NOVA IMS
Information Management School

# HETEROGENEOUS TREATMENT EFFECTS IN LOYALTY PROGRAMS

## A Study Case of Causal Inference Approach to Understanding Customer Behaviour

**Jaime Kiyoshi Kuei**

Dissertation presented as partial requirement for obtaining the Master's degree in Master of Data Science and Advanced Analytics

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão da Informação**
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade NOVA de Lisboa

# HETEROGENEOUS TREATMENT EFFECTS IN LOYALTY PROGRAMS

by

Jaime Kiyoshi Kuei

Dissertation presented as partial requirement for obtaining the Master's degree in Master of Data Science and Advanced Analytics

**Adviser**: Bruno Damásio

November, 2024

**Heterogeneous Treatment Effects in Loyalty Programs**
**A Study Case of Causal Inference Approach to Understanding Customer Behaviour**



This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.18) (Lourenço, 2021).

# Acknowledgements

I would like to express my sincere gratitude to everyone who supported me during the development of this project.

Firstly, I thank Prof. Dr. Bruno Damásio for all the guidance and support throughout this journey. His advice and encouragement were essential for the completion of this work.

I am also grateful to my colleague Maxwell Gabriel Marcos, whose help and collaboration were key in overcoming several challenges during the research process.

To NOVA IMS, especially the Department of Data Science, I extend my thanks for the structure and support provided, which made this project possible.

Finally, I am thankful to iFood for their partnership and contribution to the development of this research.

# Abstract

The study explores the heterogeneous effects of a loyalty program on user purchasing behavior, addressing the challenge of estimating causal impacts from observational data. This is a significant problem in the domain of customer behavior analysis, as understanding individual-level responses is essential for optimizing marketing strategies and resource allocation.

The proposed approach employs causal inference methods, specifically metalearners (S-Learner, T-Learner, and X-Learner), to estimate Conditional Average Treatment Effects (CATE) from a real-world dataset. To ensure robust analysis, a propensity score matching (PSM) technique was applied to create a debiased training dataset, mitigating covariate imbalance between treatment and control groups. The models were trained and evaluated using advanced metrics, including Qini curves, to assess their effectiveness in identifying high-value subpopulations.

The results demonstrate that the X-Learner outperformed other models, particularly when trained on the debiased dataset. However, performance variations between the training and test datasets highlighted challenges related to overfitting and dataset imbalance. These findings emphasize the importance of balancing datasets and incorporating propensity scores to improve the generalizability and reliability of causal estimates.

This research contributes to the field by demonstrating the applicability of causal inference techniques in marketing analytics, providing insights into the design and evaluation of loyalty programs. The results underscore the value of advanced modeling techniques for strategic decision-making in customer segmentation and targeting.

**Keywords:** causal inference, metalearners, loyalty program, propensity score matching, customer behavior analysis

# Resumo

Este estudo explora os efeitos heterogêneos de um programa de fidelidade no comportamento de compra dos usuários, abordando o desafio de estimar impactos causais a partir de dados observacionais. Este é um problema relevante na análise de comportamento do cliente, uma vez que compreender as respostas individuais é essencial para otimizar estratégias de marketing e alocação de recursos.

A abordagem proposta utiliza métodos de inferência causal, especificamente *meta-learners* (S-Learner, T-Learner e X-Learner), para estimar os Efeitos Médios Condicionais do Tratamento (CATE) em um conjunto de dados do mundo real. Para garantir uma análise robusta, foi aplicada a técnica de *propensity score matching* (PSM) para criar um conjunto de treino balanceado, reduzindo o desbalanceamento das covariáveis entre os grupos de tratamento e controle. Os modelos foram treinados e avaliados utilizando métricas avançadas, como curvas Qini, para avaliar sua eficácia em identificar subpopulações de alto valor.

Os resultados demonstraram que o X-Learner superou os demais modelos, especialmente quando treinado no conjunto de dados balanceado. No entanto, variações de desempenho entre os conjuntos de treino e teste destacaram desafios relacionados ao sobreajuste e ao desbalanceamento do dataset. Esses achados ressaltam a importância de balancear os conjuntos de dados e incorporar *propensity scores* para melhorar a generalização e a confiabilidade das estimativas causais.

Esta pesquisa contribui para o campo ao demonstrar a aplicabilidade de técnicas de inferência causal na análise de marketing, fornecendo insights sobre o design e a avaliação de programas de fidelidade. Os resultados reforçam o valor de técnicas avançadas de modelagem para decisões estratégicas em segmentação e direcionamento de clientes.

**Palavras-chave:** inferência causal, meta-learners, programa de fidelidade, *propensity score matching*, análise de comportamento do cliente

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Understanding causal relationships is essential for evaluating the performance of companies, as it enables a thorough analysis of the impact generated by different strategies. At iFood, a leader in the Brazilian food delivery market, strategies such as marketing campaigns, personalized recommendations, interface improvements, and A/B testing are frequently used. Like many other companies in the market, iFood also relies on subscription-based programs as part of its approach. However, measuring causal effects is not straightforward because the process involves addressing challenges such as confounding variables, selection biases, and the absence of controlled experimental conditions in real-world scenarios. Without careful consideration of these factors, the results may be misleading, leading to incorrect conclusions about the impact of interventions. As this study will show, this type of analysis requires careful planning, specific methodologies, and detailed attention to ensure reliable results.

## 1.1 iFood's Company Overview

iFood is a rapidly growing Brazilian technology company that has become a leader in delivery services across Latin America. Since its inception, the company has expanded from 12,000 orders per month in 2012 to over 100 million orders per month by 2024 (iFood, 2024b). As a digital platform, iFood connects consumers, delivery drivers, restaurants, and retailers, encompassing over 350,000 registered establishments and 310,000 delivery drivers across more than 1,500 cities in Brazil. The platform's mission is to optimize meal and grocery deliveries while expanding its services to include market deliveries and fintech solutions, such as iFood Benefits, aiming to lead the digitization of the food industry.

One of the company's innovative strategies is the *iFood Clube*, a subscription program that offers customers exclusive benefits, such as free delivery on eligible orders and discounts on selected meals. This program has become an essential tool for fostering customer loyalty and increasing order frequency, reinforcing iFood's position as a market leader (iFood, 2022).

As a key player in Brazil's economy, iFood contributes to 0.55% of the national GDP and supports approximately 909,000 jobs—equivalent to 0.91% of the country's population—while circulating about R$ 110 billion in 2023 (iFood, 2024a). Founded in 2011 by Patrick Sigrist, Eduardo Baer, Guilherme Bonifácio, and Felipe Fioravante, iFood evolved from a printed menu guide to one of the most prominent online platforms in the region. Movile, founded by Fabricio Bloisi, became the majority shareholder in 2012, with Bloisi serving as iFood's CEO until 2024 (iFood, 2023).

The company operates under two main business models: *Marketplace*, which accounts for 61% of app sales and relies on establishments to manage their deliveries; and *Full Service* (or Partner Delivery), responsible for 39% of sales, where iFood coordinates logistics and delivery drivers to improve the overall customer and restaurant experience (iFood, 2023).

## 1.2 Problem Statement

Subscription programs, such as Amazon Prime, Uber One, DashPass, and iFood Clube, have gained popularity as key strategies for increasing customer engagement. These programs encourage loyalty and recurring consumption by offering benefits like free delivery, exclusive discounts, and vouchers. By creating consistent purchasing habits, they aim to strengthen the relationship between customers and companies.

Measuring the actual impact of these programs, however, involves significant complexities. One of the main challenges is self-selection bias, where customers who are already more likely to spend also tend to subscribe. This makes it difficult to distinguish the program's incremental impact from the customers' pre-existing behaviours. Another important challenge is understanding the difference between short-term and long-term effects. For example, while promotional incentives may temporarily boost purchases, they might not lead to sustained loyalty or increased customer lifetime value (CLV).

This study focuses on overcoming these challenges within the context of iFood Clube. By applying advanced causal inference methods, such as metalearners, the research aims to quantify the program's incremental impact and identify behavioural changes in short-term effects among subscribers. Additionally, understanding which customer profiles benefit most from the subscription program is essential for designing strategies that maximise engagement and retention.

Based on these objectives, this study seeks to address the following key research questions:

1. **Causal Impact:** How much of the observed increase in customer spending is truly attributable to the iFood Clube subscription program, rather than pre-existing customer behaviours or external factors?

2. **Customer Profiles:** Which customer profiles experience the greatest behavioural changes as a result of subscribing to the program, and how do these profiles inform strategies to maximise engagement and retention?

## 1.3 Thesis Structure

This thesis is divided into five chapters, each addressing a specific aspect of the research. The second chapter, the Theoretical Framework, explores the theoretical foundation and existing studies on subscription programs, their challenges, and the fundamentals of causal inference. It discusses the assumptions underlying causal models, the difficulties of working with observational data, the concept of uplift modeling, the structure of metalearners, and the metrics used for causal model evaluation.

The third chapter, Empirical Strategy, details the approach used in the study, including the data generation process, the models employed, and the procedures adopted throughout the modeling process. It covers the preparation and evaluation of datasets, the assessment of base learners, and the evaluation of metalearners. The fourth chapter, Results, presents the findings of the study, ranging from data exploration to the testing of causal assumptions, as well as the performance assessments of both base learners and metalearners. This chapter also includes a discussion of the key insights obtained from the analysis. Finally, the fifth chapter, Conclusion, summarizes the main findings, reflects on their implications, and provides recommendations for future research directions.

# 2

# Theoretical Framework

This chapter provides the theoretical framework necessary to support the analysis conducted in this study. It begins by examining the role of subscription programs in customer loyalty and their influence on purchasing behaviour. Next, it addresses the challenges in measuring the incremental impact of these programs.

The discussion then introduces causal inference as a fundamental approach to estimating treatment effects, exploring key concepts such as potential outcomes, average treatment effects, and conditional treatment effects, as well as the assumptions required for robust causal analysis. Finally, the chapter presents advanced methodologies, including uplift models and meta-learners, which are instrumental for identifying heterogeneous treatment effects and measuring incrementality in real-world observational data.

This theoretical foundation establishes the basis for the empirical strategy in the next chapter, ensuring a rigorous and comprehensive analysis of the impact of subscription programs.

## 2.1 Subscription Programs and Customer Loyalty

The growing adoption of subscription programs reflects their strategic importance in customer engagement and retention efforts. Platforms like Amazon Prime, Uber One, DashPass, and iFood's subscription program exemplify this trend, offering exclusive benefits in exchange for a subscription fee. These programs aim to establish recurring consumption habits by providing economic incentives, such as free delivery and discounts, as well as non-economic benefits like a sense of exclusivity or habit formation. By doing so, they strengthen the relationship between customers and the company.

The iFood subscription program, for example, combines economic and behavioural incentives to enhance customer loyalty. Subscribers benefit from reduced delivery fees, exclusive vouchers, and discounts on special items, which not only make the platform more cost-effective but also encourage repeated use. This aligns with the findings of Iyengar et al. (2022), who highlight two primary mechanisms through which

subscription programs influence customer purchasing behaviour: economic effects and non-economic effects. Economic effects include tangible benefits like price reductions, while non-economic effects involve psychological factors, such as habit formation, a sense of exclusivity, and the "sunk cost" effect, where the initial payment creates a psychological commitment to justify continued use.

Despite their widespread adoption, the long-term effectiveness of these programs remains debated. Lin and Bowman (2022) argue that subscription programs often attract already engaged customers, redistributing spending across categories rather than generating significant new revenue. These findings underscore the need for deeper analysis of the actual behavioural shifts induced by such programs.

## 2.2 Challenges in Measuring the Impact of Subscription Programs

Measuring the incremental impact of subscription programs on consumer purchasing behaviour presents significant challenges, primarily due to self-selection bias. Customers who are naturally more inclined to spend frequently are more likely to subscribe, complicating efforts to isolate the true impact of the program. Iyengar et al. (2022) addressed this issue by combining difference-in-differences with random forests in a quasi-experimental framework. This approach accounts for unobservable characteristics and enables precise comparisons between subscribers and non-subscribers, simulating an experimental scenario. Their research revealed that subscription programs can produce significant and persistent effects on purchasing behaviour, though these effects vary across customer profiles.

This challenge is particularly relevant in the case of iFood. As a platform where subscribers are typically frequent users, finding the true incremental impact of the program from the behaviour of these already engaged customers becomes critical. Addressing this issue is essential for understanding the true behavioural shifts caused by the subscription program and for identifying customer segments most influenced by the program's incentives.

Another major challenge is distinguishing between short-term and long-term effects. Nishio and Hoshino (2024) explored this distinction in loyalty programs by examining birthday reward promotions. Their findings showed a temporary boost in purchasing behaviour due to the "points pressure effect," where customers increase their spending to achieve reward goals. However, these incentives had limited impact on long-term customer lifetime value (CLV), highlighting that temporary rewards may not lead to sustained loyalty.

These challenges evidentiate the importance of robust causal inference methodologies to measure the true impact of subscription programs. For iFood, addressing these issues means not only controlling for self-selection bias but also analysing the

heterogeneity of effects across customer segments. Insights from such analyses are crucial for refining the program's strategies and maximising its impact on engagement and retention.

## 2.3 Causal Inference

Given the challenges of self-selection bias and distinguishing between short- and long-term effects, robust methodologies are required to measure the true impact of subscription programs. Causal inference provides a framework to estimate these effects by focusing on the causal relationships between interventions and outcomes. As highlighted by Cunningham (2021), causal inference is not a recent discipline. Pioneering work dates back to Fisher (1935), Haavelmo (1943), and Rubin (1974) , who developed foundational frameworks for estimating causal effects. Recent advancements continue to refine these methods, enabling more precise measurements of how interventions influence outcomes in real-world settings.

In the context of loyalty and subscription programs, causal inference has proven important in evaluating the incremental impact of interventions. By addressing biases and isolating causal effects, these methods provide valuable insights for designing and optimising strategies, such as iFood's subscription loyalty program.

### 2.3.1 Fundamental Problem of Causal Inference

The main goal of causal inference is estimating the causal effect of an intervention. Using the framework introduced by Rubin (1974), the causal effect for a specific observation is defined as:

$$\tau_i = Y_i(1) - Y_i(0) \tag{2.1}$$

Here, $i$ stands for the specific case or observation being studied. $Y$ represents the outcome after a particular action or treatment is applied. For instance, in the context of iFood's subscription program, $Y_i(1)$ could represent the total spending of a customer enrolled in the program, while $Y_i(0)$ represents the spending of the same customer if they were not enrolled. The difference, $\tau_i$, reflects the causal effect of the subscription on the customer's spending behaviour.

However, there is a fundamental challenge: it is impossible to observe both $Y_i(1)$ and $Y_i(0)$ for the same individual at the same time. This issue, known as the Fundamental Problem of Causal Inference, highlights the complexity of measuring how one factor truly influences another in real-world scenarios. As Facure (2023) notes, addressing this problem requires methodological rigor and innovative approaches to approximate the unobservable counterfactual scenario.

### 2.3.2 Potential Outcomes

Building on the fundamental problem of causal inference, researchers often turn to the framework of potential outcomes to estimate the impact of interventions. Two key concepts within this framework are the Average Treatment Effect (ATE) and the Conditional Average Treatment Effect (CATE). These measures provide complementary perspectives on how treatments influence outcomes, offering insights at both the population and individual levels.

#### 2.3.2.1 Average Treatment Effect (ATE)

The Average Treatment Effect (ATE) quantifies the general impact of a treatment across all subjects. As Facure (2023) explains, the ATE is defined as:

$$ATE = \mathbb{E}[Y_{1i} - Y_{0i}] \tag{2.2}$$

Here, $Y_{1i}$ represents the outcome for individual $i$ when exposed to the treatment, while $Y_{0i}$ represents the outcome for the same individual without the treatment. The ATE, therefore, captures the average effect of the treatment across the entire population. Conceptually, it measures the difference in outcomes between treated and untreated groups, offering a broad perspective on the effectiveness of the intervention.

In the context of iFood's subscription program, the ATE can be used to estimate the overall impact of the program on customer behaviour, such as the average increase in order volume or spending among subscribers compared to non-subscribers. While the ATE provides valuable insights at the population level, it does not account for variations in treatment effects among different customer segments.

#### 2.3.2.2 Conditional Average Treatment Effect (CATE)

To address this limitation, the Conditional Average Treatment Effect (CATE) focuses on the heterogeneity of treatment effects by conditioning on the characteristics of each individual. Formally, the CATE is defined as:

$$CATE = \mathbb{E}[Y_1 - Y_0 \mid \mathbf{X}] \tag{2.3}$$

In this equation, $\mathbf{X}$ represents the covariates or individual characteristics used to condition the treatment effect. As highlighted by Facure (2023), the CATE enables a more granular analysis, revealing how treatment effects vary depending on the attributes of specific customer.

For example, in the case of iFood's subscription program, the CATE can identify how the program's impact differs among new customers, frequent users, or those with high sensitivity to discounts. By estimating the CATE, companies can personalise their strategies, focusing interventions to customer segments that are most likely to

benefit from the program. This approach not only enhances the effectiveness of the intervention but also optimises resource allocation.

### 2.3.3 Assumptions for Causal Inference

To conduct causal inferences and ensure that the results reflect the causal impact of an intervention, certain assumptions are necessary in both randomised experiments and observational studies. In the context of iFood's subscription program, these assumptions are particularly important for addressing challenges such as confounding variables and the lack of randomisation. Observational data often require additional care to approximate experimental conditions, making these assumptions foundational for robust causal analysis.

As mentioned by Vonk et al. (2023), five primary assumptions have the causal inference: consistency, no-interference, ignorability, conditional ignorability, and positivity. These assumptions ensure that causal estimates accurately reflect the intervention's impact while accounting for potential biases. Below, we detail each assumption and its relevance to this study.

#### 2.3.3.1 Consistency Assumption

The consistency assumption ensures that the observed outcomes match the potential outcomes for a unit subjected to a specific treatment. Formally, this is expressed as:

$$Y = Y(T = t) \tag{2.4}$$

Here, $Y$ represents the observed outcome for a unit subjected to treatment $T = t$. This assumption guarantees that the effect being measured corresponds directly to the treatment applied. For instance, in the iFood context, if a customer subscribes to the loyalty program, their observed spending ($Y$) should accurately reflect the potential outcome under that subscription. This consistency is critical for isolating the program's impact on customer behaviour.

#### 2.3.3.2 No-Interference Assumption

The no-interference assumption states that the outcome of one unit should not be influenced by the treatment assignment of another. Mathematically:

$$Y_i(T_1, T_2, \ldots, T_n) = Y_i(T_i) \tag{2.5}$$

In the context of iFood, this implies that a customer's spending behaviour should not be affected by whether other customers have subscribed to the program. This assumption ensures that spillover effects, such as network effects where one customer's behaviour influences another, do not bias the causal estimates.

However, in practice, it is unlikely that this assumption is fully satisfied. For example, a customer who subscribes to the program might share their positive experiences with friends or family, encouraging them to subscribe as well. This could indirectly influence the spending behaviour of non-subscribers, creating a small degree of interference. Although such effects may be limited in the case of iFood, they highlight a potential limitation of the analysis. It is important to acknowledge these possibilities and consider their impact on the results.

Stable Unit-Treatment Value Assumption (SUTVA) relies on both consistency and no-interference assumptions to ensure that causal inference methods can yield valid estimates of the treatment effect without external contamination or instability in outcomes.

#### 2.3.3.3 Ignorability Assumption

The ignorability assumption requires that potential outcomes are independent of treatment assignment in randomised experiments:

$$Y(0), Y(1) \perp T \tag{2.6}$$

In a randomised controlled trial (RCT), this assumption is naturally satisfied because randomisation ensures that treatment and control groups are comparable. However, in observational studies like this one, randomisation is not applied, making this assumption difficult to satisfy without additional adjustments.

#### 2.3.3.4 Conditional Ignorability Assumption

In observational studies, the conditional ignorability assumption is crucial. It states that by conditioning on confounding variables ($Z$), treatment and control groups become comparable:

$$Y(0), Y(1) \perp T \mid Z \tag{2.7}$$

For iFood, this means adjusting for variables such as customer demographics, historical spending behaviour, and order frequency. By controlling for these covariates, it is possible to estimate the program's impact more reliably, minimising bias caused by confounding factors.

#### 2.3.3.5 Positivity Assumption

The positivity assumption ensures that all subgroups defined by the covariates have a non-zero probability of receiving either treatment or control:

$$P(T = t \mid Z = z) > 0 \quad \text{for all values of } T \text{ and } Z \tag{2.8}$$

In the context of iFood, this assumption guarantees that there are sufficient customers in both the treatment (subscribed) and control (non-subscribed) groups across all subgroups defined by their characteristics ($Z$). Without positivity, causal effects cannot be estimated for certain subgroups, limiting the generalisability of the results.

Both the conditional ignorability and positivity assumptions are particularly important in observational studies. As noted by Vonk et al. (2023), there is an inherent trade-off between these two assumptions when adjusting for covariates. Specifically, the process of conditioning on more covariates can create smaller and more homogeneous subgroups, potentially leading to cases where some subgroups are entirely assigned to either the treatment or control group. This violates the positivity assumption. Conversely, insufficient adjustment for covariates may lead to violations of the conditional ignorability assumption, as confounders remain unaddressed.

To navigate this trade-off, parametric approaches are often preferred over non-parametric ones in high-dimensional settings. Parametric methods can balance the need to control for confounders while preserving enough variability in the data to satisfy positivity.

## 2.4 Challenges for Observational Studies

While the assumptions for causal inference provide a framework for estimating causal effects, applying these principles to observational data introduces additional challenges. Observational data, unlike randomized controlled trials, is collected without manipulating the study environment or subjects. Researchers simply observe and record information as it naturally occurs, often in complex real-world settings. This makes observational data useful for understanding natural behaviors and processes but also prone to biases and confounding factors that complicate causal analysis.

In the context of iFood, although randomized controlled trials (RCTs) could be implemented, there are practical challenges that make them difficult to apply. The main issue is that the subscription program is openly advertised to all users through television, online platforms, and other channels. This means that all customers have the opportunity to join, and it is not possible to prevent certain groups from subscribing after the program has been widely promoted. Because of this, observational data becomes the most practical way to study the program's impact. It also allows researchers to analyze customer behaviour in real-world settings, providing insights that are directly useful for decision-making.

However, these benefits come with significant challenges. Establishing causality in obesrvational data can be difficult due to the susceptibility to biases such as confounding bias, selection bias, and measurement bias, as described by Hammerton and Munafò (2021). For example, in the case of iFood, selection bias may arise if customers who are already frequent users are more likely to subscribe, making it harder to isolate

the true impact of the subscription program. Measurement bias could occur if certain customer behaviors, such as offline orders, are not adequately captured in the dataset.

This study relies on observational data provided by the company, as conducting randomized tests is expensive and often impractical. Despite these limitations, the richness of observational data provides an opportunity to uncover valuable insights about customer behavior and treatment effects. By employing advanced statistical methods, such as propensity score matching and causal machine learning models, this research seeks to mitigate biases and confounding factors, enhancing the reliability of the findings.

While challenges remain, leveraging observational data within these constraints allows the company to make informed decisions based on empirical evidence. For iFood, this approach is particularly valuable for understanding the heterogeneous effects of its subscription program and for refining its customer engagement strategies.

## 2.5 Uplift Models

In observational studies, such as the analysis of iFood's subscription program, estimating the causal impact of treatments on outcomes is challenging due to biases. Uplift models provide a way to address this by focusing on identifying the incremental impact of a treatment at the individual level.

Unlike traditional models, which predict overall outcomes (such as whether a customer will make a purchase), uplift models compare two scenarios for the same individual: one where the person receives the treatment (for example, subscribes to the program) and another where they do not. The goal is to estimate the difference in behaviour between these two scenarios, which represents the treatment's true impact on that individual.

As reviewed by Zhang et al. (2022), estimating the Conditional Average Treatment Effect (CATE) is a key task in causal inference, as it measures how the effect of a treatment varies across individuals or subgroups. This is particularly useful in areas like targeted marketing, where knowing which customers are most responsive to a campaign can help optimize resources and strategies.

Several methods for CATE estimation are discussed in the literature, but this study focuses on three widely used metalearners: Single Model (S-Learner), Two Models (T-Learner), and Cross Models (X-Learner). These methods were selected for their simplicity, theoretical foundation, and widespread use in causal inference tasks. Among them, the X-Learner stands out for its ability to handle imbalanced datasets and incorporate propensity scores, making it particularly suitable for observational data. The following provides a brief overview of these methods:

1. **Single Model Approach (S-Learner)**: This method uses a single supervised learning model, where the treatment indicator is included as an additional input

feature. While simple to implement, it may produce biased estimates when the relationship between treatment and outcome differs significantly between treated and control groups.

2. **Two Models Approach (T-Learner)**: This approach trains separate models for the treated and control groups, with the CATE estimated as the difference between their predicted outcomes. Although it allows flexibility in modeling each group, it may fail to leverage shared information between the groups effectively.

3. **Cross Models Approach (X-Learner)**: This method builds on the T-Learner by incorporating information from both groups into the estimation process, making it particularly effective when there is an imbalance between treated and control samples. It often achieves better accuracy in observational studies.

The next section will detail the implementation of these meta-learners in this study, including their setup and evaluation. These methods were selected based on their strong theoretical foundation and practical applicability to the dataset used in this research.

### 2.5.1 Metalearners

Meta-learners are flexible approaches widely used in causal inference to estimate the conditional average treatment effect (CATE). These methods decompose the task of estimating CATE into multiple prediction problems, which are solved using machine learning models, known as base-learners. The results are then combined to generate CATE estimates (Kunzel et al., 2019). Below are three common meta-learners: S-Learner, T-Learner, and X-Learner.

#### 2.5.1.1 Single Model (S-Learner)

The S-Learner uses a single model to estimate the observed outcome ($Y$) as a function of the covariates ($X$) and the treatment indicator ($T$). The model generates two predictions: one assuming $T = 1$ and another with $T = 0$. The difference between these predictions provides the CATE:

$$\mu(x, t) = \mathbb{E}[Y \mid X = x, T = t] \tag{2.9}$$

$$\tau(x) = \mu(x, 1) - \mu(x, 0) \tag{2.10}$$

In the context of iFood, the S-Learner could estimate the impact of the subscription program on customer spending by predicting spending under both treatment and control scenarios for each customer. While simple to implement, the S-Learner may introduce bias when the treatment ($T$) is not highly predictive, particularly when regularized models are used (Kunzel et al., 2019).

#### 2.5.1.2 Two Models (T-Learner)

The T-Learner fits two separate models to estimate potential outcomes for the treatment ($T = 1$) and control ($T = 0$) groups. These estimates are then used to calculate the CATE:

$$\mu^1(x) = \mathbb{E}[Y \mid X = x, T = 1] \tag{2.11}$$

$$\mu^0(x) = \mathbb{E}[Y \mid X = x, T = 0] \tag{2.12}$$

$$\tau(x) = \mu^1(x) - \mu^0(x) \tag{2.13}$$

For iFood, the T-Learner could be effective in modelling customer behaviour separately for subscribers and non-subscribers, especially when the CATE has a complex form. However, it may face instability in scenarios with imbalanced groups or small sample sizes (Salditt et al., 2024).

#### 2.5.1.3 Cross Models (X-Learner)

The X-Learner, proposed by Kunzel et al. (2019), is a robust approach that combines initial estimates of potential outcomes (as in the T-Learner) with pseudo-outcomes. It is especially useful in scenarios where treatment and control groups are imbalanced. Its main steps include:

1. Estimation of potential outcomes:

$$\mu^1(x), \quad \mu^0(x) \tag{2.14}$$

2. Calculation of pseudo-outcomes:

$$\psi(X) = \begin{cases} Y - \mu^0(X), & \text{if } T = 1 \\ \mu^1(X) - Y, & \text{if } T = 0 \end{cases} \tag{2.15}$$

3. Modelling pseudo-outcomes for each group:

$$\tau^0(x), \quad \tau^1(x) \tag{2.16}$$

4. Weighted combination using the propensity score ($\pi(x)$):

$$\tau(x) = \pi(x)\tau^0(x) + (1 - \pi(x))\tau^1(x) \tag{2.17}$$

The X-Learner is particularly effective in scenarios with partial violations of the positivity assumption or highly imbalanced groups, as it weights the most reliable estimates (Kunzel et al., 2019; Salditt et al., 2024).

These three meta-learners present complementary approaches and should be chosen based on the specific characteristics of the data and application context. The literature highlights that more complex methods, such as the X-Learner, often outperform simpler models in observational settings (Kunzel et al., 2019; Salditt et al., 2024).

## 2.6 CATE Evaluation

Evaluating uplift models is challenging because the true values of heterogeneous treatment effects (CATE) are not directly observable. Therefore, indirect metrics are necessary to assess the model's effectiveness in identifying individuals most likely to respond to treatment. Metrics such as **buckets**, **uplift curves**, and **Qini measures** are widely used and can be adapted to incorporate $\hat{\tau}(x)$ estimates provided by causal inference models, as discussed in Belbahri et al., 2020.

### 2.6.1 Uplift Buckets for CATE

Uplift buckets evaluate model performance by dividing individuals into groups (buckets) sorted by their estimated $\hat{\tau}(x)$, typically using quantiles like deciles or quartiles. For each bin, the **uplift** is calculated as the difference between the average outcomes of the treatment and control groups within that bin:

$$u_k = \frac{1}{n_{tk}} \sum_{i \in T_k} Y_i - \frac{1}{n_{ck}} \sum_{i \in C_k} Y_i, \tag{2.18}$$

where:

1. $n_{tk}$ and $n_{ck}$ are the number of individuals in the treatment and control groups in bin $k$, respectively;
2. $T_k$ and $C_k$ are the sets of individuals in the treatment and control groups for bin $k$;
3. $Y_i$ represents the observed outcome for an individual $i$.

This approach provides a direct measure of the difference in observed outcomes (uplift) between the treatment and control groups within each bin.

By aggregating results across bucket, this method highlights how effectively the model prioritizes individuals who benefit the most from the treatment. For example, higher uplift values in the top buckets indicate that the model successfully identifies individuals with the highest treatment effect.

### 2.6.2 Uplift Curves and Cumulative Increment

**Uplift curves** are constructed by accumulating $\hat{\tau}(x)$ estimates for individuals sorted in descending order by their estimated effect. The curve displays the cumulative treatment effect as a function of the proportion of the treated population ($\phi$):

$$g(\phi) = \frac{\sum_{i \in N_\phi} \hat{\tau}(x_i)}{\sum_{i \in N_\phi} t_i}, \tag{2.19}$$

where $N_\phi$ represents the set of individuals with the highest $\hat{\tau}(x)$ values for the top $\phi$-fraction of the population.

The area under the curve compares the model's performance to a random targeting strategy, with more effective models achieving greater cumulative gains.

### 2.6.3 Qini Coefficient

The **Qini curve** is a generalization of the uplift curve, incorporating normalized relative increments. The Qini coefficient is defined as the area between the model's curve and the random targeting line:

$$q = \int_0^1 \big(g(\phi) - \phi \cdot g(1)\big)\, d\phi, \tag{2.20}$$

where $g(1)$ represents the total cumulative treatment effect. Models with higher Qini coefficients better identify persuadable individuals, enabling more effective targeting decisions (Radcliffe & Surry, 2012).

## 2.7 Summary of Literature Review

This chapter provided a comprehensive review of the key concepts and methodologies relevant to this study. It began by discussing the role of subscription programs in fostering customer loyalty and engagement, highlighting their growing importance in modern business strategies. The challenges associated with measuring their impact, particularly in observational settings, were then explored in detail.

The discussion introduced the framework of causal inference, emphasizing the foundational assumptions such as consistency, no-interference, conditional ignorability, and positivity. The interplay between these assumptions, particularly the trade-off between conditional ignorability and positivity, was highlighted as critical for ensuring robust causal estimates. Additionally, the challenges posed by observational data, including confounding biases and selection issues, were addressed, alongside methods to mitigate these issues using parametric and non-parametric approaches.

Advanced modeling techniques, such as uplift models and meta-learners, were also reviewed, showcasing their applicability for estimating heterogeneous treatment effects (CATE) and their relevance to the context of iFood's subscription program. Lastly, evaluation metrics like uplift bins, uplift curves, and Qini measures were discussed as essential tools for assessing the performance of causal models.

Overall, this chapter established the theoretical foundation for the methodological approach presented in the next chapter. By bridging the theoretical and practical aspects of causal inference, it set the stage for a deeper exploration of the methods and data used to measure the impact of subscription programs.

# 3

# Empirical Strategy

In this section, the empirical strategy applied in the development of the research will be described. The following points will be addressed: the process of generating the data used for modelling, the modelling workflow, and the evaluation metrics employed. In the subsequent section, the results obtained from the application of the methodology will be presented.

As previously mentioned, this study aims to understand and analyse the heterogeneous effects on users' purchasing behaviour when participating in a loyalty program. The main goal is to identify characteristics that allow for the segmentation of users into distinct groups, enabling the identification of those who respond best to the program. This approach seeks to try differentiated strategies for these groups, optimising the investments made by both the company and the users in the context of the loyalty program.

## 3.1 Data

The data used in this study is observational, meaning it wasn't obtained from a randomised controlled experiment. The goal is to measure the influence of a loyalty program on the customer's purchase behaviour. To separate the dataset into control and test groups, the following approach was adopted:

1. a user is considered part of the treatment group if it has an active subscription in the analysis month and did not have any subscriptions in the previous three months.

2. a user is considered part of the control group if it does not have a subscription in the analysis month and did not have any subscriptions in the previous three months (Figure 3.1).

This method allows for a more effective separation of the two users groups, resulting in an observational dataset suitable for causal analysis with a clear distinction between the control and variant groups.

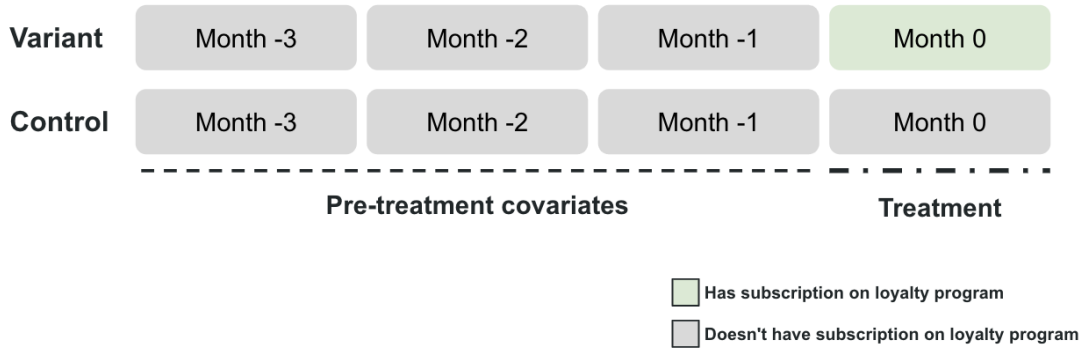



Figure 3.1: Diagram of Variant and Control Data

The users selected for analysis are those who are active on the platform, meaning they have some interaction with the application in the analysis month. Additionally, only users who have been on the platform for at least three months are considered. This choice ensures that the necessary variables are complete for analysis, as users with less than three months of activity would not have values available for all analysed variables.

The pre-treatment data include different information on user consumption habits and to ensure data security, the names of the variables will be omitted, and the results will be rescaled to mask the real effect and protect the company's data.

The dataset is divided into two main sets: training and test. The training set have 80% of the data available and is used to train the models. Within this set, a 50 fold cross-validation procedure is applied to ensure a more robust validation of the model. The testing set represents 20% of the data and is used to make the final evaluation of the models.

The dataset contains a significant volume of data for both control and treatment groups. However, an imbalance is observed, with fewer users adhering to the loyalty program compared to the total population. This imbalance introduces challenges in estimating treatment effects accurately, as it can lead to biases and reduce the reliability of causal inferences. Specifically, imbalances may increase the likelihood of overfitting, as models might learn patterns dominated by the control group, and reduce the generalizability of findings to the treatment group. Table 3.1 summarises the distribution across the total dataset, as well as the training and testing datasets:

| Dataset | Control | Treatment | Treatment Percentage [%] |
|---|---|---|---|
| Total | 15,898,431 | 729,555 | 4.38 |
| Training | 12,719,249 | 583,139 | 4.38 |
| Test | 3,179,182 | 146,416 | 4.4 |

Table 3.1: Dataset distribution across control and treatment groups.

The dataset consists of 13 numerical covariates, and after preprocessing, no missing

values remain. To ensure the quality and representativeness of the training set, additional checks were conducted, including an analysis of covariate balance between the treatment and control groups.

## 3.2 Modelling Process

The modelling process utilised the `causalml`, `lightgbm`, `scikit-learn` and `fklearn` libraries for the implementation and validation of the models. The meta-learners described in the literature review were selected to estimate the Conditional Average Treatment Effect (CATE). The three meta-learners employed were:

1. **S-Learner**: Utilises a single regression model (`LightGBM Regressor`) that incorporates the treatment indicator as a variable to estimate the potential outcomes.
2. **T-Learner**: Employs two regression models (`LightGBM Regressor`), one for estimating the potential outcomes of the treatment group and another for the control group.
3. **X-Learner**: Combines two regression models (`LightGBM Regressor`) to estimate the potential outcomes for the treatment and control groups, along with a classification model (`LightGBM Classifier`) to estimate propensity scores, which are used for weighting in the final adjustments.

### 3.2.1 Steps in the Modelling Process

The modelling process was structured into six main steps, described as follows:

**1. Data Preparation** This step involved the following procedures:

- **Initial dataset analysis:** Identification and removal of null or inconsistent values.
- **Data scaling:** Normalisation and standardisation of values to ensure consistency during model training.
- **Variable anonymisation:** Replacement of original variable names with generic identifiers to ensure data privacy.
- **Data splitting:** Division of the dataset into training and testing subsets, ensuring representative samples for both sets.

**2. Propensity Score Matching** To reduce potential bias in the covariates between the treatment and control groups, propensity score matching (PSM) was applied to the training dataset. This process created a debiased training dataset with improved balance between the groups, ensuring that causal inference assumptions were better met. This step is crucial for the validity of the results and will be discussed in detail in the following chapter.

**3. Data Exploration** The dataset was explored to verify the balance of covariates between the treatment and control groups, ensuring that the assumptions required for causal inference were not violated. The positivity assumption was also assessed by analysing the distribution of propensity scores. When violations were identified, adjustments were made to ensure compliance with this assumption.

**4. Base-Learner Evaluation** Before proceeding to model training, an additional step was conducted to evaluate different base learners across datasets designed for use within the meta-learner architectures. This evaluation was performed to identify the best-performing base learner for each meta-learner. The base learners considered included `LightGBM Regressor`, `LightGBM Classifier`, `Linear Regression`, `Logistic Regression`, and `Random Forests`. The performance of these models was assessed using metrics such as mean squared error and accuracy, depending on the learner type. The results of this step guided the selection of the most suitable base learner for each meta-learner architecture.

**Metrics for Evaluation:** The evaluation of base learners was based on specific metrics suited to the tasks performed by the models—regression or classification. Each metric was chosen for its importance in measuring the accuracy and reliability of the models within the metalearner framework:

- **MSE (Mean Squared Error)**: Measures the average squared error between predicted and actual values. It is sensitive to large errors, making it effective for regression tasks.
- **$R^2$ (Coefficient of Determination)**: Indicates the proportion of variance in the target variable explained by the model. A higher $R^2$ shows better fit.
- **AUC (Area Under the Curve)**: Evaluates the ability of classification models to distinguish between classes, especially useful for imbalanced datasets.
- **F1 Score**: Balances precision and recall into a single metric, providing insight into classification performance where both false positives and false negatives are important.

The combination of these metrics ensures that the selected base learners provide reliable predictions for both regression and classification tasks.

**5. Model Training** With the base learner chosen in the previous step, the training process was conducted using 50-fold cross-validation. This approach leveraged the dataset size to minimise the risk of overfitting and enhance generalisation. The selected base learners were integrated into the meta-learner frameworks (S-Learner, T-Learner, and X-Learner) and trained across the folds. The main evaluation metric during training

was the Qini coefficient, ensuring that the models were optimised for uplift modelling. This robust training procedure ensured the reliability of the estimated Conditional Average Treatment Effects (CATE).

**6. Model Evaluation** The models were evaluated in two stages:

- **Training sets evaluation:** The Qini metric was used to assess initial model performance across the cross-validation folds, providing a preliminary comparison between the meta-learners.
- **Test set evaluation:** The final evaluation was conducted on the test dataset. In addition to the Qini metric, the following analyses were performed:
  - **Bucket plots:** Used to assess cumulative gains across user segments ranked by $\hat{\tau}(x)$.
  - **Uplift curves:** Applied to evaluate the cumulative performance of the models in identifying users who benefit most from the treatment.

### 3.2.2 Rationale for Modelling Choices

The modelling process was designed to follow best practices in causal inference and machine learning. The application of propensity score matching aimed to reduce biases in the covariates and improve the comparability between treatment and control groups. Additionally, testing multiple base learners allowed the selection of the most effective combinations for the meta-learners. The chosen evaluation metrics ensured a comprehensive analysis of both the predictive performance and the practical impact of the models. These choices align the methodology with the study's goal of understanding the heterogeneous impact of the loyalty program on user behaviour.

# 4

# Results

In this chapter, the results of the analysis are presented and discussed. The chapter is divided into sections that explore different aspects of the data and modelling outcomes. First, an assesment of the dataset is provided, focusing on the distribution of the control and treatment groups and the challenges posed by their imbalance. Next, the performance of the selected metalearners is evaluated using various metrics, including the Qini coefficient and uplift curves. Finally, the implications of the findings are discussed in the context of the loyalty program and its heterogeneous effects on user behaviour.

The results presented in this chapter aim to validate the methodology and provide insights into the effectiveness of the loyalty program, supporting the study's main goal of identifying user segments with distinct responses to the program.

## 4.1 Assessment of the Data

Figure 4.1 illustrates the standardised mean differences of the covariates between the treatment and control groups. As recommended by Salditt et al. (2024), standardised mean differences below 0.1 are generally considered acceptable for indicating good balance. However, eight covariates (X01, X02, X05, X07, X08, X11, X12 and X13) exceed this threshold, highlighting potential risks that could introduce bias into the estimation of treatment effects.

The covariates include variables representing user characteristics, as well as behavioral data related to purchasing patterns and session interactions. While the exact nature of these variables cannot be disclosed for privacy reasons, they collectively provide a comprehensive view of user activity and profile differences between the treatment and control groups.

To address this issue, a propensity score matching (PSM) was applied to reduce the differences between the treatment and control groups. Figure 4.2 demonstrates that, after matching, all 13 covariates have normalised differences below the 0.1 threshold. This indicates that the matched dataset has achieved sufficient balance, mitigating risks

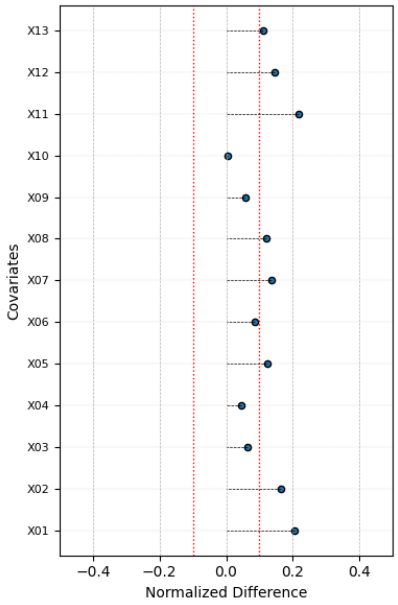



Figure 4.1: Normalised Standardised Mean Differences of Covariates

associated with imbalanced covariates.

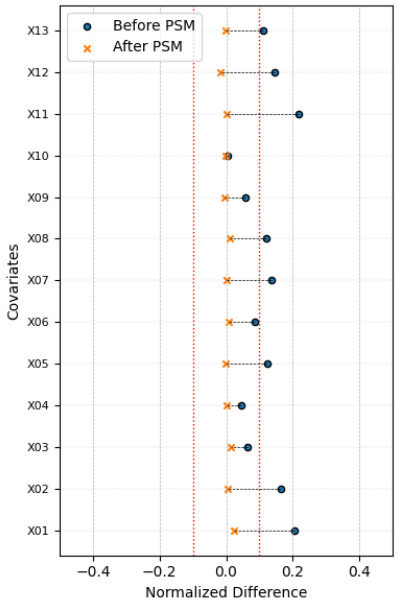



Figure 4.2: Normalised Standardised Mean Differences of Covariates After Propensity Score Matching

As highlighted in the literature review, the positivity assumption is a crucial condition for valid causal inference. This assumption ensures that all subjects have a non-zero probability of receiving either the treatment or control, given their covariates. Zhu et al. (2021) discusses a technique to address violations of the positivity assumption, known as trimming. This approach identifies subjects whose propensity scores fall outside a specified range, such as $[0.1, 0.9]$, and removes them from the dataset.

Although this reduces the effective sample size and may increase variance, it improves the validity of causal inferences by focusing on a subset of the population where the positivity assumption holds.

In the present study, the positivity assumption was tested, and the results are shown in Figure 4.3. In the original training dataset, all propensity score values were below the threshold of 0.1, indicating a clear violation of the positivity assumption. However, after applying PSM, the propensity score distribution shifted significantly, with values clustering around 0.5. This adjustment reduces bias and ensures the dataset adheres more closely to the positivity assumption, enabling more reliable causal estimates.

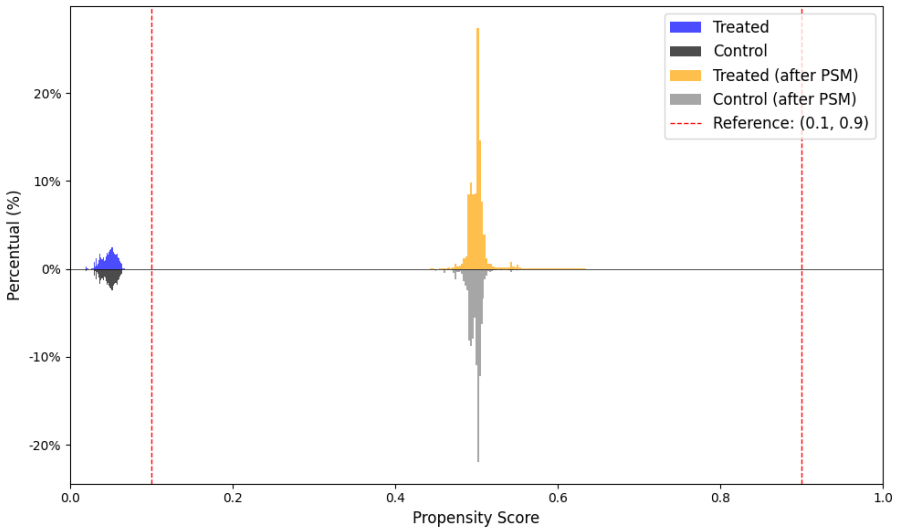
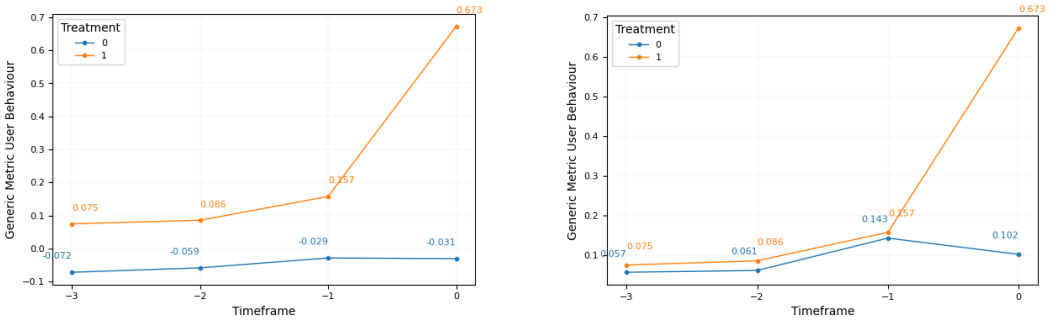



Figure 4.3: Positivity Assumption Check Before and After Propensity Score Matching

Finally, an analysis of the outcome variable was performed to evaluate bias between the treatment and control groups. Figure 4.4 compares the outcome bias before and after PSM. Before matching, a clear bias was evident between the groups during the pre-treatment periods. After matching, the reduction in outcome bias is apparent, demonstrating the effectiveness of PSM in addressing this issue.

Such imbalances underscore the importance of carefully addressing disparities between groups during the analysis process. Imbalances in covariates can introduce significant bias, compromising the validity of causal inferences. Methods such as propensity score matching (PSM) play a crucial role in mitigating these risks and ensuring robust and reliable conclusions.

After addressing the imbalance using PSM, a new denoised training dataset was created. This adjusted dataset achieves better balance between the control and treatment groups, substantially reducing bias and providing a more reliable foundation for analysis. The denoised dataset is especially critical for improving the reliability of causal estimates by addressing both covariate imbalances and violations of assumptions such as positivity.




(a) Outcome Bias Before Propensity Score Matching

(b) Outcome Bias After Propensity Score Matching

Figure 4.4: Outcome Bias Before and After Propensity Score Matching: (a) Before PSM, and (b) After PSM.

Table 4.1 summarises the distribution across the total dataset, the original training and testing datasets, and the denoised training dataset:

| Dataset | Control | Treatment | Treatment Percentage [%] |
|---|---|---|---|
| Total | 15,898,431 | 729,555 | 4.38 |
| Training | 12,719,249 | 583,139 | 4.38 |
| Training Denoised | 583,139 | 583,139 | 50.0 |
| Test | 3,179,182 | 146,416 | 4.4 |

Table 4.1: Dataset distribution across control and treatment groups (After PSM).

This distribution highlights the effectiveness of PSM in creating a balanced training dataset. By equalizing the number of observations in the treatment and control groups within the denoised dataset, the methodology mitigates potential biases and improves the reliability of subsequent causal analyses. Furthermore, the balanced dataset provides a robust basis for testing causal hypotheses while adhering to key statistical assumptions.

## 4.2 Assessment of Base Learners for Metalearners

As discussed previously, the architecture of metalearners relies on machine learning models in each of their stages. These base learners are fundamental for estimating outcomes and treatment effects, and their performance directly impacts the reliability and accuracy of the metalearners. This section evaluates the performance of various machine learning models used as base learners for the S-Learner, T-Learner, and X-Learner.

The **S-Learner** models the outcome $y$ directly by incorporating the covariates and the treatment indicator ($T$) as input variables. Since the target variable $y$ is numerical, regression models are employed as the base learners.

The **T-Learner**, in contrast, uses two separate regression models: one to estimate $y$ for the treatment group ($T = 1$) and another for the control group ($T = 0$).

The **X-Learner** builds upon the T-Learner by introducing an additional stage where a regression model is trained to estimate the Conditional Average Treatment Effect (CATE) from the outputs of the first stage. Additionally, a classification model is used to estimate the propensity score ($P(T = 1|X)$), which is incorporated into the final calculation of CATE.

### 4.2.1 Datasets for Training Base Learners

To evaluate the performance of base learners, models were trained on distinct datasets designed for regression and classification tasks:

- **Regression datasets**:
  - $Y(T = 0)|X$: Outcome for the control group conditioned on covariates.
  - $Y(T = 1)|X$: Outcome for the treatment group conditioned on covariates.
  - $Y(X, T)$: Outcome modeled using both covariates and the treatment indicator as inputs.
  - $Y|X$: Outcome modeled using covariates without the treatment indicator.
- **Classification dataset**:
  - $T|X$: Treatment assignment modeled as a classification problem using covariates.

### 4.2.2 Model Selection and Evaluation

For each dataset, three families of models were evaluated to compare performance:

- **Regression models**: LightGBM Regressor, Linear Regression, and Random Forest Regressor.
- **Classification models**: LightGBM Classifier, Logistic Regression, and Random Forest Classifier.

The results for the regression datasets are presented in Table 4.2. It is evident that the LightGBM Regressor outperformed other models across all datasets.

The results for the classification task of estimating the propensity score ($P(T = 1|X)$) are shown in Table 4.3. The LightGBM Classifier achieved the best performance, surpassing both Logistic Regression and Random Forest Classifier.

| Model | Dataset | MSE | R2 |
|---|---|---|---|
| LightGBM | $Y(T=0)\|X$ | **3.638** (0.027) | **0.707** (0.002) |
| Linear Regression | $Y(T=0)\|X$ | 4.755 (0.060) | 0.617 (0.003) |
| Random Forest | $Y(T=0)\|X$ | 4.893 (0.066) | 0.606 (0.003) |
| LightGBM | $Y(T=1)\|X$ | **7.586** (0.050) | **0.577** (0.003) |
| Linear Regression | $Y(T=1)\|X$ | 8.305 (0.069) | 0.537 (0.003) |
| Random Forest | $Y(T=1)\|X$ | 8.701 (0.076) | 0.515 (0.003) |
| LightGBM | $Y(X,T)$ | **5.839** (0.027) | **0.635** (0.002) |
| Linear Regression | $Y(X,T)$ | 6.562 (0.042) | 0.589 (0.002) |
| Random Forest | $Y(X,T)$ | 6.931 (0.048) | 0.566 (0.002) |
| LightGBM | $Y\|X$ | **6.668** (0.029) | **0.583** (0.002) |
| Linear Regression | $Y\|X$ | 7.402 (0.044) | 0.537 (0.002) |
| Random Forest | $Y\|X$ | 7.639 (0.047) | 0.522 (0.002) |

Table 4.2: Regression base learners results.

| Model | AUC | F1 Score |
|---|---|---|
| LightGBM | **0.635** (0.001) | **0.586** (0.001) |
| Logistic Regression | 0.533 (0.001) | 0.523 (0.001) |
| Random Forest | 0.550 (0.002) | 0.526 (0.003) |

Table 4.3: Classification base learners results for propensity score estimation.

### 4.2.3 Final Selection of Base Learners

Based on these evaluations, the LightGBM Regressor and LightGBM Classifier were selected as the base learners for the metalearners due to their consistent superior performance across all datasets. These models offer the advantage of handling complex relationships and large datasets efficiently, making them well-suited for the tasks required by the S-Learner, T-Learner, and X-Learner architectures. The adoption of LightGBM as the base learner is expected to enhance the accuracy and robustness of the causal effect estimates provided by the metalearners.

## 4.3 Assessment of Metalearners

After training the three metalearners using 50-fold cross-validation on the two training datasets (original and debiased), the Figure 4.5 indicate that the models trained on the debiased dataset performed better during cross-validation. This suggests that the removal of biases from the dataset improved the internal validation process. However, this improvement raises concerns about potential overfitting. The original dataset exhibited a significant imbalance between the treatment and control groups. While the application of Propensity Score Matching (PSM) reduced this imbalance by creating more balanced groups, it also resulted in a smaller training dataset. This reduction in

dataset size may limit the models' ability to generalise to unseen data, as fewer samples were available during training.

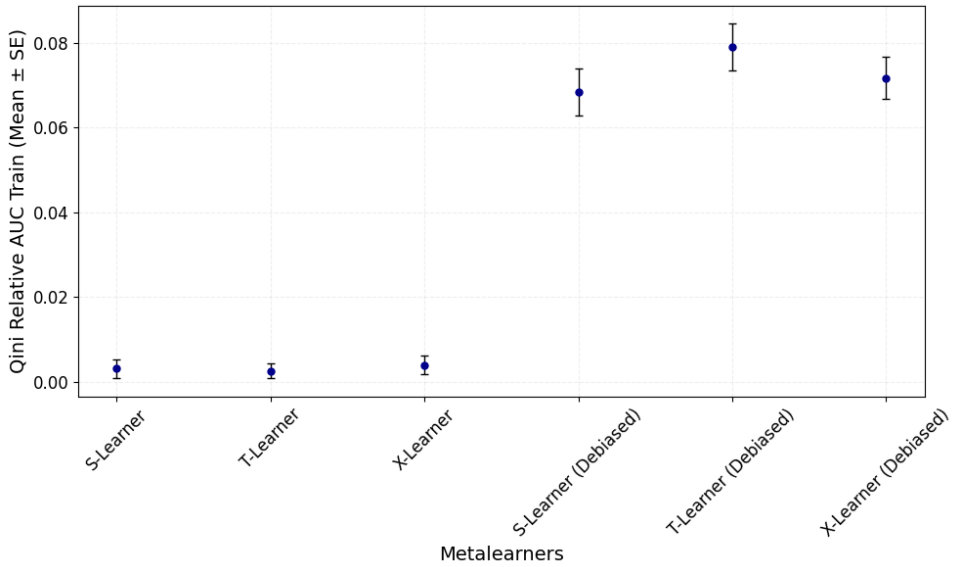



Figure 4.5: Training Set AUC Gain for Metalearners

To further validate the performance of the models, evaluations were conducted on the test dataset. The results, as shown in the Figure 4.6, indicate that the X-Learner trained on the debiased dataset achieved the best performance. Interestingly, the second-best model was the X-Learner trained on the original (biased) dataset.

This result can be explained by the design of the X-Learner, which is effective at using information from both the treatment and control groups. Its second stage adjusts the CATE predictions using propensity scores, helping to reduce the impact of imbalance in the data. When trained on the debiased dataset, the improved balance between the groups likely allowed the model to make more accurate and generalizable predictions, leading to its strong performance.

The fact that the X-Learner trained on the original dataset also performed well suggests that its design can handle some level of imbalance. This shows that the X-Learner is more resilient than other meta-learners when working with datasets that are not perfectly balanced, making it suitable for observational studies where such challenges are common.

However, this behaviour was not consistent in the test dataset when compared to the training dataset. This suggests a potential risk of overfitting, particularly for the X-Learner, which may have captured patterns specific to the training data. The high imbalance between the treatment and control groups in the test dataset likely contributed to this issue, making it more difficult for the models to perform well across all subgroups.

This discrepancy highlights the importance of having a test dataset that reflects the target population. When the test dataset is heavily imbalanced, as in this case,

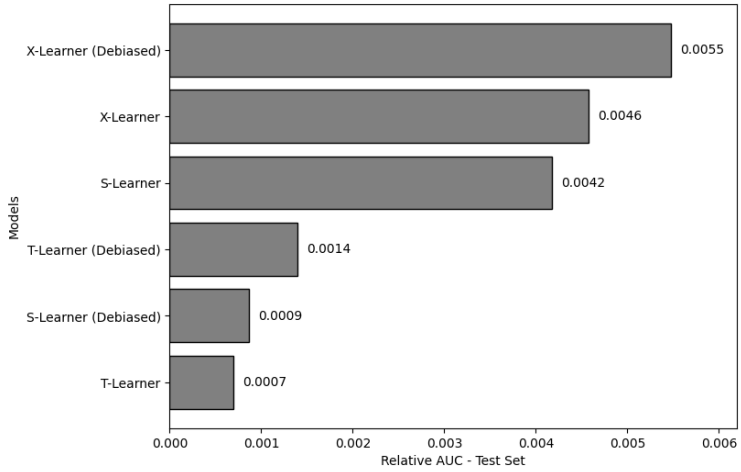



Figure 4.6: Test Set AUC Gain for Metalearners

models may fail to generalise effectively, despite strong performance on the training data. Addressing this issue could involve strategies such as rebalancing the test dataset or using simpler models less prone to overfitting.

The results also reveal that metalearners without propensity scores in their architecture consistently performed worse. This underscores the critical role of propensity scores in accounting for confounding factors in observational data. By incorporating propensity score information, models can better estimate CATE, even in datasets where treatment and control groups differ significantly.

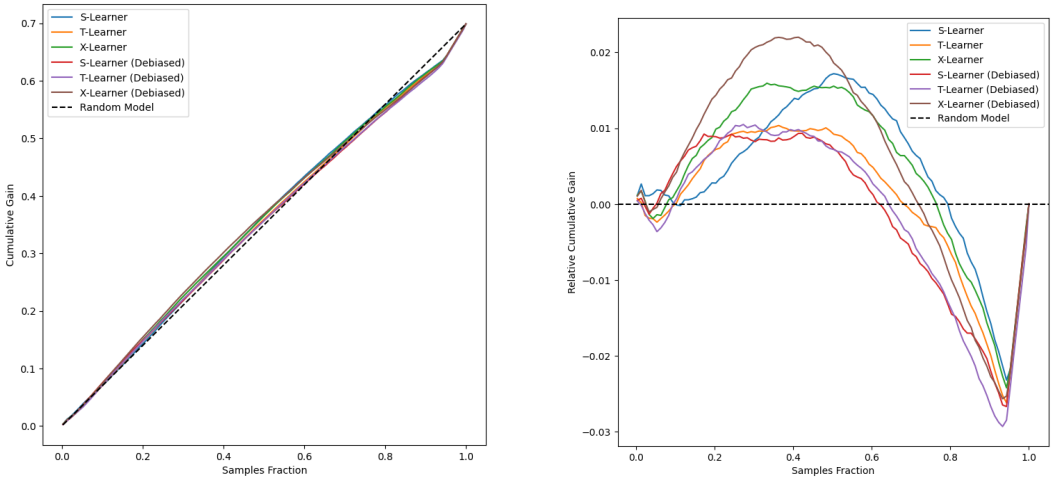
While PSM improves balance between treatment and control groups, it introduces a trade-off. The exclusion of data points that do not meet balancing criteria reduces the dataset size, potentially leading to a loss of diversity and information. This reduction may affect the generalisation capacity of the models, especially in datasets with limited observations.

Interestingly, the performance of the X-Learner trained on the original dataset reinforces its ability to address some of these challenges. This can be attributed to its second stage, which integrates propensity scores to refine CATE predictions. However, the results also suggest that using a well-balanced dataset further enhances its performance, as observed with the debiased dataset. This finding emphasises the importance of both robust model architecture and careful dataset preparation in improving causal inference outcomes.

For further analysis, the test dataset was evaluated using Qini curves. Figure 4.7 presents both the cumulative gain curve and the relative cumulative gain curve. These curves are important tools for understanding the effectiveness of the model in identifying heterogeneous treatment effects across the population.

It is noticeable that in the final portions of the curves, a negative cumulative gain is observed. This behaviour indicates that the model struggles to differentiate effectively between the treatment and control groups in the least represented portions of the

dataset. This limitation is likely due to the imbalance in the original dataset, which affects the model's ability to generalise to subgroups with fewer samples.




(a) Cumulative Gain Curve (b) Relative Cumulative Gain Curve

Figure 4.7: Qini Curves for the Test Dataset: (a) Cumulative Gain Curve and (b) Relative Cumulative Gain Curve.

This pattern is further supported by the bins plot in Figure 4.8. Each bin represents a subset of the data ordered by the model's predicted heterogeneous treatment effect. The plot shows the estimated ATE decreasing consistently from buckets 1 to 9, which highlights the presence of heterogeneous treatment effects. This behavior indicates that the model captures meaningful variations in the treatment effect across subpopulations. However, bucket 10 presents an anomalous result, with the highest ATE value, contrary to the decreasing trend. This inconsistency aligns with the observations in the relative cumulative gain curve, where the final samples demonstrate poor estimation performance.

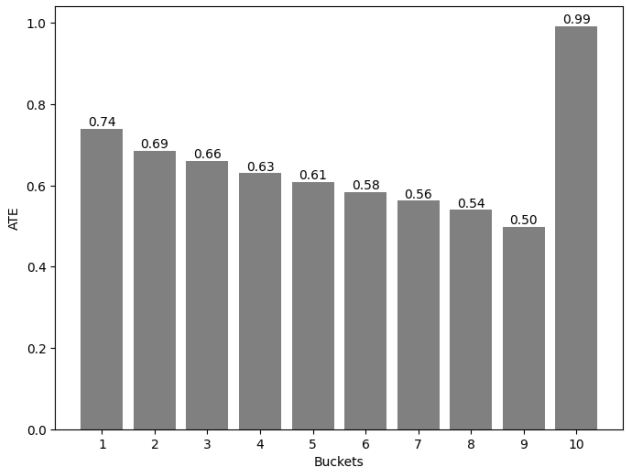



Figure 4.8: Uplift Bucket Plot

Figure 4.9 illustrates the proportion of treatment samples across the 10 buckets.

As the bucket number increases, the proportion of treatment samples decreases, with bucket 10 showing the lowest proportion. This imbalance between treatment and control groups in the higher buckets limits the model's ability to estimate treatment effects reliably, as there is insufficient overlap between the groups in these regions.

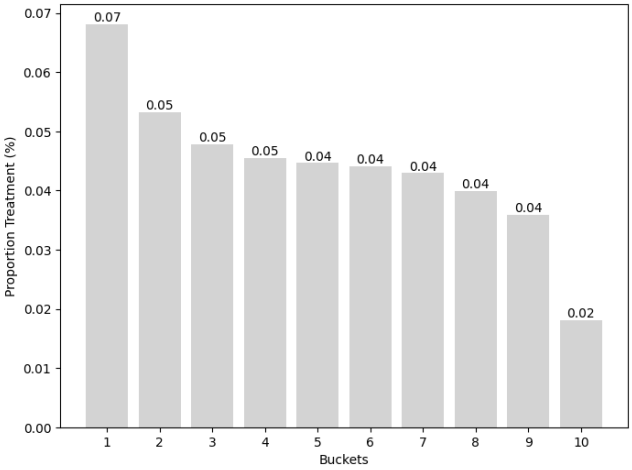



Figure 4.9: Proportion of Treatment Samples in Buckets

This trend is consistent with the negative cumulative gain observed in the Qini curve for the final samples. The reduced treatment representation in these subpopulations likely contributes to the model's declining performance, as it struggles to generalize effectively in regions with low treatment overlap.

#### 4.3.0.1 Discussion of Key Limitations

The poor performance in the final portions of the Qini and bucket plots can be attributed to several factors:

- **Dataset imbalance:** The treatment group is less represented, particularly in the tail regions of the dataset. This imbalance reduces the model's ability to reliably estimate the treatment effect for these samples, leading to higher uncertainty.

- **Limited support:** As the sample size in these regions decreases, the overlap between the treatment and control groups becomes limited, increasing the sensitivity of the estimates to noise.

- **Overfitting risks:** The application of Propensity Score Matching (PSM), while effective in balancing the dataset, reduces the overall sample size. This reduction may lead to less generalizable models, particularly for subpopulations at the tails of the distribution.

#### 4.3.0.2 Interpretation and Recommendations

The negative cumulative gain and the behavior in bucket 10 highlight the need for caution when interpreting the model's estimates for subgroups with lower representation:

- **Interpretation of negative values:** The negative values in the cumulative gain curve indicate that, for some subpopulations, the treatment effect is less favorable or even harmful. Alternatively, it may reflect higher uncertainty in the estimates due to limited comparability between treatment and control groups in these regions.
- **Focus on reliable regions:** The regions with higher density and more balanced treatment and control groups (e.g., buckets 1 to 9) should be prioritized for actionable insights. These regions provide more reliable estimates of the treatment effect.
- **Future improvements:** Future studies could address these limitations by applying stratification or alternative sampling techniques to improve the balance and representativeness of the dataset.

#### 4.3.0.3 Conclusion

Despite the limitations observed in the tail regions of the curves, the analysis demonstrates that the model captures meaningful heterogeneous treatment effects for the majority of the dataset. The decreasing trend in ATE across the first buckets aligns with expectations and reflects the model's ability to differentiate treatment effects across subpopulations. However, the negative cumulative gain and anomalous values in bucket 10 underline the importance of interpreting the results with caution, especially for less represented subgroups. These findings emphasize the need for further methodological refinements to enhance the robustness and generalizability of the model.

# 5

# Conclusion

This study investigated the evaluation of heterogeneous treatment effects in loyalty programs using methods like meta-learners and propensity score matching (PSM). The results provide valuable insights for improving strategies aimed at better understanding and serving customers, especially when using observational data.

The use of PSM was essential to reduce bias and create a more balanced dataset, ensuring that the analysis was based on reliable data. This adjustment made the results more trustworthy and helped meet important requirements for causal inference. Additionally, the analysis showed that models like LightGBM Regressor and Classifier performed well in estimating the differences in treatment effects across customer groups.

Meta-learners, especially the X-Learner, showed strong performance in identifying how the effects of loyalty programs vary for different groups. However, there were challenges in analysing groups with fewer data points, which highlights the need to be cautious when interpreting results for less-represented groups.

Combining PSM with meta-learners appears to be a promising approach, especially in cases where experiments are not feasible. While balancing the data improved the reliability of the results, it also reduced the dataset size, which made it harder to apply the models to all groups. These findings highlight the importance of carefully preparing data to ensure the models provide meaningful and accurate results.

The practical applications of this study are clear. Businesses can use these findings to focus on customer groups that respond more positively to loyalty programs, helping to design more effective strategies. While this research focused on loyalty programs, the methods can also be applied to other areas, such as assessing incentive programs or policy initiatives.

Future studies could explore advanced techniques, such as Double Machine Learning (DML), which applies machine learning models to adjust more flexibly for nonlinear confounding effects, reducing the need to assume specific data structures. This approach is particularly useful for causal analysis in observational datasets (Fuhr et al., 2024). DML could also help address challenges like the reduced dataset size caused by

balancing and maintain model robustness even with high-dimensional or sparse data. Additionally, common issues like class imbalance can be handled with methods such as stratified undersampling and calibration, which correct distortions and improve model accuracy (Nyberg and Klami, 2023). Testing these strategies in other contexts and datasets could confirm their benefits and broaden their use.

In conclusion, while this study highlights the value of combining PSM and meta-learners, it also points to the importance of addressing limitations such as class imbalance and dataset reduction. Future research should explore hybrid approaches, integrating techniques like DML and advanced balancing methods, to improve the accuracy and generalization of causal estimates. These advancements would not only strengthen the foundation for evaluating loyalty programs but also provide businesses with actionable insights to make more informed, data-driven decisions.

# Bibliography


Belbahri, M., Murua, A., Gandouet, O., & Nia, V. P. (2020, September). Qini-based Uplift Regression [arXiv:1911.12474 [stat]]. Retrieved 2024-11-21, from http://arxiv.org/abs/1911.12474 (cit. on p. 15).

Cunningham, S. (2021). *Causal inference: The mixtape* (First). Yale University Press. (Cit. on p. 7).

Facure, M. (2023). *Causal inference in python: Applying causal inference in the tech industry* (Fifth). O'Reilly Media. (Cit. on pp. 7, 8).

Fisher, R. A. (1935). *The design of experiments*. Hafner Press. (Cit. on p. 7).

Fuhr, J., Berens, P., & Papies, D. (2024, April). Estimating Causal Effects with Double Machine Learning – A Method Evaluation [arXiv:2403.14385 [stat]]. https://doi.org/10.48550/arXiv.2403.14385 (cit. on p. 35).

Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, *11*(1), 1. https://doi.org/10.2307/1905714 (cit. on p. 7).

Hammerton, G., & Munafò, M. R. (2021). Causal inference with observational data: The need for triangulation of evidence. *Psychological Medicine*, *51*(4), 563–578. https://doi.org/10.1017/S0033291720005127 (cit. on p. 11).

iFood. (2022). *Clube ifood: Descubra os benefícios!* Retrieved 2022-04-11, from https://blog-parceiros.ifood.com.br/clube-ifood/ (cit. on p. 1).

iFood. (2023). *O que é o ifood? conheça a história e a operação da empresa*. Retrieved 2023-03-03, from https://institucional.ifood.com.br/noticias/o-que-e-o-ifood/ (cit. on p. 2).

iFood. (2024a). *Efeito ifood na economia cresce e chega a 0,55% do pib*. Retrieved 2024-09-25, from https://institucional.ifood.com.br/estudos-e-pesquisas/pesquisa-fipe-ifood-2024/ (cit. on p. 2).

iFood. (2024b). *Sobre o ifood. nós entregamos mais do que pedidos. nosso propósito é alimentar o futuro.* Retrieved 2023-03-03, from https://institucional.ifood.com.br/sobre/ (cit. on p. 1).

Iyengar, R., Park, Y.-H., & Yu, Q. (2022). The Impact of Subscription Programs on Customer Purchases. *Journal of Marketing Research*, *59*(6), 1101–1119. https://doi.org/10.1177/00222437221080163 (cit. on pp. 5, 6).

Kunzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning [arXiv:1706.03461 [math, stat]]. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116 (cit. on pp. 13, 14).

Lin, C., & Bowman, D. (2022). The impact of introducing a customer loyalty program on category sales and profitability. *Journal of Retailing and Consumer Services*, *64*, 102769. https://doi.org/10.1016/j.jretconser.2021.102769 (cit. on p. 6).

Lourenço, J. M. (2021). *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

Nishio, K., & Hoshino, T. (2024). Quantifying the short- and long-term effects of promotional incentives in a loyalty program: Evidence from birthday rewards in a large retail company. *Journal of Retailing and Consumer Services*, *81*, 103957. https://doi.org/10.1016/j.jretconser.2024.103957 (cit. on p. 6).

Nyberg, O., & Klami, A. (2023). Exploring uplift modeling with high class imbalance. *Data Mining and Knowledge Discovery*, *37*(2), 736–766. https://doi.org/10.1007/s10618-023-00917-9 (cit. on p. 36).

Radcliffe, N. J., & Surry, P. D. (2012). Real-World Uplift Modelling with Significance-Based Uplift Trees (cit. on p. 16).

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350 (cit. on p. 7).

Salditt, M., Eckes, T., & Nestler, S. (2024). A Tutorial Introduction to Heterogeneous Treatment Effect Estimation with Meta-learners. *Administration and Policy in Mental Health and Mental Health Services Research*, *51*(5), 650–673. https://doi.org/10.1007/s10488-023-01303-9 (cit. on pp. 14, 23).

Vonk, M. C., Malekovic, N., Bäck, T., & Kononova, A. V. (2023). Disentangling causality: Assumptions in causal discovery and inference. *Artificial Intelligence Review*, *56*(9), 10613–10649. https://doi.org/10.1007/s10462-023-10411-9 (cit. on pp. 9, 11).

Zhang, W., Li, J., & Liu, L. (2022). A Unified Survey of Treatment Effect Heterogeneity Modelling and Uplift Modelling. *ACM Computing Surveys*, *54*(8), 1–36. https://doi.org/10.1145/3466818 (cit. on p. 12).

Zhu, Y., Hubbard, R. A., Chubak, J., Roy, J., & Mitra, N. (2021). Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiology and Drug Safety*, *30*(11), 1471–1485. https://doi.org/10.1002/pds.5338 (cit. on p. 24).

| | | |
|---|---|---|
| 2024 | Heterogeneous Treatment Effects in Loyalty Programs: A Study Case of Causal Inference Approach to Understanding Customer Behaviour | Jaime Kuei |