\mu))|.$$

We need to compute the determinant of the Jacobian, $\det(J_\phi(\mu))$. Note that since we have three independent beta random variables $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3)$ on one side and a vector of four random variables summing up to one on the other side $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta})$, it is convenient to forget about $\boldsymbol{\alpha}$, and replace it by $1 - \boldsymbol{\beta} - \boldsymbol{\gamma} - \boldsymbol{\delta}$. This ensures that the Jacobian matrix is square, and hence its determinant well-defined. First, Let us express $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$ in terms of $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$. From Eqs. (D.25) to (D.28), we find

$$\boldsymbol{z}_1 = 1 - \boldsymbol{\beta} - \boldsymbol{\gamma} - \boldsymbol{\delta} \qquad 1 - \boldsymbol{z}_1 = \boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\delta} \tag{D.35}$$
$$\boldsymbol{z}_2 = \frac{\boldsymbol{\beta}}{\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\delta}} \qquad 1 - \boldsymbol{z}_2 = \frac{\boldsymbol{\gamma} + \boldsymbol{\delta}}{\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\delta}} \tag{D.36}$$
$$\boldsymbol{z}_3 = \frac{\boldsymbol{\gamma}}{\boldsymbol{\gamma} + \boldsymbol{\delta}} \qquad 1 - \boldsymbol{z}_3 = \frac{\boldsymbol{\delta}}{\boldsymbol{\gamma} + \boldsymbol{\delta}}. \tag{D.37}$$

The second column of equations will be useful later to simplify the expression of the pdf of $\boldsymbol{\mu}$. Now, let us compute the Jacobian matrix:

$$\begin{aligned}
&\frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{\beta}} = -1 && \frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{\gamma}} = -1 && \frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{\delta}} = -1 \\
&\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{\beta}} = \frac{\boldsymbol{\gamma}+\boldsymbol{\delta}}{(\boldsymbol{\beta}+\boldsymbol{\gamma}+\boldsymbol{\delta})^2} && \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{\gamma}} = \frac{-\boldsymbol{\beta}}{(\boldsymbol{\beta}+\boldsymbol{\gamma}+\boldsymbol{\delta})^2} && \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{\delta}} = \frac{-\boldsymbol{\beta}}{(\boldsymbol{\beta}+\boldsymbol{\gamma}+\boldsymbol{\delta})^2} \\
&\frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{\beta}} = 0 && \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{\gamma}} = \frac{\boldsymbol{\delta}}{(\boldsymbol{\gamma}+\boldsymbol{\delta})^2} && \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{\delta}} = \frac{-\boldsymbol{\gamma}}{(\boldsymbol{\gamma}+\boldsymbol{\delta})^2}
\end{aligned}$$

We can compute the determinant of this matrix by going over the first column (the partial derivatives with respect to $\boldsymbol{\beta}$), and using the cofactor rule:

$$\begin{aligned}
\det(J_\phi(\mu)) &= -\left(\frac{\beta\gamma+\beta\delta}{(\beta+\gamma+\delta)^2(\gamma+\delta)^2}\right) - \frac{\gamma+\delta}{(\beta+\gamma+\delta)^2}\left(\frac{\gamma+\delta}{(\gamma+\delta)^2}\right) \\
&= -\frac{\beta+\gamma+\delta}{(\beta+\gamma+\delta)^2(\gamma+\delta)} = -\frac{1}{(\beta+\gamma+\delta)(\gamma+\delta)}.
\end{aligned}$$

Hence the pdf of $\boldsymbol{\mu}$ is

$$\begin{aligned}
f_{\boldsymbol{\mu}}(\mu) &= f_{\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_3}\left(\alpha, \frac{\beta}{\beta+\gamma+\delta}, \frac{\gamma}{\gamma+\delta}\right)|\det(J_\phi(\mu))| \\
&= \frac{1}{\prod_{j=1}^3 \mathrm{B}(a_j,b_j)}\alpha^{a_1-1}(\beta+\gamma+\delta)^{b_1-1}\left(\frac{\beta}{\beta+\gamma+\delta}\right)^{a_2-1}\left(\frac{\gamma+\delta}{\beta+\gamma+\delta}\right)^{b_2-1}\left(\frac{\gamma}{\gamma+\delta}\right)^{a_3-1} \\
&\qquad\qquad \left(\frac{\delta}{\gamma+\delta}\right)^{b_3-1}\frac{1}{(\beta+\gamma+\delta)(\gamma+\delta)} \\
&= \frac{1}{\prod_{j=1}^3 \mathrm{B}(a_j,b_j)}\alpha^{a_1-1}\beta^{a_2-1}\gamma^{a_3-1}\delta^{b_3-1}(\beta+\gamma+\delta)^{b_1-a_2-b_2}(\gamma+\delta)^{b_2-a_3-b_3}.
\end{aligned}$$

The pdf of $\boldsymbol{S}_0, \boldsymbol{S}_1$ is derived similarly. Let ζ be the function that maps the values of $\boldsymbol{S}_0, \boldsymbol{S}_1$ to $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$ (again, $\boldsymbol{\alpha}$ is implicitly computed as $1 - \boldsymbol{\beta} - \boldsymbol{\gamma} - \boldsymbol{\delta}$). Since $\boldsymbol{\mu}$ is not fully determined by $\boldsymbol{S}_0, \boldsymbol{S}_1$, we must also include one of the variables in the input of ζ to fully determine its output. Including $\boldsymbol{\delta}$ leads to the simplest expression for ζ:

$$\zeta(\boldsymbol{S}_0, \boldsymbol{S}_1, \boldsymbol{\delta}) = \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\\ \boldsymbol{\delta}\end{bmatrix} = \begin{bmatrix}\boldsymbol{S}_0 - \boldsymbol{\delta}\\ \boldsymbol{S}_1 - \boldsymbol{\delta}\\ \boldsymbol{\delta}\end{bmatrix}.$$

The Jacobian matrix of ζ is

$$\begin{aligned}
&\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{S}_0} = 1 && \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{S}_1} = 0 && \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\delta}} = -1 \\
&\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{S}_0} = 0 && \frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{S}_1} = 1 && \frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{\delta}} = -1 \\
&\frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{S}_0} = 0 && \frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{S}_1} = 0 && \frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{\delta}} = 1
\end{aligned}$$

which has a determinant of 1. Therefore, we can write

$$\begin{aligned} f_{\boldsymbol{S}_0,\boldsymbol{S}_1,\boldsymbol{\delta}}(S_0,S_1,\delta) &= f_{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta}}(\zeta(S_0,S_1,\delta))|\det J_\zeta(S_0,S_1,\delta)| \\ &= f_{\boldsymbol{\mu}}(1-S_0-S_1+\delta, S_0-\delta, S_1-\delta, \delta) \\ &= \frac{1}{\prod_{j=1}^{3}\mathrm{B}(a_j,b_j)}(1-S_0-S_1+\delta)^{a_1-1}(S_0-\delta)^{a_2-1}(S_1-\delta)^{a_3-1}\delta^{b_3-1} \\ &\qquad (S_0+S_1-\delta)^{b_1-a_2-b_2}S_1^{b_2-a_3-b_3}. \end{aligned}$$

From this, the density of $\boldsymbol{S}_0, \boldsymbol{S}_1$ can be found by integrating over δ. Recall that the Fréchet bounds put constraints on the values of δ given S_0, S_1:

$$\max\{0, S_0+S_1-1\} \leq \delta \leq \min\{S_0, S_1\}. \qquad \text{(from Eq. 6.18)}$$

Therefore, the integration can be reduced to these bounds as follows.

$$\begin{aligned} f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(S_0,S_1) = \frac{1}{\prod_{j=1}^{3}\mathrm{B}(a_j,b_j)} \int_{\max\{0,S_0+S_1-1\}}^{\min\{S_0,S_1\}} &(1-S_0-S_1+\delta)^{a_1-1}(S_0-\delta)^{a_2-1}(S_1-\delta)^{a_3-1} \\ &\delta^{b_3-1}(S_0+S_1-\delta)^{b_1-a_2-b_2}S_1^{b_2-a_3-b_3}\,\mathrm{d}\delta. \end{aligned}$$

□

Result D.6. *The raw moments of $\boldsymbol{S}_0, \boldsymbol{S}_1 \sim \mathrm{GBB}(a_1, \dots, b_3)$ for integers $r, s > 0$ are*

$$R_{rs}(\boldsymbol{S}_0,\boldsymbol{S}_1) = \sum_{p=0}^{r}\sum_{q=0}^{s}\binom{r}{p}\binom{s}{q}\frac{b_1^{\overline{r+s}}}{(a_1+b_1)^{\overline{r+s}}}\frac{a_2^{\overline{r-p}}b_2^{\overline{s+p}}}{(a_2+b_2)^{\overline{r+s}}}\frac{a_3^{\overline{s-q}}b_3^{\overline{p+q}}}{(a_3+b_3)^{\overline{s+p}}}. \tag{D.38}$$

Proof. Using a similar reasoning as in the proof of Result D.2, we have

$$\begin{aligned} R_{rs}(\boldsymbol{S}_0,\boldsymbol{S}_1) &= \mathbb{E}[\boldsymbol{S}_0^r\boldsymbol{S}_1^s] = \mathbb{E}[(\boldsymbol{\beta}+\boldsymbol{\delta})^r(\boldsymbol{\gamma}+\boldsymbol{\delta})^s] \\ &= \mathbb{E}\left[\left(\sum_{p=0}^{r}\binom{r}{p}\beta^{r-p}\delta^p\right)\left(\sum_{q=0}^{s}\binom{s}{q}\gamma^{s-q}\delta^q\right)\right] \\ &= \sum_{p=0}^{r}\sum_{q=0}^{s}\binom{r}{p}\binom{s}{q}\mathbb{E}[\boldsymbol{\beta}^{r-p}\boldsymbol{\gamma}^{s-q}\boldsymbol{\delta}^{p+q}] \\ &= \sum_{p=0}^{r}\sum_{q=0}^{s}\binom{r}{p}\binom{s}{q}\mathbb{E}[(\boldsymbol{z}_2(1-\boldsymbol{z}_1))^{r-p}(\boldsymbol{z}_3(1-\boldsymbol{z}_1)(1-\boldsymbol{z}_2))^{s-q} \\ &\qquad ((1-\boldsymbol{z}_1)(1-\boldsymbol{z}_2)(1-\boldsymbol{z}_3))^{p+q}] \\ &= \sum_{p=0}^{r}\sum_{q=0}^{s}\binom{r}{p}\binom{s}{q}\mathbb{E}[(1-\boldsymbol{z}_1)^{r+s}\boldsymbol{z}_2^{r-p}(1-\boldsymbol{z}_2)^{s+p}\boldsymbol{z}_3^{s-q}(1-\boldsymbol{z}_3)^{p+q}] \\ &= \sum_{p=0}^{r}\sum_{q=0}^{s}\binom{r}{p}\binom{s}{q}\frac{b_1^{\overline{r+s}}}{(a_1+b_1)^{\overline{r+s}}}\frac{a_2^{\overline{r-p}}b_2^{\overline{s+p}}}{(a_2+b_2)^{\overline{r+s}}}\frac{a_3^{\overline{s-q}}b_3^{\overline{p+q}}}{(a_3+b_3)^{\overline{s+p}}}. \end{aligned}$$

In the last equality, we factorized the expect value using the independence of $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$, and we used the moments of the beta distribution for integer exponents (see Eq. (2.37) in Section 2.1.5):

$$\mathbb{E}[\boldsymbol{x}^r(1-\boldsymbol{x})^s] = \frac{a^{\overline{r}}b^{\overline{s}}}{(a+b)^{\overline{r+s}}} \quad \text{when } \boldsymbol{x} \sim \mathrm{Beta}(a,b).$$

□

Result D.7. *The partial derivative of the raw moments of* $(\boldsymbol{S}_0, \boldsymbol{S}_1) \sim \mathrm{GBB}(a_1, \dots, b_3)$ *for integers* $r, s > 0$ *with respect to the distribution parameters is*

$$\frac{\partial \mathbb{E}[\boldsymbol{S}_0^r \boldsymbol{S}_1^s]}{\partial a_j} = \sum_{p=0}^{r} \sum_{q=0}^{s} \binom{r}{p} \binom{s}{q} \left(\frac{\partial}{\partial a_j} \frac{a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}}}{(a_j + b_j)^{\overline{a_j^* + b_j^*}}} \right) \prod_{k \neq j} \frac{a_k^{\overline{a_k^*}} b_k^{\overline{b_k^*}}}{(a_k + b_k)^{\overline{a_k^* + b_k^*}}} \tag{D.39}$$

$$\frac{\partial \mathbb{E}[\boldsymbol{S}_0^r \boldsymbol{S}_1^s]}{\partial b_j} = \sum_{p=0}^{r} \sum_{q=0}^{s} \binom{r}{p} \binom{s}{q} \left(\frac{\partial}{\partial b_j} \frac{a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}}}{(a_j + b_j)^{\overline{a_j^* + b_j^*}}} \right) \prod_{k \neq j} \frac{a_k^{\overline{a_k^*}} b_k^{\overline{b_k^*}}}{(a_k + b_k)^{\overline{a_k^* + b_k^*}}} \tag{D.40}$$

where

$$a_1^* = 0 \qquad a_2^* = r - p \qquad a_3^* = s - q \tag{D.41}$$

$$b_1^* = r + s \qquad b_2^* = s + p \qquad b_3^* = p + q. \tag{D.42}$$

and the inner partial derivatives evaluate to

$$\frac{\partial}{\partial a_j} \frac{a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}}}{(a_j + b_j)^{\overline{a_j^* + b_j^*}}} = \frac{a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}}}{(a_j + b_j)^{\overline{a_j^* + b_j^*}}} ((a_j)_{\overline{a_j^*}} - (a_j + b_j)_{\overline{a_j^* + b_j^*}}) \tag{D.43}$$

$$\frac{\partial}{\partial b_j} \frac{a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}}}{(a_j + b_j)^{\overline{a_j^* + b_j^*}}} = \frac{a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}}}{(a_j + b_j)^{\overline{a_j^* + b_j^*}}} ((b_j)_{\overline{a_j^*}} - (a_j + b_j)_{\overline{a_j^* + b_j^*}}). \tag{D.44}$$

Proof. Equations (D.39) and (D.40) are just an application of the linearity of the partial derivative operator, starting from Eq. (D.38). Evaluating the inner partial derivative for a_j gives

$$\begin{aligned} \frac{\partial}{\partial a_j} \frac{a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}}}{(a_j + b_j)^{\overline{a_j^* + b_j^*}}} &= \frac{a_j^{\overline{a_j^*}} (a_j)_{\overline{a_j^*}} b_j^{\overline{b_j^*}} (a_j + b_j)^{\overline{a_j^* + b_j^*}} - a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}} (a_j + b_j)^{\overline{a_j^* + b_j^*}} (a_j + b_j)_{\overline{a_j^* + b_j^*}}}{\left((a_j + b_j)^{\overline{a_j^* + b_j^*}} \right)^2} \\ &= \frac{a_j^{\overline{a_j^*}} b_j^{\overline{b_j^*}}}{(a_j + b_j)^{\overline{a_j^* + b_j^*}}} ((a_j)_{\overline{a_j^*}} - (a_j + b_j)_{\overline{a_j^* + b_j^*}}). \end{aligned}$$

The derivation is similar for b_j. □

Result D.8. *Let* $f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1, a_2, a_3 \\ b_1, b_2, b_3 \end{smallmatrix}; S_0, S_1\right)$ *be the pdf of* $\boldsymbol{S}_0, \boldsymbol{S}_1$ *following a generalized bivariate beta distribution with parameter vector* $[a_1, a_2, a_3, b_1, b_2, b_3]$.[1] *The expected value of the counterfactuals* $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$ *given* $\boldsymbol{S}_0 = S_0, \boldsymbol{S}_1 = S_1$ *can be expressed as*

$$\mathbb{E}[\boldsymbol{\alpha} \mid S_0, S_1] = \frac{f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1+1, a_2, a_3 \\ b_1, b_2, b_3 \end{smallmatrix}; S_0, S_1\right)}{f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1, a_2, a_3 \\ b_1, b_2, b_3 \end{smallmatrix}; S_0, S_1\right)} \qquad \mathbb{E}[\boldsymbol{\beta} \mid S_0, S_1] = \frac{f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1, a_2+1, a_3 \\ b_1+1, b_2, b_3 \end{smallmatrix}; S_0, S_1\right)}{f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1, a_2, a_3 \\ b_1, b_2, b_3 \end{smallmatrix}; S_0, S_1\right)} \tag{D.45}$$

$$\mathbb{E}[\boldsymbol{\gamma} \mid S_0, S_1] = \frac{f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1, a_2, a_3+1 \\ b_1+1, b_2+1, b_3 \end{smallmatrix}; S_0, S_1\right)}{f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1, a_2, a_3 \\ b_1, b_2, b_3 \end{smallmatrix}; S_0, S_1\right)} \qquad \mathbb{E}[\boldsymbol{\delta} \mid S_0, S_1] = \frac{f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1, a_2, a_3 \\ b_1+1, b_2+1, b_3+1 \end{smallmatrix}; S_0, S_1\right)}{f_{\boldsymbol{S}_0, \boldsymbol{S}_1}\left(\begin{smallmatrix} a_1, a_2, a_3 \\ b_1, b_2, b_3 \end{smallmatrix}; S_0, S_1\right)}. \tag{D.46}$$

[1] This notation is inspired by that of the generalized hypergeometric function.

Proof. We will show the proof for $\boldsymbol{\beta}$. The three other counterfactuals follow a similar pattern. Let $f_{\boldsymbol{\beta},\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\beta, S_0, S_1)$ be the joint pdf of $\boldsymbol{\beta}, \boldsymbol{S}_0, \boldsymbol{S}_1$ with parameter vector $[a_1, a_2, a_3, b_1, b_2, b_3]$. By the definition of the conditional expectation (Definition 2.2), we have

$$\mathbb{E}[\boldsymbol{\beta} \mid S_0, S_1] = \frac{\int_0^1 \beta f_{\boldsymbol{\beta},\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\beta, S_0, S_1)\,\mathrm{d}\beta}{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(S_0, S_1)}. \tag{D.47}$$

We can marginalize $f_{\boldsymbol{\beta},\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\beta, S_0, S_1)$ by integrating over the others components of μ (that is, α, γ and δ) as follows:

$$\int_0^1 \beta f_{\boldsymbol{\beta},\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\beta, S_0, S_1)\,\mathrm{d}\beta = \int_\Lambda \beta f_{\boldsymbol{\mu},\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\mu, S_0, S_1)\,\mathrm{d}\mu. \tag{D.48}$$

On the right-hand side, the integration domain is the 4-dimensional unit simplex $\Lambda = \{\mu \mid \mu_j \geq 0 \text{ and } \sum_j \mu_j = 1\}$. Since $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ are defined as $\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta}$ and $\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta}$, the pdf $f_{\boldsymbol{\mu},\boldsymbol{S}_0,\boldsymbol{S}_1}$ is equal to $f_{\boldsymbol{\mu}}$ whenever these two equalities are respected, and is zero everywhere else. The set of values of $\boldsymbol{\mu}$ that satisfies these two equalities is noted $\Lambda(S_0, S_1)$ (see Eq. 6.76). Using this, the numerator of Eq. (D.47) can be developed as

$$\int_\Lambda \beta f_{\boldsymbol{\mu},\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\mu, S_0, S_1)\,\mathrm{d}\mu = \int_{\Lambda(S_0,S_1)} \beta f_{\boldsymbol{\mu}}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\mu)\,\mathrm{d}\mu. \tag{D.49}$$

This looks similar to the pdf of $\boldsymbol{S}_0, \boldsymbol{S}_1$ in Eq. (D.33) of Result D.5. In fact, we can adjust the distribution parameters so that the expression above exactly matches the pdf of $\boldsymbol{S}_0, \boldsymbol{S}_1$ with the adjusted parameters. If we increase a_2 and b_1 by one, we obtain

$$\begin{aligned}
f_{\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2+1,a_3}{b_1+1,b_2,b_3}(S_0, S_1) &= \int_{\Lambda(S_0,S_1)} f_{\boldsymbol{\mu}}\binom{a_1,a_2+1,a_3}{b_1+1,b_2,b_3}(\mu)\,\mathrm{d}\mu \\
&= \frac{1}{\prod_{j=1}^3 \mathrm{B}(a_j, b_j)} \int_{\Lambda(S_0,S_1)} \alpha^{a_1-1}\beta^{a_2}\gamma^{a_3-1}\delta^{b_3-1}(\beta+\gamma+\delta)^{b_1+1-a_2-1-b_2} \\
&\qquad\qquad (\gamma+\delta)^{b_2-a_3-b_3}\,\mathrm{d}\mu \\
&= \frac{1}{\prod_{j=1}^3 \mathrm{B}(a_j, b_j)} \int_{\Lambda(S_0,S_1)} \beta\alpha^{a_1-1}\beta^{a_2-1}\gamma^{a_3-1}\delta^{b_3-1}(\beta+\gamma+\delta)^{b_1-a_2-b_2}(\gamma+\delta)^{b_2-a_3-b_3}\,\mathrm{d}\mu \\
&= \int_{\Lambda(S_0,S_1)} \beta f_{\boldsymbol{\mu}}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\mu)\,\mathrm{d}\mu.
\end{aligned} \tag{D.50}$$

Chaining Eq. (D.47), Eq. (D.48), Eq. (D.49), and Eq. (D.50), we obtain

$$\mathbb{E}[\boldsymbol{\beta} \mid S_0, S_1] = \frac{\int_{\Lambda(S_0,S_1)} \beta f_{\boldsymbol{\mu},\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(\mu, S_0, S_1)\,\mathrm{d}\mu}{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(S_0, S_1)} = \frac{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2+1,a_3}{b_1+1,b_2,b_3}(S_0, S_1)}{f_{\boldsymbol{S}_0,\boldsymbol{S}_1}\binom{a_1,a_2,a_3}{b_1,b_2,b_3}(S_0, S_1)}.$$

□

D.3 Noisy bivariate beta distribution

The noisy bivariate beta distribution, used in Section 6.4.5, is sampled as follows:

$$\boldsymbol{\mu} \sim \mathrm{Dir}(m) \tag{D.51}$$
$$\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta} \tag{D.52}$$
$$\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta} \tag{D.53}$$
$$\hat{\boldsymbol{S}}_t \sim \mathrm{Beta}(\lambda_t \boldsymbol{S}_t, \lambda_t(1-\boldsymbol{S}_t)) \quad \text{for } t = 0, 1. \tag{D.54}$$

We note $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NBB}(m, \lambda_0, \lambda_1)$. As for the generalized bivariate beta distribution, let us derive the properties of this new distribution that are needed in Section 6.4.5: its probability density function, its moments, the partial derivatives of its moments, and the expected value of the counterfactuals.

Result D.9. *Let $f_{\boldsymbol{x}}(\substack{a\\b}; x)$ be the pdf of a random variable $\boldsymbol{x}$ following a beta distribution* $\mathrm{B}(a, b)$*, given in Eq. (2.32). The probability density function of $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NBB}(m, \lambda_0, \lambda_1)$ is*

$$f(\hat{S}_0, \hat{S}_1) = \int_0^1 \int_0^1 f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(S_0, S_1) f_{\hat{\boldsymbol{S}}_0|S_0}(\substack{\lambda_0 S_0\\ \lambda_0(1-S_0)}; \hat{S}_0) f_{\hat{\boldsymbol{S}}_1|S_1}(\substack{\lambda_1 S_1\\ \lambda_1(1-S_1)}; \hat{S}_1)\, \mathrm{d}S_0\, \mathrm{d}S_1 \tag{D.55}$$

where $f_{\boldsymbol{S}_0,\boldsymbol{S}_1}$ is the pdf of the bivariate beta distribution given in Eq. (D.4).

Proof. Since $\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1$ are both sampled following beta distributions with parameters depending on $\boldsymbol{S}_0, \boldsymbol{S}_1$, their joint probability density function can be found by marginalizing over $\boldsymbol{S}_0, \boldsymbol{S}_1$:

$$\begin{aligned} f_{\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1}(\hat{S}_0, \hat{S}_1) &= \int_0^1 \int_0^1 f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(S_0, S_1) f_{\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1|S_0,S_1}(\hat{S}_0, \hat{S}_1)\, \mathrm{d}S_1\, \mathrm{d}S_0 \\ &= \int_0^1 \int_0^1 f_{\boldsymbol{S}_0,\boldsymbol{S}_1}(S_0, S_1) f_{\hat{\boldsymbol{S}}_0|S_0}(\substack{\lambda_0 S_0\\ \lambda_0(1-S_0)}; \hat{S}_0) f_{\hat{\boldsymbol{S}}_1|S_1}(\substack{\lambda_1 S_1\\ \lambda_1(1-S_1)}; \hat{S}_1)\, \mathrm{d}S_0\, \mathrm{d}S_1. \end{aligned}$$

□

Result D.10. *The raw moments of $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NBB}(m, \lambda_0, \lambda_1)$ for integers $r, s > 0$ are*

$$R_{rs}(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) = \sum_{p=0}^{r} \sum_{q=0}^{s} \begin{bmatrix} r \\ p \end{bmatrix} \begin{bmatrix} s \\ q \end{bmatrix} \frac{\lambda_0^p \lambda_1^q}{\lambda_0^{\bar{r}} \lambda_1^{\bar{s}}} R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1) \tag{D.56}$$

where $R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1)$ is the raw moment of the bivariate beta distribution given in Result D.2, and $\left[\substack{a\\b}\right]$ is the unsigned Stirling number of the first kind (Weisstein, 2023), defined by recurrence for integers $a, b \geq 0$ as

$$\begin{bmatrix} a+1 \\ b \end{bmatrix} = a \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} a \\ b-1 \end{bmatrix} \quad \textit{with} \quad \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 1 \quad \textit{and} \quad \begin{bmatrix} a \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} = 0. \tag{D.57}$$

Proof. By the "tower" property of the expectation operator (Wolpert, 2010), and the raw moments of the beta distribution (Eq. 2.37), we can write

$$R_{rs}(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) = \mathbb{E}\left[\hat{\boldsymbol{S}}_0^r \hat{\boldsymbol{S}}_1^s\right] = \mathbb{E}_{\boldsymbol{S}_0,\boldsymbol{S}_1}[\mathbb{E}[\hat{\boldsymbol{S}}_0^r \hat{\boldsymbol{S}}_1^s \mid \boldsymbol{S}_0, \boldsymbol{S}_1]] = \mathbb{E}\left[\frac{(\boldsymbol{S}_0\lambda_0)^{\bar{r}}}{\lambda_0^{\bar{r}}} \frac{(\boldsymbol{S}_1\lambda_1)^{\bar{s}}}{\lambda_1^{\bar{s}}}\right].$$

The unsigned Stirling numbers of the first kind are equal the coefficients of polynomial expansion of the the rising factorial (Qi, 2013):

$$x^{\overline{n}} = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} x^k.$$

This leads to

$$\begin{aligned}
\mathbb{E}\left[\frac{(\boldsymbol{S}_0\lambda_0)^{\overline{r}}}{\lambda_0^{\overline{r}}}\frac{(\boldsymbol{S}_1\lambda_1)^{\overline{s}}}{\lambda_1^{\overline{s}}}\right] &= \mathbb{E}\left[\left(\sum_{p=0}^{r}\begin{bmatrix} r \\ p \end{bmatrix}\frac{\lambda_0^p}{\lambda_0^{\overline{r}}}\boldsymbol{S}_0^p\right)\left(\sum_{q=0}^{s}\begin{bmatrix} s \\ q \end{bmatrix}\frac{\lambda_1^q}{\lambda_1^{\overline{s}}}\boldsymbol{S}_1^q\right)\right] \\
&= \sum_{p=0}^{r}\sum_{q=0}^{s}\begin{bmatrix} r \\ p \end{bmatrix}\begin{bmatrix} s \\ q \end{bmatrix}\frac{\lambda_0^p\lambda_1^q}{\lambda_0^{\overline{r}}\lambda_1^{\overline{s}}}\mathbb{E}[\boldsymbol{S}_0^p\boldsymbol{S}_1^q] \\
&= \sum_{p=0}^{r}\sum_{q=0}^{s}\begin{bmatrix} r \\ p \end{bmatrix}\begin{bmatrix} s \\ q \end{bmatrix}\frac{\lambda_0^p\lambda_1^q}{\lambda_0^{\overline{r}}\lambda_1^{\overline{s}}}R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1).
\end{aligned}$$

□

Result D.11. *The partial derivative of the raw moments of* $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NBB}(m, \lambda_0, \lambda_1)$ *for integers* $r, s > 0$ *with respect to* m *is*

$$\frac{\partial R_{rs}(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1)}{\partial m_i} = \sum_{p=0}^{r}\sum_{q=0}^{s}\begin{bmatrix} r \\ p \end{bmatrix}\begin{bmatrix} s \\ q \end{bmatrix}\frac{\lambda_0^p\lambda_1^q}{\lambda_0^{\overline{r}}\lambda_1^{\overline{s}}}\frac{\partial R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1)}{\partial m_i} \tag{D.58}$$

where the partial derivative of $R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1)$ *is given in Result D.3.*

Proof. This is a simple application of the linearity of the partial derivative operator on Eq. (D.56). □

Result D.12. *The expected value of the counterfactuals* $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$ *given* $\hat{\boldsymbol{S}}_0 = \hat{S}_0, \hat{\boldsymbol{S}}_1 = \hat{S}_1$ *can be expressed as*

$$\mathbb{E}[\boldsymbol{\mu}_j \mid \hat{S}_0, \hat{S}_1] = \frac{1}{f_{\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1}(\hat{S}_0, \hat{S}_1)}\int_\Lambda \mu_j f_{\boldsymbol{\mu}}(\mu) f_{\hat{\boldsymbol{S}}_0}\left(\tfrac{\lambda_0(\beta+\delta)}{\lambda_0(\alpha+\gamma)}; \hat{S}_0\right) f_{\hat{\boldsymbol{S}}_1}\left(\tfrac{\lambda_1(\gamma+\delta)}{\lambda_1(\alpha+\beta)}; \hat{S}_1\right) \mathrm{d}\mu. \tag{D.59}$$

Proof. The proof is similar to that of Result D.8. First, we develop

$$\mathbb{E}[\boldsymbol{\mu}_j \mid \hat{S}_0, \hat{S}_1] = \frac{\int_0^1 \mu_j f_{\mu_j,\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1}(\mu_j\hat{S}_0, \hat{S}_1)\,\mathrm{d}\mu_j}{f(\hat{S}_0, \hat{S}_1)}.$$

We can marginalize $f_{\mu_j,\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1}(\mu_j\hat{S}_0, \hat{S}_1)$ by integrating over the others terms in $\boldsymbol{\mu}$, leading to

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_j \mid \hat{S}_0, \hat{S}_1] &= \frac{1}{f_{\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1}(\hat{S}_0, \hat{S}_1)}\int_\Lambda \mu_j f_{\boldsymbol{\mu},\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1}(\mu, \hat{S}_0, \hat{S}_1)\,\mathrm{d}\mu \\
&= \frac{1}{f_{\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1}(\hat{S}_0, \hat{S}_1)}\int_\Lambda \mu_j f_{\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1|\mu}(\hat{S}_0, \hat{S}_1) f_{\boldsymbol{\mu}}(\mu)\,\mathrm{d}\mu \\
&= \frac{1}{f_{\hat{\boldsymbol{S}}_0,\hat{\boldsymbol{S}}_1}(\hat{S}_0, \hat{S}_1)}\int_\Lambda \mu_j f_{\boldsymbol{\mu}}(\mu) f_{\hat{\boldsymbol{S}}_0}\left(\tfrac{\lambda_0(\beta+\delta)}{\lambda_0(\alpha+\gamma)}; \hat{S}_0\right) f_{\hat{\boldsymbol{S}}_1}\left(\tfrac{\lambda_1(\gamma+\delta)}{\lambda_1(\alpha+\beta)}; \hat{S}_1\right) \mathrm{d}\mu.
\end{aligned}$$

□

D.4 Noisy generalized bivariate beta distribution

The noisy generalized bivariate beta distribution, used in Section 6.4.6, is sampled as follows:

$$\boldsymbol{\mu} \sim \mathrm{GD}(a_1, \dots, b_3) \tag{D.60}$$
$$\boldsymbol{S}_0 = \boldsymbol{\beta} + \boldsymbol{\delta} \tag{D.61}$$
$$\boldsymbol{S}_1 = \boldsymbol{\gamma} + \boldsymbol{\delta} \tag{D.62}$$
$$\hat{\boldsymbol{S}}_t \sim \mathrm{Beta}(\lambda_t \boldsymbol{S}_t, \lambda_t(1 - \boldsymbol{S}_t)) \quad \text{for } t = 0, 1. \tag{D.63}$$

We note $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NGBB}(a_1, \dots, b_3, \lambda_0, \lambda_1)$. None of the properties of the NBB distribution derived in Appendix D.3 depend on the exact distribution of $\boldsymbol{\mu}$. Therefore, Results D.9 to D.12 are easily adapted to the NGBB distribution by changing the references in the statement of the results. We state these properties for the sake of completeness, but we do not give the proofs, since they are identical to those in Appendix D.3.

Result D.13. *Let $f_{\boldsymbol{x}}(\binom{a}{b}; x)$ be the pdf of a random variable $\boldsymbol{x}$ following a beta distribution $\mathrm{B}(a, b)$, given in Eq. (2.32). The probability density function of $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NBB}(m, \lambda_0, \lambda_1)$ is*

$$f(\hat{S}_0, \hat{S}_1) = \int_0^1 \int_0^1 f_{\boldsymbol{S}_0, \boldsymbol{S}_1}(S_0, S_1) f_{\hat{\boldsymbol{S}}_0 | S_0}\left(\tbinom{\lambda_0 S_0}{\lambda_0(1-S_0)}; \hat{S}_0\right) f_{\hat{\boldsymbol{S}}_1 | S_1}\left(\tbinom{\lambda_1 S_1}{\lambda_1(1-S_1)}; \hat{S}_1\right) \mathrm{d}S_0 \, \mathrm{d}S_1 \tag{D.64}$$

where $f_{\boldsymbol{S}_0, \boldsymbol{S}_1}$ is the pdf of the generalized bivariate beta distribution given in Eq. (D.33).

Proof. See Result D.9. □

Result D.14. *The raw moments of $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NBB}(m, \lambda_0, \lambda_1)$ for integers $r, s > 0$ are*

$$R_{rs}(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) = \sum_{p=0}^{r} \sum_{q=0}^{s} \begin{bmatrix} r \\ p \end{bmatrix} \begin{bmatrix} s \\ q \end{bmatrix} \frac{\lambda_0^p \lambda_1^q}{\lambda_0^{\bar{r}} \lambda_1^{\bar{s}}} R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1) \tag{D.65}$$

where $R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1)$ is the raw moment of the generalized bivariate beta distribution given in Result D.6, and $\left[{a \atop b}\right]$ is the unsigned Stirling number of the first kind (see Result D.10).

Proof. See Result D.10. □

Result D.15. *The partial derivative of the raw moments of $(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1) \sim \mathrm{NBB}(m, \lambda_0, \lambda_1)$ for integers $r, s > 0$ with respect to m is*

$$\frac{\partial R_{rs}(\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1)}{\partial m_i} = \sum_{p=0}^{r} \sum_{q=0}^{s} \begin{bmatrix} r \\ p \end{bmatrix} \begin{bmatrix} s \\ q \end{bmatrix} \frac{\lambda_0^p \lambda_1^q}{\lambda_0^{\bar{r}} \lambda_1^{\bar{s}}} \frac{\partial R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1)}{\partial m_i} \tag{D.66}$$

where the partial derivative of $R_{pq}(\boldsymbol{S}_0, \boldsymbol{S}_1)$ is given in Result D.7.

Proof. See Result D.11 □

Result D.16. *The expected value of the counterfactuals $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$ given $\hat{\boldsymbol{S}}_0 = \hat{S}_0, \hat{\boldsymbol{S}}_1 = \hat{S}_1$ can be expressed as*

$$\mathbb{E}[\boldsymbol{\mu}_j \mid \hat{S}_0, \hat{S}_1] = \frac{1}{f_{\hat{\boldsymbol{S}}_0, \hat{\boldsymbol{S}}_1}(\hat{S}_0, \hat{S}_1)} \int_\Lambda \mu_j f_{\boldsymbol{\mu}}(\mu) f_{\hat{\boldsymbol{S}}_0}\left(\tbinom{\lambda_0(\beta+\delta)}{\lambda_0(\alpha+\gamma)}; \hat{S}_0\right) f_{\hat{\boldsymbol{S}}_1}\left(\tbinom{\lambda_1(\gamma+\delta)}{\lambda_1(\alpha+\beta)}; \hat{S}_1\right) \mathrm{d}\mu. \tag{D.67}$$

Proof. See Result D.12. □

Bibliography

Alaa, Ahmed and Mihaela van der Schaar (July 2018). "Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design". In: *Proceedings of the 35th international conference on machine learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of machine learning research. PMLR, pp. 129–138. URL: https://proceedings.mlr.press/v80/alaa18a.html.

Ascarza, Eva (2018). "Retention futility: Targeting high-risk customers might be ineffective". In: *Journal of Marketing Research* 55.1. Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 80–98.

Aspect, Alain, Philippe Grangier, and Gérard Roger (1982). "Experimental realization of einstein-podolsky-rosen-bohm gedankenexperiment: A new violation of bell's inequalities". In: *Physical review letters* 49.2. Publisher: APS, p. 91.

Athey, Susan and Guido Imbens (July 2016). "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27. Publisher: National Academy of Sciences, pp. 7353–7360. ISSN: 0027-8424. DOI: 10.1073/PNAS.1510489113. URL: https://www.pnas.org/content/113/27/7353 (visited on 05/06/2020).

Balke, Alexander and Judea Pearl (1994). "Counterfactual probabilities: Computational methods, bounds and applications". In: *Uncertainty Proceedings 1994*. Elsevier, pp. 46–54.

Barbe, Philippe et al. (2007). *Probabilité (L3M1)*. EDP Sciences.

Bareinboim, Elias et al. (2020). "On Pearl's Hierarchy and the Foundations of Causal Inference". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. New York, NY, USA: Association for Computing Machinery, pp. 1–62.

Barp, Alessandro et al. (2019). "Minimum stein discrepancy estimators". In: *Advances in Neural Information Processing Systems* 32.

Basiri, Mohammad Ehsan et al. (2021). "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets". In: *Knowledge-Based Systems* 228. Publisher: Elsevier, p. 107242.

Batista, Gustavo EAPA, Ronaldo C Prati, and Maria Carolina Monard (2004). "A study of the behavior of several methods for balancing machine learning training data". In: *ACM SIGKDD explorations newsletter* 6.1. Publisher: ACM, pp. 20–29. DOI: 10.1145/1007730.1007735.

Bell, John S (1964). "On the einstein podolsky rosen paradox". In: *Physics Physique Fizika* 1.3. Publisher: APS, p. 195.

Belov, Dmitry I and Ronald D Armstrong (2011). "Distributions of the Kullback–Leibler divergence with applications". In: *British Journal of Mathematical and Statistical Psychology* 64.2. Publisher: Wiley Online Library, pp. 291–309.

Bontempi, Gianluca (2017). *Handbook on "Statistical foundations of machine learning"*. Université libre de Bruxelles.

Bontempi, Gianluca and Maxime Flauder (2015). "From dependency to causality: a machine learning approach". In: *The Journal of Machine Learning Research* 16.1. Publisher: JMLR. org, pp. 2437–2457.

Bontempi, Gianluca and Patrick E. Meyer (Aug. 2010). "Causal filter selection in microarray data". In: *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pp. 95–102.

Bose, Indranil and Xi Chen (2009). "Quantitative models for direct marketing: A review from systems perspective". In: *European Journal of Operational Research* 195.1. Publisher: Elsevier, pp. 1–16.

Branch, Mary Ann, Thomas F Coleman, and Yuying Li (1999). "A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems". In: *SIAM Journal on Scientific Computing* 21.1. Publisher: SIAM, pp. 1–23.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1. Publisher: Springer, pp. 5–32.

Breiman, Leo et al. (1984). "Classification and regression trees. Wadsworth & Brooks". In: *Cole Statistics/Probability Series*.

Brown, Tom et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Buolamwini, Joy and Timnit Gebru (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR, pp. 77–91.

Chakraborty, Joymallya, Suvodeep Majumder, and Tim Menzies (2021). "Bias in machine learning software: Why? how? what to do?" In: *Proceedings of the 29th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pp. 429–440.

Connor, Robert J and James E Mosimann (1969). "Concepts of independence for proportions with a generalization of the Dirichlet distribution". In: *Journal of the American Statistical Association* 64.325. Publisher: Taylor & Francis, pp. 194–206.

Cooil, Bruce, Lerzan Aksoy, and Timothy L Keiningham (2008). "Approaches to customer segmentation". In: *Journal of Relationship Marketing* 6.3-4. Publisher: Taylor & Francis, pp. 9–39.

Correa, Juan and Elias Bareinboim (2020). "A calculus for stochastic interventions: Causal effect identification and surrogate experiments". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. Issue: 06, pp. 10093–10100.

Correa, Juan, Sanghack Lee, and Elias Bareinboim (2021). "Nested counterfactual identification from arbitrary surrogate experiments". In: *Advances in Neural Information Processing Systems* 34, pp. 6856–6867.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20. Publisher: Springer, pp. 273–297.

Coussement, Kristof, Stefan Lessmann, and Geert Verstraeten (2017). "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry". In: *Decision Support Systems* 95. Publisher: Elsevier, pp. 27–36. DOI: 10.1016/j.dss.2016.11.007.

Cover, Thomas and Peter Hart (1967). "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1. Publisher: IEEE, pp. 21–27.

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of information theory*. Publication Title: Elements of Information Theory. John Wiley & Sons. ISBN: 0-471-06259-6. DOI: 10.1002/0471200611.

Csiszár, Imre and Paul C Shields (2004). *Information theory and statistics: A tutorial*. Now Publishers Inc.

Dal Pozzolo, Andrea and Gianluca Bontempi (2015). "Adaptive machine learning for credit card fraud detection". PhD thesis. Université libre de Bruxelles. (Visited on 01/08/2021).

Dal Pozzolo, Andrea, Olivier Caelen, Reid A Johnson, et al. (2015). "Calibrating probability with undersampling for unbalanced classification". In: *2015 IEEE Symposium Series on Computational Intelligence*. IEEE, pp. 159–166.

Dal Pozzolo, Andrea, Olivier Caelen, Yann-Ael Le Borgne, et al. (2014). "Learned lessons in credit card fraud detection from a practitioner perspective". In: *Expert systems with applications* 41.10. Publisher: Elsevier, pp. 4915–4928.

De Stefani, Jacopo (2022). "Towards multivariate multi-step-ahead time series forecasting: A machine learning perspective". English. Publisher: Université libre de Bruxelles. Thesis. Brussels, Belgium: Université libre de Bruxelles.

Devriendt, Floris, Jeroen Berrevoets, and Wouter Verbeke (2021). "Why you should stop predicting customer churn and start using uplift models". In: *Information Sciences* 548. Publisher: Elsevier, pp. 497–515.

Devriendt, Floris, Darie Moldovan, and Wouter Verbeke (2018). "A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics". In: *Big data* 6.1. Publisher: Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, pp. 13–41.

Devriendt, Floris, Jente Van Belle, et al. (2020). "Learning to rank for uplift modeling". In: *IEEE Transactions on Knowledge and Data Engineering*. Publisher: IEEE.

Diemert Eustache, Betlei Artem, Christophe Renaudin, and Amini Massih-Reza (2018). "A Large Scale Benchmark for Uplift Modeling". In: *Proceedings of the AdKDD and TargetAd Workshop, KDD, London,United Kingdom, August, 20, 2018*. ACM.

Ding, Ming (2022). "The road from MLE to EM to VAE: A brief tutorial". In: *AI Open* 3. Publisher: Elsevier, pp. 29–34.

Durrett, Rick (2019). *Probability: theory and examples*. Vol. 49. Cambridge university press.

Edwards, J H (1957). "A note on the practical interpretation of 2× 2 tables". In: *British journal of preventive & social medicine* 11.2. Publisher: BMJ Publishing Group, p. 73.

Einstein, Albert et al. (1969). "Briefwechsel: 1916-1955". In: Publisher: Nymphenburger Verlagshandlung.

Emmanuel, Tlamelo et al. (2021). "A survey on missing data in machine learning". In: *Journal of Big Data* 8.1. Publisher: SpringerOpen, pp. 1–37.

Farris, Paul W et al. (2010). *Marketing metrics: The definitive guide to measuring marketing performance*. Pearson Education.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8. Publisher: Elsevier, pp. 861–874.

Fernández-Loria, Carlos and Foster Provost (2022a). "Causal Classification: Treatment Effect Estimation vs. Outcome Prediction". In: *Journal of Machine Learning Research* 23.59, pp. 1–35.

— (2022b). "Causal decision making and causal effect estimation are not the same… and why it matters". In: *INFORMS Journal on Data Science*. Publisher: INFORMS.

Fiegener, Mark K (2010). "Numbers of doctorates awarded continue to grow in 2009; indicators of employment outcomes mixed. InfoBrief. NSF 11-305." In: *National Science Foundation*. Publisher: ERIC.

Fréchet, Maurice (1935). "Généralisation du théoreme des probabilités totales". In: *Fundamenta mathematicae* 1.25, pp. 379–387.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*. Publisher: JSTOR, pp. 1189–1232.

Ganin, Yaroslav et al. (2016). "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 17.1. Publisher: JMLR. org, pp. 2096–2030.

Gao, Wei et al. (2010). "Stability analysis for ranking algorithms". In: *2010 IEEE international conference on information theory and information security*. IEEE, pp. 973–976.

García-Martín, Eva et al. (2019). "Estimation of energy consumption in machine learning". In: *Journal of Parallel and Distributed Computing* 134. Publisher: Elsevier, pp. 75–88.

Geenens, Gery (2020). "Copula modeling for discrete random vectors". In: *Dependence Modeling* 8.1. Publisher: De Gruyter Open, pp. 417–440.

Geiler, Louis, Séverine Affeldt, and Mohamed Nadif (2022). "A survey on machine learning methods for churn prediction". In: *International Journal of Data Science and Analytics* 14.3. Publisher: Springer, pp. 217–242.

Ghahramani, Zoubin and Michael Jordan (1993). "Supervised learning from incomplete data via an EM approach". In: *Advances in neural information processing systems* 6.

Goldberg, David E (1989). "Genetic algorithms in search". In: *Optimization, Machine Learning*. Publisher: Addison-wesley.

Gubela, Robin M and Stefan Lessmann (2021). "Uplift modeling with value-driven evaluation metrics". In: *Decision Support Systems*. Publisher: Elsevier, p. 113648.

Gubela, Robin M, Stefan Lessmann, and Szymon Jaroszewicz (2020). "Response transformation and profit decomposition for revenue uplift modeling". In: *European Journal of Operational Research* 283.2. Publisher: Elsevier, pp. 647–661.

Guelman, Leo, Montserrat Guillén, and Ana M. Pérez-Marín (2015). "Uplift random forests". In: *Cybernetics and Systems* 46.3-4. Publisher: Taylor & Francis, pp. 230–248. ISSN: 10876553. DOI: 10.1080/01969722.2015.1012892.

Guido, Gianluigi et al. (2011). "Targeting direct marketing campaigns by neural networks". In: *Journal of Marketing Management* 27.9-10. Publisher: Taylor & Francis, pp. 992–1006.

Gupta, Sunil et al. (2006). "Modeling customer lifetime value". In: *Journal of service research* 9.2. Publisher: Sage Publications Sage CA: Thousand Oaks, CA, pp. 139–155.

Gutierrez, Pierre and Jean-Yves Gérardy (Jan. 2016). "Causal Inference and Uplift Modelling: A Review of the Literature". In: *Proceedings of The 3rd International Conference on Predictive Applications and APIs*. Ed. by Claire Hardgrove et al. Vol. 67. Series Title: Proceedings of Machine Learning Research. Microsoft NERD, Boston, USA: PMLR, pp. 1–13. URL: http://proceedings.mlr.press/v67/gutierrez17a.html.

Hahn, P Richard, Jared S Murray, and Carlos M Carvalho (2020). "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)". In: *Bayesian Analysis* 15.3. Publisher: International Society for Bayesian Analysis, pp. 965–1056.

Hand, DJ (1981). "Discrimination and classification john wiley & sons". In: *Inc., New York*.

Hansotia, Behram and Brad Rukstales (2002). "Incremental value modeling". In: *Journal of Interactive Marketing* 16.3. Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 35–46.

Hansotia, Behram J and Bradley Rukstales (2002). "Direct marketing for multichannel retailers: Issues, challenges and solutions". In: *Journal of Database Marketing & Customer Strategy Management* 9.3. Publisher: Springer, pp. 259–266.

Haupt, Johannes and Stefan Lessmann (2022). "Targeting customers under response-dependent costs". In: *European Journal of Operational Research* 297.1. Publisher: Elsevier, pp. 369–379.

Heckman, James Joseph (1991). *Randomization and social policy evaluation*. National Bureau of Economic Research Cambridge, MA.

Hewitt, Edwin and Karl Stromberg (2013). *Real and abstract analysis: a modern treatment of the theory of functions of a real variable*. Springer-Verlag.

Hillstrom, Kevin (2008). "The minethatdata e-mail analytics and data mining challenge, 2008". In: *URL https://blog. minethatdata. com/2008/03/minethatdata-e-mail-analytics-and-data. html.*

Holland, Paul W. (Dec. 1986). "Statistics and Causal Inference". en. In: *Journal of the American Statistical Association* 81.396, pp. 945–960. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1986.10478354. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354 (visited on 06/06/2023).

Huber, Peter J (2004). *Robust statistics*. Vol. 523. John Wiley & Sons.

Ibrahim, Muhammad Sohail, Wei Dong, and Qiang Yang (2020). "Machine learning driven smart electric power systems: Current trends and new perspectives". In: *Applied Energy* 272. Publisher: Elsevier, p. 115237.

Idris, Adnan and Asifullah Khan (2014). "Ensemble based efficient churn prediction model for telecom". In: *Frontiers of Information Technology (FIT), 2014 12th International Conference on*, pp. 238–244. DOI: 10.1109/fit.2014.52.

Imbens, Guido W and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jain, Hemlata, Ajay Khunteta, and Sumit Srivastava (2021). "Telecom churn prediction and used techniques, datasets and performance measures: a review". In: *Telecommunication Systems* 76. Publisher: Springer, pp. 613–630.

Jalaldoust, Kasra and Elias Bareinboim (May 2023). *Transportable representations for out-of-distribution generalization*. Tech. rep. Issue: R-99. Causal Artificial Intelligence Lab, Columbia University.

Jaroszewicz, Szymon and Dan A Simovici (2001). "A general measure of rule interestingness". In: *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 253–265.

Jaskowski, Maciej and Szymon Jaroszewicz (2012). "Uplift modeling for clinical trial data". In: *ICML Workshop on Clinical Data Analysis*.

Johannemann, Jonathan et al. (2019). "Sufficient representations for categorical variables". In: *arXiv preprint arXiv:1908.09874*.

Jović, Alan, Karla Brkić, and Nikola Bogunović (2015). "A review of feature selection methods with applications". In: *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. Ieee, pp. 1200–1205.

Jung, Yonghan, Jin Tian, and Elias Bareinboim (2021). "Estimating Identifiable Causal Effects through Double Machine Learning". In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Kane, Kathleen, Victor SY Lo, and Jane Zheng (2014). "Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods". In: *Journal of Marketing Analytics* 2. Publisher: Springer, pp. 218–238.

Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera (2021). "Algorithmic recourse: from counterfactual explanations to interventions". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362.

Kayaalp, Fatih (2017). "Review of Customer Churn Analysis Studies in Telecommunications Industry". In: *Karaelmas Fen ve Mühendislik Dergisi* 7.2, pp. 696–705. DOI: 10.7212/zkufbd.v7i2.875.

Kennedy, James and Russell Eberhart (1995). "Particle swarm optimization". In: *Proceedings of ICNN'95-international conference on neural networks*. Vol. 4. tex.organization: IEEE, pp. 1942–1948.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Kirkpatrick, Scott, C Daniel Gelatt Jr, and Mario P Vecchi (1983). "Optimization by simulated annealing". In: *Science (New York, N.Y.)* 220.4598. Publisher: American association for the advancement of science, pp. 671–680.

Kleinberg, Samantha (2015). *Why: A guide to finding and using causes*. O'Reilly Media, Inc.

Kolmogoroff, Andrey (1933). "Grundbegriffe der wahrscheinlichkeitsrechnung". In: Publisher: Springer.

Künzel, Sören R. et al. (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.10. arXiv: 1706.03461 Publisher: National Acad Sciences, pp. 4156–4165. ISSN: 10916490. DOI: 10.1073/pnas.1804597116.

Lebichot, Bertrand et al. (2021). "Transfer Learning Strategies for Credit Card Fraud Detection". In: *IEEE Access* 9, pp. 114754–114766. DOI: 10.1109/ACCESS.2021.3104472.

Lee, Lillian (2000). "Measures of distributional similarity". In: *arXiv preprint cs/0001012*.

Li, Ang and Judea Pearl (2019). "Unit Selection Based on Counterfactual Logic". In: *IJCAI*. International Joint Conferences on Artificial Intelligence Organization, pp. 1793–1799. DOI: 10.24963/ijcai.2019/248. URL: https://doi.org/10.24963/ijcai.2019/248.

— (2022). "Unit selection with causal diagram". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. Number: 5, pp. 5765–5772.

Li, Xinran and Peng Ding (2016). "Exact confidence intervals for the average causal effect on a binary outcome". In: *Statistics in Medicine* 35.6. Publisher: Wiley Online Library, pp. 957–960.

Lin, Jiayu (2016). "On the dirichlet distribution". In: *Master's Report*. Publisher: Queen's University Kingston Ontario, Canada.

Liu, Xu-Ying, Jianxin Wu, and Zhi-Hua Zhou (2009). "Exploratory undersampling for class-imbalance learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2. Publisher: IEEE, pp. 539–550. DOI: 10.1109/tsmcb.2008.2007853.

Livieris, Ioannis E et al. (2019). "A weighted voting ensemble self-labeled algorithm for the detection of lung abnormalities from X-rays". In: *Algorithms* 12.3. Publisher: MDPI, p. 64.

Lo, Victor S Y (2002). "The true lift model: a novel data mining approach to response modeling in database marketing". In: *ACM SIGKDD Explorations Newsletter* 4.2. Publisher: ACM New York, NY, USA, pp. 78–86.

Louizos, Christos et al. (2017). *Causal Effect Inference with Deep Latent-Variable Models*. arXiv: stat.ML/1705.08821.

Majeed, Abdul and Sungchang Lee (2020). "Anonymization techniques for privacy preserving data publishing: A comprehensive survey". In: *IEEE access : practical innovations, open solutions* 9. Publisher: IEEE, pp. 8512–8545.

Massimetti, Giulia (2021). "Uplift modeling for marketing campaigns in the telecommunication industry". English. MA thesis. Brussels, Belgium: Université libre de Bruxelles.

Mehrabi, Ninareh et al. (2021). "A survey on bias and fairness in machine learning". In: *ACM computing surveys (CSUR)* 54.6. Publisher: ACM New York, NY, USA, pp. 1–35.

Micci-Barreca, Daniele (2001). "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems". In: *ACM SIGKDD Explorations Newsletter* 3.1. Publisher: ACM New York, NY, USA, pp. 27–32.

Michel, René, Igor Schnakenburg, and Tobias Von Martens (2019). *Targeting uplift: An introduction to net scores*. Springer Nature.

Ming, Yifei, Hang Yin, and Yixuan Li (2022). "On the impact of spurious correlation for out-of-distribution detection". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. Issue: 9, pp. 10051–10059.

Mintz, Yoav and Ronit Brodie (2019). "Introduction to artificial intelligence in medicine". In: *Minimally Invasive Therapy & Allied Technologies* 28.2. Publisher: Taylor & Francis, pp. 73–81.

Mitrović, Sandra et al. (2018). "On the operational efficiency of different feature types for telco Churn prediction". In: *European Journal of Operational Research* 267.3. Publisher: Elsevier, pp. 1141–1155. ISSN: 03772217. DOI: 10.1016/j.ejor.2017.12.015. URL: https://doi.org/10.1016/j.ejor.2017.12.015.

Moraffah, Raha et al. (2020). "Causal interpretability for machine learning-problems, methods and evaluation". In: *ACM SIGKDD Explorations Newsletter* 22.1. Publisher: ACM New York, NY, USA, pp. 18–33.

Mueller, Scott, Ang Li, and Judea Pearl (2021). "Causes of Effects: Learning individual responses from population data". In: *arXiv preprint arXiv:2104.13730*.

Mueller, Scott and Judea Pearl (2022). "Personalized decision Making–A conceptual introduction". In: *arXiv preprint arXiv:2208.09558*.

Olkin, Ingram and Thomas A Trikalinos (2015). "Constructions for a bivariate beta distribution". In: *Statistics & Probability Letters* 96. Publisher: Elsevier, pp. 54–60.

Ongaro, Andrea and Sonia Migliorati (2013). "A generalization of the Dirichlet distribution". In: *Journal of Multivariate Analysis* 114. Publisher: Elsevier, pp. 412–426.

Óskarsdóttir, María, Cristián Bravo, et al. (2017). "Social network analytics for churn prediction in telco: Model building, evaluation and network architecture". In: *Expert Systems with Applications* 85. Publisher: Elsevier, pp. 204–220. DOI: 10.1016/j.eswa.2017.05.028.

Óskarsdóttir, María, Tine Van Calster, et al. (2018). "Time series for early churn detection: Using similarity based classification for dynamic networks". In: *Expert Systems with Applications* 106. Publisher: Elsevier, pp. 55–65. DOI: 10.1016/j.eswa.2018.04.003.

Parsons, Simon and Anthony Hunter (1998). "A review of uncertainty handling formalisms". In: *Applications of uncertainty formalisms*. Publisher: Springer, pp. 8–37.

Pearl, Judea (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

Pearl, Judea (2009). *Causality: models, reasoning, and inference*. Cambridge university press. ISBN: 978-0-521-89560-6.

— (2017). "Detecting latent heterogeneity". In: *Sociological Methods & Research* 46.3. Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 370–389.

— (July 2020). *Data versus Science: Contesting the Soul of Data-Science*. English. Blog. URL: http://causality.cs.ucla.edu/blog/index.php/category/data-fusion/ (visited on 12/12/2023).

Pearl, Judea and Dana Mackenzie (2018). *The book of why: the new science of cause and effect*. Basic books.

Pearl, Judea and James M Robins (1995). "Probabilistic evaluation of sequential plans from causal models with hidden variables." In: *UAI*. Vol. 95. Citeseer, pp. 444–453.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. ISSN: 1533-7928. URL: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 04/27/2023).

Perini, Lorenzo, Connor Galvin, and Vincent Vercruyssen (2020). "A ranking stability measure for quantifying the robustness of anomaly detection methods". In: *ECML PKDD 2020 workshops: Workshops of the european conference on machine learning and knowledge discovery in databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, ghent, belgium, september 14–18, 2020, proceedings*. Springer, pp. 397–408.

Perrone, Michael P and Leon N Cooper (1995). "When networks disagree: Ensemble methods for hybrid neural networks". In: *How we learn; how we remember: Toward an understanding of brain and neural systems: Selected papers of leon N cooper*. World Scientific, pp. 342–358.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Plečko, Drago and Elias Bareinboim (July 2022). *Causal fairness analysis*. Tech. rep. Issue: R-90. Causal Artificial Intelligence Lab, Columbia University.

— (2023). "Causal fairness for outcome control". In: *arXiv preprint arXiv:2306.05066*.

Portugal, Ivens, Paulo Alencar, and Donald Cowan (2018). "The use of machine learning algorithms in recommender systems: A systematic review". In: *Expert Systems with Applications* 97. Publisher: Elsevier, pp. 205–227.

Qi, Feng (2013). "Integral representations and properties of Stirling numbers of the first kind". In: *Journal of Number Theory* 133.7, pp. 2307–2319.

Radcliffe, Nicholas J and Patrick D Surry (2011). "Real-world uplift modelling with significance-based uplift trees". In: *White Paper TR-2011-1, Stochastic Solutions*. Publisher: Citeseer, pp. 1–33.

Radclifte, Nicholas J and Rob Simpson (2008). "Identifying who can be saved and who will be driven away by retention activity." In: *Journal of Telecommunications Management* 1.2.

Ramesh, Aditya et al. (2021). "Zero-shot text-to-image generation". In: *International conference on machine learning*. PMLR, pp. 8821–8831.

Rish, Irina et al. (2001). "An empirical study of the naive Bayes classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. Number: 22, pp. 41–46.

Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1. Publisher: Oxford University Press, pp. 41–55.

Rößler, Jannik and Detlef Schoder (2022). "Bridging the gap: A systematic benchmarking of uplift modeling and heterogeneous treatment effects methods". In: *Journal of Interactive Marketing* 57.4. Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 629–650.

Rubin, Donald B (2005). "Causal inference using potential outcomes: Design, modeling, decisions". In: *Journal of the American Statistical Association* 100.469. Publisher: Taylor & Francis, pp. 322–331.

— (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology*. ISSN: 00220663. DOI: 10.1037/h0037350.

Rzepakowski, Piotr and Szymon Jaroszewicz (2012). "Decision trees for uplift modeling with single and multiple treatments". In: *Knowledge and Information Systems* 32.2. Publisher: Springer, pp. 303–327.

Sagi, Omer and Lior Rokach (2018). "Ensemble learning: A survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4. Publisher: Wiley Online Library, e1249.

Schölkopf, Bernhard et al. (2012). "On causal and anticausal learning". In: *arXiv preprint arXiv:1206.6471*.

Shalit, Uri, Fredrik D. Johansson, and David Sontag (2017). "Estimating individual treatment effect: generalization bounds and algorithms". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. Vol. 6. arXiv: 1606.03976. JMLR. org, pp. 3076–3085. ISBN: 978-1-5108-5514-4.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *Bell system technical journal* 27.3. Publisher: Wiley Online Library, pp. 379–423.

Spirtes, Peter, Clark Glymour, and Computer Review (1991). "An algorithm for fast recovery of sparse causal graphs". In: *Social science computer review* 9.1. Publisher: Sage Publications Sage CA: Thousand Oaks, CA, pp. 62–72.

Tian, Jin and Judea Pearl (2000). "Probabilities of causation: Bounds and identification". In: *Annals of Mathematics and Artificial Intelligence* 28.1. Publisher: Springer, pp. 287–313.

— (2002). "A general identification condition for causal effects". In: *Aaai/iaai*, pp. 567–573.

Tianyuan, Zhang and Sérgio Moro (2021). "Research trends in customer churn prediction: a data mining approach". In: *World conference on information systems and technologies*. Springer, pp. 227–237.

United States Census Bureau, Economics and Statistics Administration (2011). *Statistical Abstract of the United States: 2011*. English. Tech. rep. 130. United States: US Government Printing Office. URL: https://www.census.gov/library/publications/2010/compendia/statab/130ed.html (visited on 12/12/2023).

Verbeke, Wouter, David Martens, and Bart Baesens (2014). "Social network analysis for customer churn prediction". In: *Applied Soft Computing* 14. Publisher: Elsevier, pp. 431–446. DOI: 10.1016/j.asoc.2013.09.017.

Verbeke, Wouter, Diego Olaya, Jeroen Berrevoets, et al. (2021). "The foundations of cost-sensitive causal classification". In: *European Journal of Operational Research*. Publisher: Elsevier, pp. 1–20.

Verbeke, Wouter, Diego Olaya, Marie-Anne Guerry, et al. (2022). "To do or not to do? Cost-sensitive causal classification with individual treatment effect estimates". In: *European Journal of Operational Research*. Publisher: Elsevier.

Verhelst, Théo (2018). "Churn prediction and causal analysis on telecom customer data". English. MA thesis. Brussels, Belgium: Université libre de Bruxelles. URL: https://theoverhelst.me/documents/thesis.pdf (visited on 09/28/2023).

Verhelst, Théo and Gianluca Bontempi (2024). "Identifying counterfactual probabilities using bivariate distributions and uplift modeling". In: *to be submitted*.

Verhelst, Théo, Olivier Caelen, et al. (2020). "Understanding Telecom Customer Churn with Machine Learning: From Prediction to Causal Inference". In: *Artificial Intelligence and Machine Learning*. Ed. by Bart Bogaerts et al. ISSN: 16130073. Springer International Publishing, pp. 182–200. ISBN: 978-3-030-65154-1.

Verhelst, Théo, Mercier Denis, et al. (2024). "Customer segmentation from counterfactual probabilities: new insights for the telecom sector". In: *to be submitted*.

Verhelst, Théo, Denis Mercier, et al. (2023a). "A churn prediction dataset from the telecom sector: a new benchmark for uplift modeling". In: *ECML PKDD 2023 Workshops - Workshop on Uplift Modeling and Causal Machine Learning for Operational Decision Making*.

— (Mar. 2023b). "Partial counterfactual identification and uplift modeling: theoretical results and real-world assessment". en. In: *Machine Learning*. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-023-06317-w. URL: https://link.springer.com/10.1007/s10994-023-06317-w (visited on 05/03/2023).

Verhelst, Théo, Jeevan Shrestha, et al. (2021). "Predicting reach to find persuadable customers: Improving uplift models for churn prevention". In: *Discovery science*. Ed. by Carlos Soares and Luis Torgo. Cham: Springer International Publishing, pp. 44–54. ISBN: 978-3-030-88942-5.

Verhelst, Théo, Wouter Verbeke, et al. (2023). "Uplift vs. Predictive Modeling: a Theoretical Analysis". In: *Submitted to Journal of Machine Learning Research*.

Webb, Geoffrey I (2000). "Multiboosting: A technique for combining boosting and wagging". In: *Machine learning* 40.2. Publisher: Springer, pp. 159–196.

Weinberger, Naftali (2021). "Comparing Rubin and Pearl's causal modelling frameworks: a commentary on Markus (2021)". In: *Economics & Philosophy*. Publisher: Cambridge University Press, pp. 1–9.

Weisstein, Eric (2023). *Stirling Number of the First Kind*. English. URL: https://mathworld.wolfram.com/StirlingNumberoftheFirstKind.html (visited on 11/23/2023).

Wijaya, Davin et al. (2021). "Uplift modeling VS conventional predictive model: A reliable machine learning model to solve employee turnover". In: *International Journal of Artificial Intelligence Research* 5.1, pp. 53–64.

Winer, Russell S (2001). "A framework for customer relationship management". In: *California management review* 43.4. Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 89–105.

Wolpert, Robert L (2010). "Conditional expectation". In: *University Lecture*.

Zaniewicz, Lukasz and Szymon Jaroszewicz (2013). "Support vector machines for uplift modeling". In: *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, pp. 131–138.

Zhang, Junzhe, Jin Tian, and Elias Bareinboim (2022). "Partial counterfactual identification from observational and experimental data". In: *International conference on machine learning*. tex.organization: PMLR, pp. 26548–26558.

Zhang, Weijia, Jiuyong Li, and Lin Liu (2021). "A unified survey of treatment effect heterogeneity modelling and uplift modelling". In: *ACM Computing Surveys (CSUR)* 54.8. Publisher: ACM New York, NY, pp. 1–36.

Zhu, Bing, Bart Baesens, and Seppe K L M vanden Broucke (2017). “An empirical comparison of techniques for the class imbalance problem in churn prediction”. In: *Information sciences* 408. Publisher: Elsevier, pp. 84–99. DOI: 10.1016/j.ins.2017.04.015.