Subpopulation treatment effect Modeling : machine learning approaches to model treatment effect heterogeneity

Atef Shaar

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0058

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Informatique et Réseaux »

présentée et soutenue publiquement par

Atef SHAAR

le 14 December 2018

Subpopulation Treatment Effect Modeling : machine learning approaches to model treatment effect heterogeneity

Directeur de thèse : **Pr. Talel ABDESSALEM**

Jury :

Pr. Albert BIFET, Professeur à Télécom ParisTech	Président/Examinateur
Pr. Rokia MISSAOUI, Professeure à l'Univérsité du Quebec en Outaouais	Rapporteur
Pr. Stéphane BRESSAN, Professeur à National University of Singapore	Rapporteur
Pr. Olivier SEGARD, Directeur du département MMS, Télécom Business School	Examinateur
Pr. Talel ABDESSALEM, Professeur à Télécom ParisTech	Eencadrant de thèse
Pr. Hajer KEFI, Professeure à Paris School of Business	Co-encadrant de thèse

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

T H È S E

PhD Thesis

submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

of the École doctorale Informatique, Télécommunications et Électronique

(Paris)

Specialty: Computer Science and Network

Atef SHAAR

Subpopulation Treatment Effect Modeling: machine learning approaches to model treatment effect heterogeneity

Defence date: 14^{th} December 2018

Committee in charge:

President/Examiner:	Albert BIFET	-	Professor, Télécom ParisTech
Reviewers:	Rokia MISSAOUI	-	Professor, Univérsité du Quebec en Outaouais
	Stéphane BRESSAN	-	Professor, National University of Singapore
Examiner:	Olivier SEGARD	-	Professor, Télécom Business School
Supervisor:	Talel ABDESSALEM	-	Professor, Télécom ParisTech
Co-supervisor:	Hajer KEFI	-	Professor, Paris School of Business

Thèse

présentée pour obtenir le grade de docteur

de l'École doctorale Informatique, Télécommunications et Électronique

(Paris)

Spécialité : Informatique et Réseaux

Atef SHAAR

Approches d'apprentissage automatique pour la modélisation des effets d'un traitement sur une sous-population

Date de la soutenance : 14 Décembre 2018

Composition du jury :

Président/Examinateur :	Albert BIFET	-	Professeur, Télécom ParisTech
Rapporteurs :	Rokia MISSAOUI	-	Professeure, Univérsité du Quebec en Outaouais
	Stéphane BRESSAN	-	Professeur, National University of Singapore
Examinateur :	Olivier SEGARD	-	Professeur, Télécom Business School
Directeur de thèse :	Talel ABDESSALEM	-	Professeur, Télécom ParisTech
Co-directeur de thèse :	Hajer KEFI	-	Professeure, Paris School of Business

Abstract

Subpopulation treatment effect modeling (STEM) is a machine learning technique that is used to choose the optimal treatment (i.e., stimulus) for each subgroup. A critical challenge facing the STEM is information uncertainty. Data uncertainty exists due to the fundamental problem of causal inference, i.e., only a subset of treatments' responses are observed. In machine learning domain, specific binning techniques are applied to bypass the problem of uncertainty. However, one drawback of current STEM binning approaches is the poor handling of continuous, ordered, and time-series data variables, leading to unreliable and non-interpretable results.

In this thesis, first, we fill the gaps in the literature and propose a detailed study of current techniques. Second, we solve STEM shortcomings regarding uncertainty in the data by proposing subpopulation treatment effect sliding trees. Third, we propose the subpopulation treatment effect neighborhood random forests to minimize the effect of noise in data. Fourth, we address the problem of disturbance in data by proposing the balanced reflective uplift modeling technique. We evaluate the performance of the proposed solutions using simulated and real datasets, and we show how our approaches outperform other methods in terms of Qini and Spearman's rank correlated coefficient.

Résumé

La modélisation des effets de traitement de sous-population (STEM) est une technique d'apprentissage automatique utilisée pour choisir le traitement optimal (c'est-à-dire un stimulus) pour chaque sous-groupe. L'incertitude de l'information est un problème critique pour le STEM. L'incertitude sur les données existe en raison du problème fondamental de l'inférence causale, c'est-à-dire que seul un sous-ensemble des réponses des traitements est observé. Dans le domaine de l'apprentissage automatique, des techniques de tri spécifiques sont appliquées pour contourner le problème de l'incertitude. Cependant, l'un des inconvénients des méthodes de tri STEM actuelles est le traitement médiocre des variables de données continues, ordonnées et chronologiques, ce qui conduit à des résultats peu fiables et non interprétables.

Dans cette thèse, nous avons d'abord comblé les lacunes de la littérature et proposé une étude détaillée des techniques actuelles. Deuxièmement, nous résolvons les insuffisances en STEM concernant l'incertitude dans les données en proposant des arbres à effet de traitement de sous-population glissant. Troisièmement, nous proposons les forêts aléatoires de voisinage avec effet de traitement des sous-populations afin de minimiser l'effet du bruit dans les données. Quatrièmement, nous abordons le problème de la perturbation dans les données en proposant la technique de modélisation équilibrée du soulèvement par réflexion. Nous évaluons la performance des solutions proposées en utilisant des jeux de données simulés et réels, et nous montrons comment nos approches surpassent les autres méthodes en termes de coefficient de corrélation de rang de Qini et Spearman.

"The best thing for being sad," replied Merlin, beginning to puff and blow, "is to learn something. That's the only thing that never fails. You may grow old and trembling in your anatomies, you may lie awake at night listening to the disorder of your veins, you may miss your only love, you may see the world about you devastated by evil lunatics, or know your honour trampled in the sewers of baser minds. There is only one thing for it then - to learn. Learn why the world wags and what wags it. That is the only thing which the mind can never exhaust, never alienate, never be tortured by, never fear or distrust, and never dream of regretting. Learning is the only thing for you. Look what a lot of things there are to learn."

T.H. White,
The Once and Future King

Dedicated to my family

Acknowledgments

The work presented in this thesis have been realized in the Network and Computer Science Department (INFRES) of Telecom ParisTech located at Paris, France.

First of all, I would like to thank my supervisor Prof. Talel Abdessalem and my co-supervisor Prof. Hajer Kefi for having provided me the opportunity to do this PhD thesis. During these three years, they provided me a flawless support, both technical and personal, as well as a supervision of excellent quality. By working with them, I learnt many things, most notably professionalism and pragmatism.
I would also like to thank Prof. Olivier Segard for his support, advices and availability.
I would like to thank Professors Rokia Missaoui, Stéphane Bressan and Albert Bifet for having accepted to be part of the jury committee.

I would like to thank Prof. Claus Rautenstrauch, who, although no longer with us, continues to inspire by his example and dedication to the students he served over the course of his career.

I would also like to thank teachers Samia Latch, Souad Barodi and Najwa Yunis for the motivation and support. Also, I would like to thank Prof. Jorge Marx Gómez, Prof. Rasha Massoud, and Prof. Federico Pigni for their support during my studies.

Finally, I would like to thank a special group of people with whom I shared many great moments during the last three years, both inside and outside of Telecom. These special thanks go to the following list of doctors: Mostafa Haghir Chehreghani, Jacob Montiel, Marie Al-Ghossein, Thomas Rebele, Quentin Lobbé, Maximilien Danisch, Ziad Ismail, Oana Balalau, Jean-Benoît Griesner and Ashish Dandekar.
I would like to thank my friends Nedeljko Radulović, Alaa Alzaibak, Simon Semaan and Ali Bishani who always stood besides me.

Special thank to my best friend and my fiance Dr. Hana Baccouch for her motivation, support and patience.

Last but not least, I would like to thank my family, aunt Bushra, Prof. Malek Sibai, uncle Bashar. Special thank to my parents, Youssef and Fayrouz, and to my siblings, Dr. Mohammed, Dr. Effat, and Eng. Amer for their continuous support and presence.

Contents

List of Figures

List of Tables

Abbreviations

ATE Average treatment effect

AUUC Area under uplift curve

BMT Bone marrow transplant

BRUM Balanced reflective uplift modeling

CATEF Conditional average treatment effect function

CCIF Causal conditional inference forests

CDAGs Causal direct acyclic graphs

CF Causal forests

Comb_RF Combined subpopulation treatment effect random forests technique

CPS Current population survey

DAGs Direct acyclic graphs

ECE Estimated causal effect

ED_RF Euclidean distance based subpopulation treatment effect random forests technique

ICE Individual causal effect

KNN K-nearest neighbors

LASSO Least absolute shrinkage and selection operator

MSE Mean squared error

MTS-STEM Model then split subpopulation treatment effect modeling

NSW National supported work demonstration

OWL Outcome weight learning

PSID Panel study of income dynamics

PTE Personal treatment effect

PTL Personalized treatment learning

QC Qini coeffecient

RHC Right heart catheterization

SKNN Sequential k-nearest neighbors

SNCF Société nationale des chemins de fer français (French National Railway Company)

SSE Sum of squared errors

STE Subpopualtion treatment effect

STE-NBP Subpopualtion treatment effect neighborhood based binning

STE-NRF Subpopualtion treatment effect neighborghood random forests

STE-STrees Subpopualtion treatment effect sliding trees

STEM Subpopualtion treatment effect modeling

STM-STEM Split then model subpopulation treatment effect modeling

SVMs Support vector machines

SW Sliding window

TTM-STEM Transform then model subpopulation treatment effect modeling

Two_Models_RF Two model subpopulation treatment effect random forests technique

UBI Universal basic income

List of publications

Journal papers

— Atef Shaar, Hajer Kefi, and Talel Abdessalem. Subpopulation Treatment Effect Sliding Trees: New Method to Model Consumer's Response Heterogeneity. *Decision Support Systems*, 2018. (submitted)

— Atef Shaar, Hajer Kefi, and Talel Abdessalem. Subpopulation Treatment Effect Neighborhood Random Forest: New method for Subpopulation Treatment Effect Modeling. *ACM Transactions on Management Information Systems*, 2018. (submitted)

Conference papers

— Atef Shaar, Hajer Kefi, and Talel Abdessalem. The Balanced Reflective Uplift Modeling: Presentation of a New Model and Experimental Application in the Healthcare Sector. *Global Information Technology Management Association World Conference*, 2018. (Honorable Mention Award)

— Atef Shaar, Hajer Kefi, Talel Abdessalem, and Olivier Segard. Proposing a New Approach to Uplift Modeling: The Balanced Reflective Uplift Modeling. *The Conference on Digital Experimentation at MIT*, 2016

Demonstration papers and posters

— Atef Shaar and Talel Abdessalem. Pessimistic Uplift Modeling. *Machine Learning and Data Analytics Symposium*, 2016

— Atef Shaar, Talel Abdessalem, and Olivier Segard. Uplift Modeling for Recommendation Systems. *Big Data and Market Insights*, 2015

Introduction

"Begin at the beginning" the King said, very gravely,
"and go on till you come to the end: then stop."

Lewis Carrol,
Alice in Wonderland

In many applications, it is essential to measure how a specific treatment affects the population, for example, to study the impact of an advertisement on consumers' purchase behavior, to measure the effect of a specific drug on patients' health, or to compute the impact of new website design on the visitors' click rate. In order to estimate the effect of each of the previous treatments, a randomized experiment has to be conducted, only then the average causal effect of the treatment (ATE) can be calculated. However, besides the ATE, the resulting data of the experiment contains additional information about the variation of the causal effect across the population. This information can be used to draw inferences about the treatment effect on a specific group of the population, i.e., to measure the subpopulation treatment effect (STE).

With the increasing size and complexity in the data structures in the big data era, and with the rapid advances in data collection and storage technology, the concept of learning from data using statistics and machine learning algorithms have emerged in various domains [2]. Machine learning algorithms have been developed to understand the patterns in data and to discover the relationships that are hidden from the human perception. Also, the vast developments in computer processing allow us to model complicated systems which will enable a better recommendation system, accurate disease diagnosis software, fast language translations, proactive fraud detection, and effective marketing strategies. However, a little attention in data science research has been directed towards problems that concern using experimental data to discover causal inferences.

We name the field of science that is concerned with how to apply machine learning to model the effect of a treatment on a subpopulation, subpopulation treatment effect modeling (STEM). STEM is one of the emerging approaches in data science and machine learning domains. It is also known as uplift modeling, net-lift modeling, differential response modeling, and incremental-value modeling [3, 4, 5, 6]. STEM is not applied to predict the class variable, but to predict the variation of the class variable [7]. In most cases, one treatment is rarely beneficial for the overall

population. Therefore, STEM is used to model the treatment effect based on subgroup characteristics. Thus, STEM has become an excellent tool for personalizing treatment.

STEM techniques have been applied in many domains. In the healthcare sector, STEM has been used to maximize treatment benefits for each patient [8, 9, 10]. Also, it has been used for cancer treatment, by estimating the benefits of chemotherapy based on the stage of cancer [11, 12]. Besides, using survival data, STEM has been used to predict the probability of patients' recovery [13]. STEM was also used to explore the role of drug use in sexual risk [10].

In marketing, the gain of using a customized advertisement was the main reason to adopt STEM [14, 6]. Another application in marketing is the detection of advertisements' negative impact, and it helps in managing marketing spending [15, 16, 17]. Moreover, STEM has been used in economics to test the effect of labor training program on post-intervention earnings; also it was used to select the most effective voter mobilization strategies [18, 19]. In finance, STEM helped insurance companies to maximize their profit by enhancing personalized rates change for policyholders [14].

The literature divided STEM approaches into two basic categories [6]. The first approach, called indirect STEM, combines multiple models' scores to generate the final STEM score. The second approach, called direct STEM, utilizes one model only to generate the final STE score [14, 20]. Indirect STEM models are more straightforward and easy to implement. However, the indirect STEM approaches show more sensitivity to noise than the direct STEM methods [21]. The differences between direct and indirect STEM have been addressed in [21, 14, 22, 20].

However, the literature of STEM is scattered due to the multiple used terms (e.g., uplift modeling, net-lift modeling, differential prediction ...). Also, there are several definitions for STEM that differ based on the domain and the used technique. This situation creates a domain based STEM literature, which leads to a problem of repetitive researches and it slows down the scientific development of STEM. Furthermore, there is no consistent evaluation measure for STEM models, which creates another challenge for advancement in STEM research.

Despite its significance, STEM shows a deficiency in reliability for real-life applied problems. Mainly because of the causal inference framework of STEM, we do not know the true value of any action (per person). The missing information in STE data is the main reason for the uncertainty problem. Customarily, scientists conduct experiments combined with grouping techniques to bypass this problem. In the machine learning domain, and mainly when dealing with observational data, specific binning techniques have to be applied to formulate the needed groups that will allow using the data for inference.

The current binning techniques are based on threshold binning, i.e., finding the optimal partition-point. Threshold-binning will try to construct a new child node (subgroup) in a way that maximizes the STE difference between the child node and the original node. By doing so, the threshold-binning guarantees more STE-homogeneous subgroups after each binning. For the nominal type of features, the threshold-binning works flawlessly. Nevertheless, for the continuous type of at-

tributes, the threshold-binning suffers from uncertainty magnification and higher noise sensitivity because of three points:

— The fundamental problem of causal inference, we only know part of the true value, which results in uncertain true classes.

— The rigid binning procedure, which depends on these uncertain true classes, will lead to uncertain partition points, hence increase over-fitting.

— The consequences of one error partition-point will affect all the other child nodes, which will later lead to incorrect results.

For this reason, it is essential to create a customized solution for STEM that can handle continuous attributes and minimize the uncertainty and over-fitting of STE models.

In this thesis, we first reintroduce the STEM problematic from the causal point of view. In addition, we present all the terms that are used for STEM. We propose a new taxonomy for STEM approaches. And we list all the measurements that evaluate STEM models performance.

Second, we explain the problems of uncertainty and disturbances that faces STEM. We analyze the current binning techniques typically used by STEM approaches. Next, we show the limitations of the current binning techniques. Then, we introduce a new binning approach called neighborhood-based STE binning (STE-NBP). Later, we compare between various binning techniques taking into consideration all potential outcomes, and we present the benefits of using our new binning technique.

Third, we propose a new approach to model STE using decision trees and the STE-NBB binning technique called subpopulation treatment effect sliding trees (STE-STrees). Utilizing the STE-STrees approach, researchers will be able to make better decisions on time using updated information, leading to predictions that are more accurate and results that can be employed in decision support systems. Moreover, in the marketing domain, STE-STrees can be applied as a dynamic A\B testing techniques, which will assist in maintaining relationships with customers in this era of big data.

Fourth, we propose the subpopulation treatment effect neighborhood random forests (STE-NRF)as an improvement of STE-STress to better handle bias and noise in the data. We demonstrate how our method employs the neighborhood-based binning technique to improve the performance of STEM.

Fifth, we introduce a new STEM approach named the balanced reflective uplift modeling (BRUM), which helps the standard STE in high disturbances environments. Afterward, we compare different STEM approaches using a thorough simulation study of various scenarios.

And later on, using real datasets from the medical and the business domains. We conduct an experiment with the marketing team in the French national railway company (SNCF) to improve the decision-making for advertisement campaign. Using data gather through an E-mail A\B testing, we applied the STEM models in three different cases, and we showed that our approach can be a better targeting strategy

to maximize response rate and minimize marketing spending (see Figure 1). We show how our approach outperforms all other STEM methods in terms of *Qini* coefficient and *Spearman's* rank correlation coefficient.

Finally, we conclude by presenting the limitations and the future perspectives of our research.

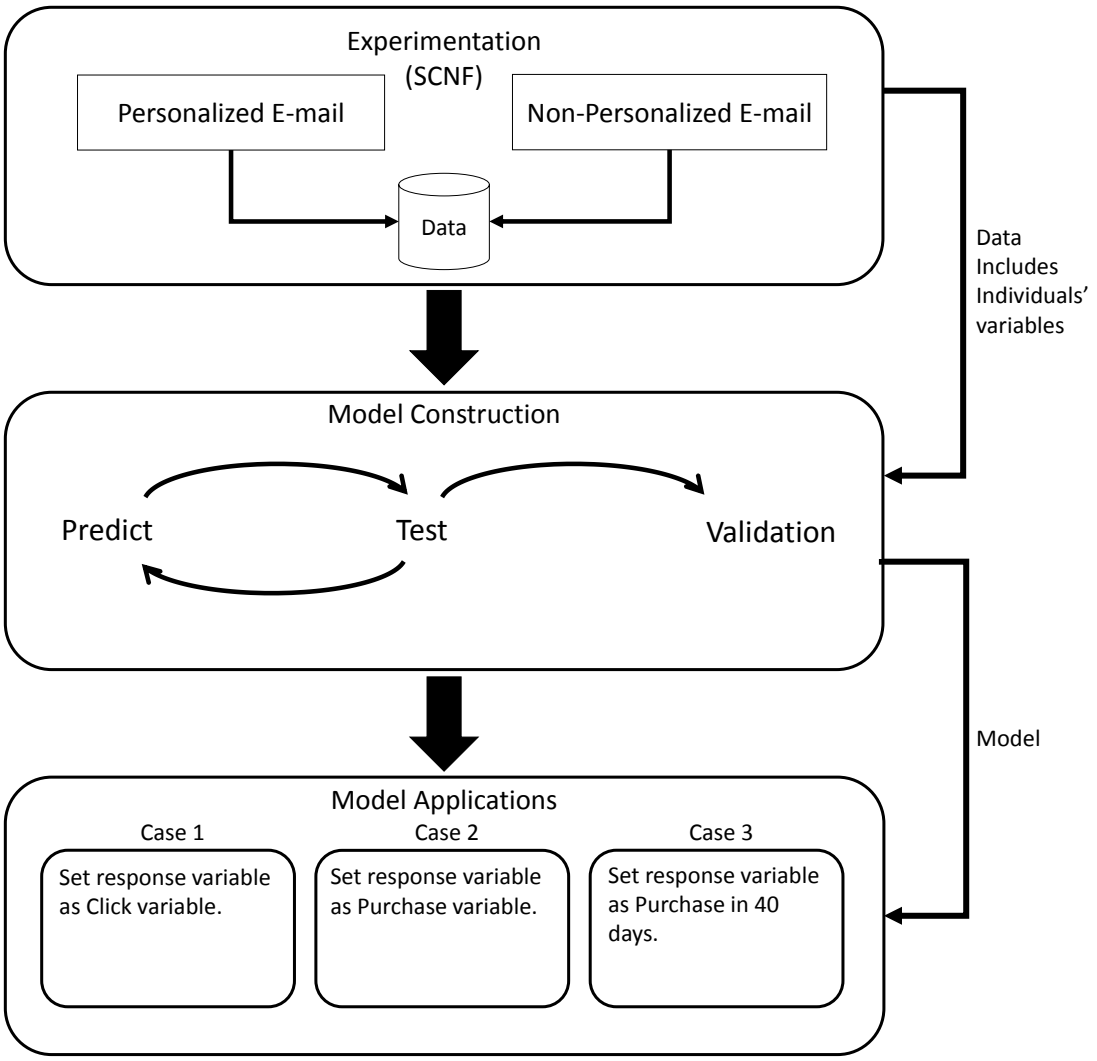


Figure 1 – Experiment methodology.

Thesis contributions

— Revisit STEM problem. We introduce the idea of STEM from a causal point of

view. In addition, we gathered all the terms that are used for STEM. Finally, we list all the measurements that evaluate STEM models performance.

— Propose a new taxonomy for the STEM approaches based on three main actions (split, model, transform).

— Propose new binning technique that is based on neighborhood cases (STE-NBB). This leads to better binning, less bias and creates homogeneous subgroups for better prediction score.

— Introduce a new approach to model STE using decision trees and STE-NBB binning technique, we call it subpopulation treatment effect sliding trees (STE-STrees).

— Propose subpopulation treatment effect neighborhood random forests (STE-NRF) as an improvement of STE-STress to better handle bias and noise in the data.

— Introduce a new STEM approach named the balanced reflective uplift modeling (BRUM), which helps the traditional STE in high disturbances environments.

— Apply new STEM approach to optimize email marketing campaign using SNCF data.

Synopsis and outline

In this thesis, we revisit the problem of modeling treatment effect heterogeneity and reformulate it as a machine learning oriented subpopulation treatment effect modeling (STEM). We fill the gaps in the literature and propose a detailed study of current techniques, and we highlight the main challenges that face STEM. We solve STEM shortcomings regarding the sensitivity to disturbance and uncertainty in the data, by proposing several new STEM approaches. We evaluate the performance of STEM approaches using simulated and real datasets, and we show how our approach outperforms other methods in terms of Qini coefficient and Spearman's rank correlated coefficient.

This thesis is composed of four chapters (see Figure 2). The first chapter is a review for the concept of the STEM. The second chapter discusses the problem of uncertainty and proposes solutions concerning the problems of uncertainty in STEM. The third chapter proposes solutions concerning the problems of disturbances in STEM. The fourth chapter is the conclusions, limitations and future perspectives.

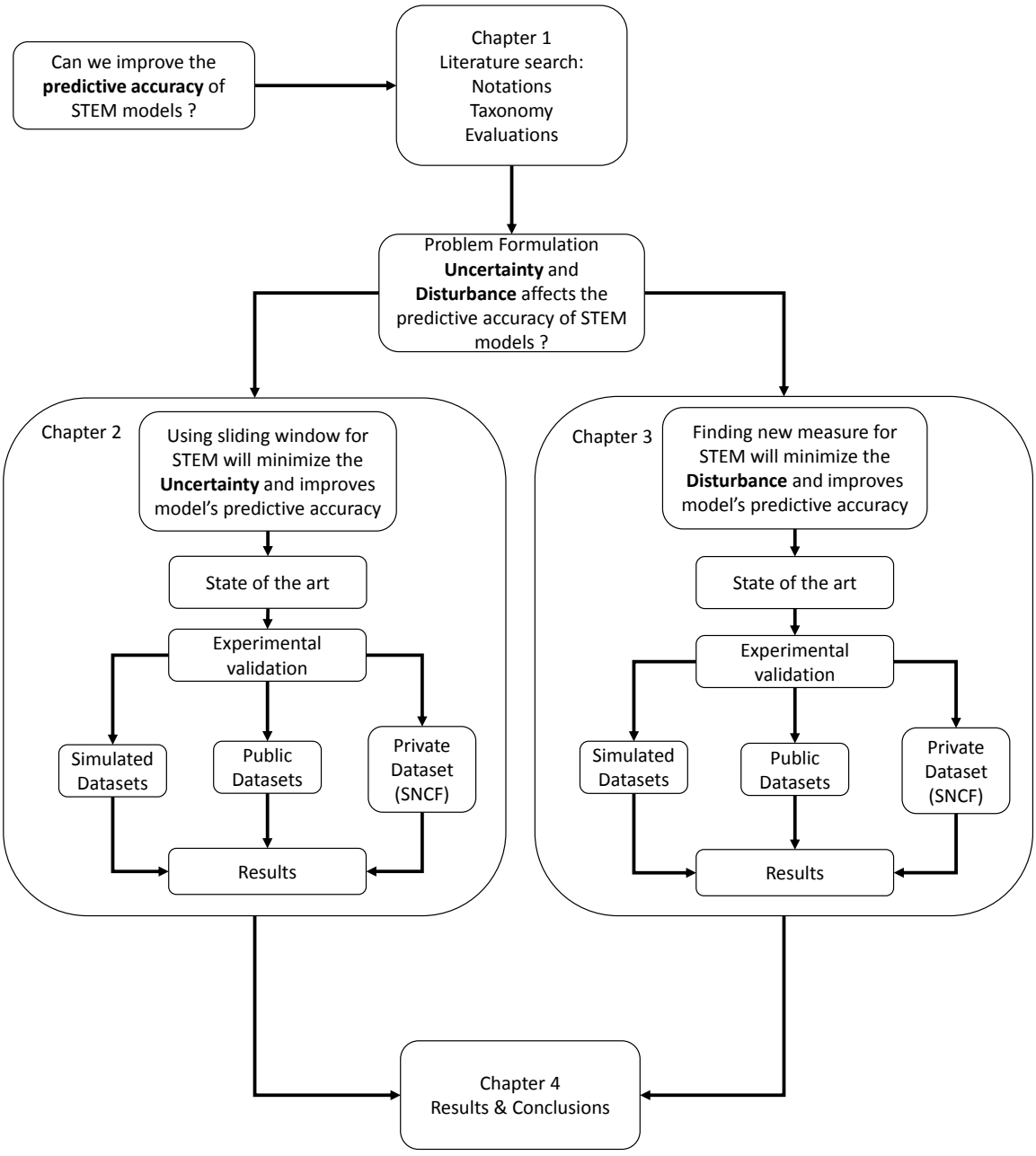


Figure 2 – Research design.

Chapter 1

Subpopulation Treatment Effect Modeling Review

> In science, read, by preference, the newest works; in literature, the oldest.
>
> Edward G. Bulwer-Lytton,
> *Caxtoniana: A Series of Essays on Life, Literature, and Manners*

In this chapter, we provide a different reading of subpopulation treatment effect modeling (STEM). We redefine STEM from a causal point of view. Then, we present state of the art in the field and discuss the existing works.

We believe that the contributions of the STEM approaches must be discussed relative to the following three research axes:

— works that are focused on new machine learning techniques for STEM approaches;

— works that are targeted towards the real-life applications of STEM approaches;

— works that are concerned in data processing for better STEM insights;

Generally speaking, works from the machine learning domain usually consider the data-driven approach and domain-agnostic framework, at the same level, therefore missing opportunities for discovering new insights.

On the other hand, works that are focused on the application of STEM are typically interested in easy-to-implement approaches that maximize specific requirements (e.g., profit, survival rates, ...). Typically, these works provide tools to explore the impact of the specific treatment given background knowledge has been considered a priori.

In this work, first, we discuss the paradigms underlying commonly used STEM approaches. Secondly, we show the main causes of STE. Thirdly, we introduce a new taxonomy for STEM approaches. Fourthly, we reveal the terminologies and names that are used to in the literature of STEM. Following, we show the different measurements of STEM models performance.

1.1 Introduction

The history of Treatment effect is the history of causality, which goes back to the history of science. From the ancient Greece when Plato described the principle of causality as "All that becomes must needs become by the agency of some cause; for without a cause nothing can come to be." (*Timaeus* 28a4–6) [23]. Aristotle stated in the Book II of Physics that "we should consider how many and what sorts of causes there are. For our inquiry aims at knowledge; and we think we know something only when we find the reason why it is so, i.e., when we find its primary cause."[24]. Aristotle explained that there are four different causes, or explanations, for any entity, those are the *formal* cause, the *material* cause, the *final* cause, and the *efficient* cause[25].

For example, for an answer to the question of "What is this?" while aiming at an Apple, a *formal* cause will be (This is an Apple). The answer to "What is Apple made of?" will be, for example, it is made of "x and y," and this will be the *material* cause of an apple. The third cause is the *final* cause, it answers the question of "What is the Apple for?," or "Why this Apple exist?" the final cause could be to transfer nutrition or energy. The *efficient* cause is the fourth cause; it considers answering the question of "What makes that Apple?", or "What causes that Apple?", and an answer could be a tree or seed (depending on your philosophical point of view). What matters for science is the efficient cause, which is what we aim when we discuss the treatment effect, i.e., the effect of the efficient cause.

The term *Treatment Effect* is an ambiguous term by its own, it lacks the object, the one that got the treatment. But usually, it refers to the *Average Treatment Effect*. The average treatment effect (ATE), which is used most of the time to reflect the impact of an action, is the most used and studied in the literature. However, in many cases, we are interested in knowing how that impact changes, what factors affecting it, and how those changes could be estimated (i.e., subpopulation treatment effect). For this reason, researchers do heterogeneous treatment effect analysis [26], conditional probability analysis [27], subpopulation treatment effect pattern plot [28], and many other analysis techniques. However, by harnessing the power of data mining and machine learning, new methods for estimating subpopulation treatment effect (STE) are on the rise. Those methods can deal with uncertain, missing, and domain-agnostic data. With the help of new machine learning techniques, the subpopulation treatment effect modeling (STEM) domain was born.

Below, we will present three real life applications of STEM. The first use case is from the business domain where STEM is used for optimized targeting. The second use case is from the healthcare sector where STEM is used for personalized medicine. The third use case is from the economic sector, where STEM is used for policy-making.

1.1.1 Marketing use case

Let's assume that there are three types of promotional advertisements, *Alpha*, *Beta*, and *Gamma*. Similar to the standard settings of an A/B testing framework [1], the advertisements' types differ in one factor only (e.g., advertisement color, font type, mean of communication, ...). Additionally, let's assume that the experiment population is composed of thirty clients distributed equally and randomly between the three advertisements' types (alpha, beta, and gamma) and results of the experiment showed that each of the advertisements' types has a 50% response rate. Usually, the optimal marketing strategy will go by choosing either of them. Suppose that X_1 and X_2 two variables that represents clients' information (i.e., age, gender, ...). By plotting each client's response into a two-axis plot, X_1 and X_2, a variance of the advertisement effect appears (see Figure 1.1).

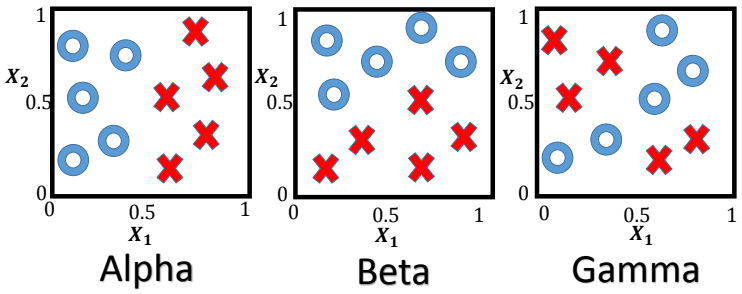


Figure 1.1 – Marketing experiment visualization for three types of advertisements plotted on two dimensions representing two features (X_1, X_2). A circle represents a client with a positive response, an X represents a client with a negative response.

We notice that there is heterogeneity in the response based on the chosen advertisement. STEM can be used to optimize the response rate of each advertisement type. The optimization is done by personalizing the advertisement based on each client. The simplest form of personalizing is to create a simple decision rule based on a specific client characteristic, for example, clients who are older than a specific age should be targeted with a specific advertisement. Using machine learning, those decision rules are generated automatically using techniques like decision trees and associate rule learners. Decision tree technique, as an example, relies on binning the population to find the subgroups with higher response rate. Most of the STEM approaches are based on the decision trees technique.

For example, the STEM will separate between consumers who prefer *Alpha* advertisement over *Beta* and *Gamma* and consumers who have no preference at all (they will respond anyway). We are interested to know which customer we should contact and using which advertisement type. Let's assume that we will partition the

1. A/B testing is type of controlled experiments that is used to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective[29].

population using the X_1 feature. By using a STEM approach, we expect to have the following rules (see Figure 1.2):

— Clients with $X_1 <= 0.5$ to be targeted by Alpha type of advertisement

— Clients with $X_1 > 0.5$ to be targeted by Gamma type of advertisement

By following the rules that are generated by the STE model, we have eight respondents out of ten, so we can reach 80% response rate instead of 50%.

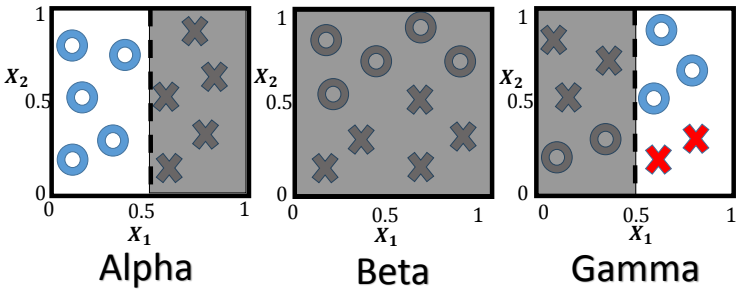


Figure 1.2 – Marketing experiment visualization for three types of advertisements plotted on two dimensions representing two features (X_1, X_2). A circle represents a client with a positive response, an X represents a client with a negative response, the lighted area represents selected instances, the gray area represents non-selected instances.

Many companies utilize STEM in their managerial decision-making. For example, one problem that faces companies and organizations is the customer churn (i.e., customer defections). Companies create retention programs (e.g., promotions, loyalty programs, etc.) to stop customer churn and to retain as many customers as possible. The large mobile operator *Telenor* has used the Uplift Modeling, which is type of STEM, to optimize customer retention programs. By targeting only 60% of potential churners, *Telenor* has been able to reduce the customers churn rate by 1.8% which represented an improvement of 36% compared to the traditional retention programs used previously [30].

Also, U.S. Bank has used the STEM to reduce the costs and improve the return-on-investment of their marketing programs. As a result, U.S. Bank has been able to eliminate the negative effect of the marketing program by targeting only 40% of their customers, which leads to increase in revenue by 327% (growth of 500K dollars per campaign) [31].

T-Mobile Austria used the STEM to optimize their marketing strategy, and it has been able to reduce churn by 20% and cut several hundred thousand Euros from software costs [32]. Also, *hp* has implemented STEM to estimate incremental sales in one of their U.S. retail chains. Their STE model has been able to predict the incremental sales by 6% error rate, which optimized their decision-making and minimized their spending costs [33].

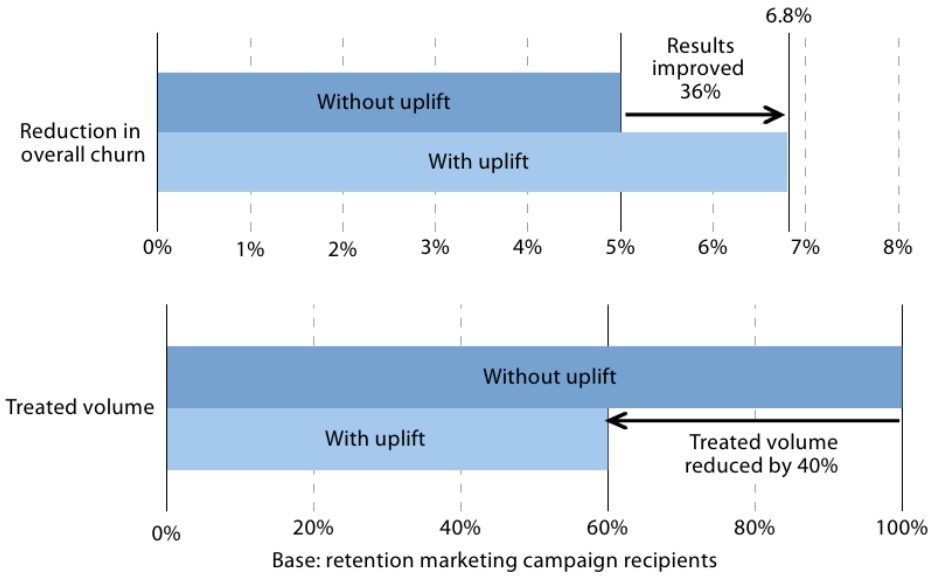


Figure 1.3 – Telenor retention program results (Clients targeted using traditional methods vs. Clients targeted using subpopulation treatment effect modeling (uplift)).

1.1.2 Healthcare use case

Given that we have a population of patients. And we have two types of medical pills that may aid in their recovery, the red pills and the blue pills. Let's say that we experimented to investigate the effect of the treatments. In Figure 1.4, we can see the results of our experiments. Based on each treatment effect, we can distinguish four types of patients, and we named them A, B, C, and D in Figure 1.4.

STEM can be applied in these cases to find the optimally personalized medicine. The STEM will maximize patient recovery rates and minimize the negative side effects. For example, the STEM will recommend a red pill for the patients with type A and a blue pill for the patients with type D. Also, the STEM will not recommend treating patients with types B and C, as they show no preference for the red not for the blue pill.

For example, [8] proposed a new STEM approach using sequential k-nearest neighbors (SKNN) analysis for personalized medicine, the authors argued that STEM improvement is necessary because there is no average patient, and all patients should have personalized medicine. Also, STEM has been used to increase the survival rate of patients [13]. STEM provided clinically interesting results when it was applied to a breast cancer dataset to differentiate between in situ and invasive cancers [12, 34]. In addition, the authors applied STEM to identify patients who are susceptible to

Patient Type	Red Pill	Blue Pill
A	Recovered	Not Recovered
B	Recovered	Recovered
C	Not Recovered	Not Recovered
D	Not Recovered	Recovered

Figure 1.4 – Red/Blue pill medical experiment example.

the risk of heart attack from taking COX-2 inhibitors [12].

1.1.3 Economics use case

In August 2018, people living in the Rheinau, a municipality in Switzerland, have been able to sign-up for an experimental project to test the Universal Basic Income (UBI) policy for over a year [35]. Suppose that the experiment is already done and we are interested in investigating the effect of the UBI on the educational progress of twenty college students (see Figure 1.5).

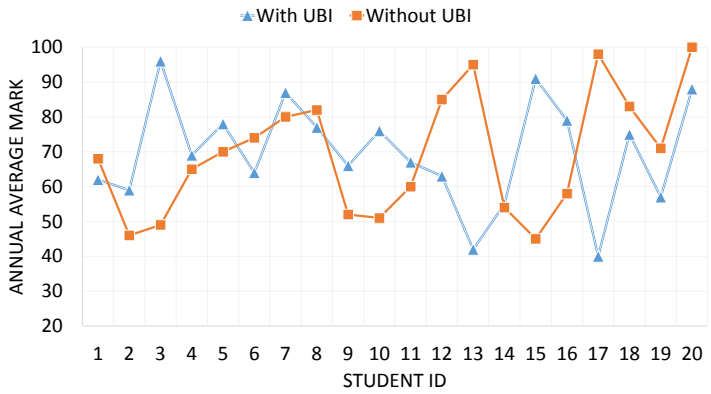


Figure 1.5 – Line charts for students average annual score (Case of universal basic income (UBI) vs. Case without universal basic income (Without UBI)).

We see in Figure 1.6 how STE change based on each student. The STEM will use this heterogeneity in the treatment effect to model those who are most benefited from the UBI.

The same concept has been applied by [36, 19] using the National Supported Work Demonstration (NSW) program dataset [19]. The NSW program is an employment program applied in the 1970s to help workers in the United States with

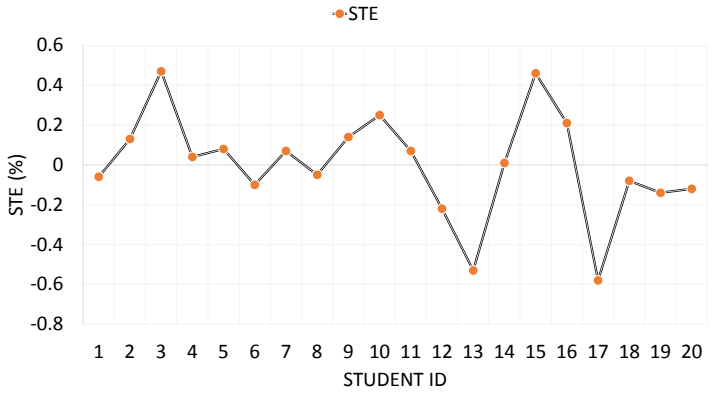


Figure 1.6 – Subpopulation treatment effect for each student.

social and economic difficulties to get employment. The authors in [36, 19] were interested in identifying the characteristics of workers for whom the program was beneficial. Using a STEM approach called Causal Conditional Inference Forests (CCIF) [17], they found patterns in subgroups that share common responses. The authors concluded that the program was most useful for workers with higher numbers of years of education and relatively lower earnings at the start of the program, as well as those married with higher earnings. Also, they discovered that the program was least effective for Hispanics with low education.

Other authors took subjects from mixed experimental and non-experimental data sets, including the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS) [18, 37], to analyze different treatment effect predictive approaches that can be applied to non-experimental datasets. The authors concluded that a STEM approach called Causal Inference Tree can be useful in non-experimental treatment effect heterogeneity detection,i.e., for personalized treatment [38].

1.2 Notation and definitions

In this section, we introduce the notation and the causal diagram adopted in our work. We define STE and STEM, and we differentiate between the interaction and the effect modification.

1.2.1 Notation

The notation adopted in our work will follow the notation of the causal inference literature. For simplification, we will use an example of advertisement experimentation alongside the notation. The experiment will have a binary treatment variable that reflects the event of receiving an advertisement. Also, the experiment will have a binary outcome variable that reflects the event of purchasing a product.

Let T be the treatment variable ($T \in \{0, 1\}$). T refers to the event of receiving an advertisement. Let Y be the outcome variable ($Y \in \{0, 1\}$). Y refers to the event of purchasing a product. Let X be the set of features that describe the subjects in our experiment ($X = \{X_1, X_2, X_3..\}$). For example, X_1 may refer to the vector of subjects' genders, and X_2 may refer to the vector of subjects' ages.

Let the subscript i of any variable V represents the individual's value of that variable v_i. For each subject in the experiment, we have individual values y_i, t_i, x_i, representing respectively the individual's outcome, the individual's treatment, and the individual's characteristics. For example, x_1 refers to a vector of features that define the first subject, and $x1_1$ relates to the gender of the first subject.

Let the superscript j of any variable V refers to the potential value of the variable V given the variable j. For example, $V^{j=k}$ is the potential value of the variable V given the variable j is equal to the value k.

It is essential to distinguish between the actual (observed) outcome and the potential outcome. In the provided example, each subject has two potential outcomes. The potential outcome had the subject been treated, i.e. $y_i^{t=1}$, and the potential outcome had the subject been untreated and $y_i^{t=0}$, while only one potential outcome is observable, which is the subject's actual outcome y_i.

In Figure 1.7, we can see a demonstration of a hypothetical randomized marketing experiment. Just for the sake of simplicity, we decided to experiment on twenty subjects. A dot in Figure 1.7 represents one subject. A black dot represents a subject who purchased a product, while the white dot represents a subject who did not purchase a product.

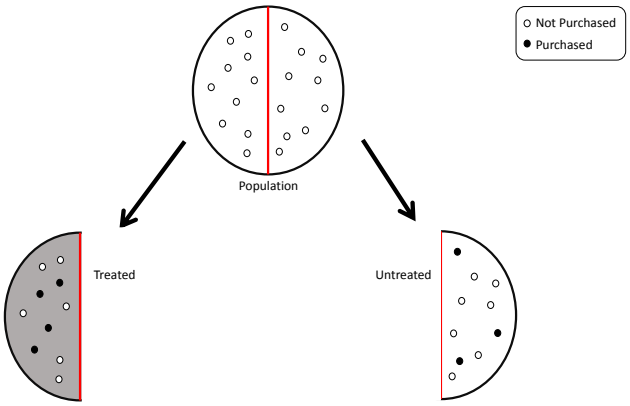


Figure 1.7 – An Illustration of a randomized marketing experiment.

There are three main ways to measure the effect (causal effect) of the advertisement (treatment) on the population purchasing behavior. They are conditional proportions' difference, conditional proportions' ratio, and conditional odds' ratio [39]. All of them compare the proportion of subjects who responded (purchased) in treatment against the proportion of subjects who responded in the control group.

Assuming the results of the hypothetical experiment are shown in the Table 1.1:

Table 1.1 – Results for the randomized marketing experiment.

# **Population**	**Treated** (T)	# **Subjects**	# **Purchased** (Y)
20	TRUE	10	4
	FALSE	10	3

We can check the causal effect (i.e., the treatment effect) of the advertisement by comparing the measured outcome between two states, one with the cause and one without it using:

— The difference between conditional proportions

$$Pr(Y = 1|T = 1) - Pr(Y = 1|T = 0)$$

$$\frac{4}{10} - \frac{3}{10} = 0.1$$

— The ratio of conditional proportions

$$\frac{Pr(Y = 1|T = 1)}{Pr(Y = 1|T = 0)}$$

$$\frac{\frac{4}{10}}{\frac{3}{10}} = 1.333$$

— The ratio of the conditional odds [2]

$$\frac{Pr(Y = 1|T = 1)/Pr(Y = 0|T = 1)}{Pr(Y = 1|T = 0)/Pr(Y = 0|T = 0)}$$

$$\frac{\frac{4}{10}/\frac{6}{10}}{\frac{3}{10}/\frac{7}{10}} = 1.555$$

All of those measures are assessing the causal effect of marketing advertisement on population purchase rate. But each measure focuses on a different scale. It is worth noting that those measures are only applicable in the condition of a randomized experiment.

Now, imagine that we can see the unobservable potential outcomes. In Figure 1.8, we illustrate the potential results, observed and unobserved, for the same marketing experiment.

And the results of the experiment with all potential outcomes are shown in the Table 1.2:

2. The odds of an event is the ratio of the probability that the event will happen to the probability that the event will not happen [40].

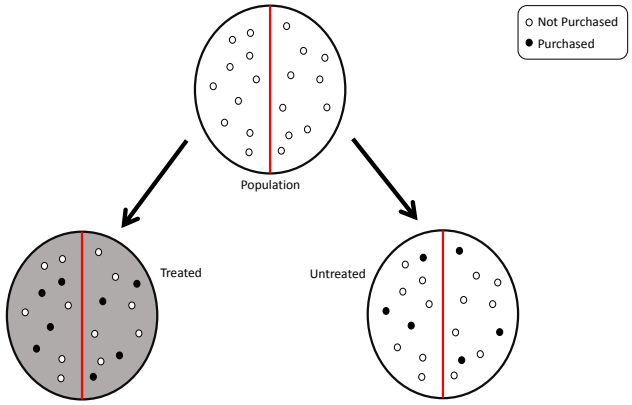


Figure 1.8 – An Illustration of a randomized marketing experiment with all potential outcomes.

Table 1.2 – Results for the randomized marketing experiment with all potential outcomes.

# **Population**	**Treated** (T)	# **Subjects**	# **Purchased** (Y)
20	TRUE	20	8
	FALSE	20	6

In this case, we can see the effect of the treatment on the whole population. And we can use the potential response proportion instead of the conditional response proportions.

For example, the potential purchase proportion had an advertisement been received $Pr(Y^{t=1} = 1)$ is the purchase rate given that all the subjects are treated. Whilst the conditional purchase proportion had the advertisement been received $Pr(Y = 1|T = 1)$ is the purchase rate of the treated subgroup only.

We can calculate the causal effect using the same causal measures:

— The difference between potential proportions

$$Pr(Y^{t=1} = 1) - Pr(Y^{t=0} = 1)$$

$$\frac{8}{20} - \frac{6}{20} = 0.1$$

— The ratio of potential proportions

$$\frac{Pr(Y^{t=1} = 1)}{Pr(Y^{t=0} = 1)}$$

$$\frac{\frac{8}{20}}{\frac{6}{20}} = 1.333$$

— The ratio of the potential odds

$$\frac{Pr(Y^{t=1}=1)/Pr(Y^{t=1}=0)}{Pr(Y^{t=0}=1)/Pr(Y^{t=0}=0)}$$

$$\frac{\frac{8}{20}/\frac{12}{20}}{\frac{6}{20}/\frac{14}{20}} = 1.555$$

In the literature, the difference between conditional proportions is called an *association*, and the difference between potential proportions is called a *causation*. Only when we create the perfect experiment that assures the three major conditions (exchangeability, positivity, and consistency) [39], we can interpret the association as a causation. The exchangeability condition refers to assumption that our subjects are randomly assigned to the treatments and the treatment groups are exchangeable, i.e., the outcome of an untreated subject will be similar to the outcome of a treated subject if he/she was assigned to a treatment group. The positivity condition refers to the assumption that the conditional probability of receiving every value of treatment is greater than zero. It means that any individual has a positive probability of receiving all values of the treatment variable, i.e., in our dataset, we should have the treated and untreated subjects. And the consistency refers to the assumption that the treatments (interventions) is well-defined, so we are sure that the outcome of a specific treatment is actually due to our intervention. That is, in the perfect randomized experimental conditions, we will have:

$$Pr(Y=1|T=1) - Pr(Y=1|T=0) = Pr(Y^{t=1}=1) - Pr(Y^{t=0}=1)$$

In our marketing experiment, it seems that we did an unrealistically perfect randomized experiment, and we successfully emulated the potential outcome worlds. However, in the real world experimentation, we cannot guarantee the ideal experimental conditions. But we may get close to it.

We define the *average treatment effect*(ATE) as the difference between the proportion of the observed outcome under the treatment condition and the proportion of the observed outcome under the control (no treatment) condition.

$$ATE = Pr(Y=1|T=1) - Pr(Y=1|T=0)$$

If we want to see the effect of the treatment on a particular person, then we use the *personal treatment effect*(PTE) which can only be expressed using the potential personal outcome:

$$PTE_i = Y_i^{t=1} - Y_i^{t=0}$$

Imagine that we chose a subject from our population, and we checked the causal impact of the advertisement on the purchase behavior. Based on the figure 1.9 the selected subject does not change his/her personal purchase decision based on

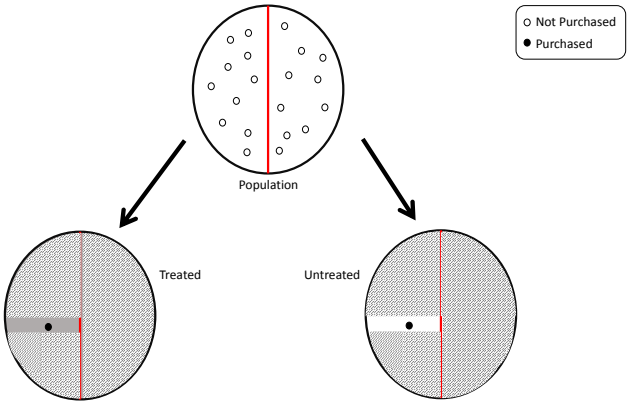


Figure 1.9 – The personal treatment effect (PTE) for a subject in an Illustration for a randomized marketing experiment.

the advertisement (the subject will purchase whether he/she has been treated or untreated), so the advertisement does not affect the subject's purchasing behavior.

However, because the information on potential outcomes will never be available in a real experiment, we have to rely on other approaches to estimate the personal treatment effect. Mainly, we can match our subject in the treatment group to another subject in the control group. In this case, we can compare the response rate between the two subpopulations of subjects that share common characteristics. We call this variance the *subpopulation treatment effect*, and we can define it mathematically as the following:

$$STE_i = Pr(Y = 1|T = 1, X = x_i) - Pr(Y = 1|T = 0, X = x_i) \tag{1.1}$$

We can see in Figure 1.10 that after matching and by using STE, we will find that the matched subject did not purchase if he/she had not been treated. It means that it is favorable to send an advertisement to the person with x_i characteristics, which is contrary to what did the personal treatment effect of that person show.

This problem is not uncommon in real experiments, it is due to the fact that finding an exact match for a subject is an incredibly difficult task. Finding a match requires first to compress the subject's characteristics, despite their volume and complexities, into data that can be matched with another subject from the other treatment. And even if we found an exact match for the required subject, matching will be harder if we consider the factor of time. This problem is known in the causal inference domain as "*the fundamental problem of causal inference*"[41].

Assuming that we can identify an individual by its infinite characteristics, and that we are limited by our knowledge and our computation capabilities, we can use STE as an approximation for PTE. However, when we have access to more data, our STE will converge to PTE.

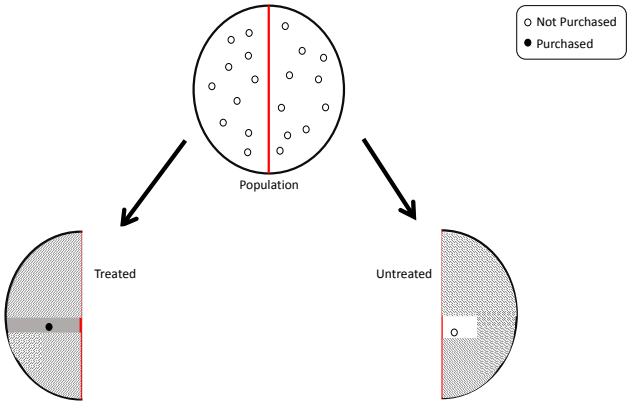


Figure 1.10 – The subpopulation treatment effect (STE) for a subject in an Illustration for a randomized marketing experiment.

Proposition 1 *STE converges to PTE when empirical data increases.*
$\lim_{x_i \to \infty} STE_i = PTE_i$

1.2.2 Causal directed acyclic graphs

Mathematical notation alone is not expressive enough for causal explanation. For this reason, we will use causal directed acyclic graphs as a visualization tool for STE cases. Directed acyclic graphs are directed graphs that has no directed circuits [42]. DAGs or *direct acyclic graphs* are visual representations of qualitative causal assumptions [43](see Figure 1.11).

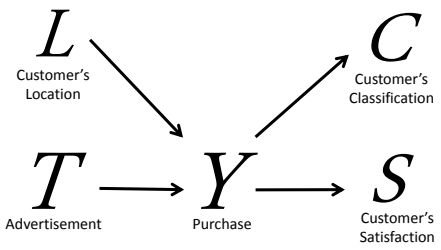


Figure 1.11 – Example of a direct acyclic graph.

DAGs contain two kinds of elements the nodes, and the edges. A node represents variable measurement. For example, node B in the Figure 1.11 could be a measurement for participants' age. A directed edge accounts for a causal relationship (non-mechanical[3]) between two nodes. For example, in the Figure 1.11, there is a

3. Mechanical causal relationship implies material relations between events.

causal relationship between the node T and node A.

Note that we intentionally define a node as a variable measurement, and not as a variable, because we want to emphasize the idea of measurement error. For each variable we measure, there is a percentage of measurement error, and ignoring these errors will lead to false results.

The CDAGs or *causal DAGs* are DAGs with one additional condition, which is that all common causes for any pair of variables, regardless of whether they are observed or unobserved, are represented in the graph [44]. We use CDAGs basically to help us decide which variable should enter the statistical model and which should not. Also, using DAGs will unify our hypotheses for the origin of bias in the data, it also clarifies our thoughts about the research topic. CDAGs are used to represent causal Bayesian networks. We define the path as a set of edges that connect two variables, no matter the directions of those edges. A path represents an association between two nodes. For example, there is an association between the node A and the node R in the Figure 1.11.

Based on the marketing experimentation in section A.2.1, we have a binary treatment and a binary purchase variables. We can draw the CDAG as seen in the Figure 1.12.

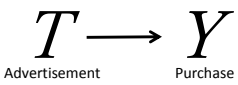


Figure 1.12 – Marketing experiment causal direct acyclic graph example.

We can express the conditioning on (or controlling) specific node by drawing a box around it. As seen in the Figure 1.13, we controlled the treatment node, i.e., we allow only one value for treatment.

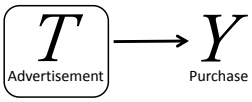


Figure 1.13 – Marketing experiment causal direct acyclic graph example after we control the treatment node.

The importance of CDAGs for STEM arises from three points:

— Provide the ability to express and transmit the hypotheses that are made based on researcher background knowledge.

— Assist in understanding associations' types between variables and the possible bias.

— Improve the process of pre-modeling feature selection.

1.2.3 Definition of subpopulation treatment effect

STE definition is obtained from the definition of effect modification, it is when the effect of a variable on another varies across strata of on or more variables [45].

The distinction between Interaction and effect modification is crucial here. Am Interaction can be done when we have two interventions, while effect modification requires only conditioning on a third variable[46].

Definition 1 *Subpopulation Treatment Effect is the disparity of treatment effect that happens when the effect of a variable on another varies across strata of one or more variables.*

1.2.4 Definition of subpopulation treatment effect modeling

The subpopulation treatment effect modeling is the science that is concerned with modeling the relations between treatment variables and effect modifications variables.

Definition 2 *Subpopulation Treatment Effect Modeling is the field of machine learning that is concerned with modeling the subpopulation treatment effect in the data.*

1.3 Why do we have a subpopulation treatment effect?

In all experiments, it's essential to study the variation of treatment effect across the population to uncover potential insights. The variance of the subpopulation treatment effect (STE) is the manifestation of having a heterogeneity in the treatment effect.

In this section, we will study the origins of subpopulation treatment effect from a causal inference point of view. We will explain the 10 causes of STE using an example use-case from the business domain.

For simplicity, the term "variable" will be used to refer to the variable's measurement.

We find STE in the dataset of any experiment because of two reasons:

— We conditioned on the wrong variable.

— We didn't condition on the right variable.

By knowing the differences between STE types, we can create a better understanding of the reasons behind having STE, how to calculate STE, how to differentiate between the true STE and the accidental STE.

Types of subpopulation treatment effect

For each type of STE, we will demonstrate how to unravel the hidden STE, and how to eliminate the accidental STE.

1. **Direct cause:** The direct cause of the response variable is the main reason for having STE in the dataset. For example, the required shipping-time for a specific product can affect the purchasing decision (see Figure 1.14). For an experiment, if the experimenter did not control the shipping time, the experiment dataset will include hidden STE. For example, even if the advertisement was very convincing for purchasing, shipping time could alter the decision later. The STE for the CDAG in the Figure 1.14 can be expressed as:

$$STE = ATE = E[Y = 1|T = 1] - E[Y = 1|T = 0] \tag{1.2}$$

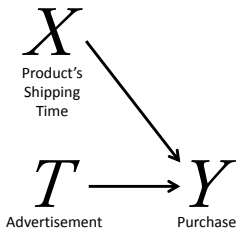


Figure 1.14 – Causal direct acyclic graph for an example of the **direct cause**.

In this case, to unravel the STE effect, we can control by conditioning on the second cause of the response variable (see Figure 1.15).

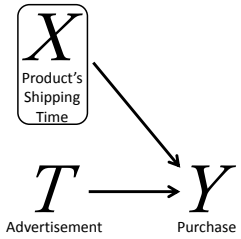


Figure 1.15 – Causal direct acyclic graph for an example of **conditional direct cause**.

The STE for the CDAG in the Figure 1.14 will become:

$$STE = E[Y = 1|T = 1, X = x] - E[Y = 1|T = 0, X = x] \tag{1.3}$$

By controlling the product's shipping time, we can calculate the main effect of advertisement on a specific stratum, which is the required STE.

2. **Direct effect:** By conditioning on the direct effect of the response variable, we might create subpopulations that have a different treatment effect. This case of STE is an accidental STE (it is actually a biased treatment effect).

   For example, by conditioning on specific personal information regarding the purchase (e.g., customer's payment method), we, unintentionally, generate an STE-like case (see Figure 1.16).

   The STE, in this case, will be:

$$STE = E[Y = 1|T = 1, X = x] - E[Y = 1|T = 0, X = x] \tag{1.4}$$

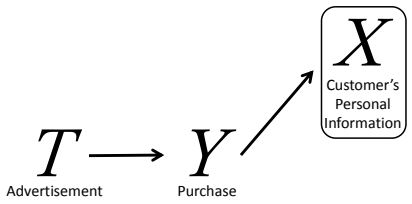


Figure 1.16 – Causal direct acyclic graph for an example of **conditional direct effect**.

However, to eliminate the accidental STE in this case, we can, if possible, remove the conditional case on the direct effect variable (see Figure 1.17). And we calculate STE as:

$$STE = ATE = E[Y = 1|T = 1] - E[Y = 1|T = 0] \tag{1.5}$$

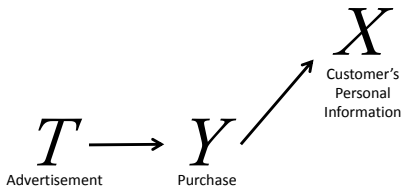


Figure 1.17 – Causal direct acyclic graph for an example of **direct effect**.

Sometimes the dataset of our experiment has already been conditioned on a specific direct effect variable, in these cases, it is worth to mention it in the STE analysis report for better interpretations.

3. **Direct intermediate:** It is very common to find this type of STE cause in experiment databases. It is when there is a variable that happens to be the effect of the treatment and the cause of the response. For example, in Figure 1.18, we can see that the customer's interest in the product affects the final decision to purchase, and also influenced by receiving the advertisement. This will create hidden STE in our dataset, that is calculated as:

$$STE = ATE = E[Y = 1|T = 1] - E[Y = 1|T = 0] \tag{1.6}$$

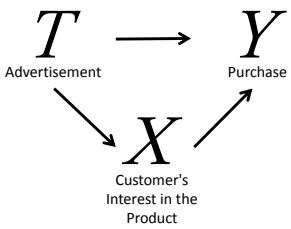


Figure 1.18 – Causal direct acyclic graph for an example of **direct intermediate**.

To calculate the STE caused by an intermediate, we control the intermediate node (see Figure 1.19). For example, we may find that the STE for customers who are interested in a product are more willing to purchase it if the advertisement reminded them.

$$STE = E[Y = 1|T = 1, X = x] - E[Y = 1|T = 0, X = x] \tag{1.7}$$

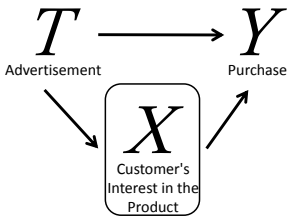


Figure 1.19 – Causal direct acyclic graph for an example of **conditional direct intermediate**.

However, it worth noting that in these cases, and because we are conditioning on the effect of the treatment, we might add bias to the results (i.e., imbalanced treatment control subgroups). So, we have to check and solve this bias in our dataset after the controlling phase.

4. **Indirect cause:** Is also common to forget about indirect causes. Those variables that affect a direct cause variable (see Figure 1.20). For example, consumer location is an indirect cause of the purchase decision. It affects the shipping time required for a specific product. Hence, it affects the purchase decision and creates hidden STE in the dataset.

$$STE = ATE = E[Y = 1|T = 1] - E[Y = 1|T = 0] \tag{1.8}$$

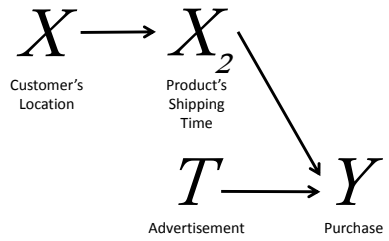


Figure 1.20 – Causal direct acyclic graph of an example of **indirect cause**.

We can control the effect of the indirect cause by conditioning on the indirect cause variable (see Figure 1.21).

$$STE = E[Y = 1|T = 1, X = x] - E[Y = 1|T = 0, X = x] \tag{1.9}$$

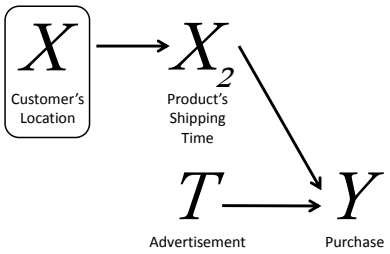


Figure 1.21 – Causal direct acyclic graph of an example of **conditional indirect cause**.

5. **Indirect effect:** The indirect effect of the response variable could create an accidental STE in our database if we conditioned on it. For example, in Figure 1.22, we can see how a variable that could represent a classification of customers (i.e., a segment of the consumer), creates hidden STE in our dataset.

$$STE = E[Y = 1|T = 1, X = x] - E[Y = 1|T = 0, X = x] \tag{1.10}$$

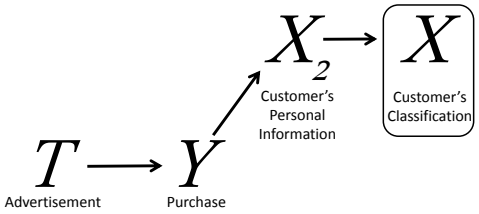


Figure 1.22 – Causal direct acyclic graph of an example of **conditional indirect effect**.

To eliminate the accidental STE, we can remove the conditioning on the indirect effect, as seen in the Figure 1.23.

$$STE = ATE = E[Y = 1|T = 1] - E[Y = 1|T = 0] \tag{1.11}$$

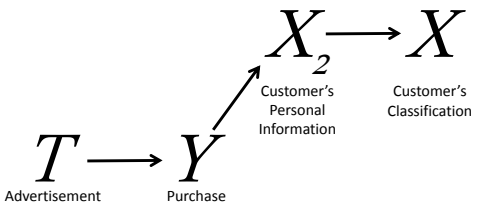


Figure 1.23 – Causal direct acyclic graph of and example of **indirect effect**.

Note that, in cases where we cannot un-condition the indirect effect, it is still important to mention that in the STE analysis report for future investigation.

6. **Indirect intermediate:** Similar to the intermediate variable, the indirect intermediate will create the same STE effect. In the Figure 1.24, we show how the consumer's budget affects his/her interest in purchasing specific products. This will, later, create hidden STE in our dataset.

$$STE = E[Y = 1|T = 1, X = x] - E[Y = 1|T = 0, X = x] \tag{1.12}$$

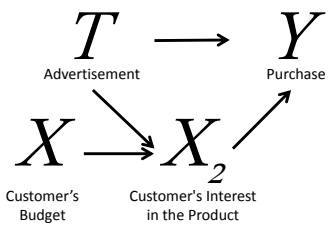


Figure 1.24 – Causal direct acyclic graph of an example of **indirect intermediate**.

To control indirect intermediate, we condition on the indirect intermediate variable (see figure 1.25). Notice that conditioning on the indirect intermediate is **not** enough to calculate STE. Actually, we have to condition on the direct intermediate and solve the generated bias.

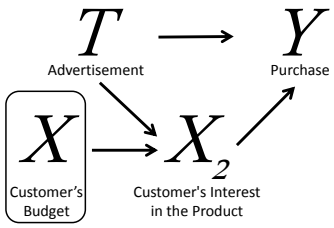


Figure 1.25 – Causal direct acyclic graph of an example of **conditional indirect intermediate**.

7. **Common cause:** The common cause is a confounder that should be dealt with. But that does not mean it does not also create hidden STE in the dataset. For example, in the Figure 1.26, we can see how information about consumers' past purchases affects the treatment and response variable. Sometimes, we send advertisements to consumers that we have in our database. Those who already purchased.

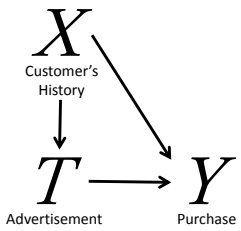


Figure 1.26 – Causal direct acyclic graph of an example of **common cause**.

Similar to confounding controlling, we control by conditioning on the common cause variable (see Figure 1.27). Also, we have to fix the bias that is generated from conditioning on the confounder [4].

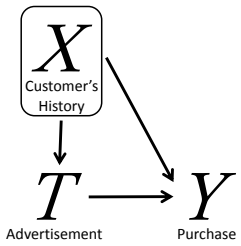


Figure 1.27 – Causal direct acyclic graph of an example of c**onditional common cause**.

8. **Common effect:** By conditioning on a common effect, we create a confounder and we create accidental STE in the dataset. For example, in Figure 1.28, we can see that customer satisfaction (e.g., good feedback) can be caused by the treatment, and by purchasing the product.

4. Confounder is a variable that influences both the dependent variable (response) and independent variable (treatment) causing a spurious association[47].

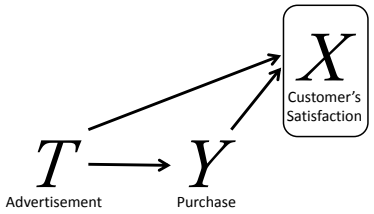


Figure 1.28 – Causal direct acyclic graph of an example of **conditional common effect**.

We can remove the accidental STE by removing the condition on the common effect variable (see Figure 1.29).

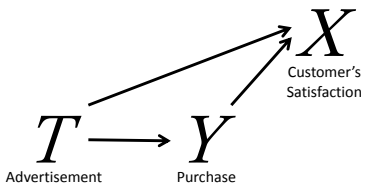


Figure 1.29 – Causal direct acyclic graph of an example of **common effect**.

9. **Cause by proxy:** The cause by proxy will create hidden STE in the dataset. A proxy could be one or more variables that have not been observed during the experiment (i.e., there is no data about proxy variables in the experiment dataset). For example, in the Figure 1.30, the device that the consumer uses (e.g., iPhone, Samsung, Nokia, etc) might reveal consumer characteristics traits, a study in 2016 shows that Android users were perceived to have greater levels of honesty-humility, agreeableness, and openness but be less extroverted than iPhone users [48]. Those characteristics traits are a proxy and they could affect the purchase decision.

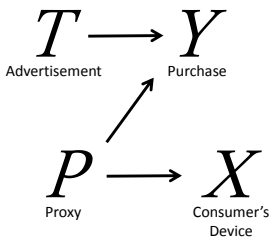


Figure 1.30 – Causal direct acyclic graph of an example of the **cause by proxy**.

By controlling on the cause-by-proxy variable, we would have controlled the proxy variable, hence, we controlled the STE in our dataset. Notice here that there is still a high chance of bias and error because of faulty assumptions regarding the proxy or because of not complete datasets.

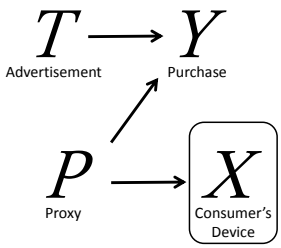


Figure 1.31 – Causal direct acyclic graph of an example of **conditional cause by proxy**.

10. **Effect by proxy:** It is when there is a variable that shares a common effect with the response variable. For example, in the Figure 1.32, we see how consumer loyalty and purchasing decision may cause a change in a third variable (e.g., product's review). By controlling on the effect-by-proxy variable, we control the proxy variable which will create accidental STE in our dataset.

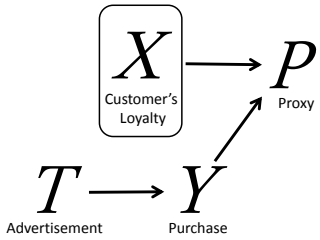


Figure 1.32 – Causal direct acyclic graph of an example of **conditional effect by proxy**.

To eliminate the accidental STE in our dataset, we can, if possible remove the control from the effect-by- proxy variable (see figure 1.33).

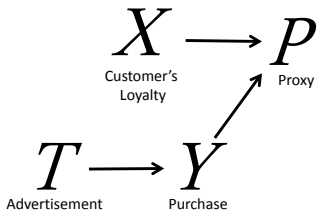


Figure 1.33 – Causal direct acyclic graph of an example of **effect by proxy**.

Figure 1.34 combines all causes of STE (real and accidental). It is important to remove the accidental STE before measuring the real STE.

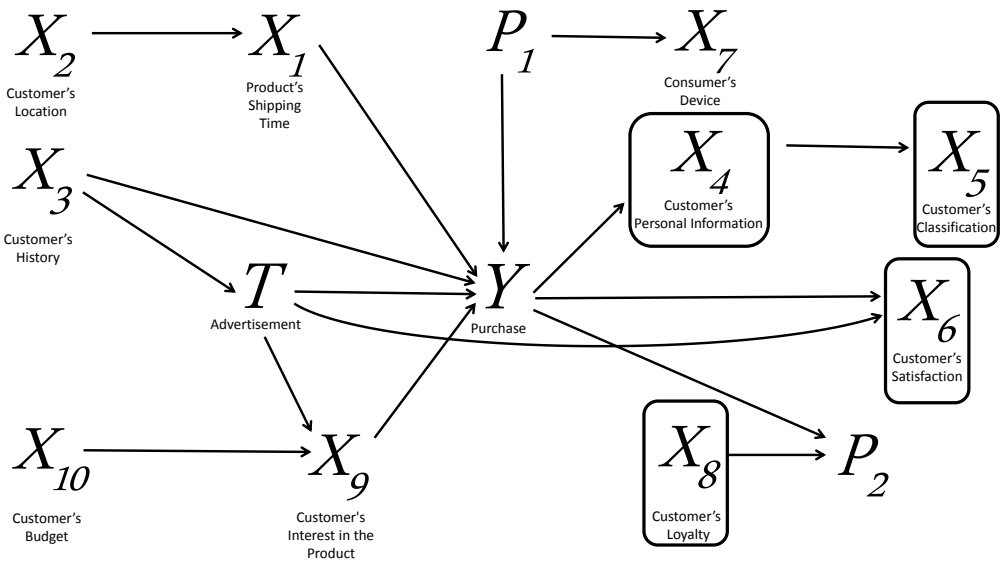


Figure 1.34 – Causal direct acyclic graph of an example of all subpopulation treatment effect causes.

Chapter 2

Subpopulation Treatment Effect Modeling Taxonomy

The STEM techniques have developed through a multi-domain environment. This situation allows for new and innovative solutions to appear as STEM techniques. We will do the taxonomy based on three main processes.

— Split: which refers to the process of splitting the dataset based on each treatment.

— Model: which refers to the process of using the data to create a model.

— Transform: which refers to the data manipulation processes that happen before modeling.

2.0.1 Split then model

The split then model subpopulation treatment effect modeling (STM-STEM) approaches split the database based on each treatment. Then, for each treatment subset, a predictive model is built. Later, the final estimation of STE is calculated based on the different response models. Due to its simplicity, STM-STEM approach is the most intuitive way for the STEM. For example, in a treatment/-control experiment, one response model will be fitted based on the treatment subset $M_{treat} = E[Y|X, T = 1]$, and another response model will be fitted based on the control subset $M_{control} = E[Y|X, T = 0]$. Then, an estimated STE will be calculated as the difference between the estimated response models.

$$\widehat{STE}_i = M_{treat}(Y|X = x_i) - M_{control}(Y|X = x_i) \tag{2.1}$$

A regression-based STM-STEM approach has been applied by [5, 49, 6, 50]. For a binary treatment experiment, let x_i is a vector of predictor variables for the case i, $Pr(Y = 1|T = 1, x_i)$ is the response probability of case i under treatment and $Pr(Y = 1|T = 0, x_i)$ is the response probability of case i under control, then

$$M_{treat}(Y = 1|X = x_i) = \frac{e^{B_{x_i}^{T=1}}}{1 + e^{B_{x_i}^{T=1}}}$$

$$M_{control}(Y=1|X=x_i) = \frac{e^{B^{T=0}_{x_i}}}{1+e^{B^{T=0}_{x_i}}}$$
$$\widehat{STE}_i = M_{treat}(Y=1|X=x_i) - M_{control}(Y=1|X=x_i)$$

Where $B^{T=1}$ and $B^{T=0}$ is a vector of logistic regression coefficients for the treatment and control model respectively[5].

or as a decision-tree based approaches [5, 51, 52, 53, 54, 22]. Also, an STM-STEM approach based on inductive logic programming (ILP) [55] has been proposed by [34]. [34] proposed building separated trees augmented Naive Bayes (TAN) Bayes-net model for each treatment, then to maximize the area under the uplift curve. STM approaches have shown high sensitivity to noise, as explained in [21]. The authors argued that the separate models in STM-STEM focus on the response variable more than the variance of the response, which will increase the effect of noise and minimize the ability to model the STE in the data. In addition, [7, 22] showed that STM-STEM approaches outperformed other STE methods. However, [50] contradicted the past argument and defended that the STM approaches are better than other approaches using a linear regression approach.

2.0.2 Transform then model

The transform then model subpopulation treatment effect modeling (TTM-STEM) approaches do not split the dataset. But they manipulate the data in a way to fit into some specific requirement. Later, they model the STE based on the modified dataset.

This is the most used TTM-STEM approach. The idea behind it is to convert the class variable to another form to be compatible with the required algorithms. For example, [56] remodeled the class variable for the conventional treatment/control datasets into a binary class. The authors in [56] combine the cases that are labeled as treated and responded with the cases that are labeled as controlled and not responded in one subgroup that is labeled as a positive subgroup. Then, the authors combine the cases that are labeled as treated and not responded with the cases that are labeled as controlled and responded in one subgroup that is labeled as a negative subgroup. Later, the authors [56] used a binary classification model to estimate the positive response, and used this estimate as a STEM approach. The simplicity of this approach enabled the researchers to use it with ensemble modeling techniques [6]. This approach has been used and later improved by [9, 6].

In addition, using support vector machines ($SVMs$) techniques to model STE has been used in [7, 19, 57]. For example, [19] proposed a modified SVMs technique, the response variable is transformed from $Y \in \{0,1\}$ to $Y' \in \{-1,1\}$. Also, the features are re-scaled based on [58] to apply two separate Least Absolute Shrinkage and Selection Operator constraints (LASSO)[58], one for the treatment effect and another for STE. To estimate STE, we calculate the difference in the values of the predicted response under the treatment and the control conditions.

$$STE(x_i) = \frac{1}{2}[(\widehat{Y'_i}|T=1, X=x_i) - (\widehat{Y'_i}|T=0, X=x_i)]$$

Where $\widehat{Y_i^{'}}$ is the predicted value of $Y^{'}$ for the subject i.

The authors in [57] maximized the area under the uplift curve to predict the highest STE subgroups. Another TTM-STEM method was proposed by [11], and it consists of transforming the features in addition to the class variable transformation, then to fit a logistic regression model to the data. SVMs techniques have also been used to detect STE by [59]. The authors named their approach *outcome weight learning (OWL)*.

In addition, [22] proposed a new STEM approach *causal trees*, they transformed the outcome then added in-sample and out-of-sample cross validation methods to choose the best penalty. The algorithm for the causal trees is described in Algorithm 1.

Where $Q^{is} = -\frac{1}{N}\sum_{i=1}^{N}(Error)^2$ is in-sample goodness of fit and $Q^{crit} = Q^{is} - \alpha.K$ is split criterion and $Q^{os} = -\frac{1}{N}\sum_{i=1}^{N}(Error)^2$ is out-of-sample goodness of fit.

2.0.3 Model then split

The model then split subpopulation treatment effect modeling MTS-STEM approaches do not split the data, nor modify it, but model the STE directly. Then, in the last step of modeling, the estimated STE is computed by doing one split for a subset of data.

This approach is mostly used with modified decision trees. The authors in [1] proposed the first model that used to calculate STE, it is called differential response tree. It is based on a modified cart tree. The idea is simple; they change the binary split criterion of each node to favor the split that maximums the STE difference between the child nodes and the parent node. In a way that it maximizes the difference between the two STEs while going down the tree. The example illustrated in Figure 2.1 shows an application of the differential response tree applied to a mail marketing campaign dataset. The authors in [1] concluded that female clients aged more than 40 years old have higher STE (uplift) and they exhibit better target for future campaign.

The authors in [60] described what is called Incremental break-even decision rule as a STEM alternative to the normal break-even decision rule. The break-even decision rule is described by [5] as a tool used in economics as an indicator to help in choosing the profitable investments. The idea is to select only clients that have *Net Expected Incremental Profit* greater than zero.

Additionally, [61, 14, 62] used ensemble techniques with STE models. For example, [36] used the random forests technique to calculate STE adds a significant advantage especially because of the uncertainty problem of STE data 3.1. The authors in [17] proposed a new STEM approach called *Uplift random forest*. It is a tree based method, it uses the standard random forest methodology [63], but using STE-based splitting criteria proposed by [64], the pseudo-code of the uplift random forest approach is presented in Algorithm 2.

Algorithm 1: Causal trees

Input: F as the features array, split criterion :(Q^{crit}), A is a set potential penalty values of α;
Set: α as the initial penalty value; $C = \{\}$ as a set of the completed nodes;
R-fold cross-validations $R = 10$;
for *each* a *in* A **do**
  for *each* r *in* R **do**
    Select the complement subpopulation r' ;
    Build a tree $CausalTree^{a,r}$ using r';
    repeat
      Calculate STE for the current node;
      for *each node* ***not*** *in* C **do**
        Set the current node t;
        Calculate the Q^{crit} for the current node $Q^{crit}_{original}$;
        for *each feature* f *in* F **do**
          Find the best split point sp_f the maximizes Q^{crit};
          Add $Q^{crit}_{sp_f}$ to the list of Q_{sp_f};
        end
        if $\max_{f=1}^{F} Q_{sp_f} <= Q^{crit}_{original}$ **then**
          Add the leaf node t to the set of completed nodes C;
        end
        else
          Split based on the feature with maximum Q_{sp_f};
        end
      end
    until *all terminal nodes are in the set* C;
    Prune the causal tree;
    for *each node in* $CausalTree^{a,r}$ **do**
      calculate ΔT as follows:;
      $$\frac{Q^{is}(STE^{before-pruning}) - Q^{is}(STE^{after-pruning})}{|STE^{before-pruning}| - |STE^{after-pruning}|} ;$$
    end
    Set $STE^{a,r}$ as the $STE^{after-pruning}$ for the $\max \Delta T$;
  end
  Set the average of goodness-of-fit measures for penalty value a as $\bar{Q^{os}}(a) = \frac{1}{R}\sum_{r=1}^{R} Q^{os}(STE^{a,r})$;
end
Choose $a*$ as the maximum $\bar{Q^{os}}(a)$;
Output: The causal tree with $a = a*$;

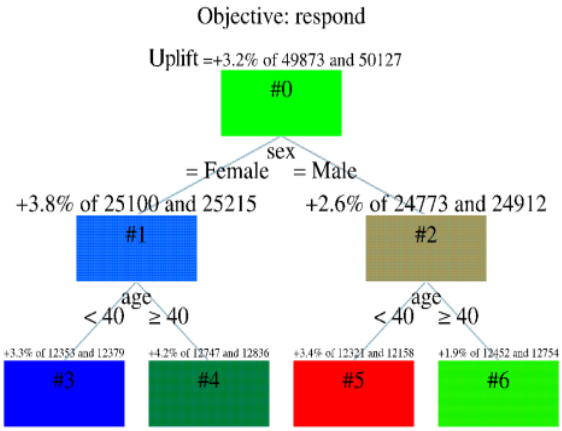


Figure 2.1 – Differential Response Tree, Source: Radcliffe(1999) [1]

Algorithm 2: Uplift Random Forest

Input: F as features array, B as the number of trees, N_{sample} as population sample ratio, F_{sample} as the number features sample, n_{min} is minimum node size, *split criteria* :$(KL_{gain}, E_{gain}, \chi^2_{gain},$ *or* $L1_{gain})$;

Set: N as number of cases;

for $b = 1$ *to* B **do**

  Sample a fraction N_{sample} of the training observations N without replacement;

  for *each node* **do**

    repeat

      Select F_{sample} feature at random from the F features;

      Select the best split-point among the F_{sample} features based on *split criteria*;

      Split the node into two branches;

    until *a minimum node size* n_{min} *is reached*;

  end

end

Output: the ensemble of Uplift trees UT_b; b ={1,...,B};

We can get the estimated STE for x_i by averaging the score of the individual trees in the ensemble

$$STE(x_i) = \frac{1}{B}\sum_{b=1}^{B} UT_b(x_i) \tag{2.2}$$

The Splitting criteria used in uplift random forest (Kullback-Leibler divergence, Squared Euclidean distance, Chi-squared divergence, and L1-norm divergence) are proposed by [65], they are based on maximizing the divergence of class distribution between treatment and control groups.

The authors in [17] proposed another STEM ensemble method called the causal conditional inference forest (CCIF), it minimizes the over-fitting and the bias in the variable selection process. The enhancement is done by separating the splitting criteria process from variable selection and using an efficient stopping criterion. In CCIF, at each terminal node (step 4), we test the null hypothesis of no interaction effect between the selected features and the treatment variable. We stop if we cannot reject the null hypothesis, the pseudo-code of CCIF is presented in Algorithm 3.

Algorithm 3: Causal conditional inference forest

Input: F as features array, B as the number of trees, N_{sample} as population sample ratio, F_{sample} as the number features sample, n_{min} is minimum node size, *split criteria* :(G^2, $KL_{gain}, E_{gain}, \chi^2_{gain}$, *or* $L1_{gain}$);
Set: N as the number of cases;
for $b = 1$ *to* B **do**
  Draw a balanced treatment/control sample from N;
  for each node **do**
    repeat
      Select n covariates at random from the p covariates;
      Test the global null hypothesis of no interaction effect between the treatment T and any of the F variables;
      if the null hypothesis cannot be rejected **then**
        Stop
      end
      else
        Select the feature k with the strongest interaction effect (i.e., the one with the smallest adjusted P value);
        Select the best split-point among based on the *split criteria*;
        Split the node into two branches;
      end
    until a minimum node size n_{min} is reached;
  end
end
Output the ensemble of causal conditional inference tree $CCIT_b$; $b = \{1,...,B\}$

Where G^2 represents the significance-based splitting criterion proposed by [21].

$$G^2 = \frac{(n-4)(STE_{righ\ node} - STE_{left\ node})^2}{(\frac{1}{N_{tr}} + \frac{1}{N_{tl}} + \frac{1}{N_{cr}} + \frac{1}{N_{cl}}) * SSE}$$

And n is equal to the total number of observations, $N_{tr}, N_{tl}, N_{cr}, N_{cl}$ are the number of observations for treatment and control in the left and the right node, the sum of squared errors (SSE) is,

$$SSE = \sum_{t \in 0,1} \sum_{y \in 0,1} N_{tr} \times Pr(Y = y|T = t) \times (1 - Pr(Y = y|T = t))$$

Then We can get the estimated STE for x_i by averaging the score of the individual

trees in the ensemble

$$STE(x_i) = \frac{1}{B}\sum_{b=1}^{B} CCIT_b(x_i) \tag{2.3}$$

Using what is called "Honest estimation," which assures less biased binning process. They solve the problem of binning by using two samples, one sample to choose how to partition and the other to estimate the STE for each node.

A treatment dummy approach was proposed by [4], by fitting one model for modeling the main response effect and adding a treatment dummy variable to detect the interaction between the features and treatments. Let $T \in \{0, 1\}$ denotes the treatment variable and Y denotes the response variable. We can fit a logistic regression model as follows:

$$\widehat{Y_i} = E[Y|X_i] = \frac{exp(\alpha + \beta' X_i + \theta T_i + \gamma' X_i T_i)}{1 + exp(\alpha + \beta' X_i + \theta T_i + \gamma' X_i T_i)}$$

Where α denotes the intercept, β is a vector of parameters denotes the effect of variables, θ denotes the main treatment effect, and γ denotes the STE effect. The final estimate of STE is calculated by setting the treatment dummy variable in the model as treated (T=1), then as controlled (T=0) as follows:

$$STE(x_i) = \frac{exp(\alpha + \beta' X_i + \theta T_i + \gamma' X_i T_i)}{1 + exp(\alpha + \beta' X_i + \theta T_i + \gamma' X_i T_i)} - \frac{exp(\alpha + \beta' X_i)}{1 + exp(\alpha + \beta' X_i)}$$

Also, authors in [8, 38] proposed K-Nearest Neighbors approaches (KNN) [66] for modeling STE. KNN, has been first used to calculate STE by [8]. The idea behind their method is simple, [8] named their algorithm "Patients-Like-Me Algorithm," and it does so by using the KNN algorithm. By applying this approach, the algorithm sequentially tests the patients that are near to our target patient (i.e., share similar features) until a significance difference from the average treatment effect (ATE) is found. The disadvantage of this method is that it is computationally intensive, especially for high dimensional data.

2.1 Subpopulation treatment effect modeling evaluation

Measuring the performance of STE models is one of the critical challenges in STEM domain. The classical techniques to evaluate machine learning methods are based on cross-validation, where the dataset is separated into a training dataset and a validation dataset, then a loss function is applied on the validation dataset to measure the error rate that reflects the ability of the model to fit the true values. However, because of the fundamental problem of causal inference, where we only know the outcome of one treatment only, so we do not have the true treatment effect value. For STEM evaluation, the problem is more complicated.

The only way to measure the STE models is by emulating the true value through a randomized experiment that respects the three conditions (exchangeability, positivity, consistency) [39]. For non-experimental datasets, the true value is emulated

by a matching algorithm, where similar cases are grouped to calculate the aggregated true value (STE) using the equation A.2.1.

In this section, we review evaluation approaches for assessing the performance of STE models.

2.1.1 Qini measures

Qini measures are rank-based measures introduced by [67] as a generalization of the Gini measures (i.e., Gains)[68]. Qini measures are based on Qini Curve (the gains chart of STE), which is similar to a gains curve. To draw the Qini curve, cases should be sorted in descending order based on their predicted STE (model's score). Then, the cases should be binned into subgroups. The purpose of binning is to get the "Actual STE" of the subgroup so that we can compare it to the estimated STE. Then, on the Y-axes of the chart, the Qini curve, or the cumulative incremental gains curve, is plotted (see Figure 2.2).

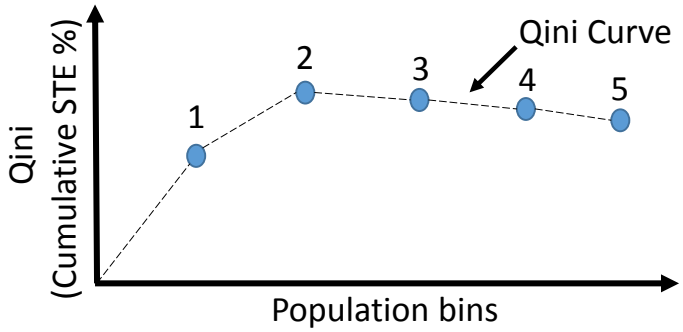


Figure 2.2 – Qini curve.

The Qini curve represents the cumulative STE gains that are obtained if we sort the cases based on their STE estimation. The motivation behind Qini curve is to compare the Gains of each STE model based on the sorting efficiency of cases, and not by comparing the predicted value with the true value (because we do not have access to true value). The Qini curve begins from the zero point and ends at the total STE gains. However, if we randomly sort the cases, we will get a cumulative STE gains that are equal to the total STE gains (see Figure 2.3). We can draw the random STE line as a diagonal line between the zero point and the total cumulative STE gains point.

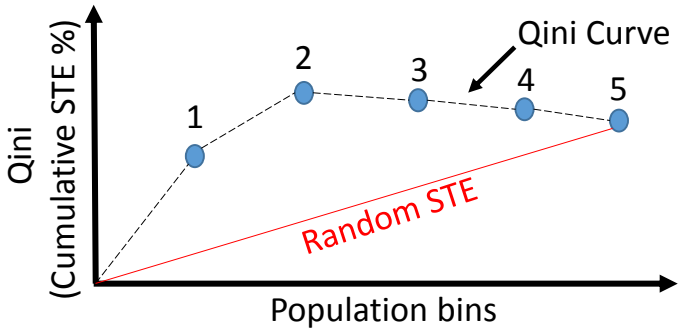


Figure 2.3 – Qini curve with random.

Also, for a binary treatment and a binary response example, we can draw the optimal Qini curve by sorting the cases in the following order, treated and responded cases, followed by controlled and not-responded cases, followed by the controlled and responded cases, and finally the treated and not-responded cases (see Figure 2.4).

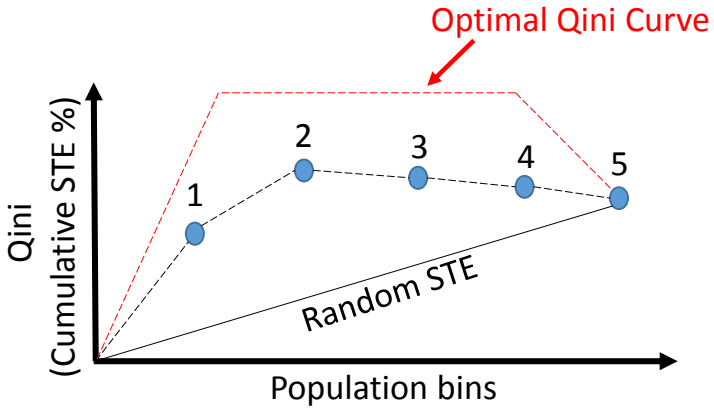


Figure 2.4 – Qini curve with random and optimal curve.

Qini measures consist of the Qini value Q, q_0, and top qini. The Qini value Q is the ratio of the area of the model's Qini curve above the random line to the optimum Qini curve above the random line. And the q_0 is the ratio of the area of the model's Qini curve above the random line to the optimum Qini curve above the random line without the negative effect.

In some domains, it is favorable to minimize the targeted population. For example, in marketing, minimizing the targeted population will reduce the campaign cost

that would maximize the return on marketing investment. In those cases, comparing the top 5%, 10%, and 20% Qini for different STE models will provide the beneficial information that is required to minimize the targeted population. For example, in future marketing campaign, they can target clients in highest 10% of Qini, those clients are predicted to have high response rate for the advertisement.

2.1.2 Gini coefficient

Authors have used the Gini coefficient [69] as a measurement of STE [70, 6]. Gini coefficient is calculated as the ratio of the area of the model's cumulative gains curve above the random line to the optimum curve above the random line (see Figure 2.5).

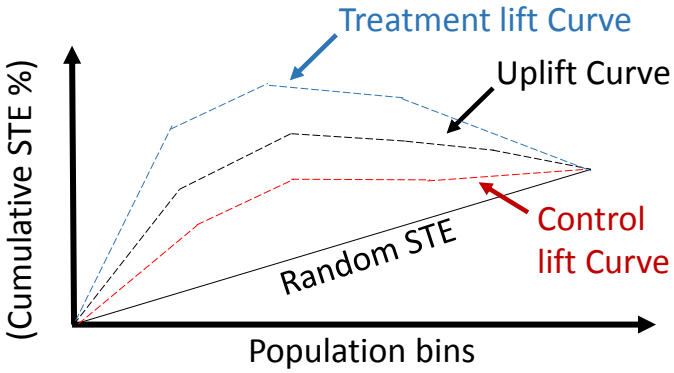


Figure 2.5 – Uplift curve.

Area under uplift curve (AUUC)

The area under the uplift curve is similar to the Qini value Q. It has been used in [7, 71, 62, 64, 9]. First, we draw separate lift curves on treatment and control data. Then, we draw the uplift curve which is equal to the difference between the lift curve on treatment data and lift curve on control data. The uplift curve shows the net gain in STE if a given percentage of the population is targeted or treated (see Figure 2.5).

2.1.3 Moment of uplift measures

Moment of uplift measures are also ranked based measures that have been developed by [72] to capture the accuracy and boldness in STE model's prediction. The authors in [72] concluded that the quadratic form of the moment of uplift can be a better replacement for Qini and other STE measures in STE model assessment. Moment of uplift measures focus on STE model's prediction qualities like monotonicity, prediction errors, spread and maximum STE.

2.2 Subpopulation treatment effect modeling names

In the literature of STEM, various terms have been used to refer to STEM in the literature, for this reason, it was important for us to study the terms and the domains. This step is an attempt to unify the terminology used for STEM in order to facilitate the search for literature for future researchers.

The main reason behind the various names of the STEM is the diverse domains that use it. In the Figure 2.6, we present the main journal and conferences that published STEM related papers.

In Figure 2.6, we summarize the domains of STEM by gathering all most of the journals and conferences that published papers contains one or more term included in Figure 2.7.

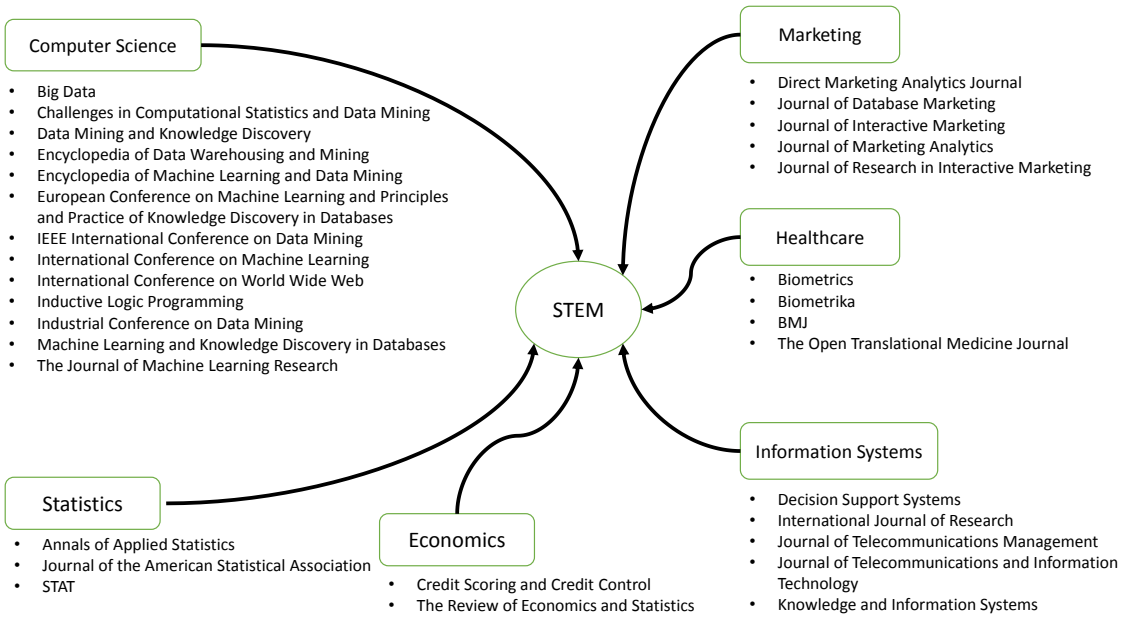


Figure 2.6 – Journals and Conferences that published about subpopulation treatment effect modeling.

Based to our research, we combined all the names of STEM in all domains and present it in the Figure 2.7.

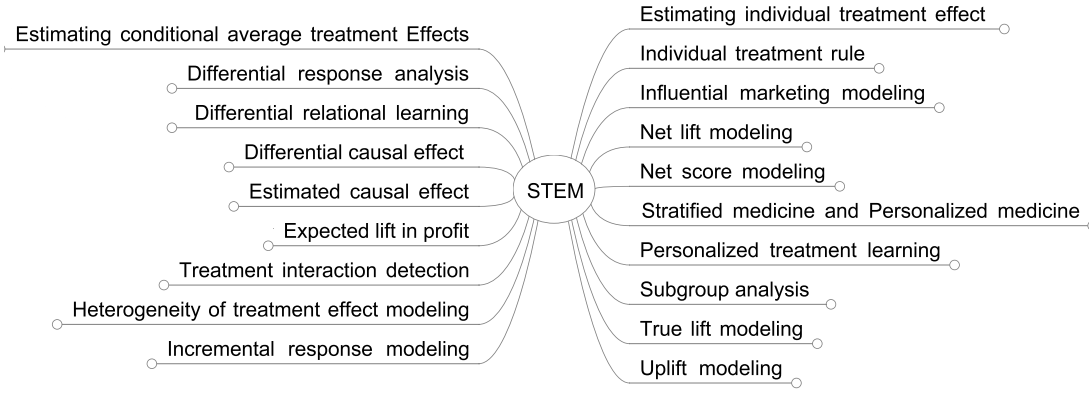


Figure 2.7 – Subpopulation treatment effect modeling names.

Table 2.1: Table of Subpopulation treatment effect modeling terms.

Begin of Table	
STEM term	Description
Estimating conditional average treatment effect	This term conditional average treatment effect (CATE) is mostly used in causality, politics, statistics, and economics domains [73, 74, 19, 27, 22, 75]. Also, [76] propose the term conditional average treatment effect function (CATEF) to refer to the methods to estimate the heterogeneity of the average treatment effect with respect to observed covariates of interest.
Differential response analysis	This term is introduced by [1] in their paper which was the first paper that proposed an *STEM* approach in the marketing domain. Same as Differential Causal Effects, the term lacks the specificity.
Differential relational learning	It was introduced in 2002 by [77]. The work is motivated by an application to use mammography for better classification of patients. The term is also used in [57].
Differential causal effects	This term was introduced in [38] to study the effect of treatment across subpopulations.

Continuation of Table 2.1

STEM term	Description
Estimated causal effect	The author [78] defined ECE as the statistical estimate of ICE *Individual Causal Effect*. The authors use the name *Cadit Modeling* in the paper [79].
Expected lift in profit	Driven by its business application, [51] introduced the *Expected lift in profit* to model the best targeting strategy in order to maximize the profit lift.
Treatment interaction detection	The term, Treatment interactions detection, has been used in [80, 11]. Their work is motivated by determining from a large set of biomarkers the subset of patients that can potentially benefit from a treatment.
Heterogeneity of treatment effect modeling	It is the most used and correct term for STEM. It is related to the statistical, bio-statistical, medical, and economic literature. Based on our knowledge, the term "heterogeneity of treatment effects" was introduced first by [81], the authors used this term to explain the heterogeneity of the treatment effects across many studies and not the heterogeneity in the same study. In 1992, [82] derives a test for the heterogeneity of treatment effects in a data consists of matched pairs. The term is used in many other papers like [83, 84, 28, 85, 43, 19, 86, 87, 88, 89, 22, 90, 91].
Incremental response modeling	Incremental response modeling and Incremental value modeling has been introduced by [60]. The motivation behind this term is to detect the customers who will positively change their behavior toward a desired action only if they are targeted by a marketing communication (i.e., promotion, coupon, advertisement, etc.). In other words, they will increase their value, as a customer, if they receive a marketing communication action. It is used by [60, 5, 92, 93, 94, 95, 96].
Estimating individual treatment effect	This term is a synonym of the widely used term *Individual causal effect*; it is used in epidemiology and causality literature, it is used in[97, 98, 14, 17, 12, 99, 100, 101, 102, 10].

Continuation of Table 2.1

STEM term	Description
Individual treatment rule	[103] introduced this term. The *Rule* in this term represents the map from the space of features to the space of the treatments, and the optimal rule is the rule that maximized the expected outcome. This term is also used by [59, 19, 14, 57].
Influential marketing modeling	Introduced by[56], driven by the marketing application of STEM. The author used the word *influential* because its implementation aims to find only customers that can be positively influenced by marketing communication.
Net lift modeling	Introduced by [49]. The term has been used later in [104, 105, 106].
Net score modeling	This term is similar to the *Net lift modeling*, using the score instead of lift is just a reflection of its marketing application. It is used in [107, 108, 109].
Stratified medicine and Personalized medicine	Stratified medicine is one of the main motivations of the STEM. Many researchers used this term, like [99, 98, 8].
Personalized treatment learning	Introduced by [14], the term PTL is called on the task of learning the optimal personalized treatment that maximized the probability of the desired outcome.
Subgroup analysis	The term *Subgroup analysis* describe correctly the goal of STEM, the authors in [37] mentioned that the purpose of their work is to detect and identify the heterogeneity of the treatment effect across subpopulations, which is precisely what STEM do.
True lift modeling	It has been introduced in 2002 by [4]. This term is named *True lift*, and is aimed to differentiate it from the regular Lift that can be modeled using response modeling. The term then has been used in [110, 6].

Continuation of Table 2.1	
STEM term	Description
Uplift modeling	The most popular term for the STEM in the marketing literature is *Uplift Modeling*. It was introduced in 2006, in a white paper by *Portrait Software TM* [111]. After that, the term *uplift model* appeared in a paper by Radcliffe [70], in their paper, the authors explains in detail what is uplift models, how it is built, and what are its boundaries and limits. The term, *uplift modeling* has been used in [65, 21, 112, 113, 114, 72, 9, 115, 7, 34, 62, 108, 13, 116, 117, 118, 119].
End of Table	

2.3 Conclusion

In this chapter, we introduced the idea of subpopulation treatment effect modeling (STEM) from the probabilistic and the causal perspectives. First, we showed the benefits of applying STEM in three use cases from the marketing, healthcare and economics domains. Then, we introduced the notations and the main definitions. Second, we presented the subpopulation treatment effect (STE) from causal point of view and we differentiate between subpopulation treatment effect and personal treatment effect. Third, we established the 10 reasons behind having STE in an experiment, then we distinguish between real STE and accidental STE. We used causal directed acyclic graphs to demonstrate between the 10 different causes of STE. Then, we showed how we can control the variables to calculate STE. Fourth, we introduced a new taxonomy for STEM approaches, our new taxonomy is based on the three main processes, split, model, and transform. Later, we showed the techniques and measures that are used to evaluate STEM approaches in the literature. Finally, we revealed the different terms that are used for the STEM in multi domains.

Chapter 3

Uncertainty and Subpopulation Treatment Effect Modeling

> In science, read, by preference, the newest works; in literature, the oldest.
>
> Edward G. Bulwer-Lytton,
> *Caxtoniana: A Series of Essays on Life, Literature, and Manners*

Subpopulation treatment effect modeling (STEM) is a branch of machine learning concerned with designing models that map treatment effect variations across the population and is widely used in business and healthcare sectors. Various difficulties arise due to the fact that STEM techniques are based on causal inference principles [38, 120]. In particular, the difficulties is originated because the standard classification machine learning algorithms are not adapted to deal with STEM difficulties.

In fact, the most difficult problem facing STEM is uncertainty, which manifests as missing information problem. Typically, binning techniques are utilized by standard STEM approaches to address uncertainly. The concept is to convert the population into subgroups, then replace the missing information by values from the subgroup. The high uncertainty characteristic of data composes STEM models more sensitive to bias and noise in data.

This chapter investigates uncertainty problems in the STE modeling context, and discusses current binning technique limitations. We propose neighborhood based STE binning and the STE sliding trees (STE-STrees) framework to model treatment effect heterogeneity using neighborhood based binning. Then, we introduced the neighborhood random forests (STE-NRF) framework which improved the ability to detect STE in noisy scenarios. Later, we propose the balanced reflective uplift modeling (BRUM), an improved STEM technique that is more tolerant to disturbances in data.

We compare our proposed approaches performances to current approaches using simulated and real datasets related to business and healthcare domains. The results

verify that the proposed approaches outperforms existing methods in terms of Qini coefficient and Spearman's rank correlation coefficient.

3.1 Context and problem statement

3.1.1 General context

The dramatic increase of accessible data and the high demand for techniques to aid decision support systems encouraged the emerging of learning from data phenomenon to provide the essential tools to maximize decision confidence, minimize side effects, and facilitate decision making processes. In many circumstances, customized decisions must be made based on each case's characteristics. Subpopulation treatment effect modeling (STEM) helps the decision making process and maximizes particular action effects.

Suppose a doctor must decide on a specific medical treatment among three available therapies. Then STE modeling can be employed to choose the most beneficial treatment for each patient and simultaneously minimize any potential side effects. In a different context, suppose a marketing manager must select the client segment to receive a specific promotion due to limited product or project budget. STEM can be applied to identify the most profitable customer subgroup, avoiding customers that will purchase the product anyway (i.e., without requiring promotion).

A critical challenge for STEM is information uncertainty. Data uncertainty always exists due to fundamental causal inference [41], i.e., only a subset of treatments can be directly observed. This problem is commonly addressed by employing grouping techniques. In machine learning with observational data, specific binning techniques are required to formulate appropriate groups to allow using the data for inference.

Those binning techniques are facing harder challenges when they encounter a continuous feature to bin. The previous STEM approaches do not properly process continuous attributes. Mainly, the current binning technique for a continuous variable in STEM decision trees (i.e., the splitting criteria) do not aid in STE discovery across subpopulations.

The current binning techniques are based on threshold partitioning, i.e., finding the optimal partition-point. Threshold partitioning will try to construct a new child node (subgroup) in a way that maximizes the STE difference between the child node and the main node. By doing so, the threshold binning guarantees a more target homogeneous subgroup after each partition. For nominal features, the partition-threshold works perfectly. However, for continuous attributes, this b-threshold suffers from uncertainty and error because of three points:

— The fundamental problem of causal inference, we only know part of the true value, which results in uncertain true classes.

— The rigid binning procedure, which depends on these uncertain true classes, will lead to uncertain partition points, and hence over-fitting.

— The consequences of one error splitting point will affect all the other child nodes, which will later lead to incorrect results.

We need a customized solution for STEM that can handle continuous attributes and minimize the uncertainty and over-fitting of STE models.

We first analyze current binning techniques typically used by STEM approaches, and identify current binning technique limitations. We then propose the neighborhood based STE binning approach and compare various binning techniques considering all potential outcomes, and identify proposed STE binning benefits.

Subsequently, we propose STE neighborhood random forests (STE-NRF) approach, employing neighborhood based binning to improve STEM performance. We compare different STEM approaches by simulating eight different scenarios and real datasets from medical and business domains. We show the proposed approach outperforms all other STE methods in the measurements and terms used in the literature. Finally, we discuss limitations and future perspectives arising from this research.

As mentioned before, STE models are based on the causal inference framework, and thrive under certain conditions that guarantee model certainty. For simplicity, let us assume an experimental dataset with perfect conditions (i.e., exchangeability, positivity, and consistency), which guarantee perfect certainty. Perfect certainty indicates that, if two members with the same attribute values receive the same treatment, T, they respond in the same way, i.e.,

$$\forall i, j \in A; if x_i = x_j \implies y_i|T = y_j|T,$$

where x_i, y_i, x_j, y_j are vectors of features and response values for the i, j members, respectively. A controlled experiment under these perfect conditions can return the true personal treatment effect, PTE, which in this case is equal to the subpopulation treatment effect, STE, directly using the causal inference or risk difference relationship

$$STE_i = PTE_i = \text{Response rate}_i \text{ for treatment} - \text{Response rate}_i \text{ for control.}$$

For example, in an experimental datasets, we randomly selected two members from the population and conducted a treatment/control experiment to estimate the treatment effect by finding the difference between the treatment and control response,

$$\widehat{STE_i} = \frac{\sum_{i=1}^{n} \widehat{PTE_i}}{n} = Pr[Y = 1|\text{Treatment}] - Pr[Y = 1|\text{Control}]$$

which is valid because the control and treatment groups were exchangeable (see Fig. 3.1).

However, for non-experimental datasets, randomly selecting two members from the dataset with similar characteristics, where one is assigned to treatment, and the other is assigned to control (post assignment), cannot guarantee exchangeability. Thus, we cannot be certain if treatment and control members are exchangeable, i.e., one of the two STE equation variables is uncertain (see Fig. 3.2),

$$\widehat{PTE_i} = Pr[Y = 1|\text{Treatment}] - \text{Unknown}$$

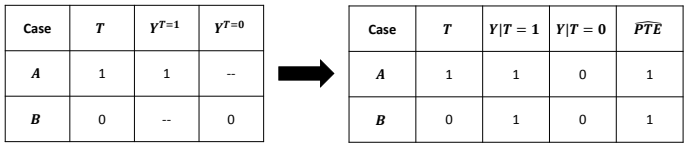


Figure 3.1 – Randomized experiment

or

$$\widehat{PTE_i} = \text{Unknown} - Pr[Y = 1|\text{Control}].$$

Therefore, we can use STEM for non-experimental datasets, but we must consider that the data has maximum 50% certainty.

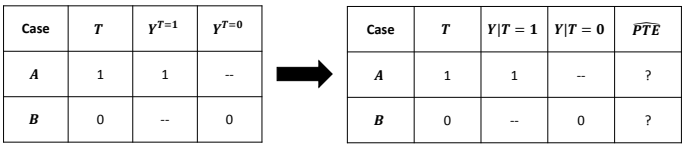


Figure 3.2 – Non-experimental results

Thus, to ensure STE models work with non-experimental datasets, we assume that members with similar features will respond similarly to the same stimuli, which injects more uncertainty into the model and hence decreases certainty even further.

Similar to the causal inference framework, STEM relies on subgroups (or bins) to construct inference between groups. Therefore, at least one member from each treatment should be present in any subgroup to be considered informative. The estimated true value can then be calculated for a treatment or control experiment as

$$\widehat{STE_i} = \text{Response rate}_i \text{ for treatment} - \text{Response rate}_i \text{ for control.} \tag{3.1}$$

Various studies have considered many datasets to model STE, focusing on increasing model accuracy in the presence of high data uncertainty. Generally, these approaches have avoided modifying the original data, because the lower certainty will cause significant impact on model inferential ability. In addition, most real datasets include missing, erroneous, and biased data.

As discussed above, data subgrouping or binning is the core process for any STEM algorithm, and low data certainty requires extreme care when processing the data. Binning problems are exacerbated when data is unsuitable for binning, i.e., not nominal data. In this case, we rely on other tools to transform the data to be more suitable for STE models, in particular, to bin the data into chunks that maximize information gained from each bin.

3.1.2 Binning

Binning simplifies the data, speeds the process, and facilitates interpretation; and has been widely investigated for data mining and exploration. Some data mining algorithms do not support continuous data as input, hence binning converts continuous to categorized data [121]. However, any binning on a continuous variable inevitably causes information loss [122, 123, 124], and the goal is to minimize this loss.

The most common binning example is histogram bin assignment, where an algorithm decides the appropriate proper number of deciles (bins) to provide a simple visualization. The same concept is applied for data mining, using binning to group members in the most informative way.

Suppose we have a dataset with N members and the binning algorithm divide the continuous variable, X, into m bins $P = p_1, p_2, p_3, ..p_m, p_i$, representing intervals of x values. Then the bin mean can be expressed as $\overline{p_i} < \overline{p_{i+1}}$ for $i = 1, 2, \ldots m$.

The most widely used technique to bin the population for STEM is the difference of the difference of STE, $\Delta\Delta STE$ [21], which calculates achieved gains from a specific binning as the difference between the STE of the original population (STE^{original}) and the difference between the two bins (i.e., $\Delta STE = |STE^{\text{first bin}} - STE^{\text{second bin}}|$). This maximizes STE difference between the original and children nodes in each tree split. Other approaches use distributional divergence binning techniques, e.g. Kullback-Leibler divergence, squared Euclidean distance, and χ^2 divergence between each STE binning [65, 115, 14, 17].

Significance based splitting criteria has also been employed as a binning technique [21, 17]. This approach is equivalent to the χ^2 test of the interaction effect [37]. Significance based splitting criteria exhibits good results compared to other binning techniques. In addition, the binning problem has also been addresses using honest estimation [125], which assures a less biased binning process and employs two samples to solve the binning problem using one sample to choose how to partition and a second sample to estimate STE for each node.

All STE related binning techniques attempt to find subgroups with heterogeneous treatment effect. Binning rules are based on the response variable, where response rate heterogeneity defines the subgroup boundaries. However, incorrect binning could produce erroneous estimates and biased models. In particular, binning continuous variables offers more splitting points than categorical variables. Therefore, continuous variables present further binning problems. STE approaches should find enhanced solutions for ordinal and continuous features. The STEM challenge is to uncover subgroups with heterogeneous responses that share common characteristics, despite noise and data uncertainty.

Subsequent sections explain the response based binning technique for STEM and its weak points. We then introduce the proposed STE neighborhood based binning approach and compare the techniques using simulated data considering all combinations of potential outcomes.

3.2 Subpopulation treatment effect response based binning

As discussed in Section 3.1, STE uses binning to avoid causality problems, generally based on the response variable. STE response based binning determines whether a given multiset of integers, S, can be binned into subsets $S1$ and $S2$ such that estimated subset STE difference is more significant than the original subset. Members with a common response are grouped into one subpopulation.

Suppose we have a sorted population with five members, and we want to bin the population such that at least two members in each bin, preserving higher subgroup homogeneity by not changing member order. Figure 3.3 shows the two possible binning solutions.

— Case A. Members with ID 1 and 2 are grouped with

$$STE_{\mathrm{A}_{left}} = Pr(\text{Response}|\text{treatment}) - Pr(\text{Response}|\text{control}) = 1 - 0 = 1,$$

and members with IDs 3–5 are grouped with $STE_{\mathrm{A}_{right}} = -0.5$. We calculate STE for each subgroup base on (Equation 3.1).

— Case B. We assemble members with IDs 1–3 together with $STE_{\mathrm{B}_{left}} = 0.5$, and ID 4 and 5 together with $STE_{\mathrm{B}_{right}} = 0$.

Case A bins the population with difference (Figure 3.3)

$$\Delta STE_{\mathrm{A}} = |STE_{\mathrm{A}_{left}} - STE_{\mathrm{A}_{right}}| = 1.5$$

, and case B has $\Delta STE_{\mathrm{B}} = 0.5$.

Response based binning prefers case A, because this has larger variance between the two subpopulations, and satisfies the homogeneous subgroup requirement.

Response based binning methods find the most significant variance between subpopulations based on a specific response measurement. In the current example, we use the difference between the two STEs, but other binning methods have also been used for STE binning (Section (A.2.2), including Euclidean distance [126], Kullback-Leibler divergence [127, 128], and χ^2 divergence [128, 129]. All those methods are response based binning techniques.

After each binning, the model saves binning locations as a rule and these binning rules are used to classify future unlabeled cases. For example, suppose a new member has similar characteristics as ID 2. Then the new member is given the same score as the subgroup that contains ID 2, i.e., $STE_{\mathrm{A}_{left}} = 1$ in this example.

Response based binning is effective for nominal variables, but problematic for continuous, interval, or ordinal variables. In these cases, the requirement for homogeneous response subgroups risks losing valuable information, effectively increasing noise [130, 131, 132, 121].

Section 3.3 details the STE neighborhood binning (STE-NBB) approach and underlying concepts. We provide a simple example to demonstrate the difference between conventional STE binning and STE-NBB and compare the techniques using simulated data, considering all potential outcome combinations.

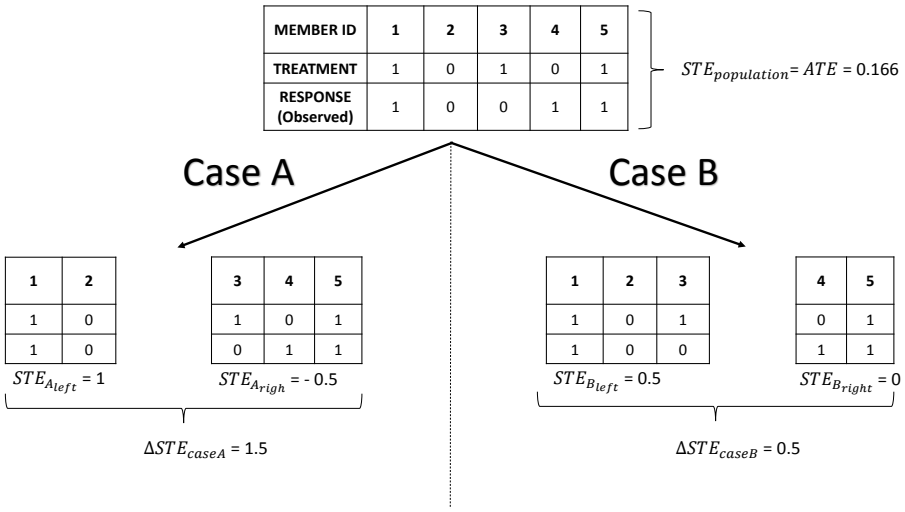


Figure 3.3 – Subpopulation treatment effect (STE) response based binning example.

3.3 Subpopulation treatment effect neighborhood based binning

3.3.1 Proposed solution

We propose neighborhood based binning (STE-NBB) to mitigate response based binning limitations discussed above. Neighborhood based binning assumes that nearby cases have similar treatment effects, hence their model scores should not vary. Therefore, neighborhood information can be used to define a bin, and bin member $\widehat{PTE}$ values are the $\widehat{STE}$ for their bin (Figure 3.8).

We use sliding windows (SW)[133] to apply STE neighborhood based binning. The sliding window limits the number of members in a window, ensuring the most appropriate member descriptions. Thus, two main parameters are required: window and step size. Window size specifies the number of members each window can include, and step size specifies the number of members each window should skip to start a new window group.

Given a numerical feature $X; x \in R$, consider a window (subgroup) of length n, where the window's length represents the number of members that are included in each window. In addition, consider a step size k, where the step size represents the number of members that the window slides (moves) across in each iteration. A sliding window algorithm with window size n and step size k applied to feature X with a number of members equal to m will create L different windows, where $L \cong \frac{m-n}{k} + 1$.

Figure 3.4 shows the application of the SW algorithm to feature X, with a window size of three members and a step size of two members. For each window w_i, we apply

a specific function $f(w_i)$, and then assign the value of the function to the associate variable SW_i. Notice that the SW algorithm converts the continuous variable to an ordinal subpopulation (ordered categorical variable). Hence, it is a suitable solution for the problem of STEM (because of the fundamental problem of causal inference, STE models require subpopulation data and not data for unique individuals).

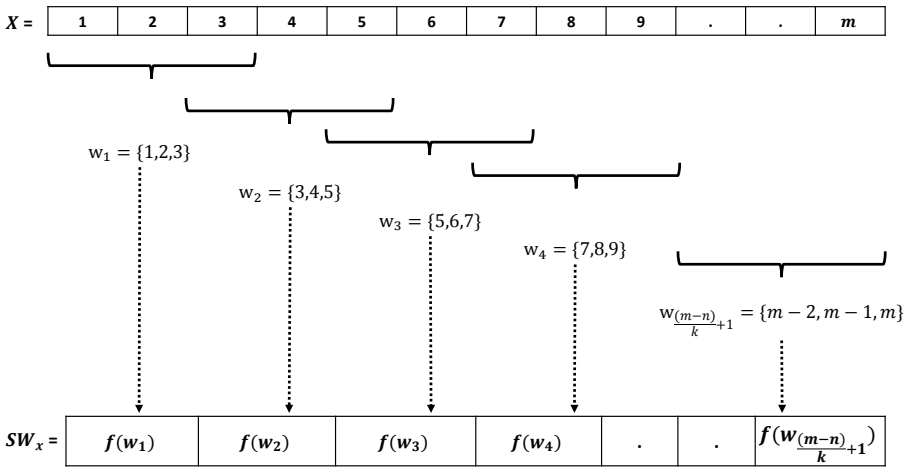


Figure 3.4 – Sliding window

Based on the assumption that nearby cases form more homogeneous subgroup, the main benefit of utilizing the sliding window technique for STEM is the emphasis maximizes the homogeneity of the subpopulations. The SW defines a fixed zone around each member (the window size), which will help ensure the localization of the estimated score. On the other hand, by taking into consideration the values surrounding each member, the SW will maximize the generality of the STE model and minimize the effect of erroneous (or outlier) values. These features of the SW provide smoothness to STE model estimations and ensure the homogeneity of the subpopulations.

Two main problems arise in the process of applying the SW in STEM.

— The first problem is the accumulation of noise from other features (features not used by the SW algorithm).

— The second problem is the evaluation (error measurement) by the STEM.

The problem of accumulating noise from other features is owing to the fact that the number of subpopulations produced by the SW algorithm is less than the number of members. The other features in the conventional SW algorithms are altered using a specified function. The transposition of these other features will alter their original values in a non-systematic way, leading to the buildup of noise, which will produce errors in the models.

An example of the first problem can be seen in *Figure* 3.5. This figure demonstrates an application of the SW algorithm on a sample of the population. The

sample contains five members (ID), two features (X_1, X_2), a binary treatment variable ($Treatment$), and a binary response variable ($Response$). After applying an SW algorithm with a three-member window size, one-member step size, and the arithmetic mean as an SW function, we notice the result of applying the SW to each variable (SW_{X_1}, SW_{X_2}), and the estimated ($\widehat{STE}$) for each window calculated by equation A.1. Notice how the other feature X_2 suffers from a loss of information due to the applied SW. The variable SW_{X_2} contains the same information in all of the windows, leading to faulty modeling and a loss of information.

ID	**1**	**2**	**3**	**4**	**5**
X_1	10	20	30	40	50
X_2	1	0	0	1	0
Treatment	1	0	1	0	1
Response	1	1	0	0	1

SW_ID	**1**	**2**	**3**
SW_{X_1}	20	30	40
SW_{X_2}	0.333	0.333	0.333
SW_{STE}	-0.5	-0.5	0.5

Figure 3.5 – Other-feature problem example (accumulation of noise)

The second problem faced by the SW application in a STEM framework is the model evaluation. The STE model's evaluation is based on the *Qini* coefficient [21], which measures how accurately the STE model predicts the ranking of the populations' STEs. *Qini* requires a comparison of the estimated $\widehat{STE}$ and the real STE, which can only be calculated using the individuals' treatment and response variables. For example, to draw the optimal *Qini* curve [21], the population should be sorted in the following order: *treatment and response* members, then *control and no-response* members, then *treatment and no-response* members, and finally *control and response* members (see Figure 3.6).

ID	**1**	**2**	**3**	**4**	**5**
X_1	10	20	30	40	50
X_2	1	0	0	1	0
Treatment	1	0	1	0	1
Response	1	1	0	0	1

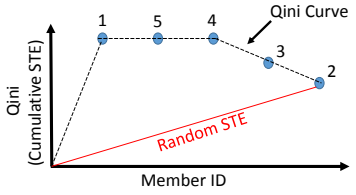


Figure 3.6 – Subpopulation treatment effect modeling Qini curve

The conventional sliding window technique does not keep the individual's treatment and response value, instead it creates new instances with aggregated values for

treatment and response. This will be problematic for drawing the Qini curve, which requires instances to be sorted in a specific order, as mentioned before.

To overcome above-mentioned obstacles, we use a reconstruction function $g(.)$ to estimate the STE for each member of the population.

Given the $SW = \{w_1, w_2, \dots w_{\frac{m-n}{k}+1}\}$, a set of all windows. The sliding function $f(w_j)$, where $j \in [1 \dots \frac{m-n}{k} + 1]$. We note v_i a vector containing the outcomes of the sliding function $f(w_j)$ applied to each window w_j in the subset of windows that includes the member i. For example, in the Figure 3.7, the vector v_3 for the member $i = 3$ is $v_3 = [f(w_1), f(w_2)]$.

We define the subpopulation treatment effect sliding window variable $STE - NBB_i$ for the member i as follows:

$$STE - NBB_i = g(v_i) \tag{3.2}$$

Where $g(v_i)$ is the reconstruction function of the member i.

We note, that the function $g(v_i)$ can be defined in various forms, based on the research problem and the required measurement. For example, $g(v_i)$ could be a summation, a product, or a mean function. In addition, note that the function $f(w_i)$ follows the same form of the conventional sliding window function.

Figure 3.7 demonstrates the application of the *STE sliding window* technique to feature X, with a window size of three members and a step size of two members. For each window w_i, we apply a specific function $f(w_i)$. After this, we assign the value of the function to the associate variable SW_i. Then, we apply the function $g(.)$ to each window to remap the sliding window function's outcomes into the populations' members. Notice in this example how the third member's ($id = 3$) estimation of STE is a product of the first and second windows (because the third member is part of the first two windows).

The outcome of STE-NBB allows us to bypass the problems of the conventional SW. We can use the outcome of our approach to draw the *Qini* curve. In addition, the noise from the other features is not aggregated.

In Figure 3.8, we use the example from section A.3.3 (Figure 3.3) to demonstrate STE-NBB approach.

Neighborhood based binning first decides window and step size. In this example, step size = 1 member, and window size = 2 members. The process begins with the first member, and we calculate $\widehat{STE_{step}}$ for each the window using (3.1). We then assign $\widehat{STE_{step}}$ to each window member. The procedure is repeated until no members remain unassigned. The example requires 4 iterations from the first member to the last. Finally, we calculate the numerical average of $\widehat{STE_{step}}$ to provide $\widehat{STE}$, which we use later for estimation (classification).

3.3.2 Performance evaluation

To evaluate binning technique accuracy, we compared member-wise binning estimation scores with the true treatment effect obtained using the potential outcome

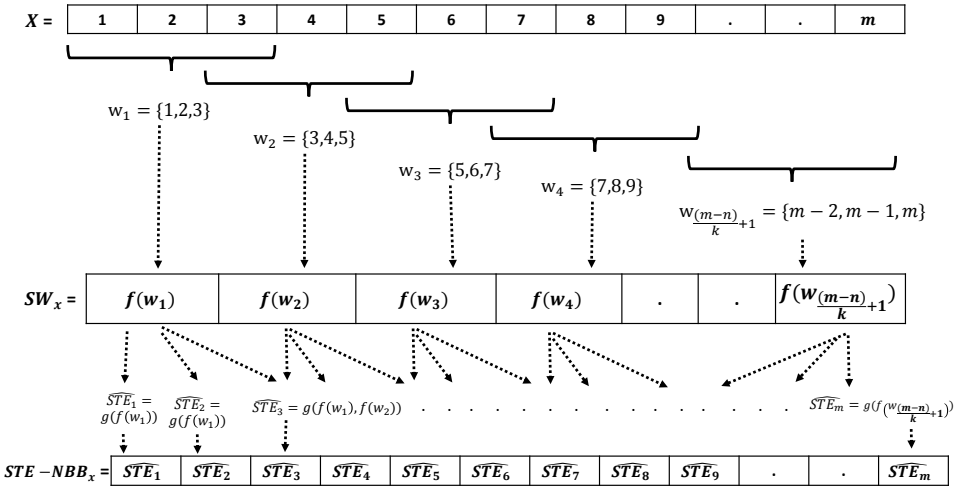


Figure 3.7 – Subpopulation treatment effect modeling neighborhood based binning.

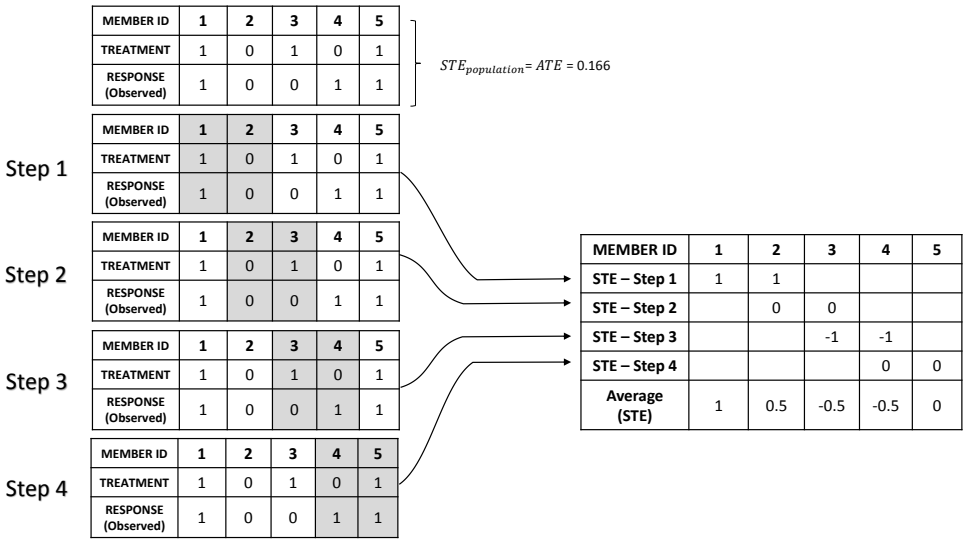


Figure 3.8 – Subpopulation treatment effect neighborhood based binning example.

framework. The population comprises 5 members, with 1 binary treatment/control variable, and 1 binary response variable (observation), as shown in Figure 3.3. We add two new variables, one for each potential outcome, potential outcomes 1 and 2. Potential outcome 1 refers to a potential positive outcome, and potential outcome 2 to a potential negative outcome for the unobserved treatment. We then calculated potential personal treatment effect, PTE, for each potential outcome, as shown in

Figure 3.9.

MEMBER ID	1	2	3	4	5
TREATMENT	1	0	1	0	1
RESPONSE (Observation)	1	0	0	1	1
Potential Outcome 1	1	1	1	1	1
Potential Outcome 2	0	0	0	0	0
Potential PTE 1	0	1	-1	0	0
Potential PTE 2	1	0	0	-1	1

Figure 3.9 – Potential personal treatment effect.

We used mean squared error, MSE, to measurement how each binning technique explains or fits bin members. We calculated MSE for each potential outcome. For example, for member ID $= 1$, we calculate the personal treatment effect (PTE) for the first potential outcome, and then estimate error as the difference between PTE and $\widehat{STE}$ for each binning technique (Figure 3.10). Finally, we calculate MSE for each binning technique.

Figure 3.10 shows that both binning techniques generate $MSE = 25\%$ for the first potential outcome. For the second potential outcome, response based binning generates $MSE = 75\%$, whereas neighborhood based binning generates $MSE = 35\%$. Thus, neighborhood based binning reduced error rate by 20%, and is critical to minimize error.

Potential Outcome 1

MEMBER ID	1	2	3	4	5
TREATMENT	1	0	1	0	1
RESPONSE (Observation)	1	0	0	1	1
Potential Outcome 1	1	1	1	1	1
Potential PTE	0	1	-1	0	0
$\widehat{STE}_{Response}$	1	1	-0.5	-0.5	-0.5
$\widehat{STE}_{Neighborhood}$	1	0.5	-0.5	-0.5	0
$Error_{Response}$	-1	0	-0.5	0.5	0.5
$Error_{Neighborhood}$	-1	0.5	-0.5	0.5	0
$MSE_{Response}$	0.35				
$MSE_{Neighborhood}$	0.35				

Potential Outcome 2

MEMBER ID	1	2	3	4	5
TREATMENT	1	0	1	0	1
RESPONSE (Observation)	1	0	0	1	1
Potential Outcome 2	0	0	0	0	0
Potential PTE	1	0	0	-1	1
$\widehat{STE}_{Response}$	1	1	-0.5	-0.5	-0.5
$\widehat{STE}_{Neighborhood}$	1	0.5	-0.5	-0.5	0
$Error_{Response}$	0	-1	0.5	-0.5	1.5
$Error_{Neighborhood}$	0	-0.5	0.5	-0.5	1
$MSE_{Response}$	**0.75**				
$MSE_{Neighborhood}$	0.35				

Figure 3.10 – Potential outcomes errors.

In this example, the response variable for the five members was $1, 0, 0, 1, 1$ respectively. Since we have a binary response variable, there are $2^5 = 32$ different possible response combinations (potential cases). Therefore, we compare binning techniques for all potential true value combinations, and then calculate estimation error aggre-

gated average. Figure 3.11 shows neighborhood and response based binning MSE over the 32 different potential response combinations. Neighborhood based binning improves personal MSE in most cases, reducing error by 10% on average.

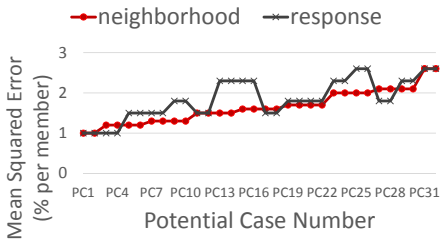


Figure 3.11 – Mean square error (MSE) for potential response cases (minimum bin size = 2).

To check how changing bin size affects average error rate, we repeated the study using minimum bin size = 3 and 4 members. Figure 3.12 shows that neighborhood based binning reduces error by 12%, and 24% respectively.

Thus, neighborhood based binning minimizes estimated STE error for each member. The main reason for the success of neighborhood based binning is its smoothing, which helps mitigate overfitting of response based binning, as discussed in the section 3.1, due to the STE data uncertainty problem.

3.4 Subpopulation treatment effect sliding trees

3.4.1 Proposed solution

By using the sliding windows for STEM, we solved one major problem facing STEM, which is handling a continuous variable without losing the homogeneity of the subpopulations. However, to implement STE-NBB in the STEM framework, we propose our second contribution, STE-STrees.

STE-STrees uses decision trees as a function for the sliding window $f(w_i)$. And we use the mean function as the reconstruction function $g(v_i)$. First, we have to sort the members based on the variable's value. Members that have features with similar values are joined together as one case, and the target is the average of their targeted values. The targeted value is then simply calculated using equation A.1. After that, we select the desired window size and step size. As a rule of thumb, the window size should not be less than 30 instances and should not exceed 50% of the population size. Step size, on the other hand, depends on the available computational power, the larger the step size, the less computation power required.

For each window, we build decision trees as a prediction model. The decision trees' depth depends on the number of features in the dataset. Typically, half of the features are selected in each tree. After that, the combined decision trees are

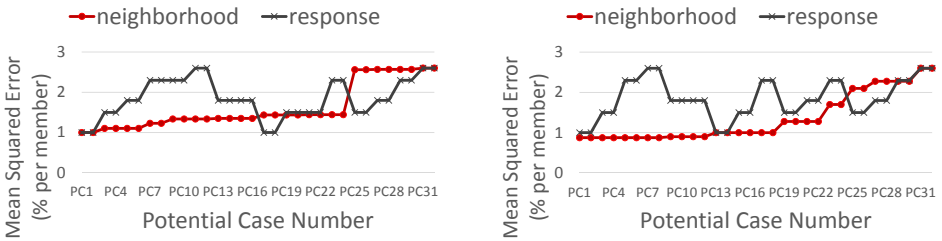


Figure 3.12 – Mean square error (MSE) for potential response with minimum bin size (left) 3; and (right) 4.

used in the prediction process. In our experiment, we used the arithmetic mean of the decision tree scores as a mapping function to calculate the estimated STE for an individual member.

The algorithm for the STE-STrees is shown in algorithm 4:

Algorithm 4: Building STE-STrees

Input: F, T, R
F represents the set of features
T represents the treatment/control binary variable
R represents the response variable
Output: STE sliding trees list.
SET case's treatment variable t_i;
SET case's response variable r_i;
SET target response variable r;
SET model's target variable Y $Y = \frac{\sum_{i=1}^{i=n}[r_i=r \wedge t_i=1]}{\sum_{i=1}^{i=n}[t_i=1]} - \frac{\sum_{i=1}^{i=n}[r_i=r \wedge t_i=0]}{\sum_{i=1}^{i=n}[t_i=0]}$;
GET window size w, step size s, features sample size q;
for *each feature* f *in features* F **do**
    RANK feature;
    SET pointer = 1;
    repeat
        GET elements IN $[pointer, pointer + w]$;
        BUILD decision tree for target variable Y USING features q;
        SAVE decision tree in STE sliding trees list;
        INCREMENT pointer by s ;
    until $(pointer + w) > f$ *size*;
end

We calculate the estimated member's STE as the average of the estimated STEs of all the windows that include the case. We call this method sliding STE estimation.

$$SlidingSTE_i = \frac{\sum_{k=1}^{n} E_{ki}(STE_i)}{n} \tag{3.3}$$

where n is the number of decision trees, and $E_{ki}(STE)$ is the estimation score of decision tree k for case i.

Our approach (STE-STrees) successfully solves the problems faced by the current STEM techniques. STE-STrees can process both continuous and categorical variables without sacrificing the noise nor the homogeneity of the subgroups.

3.4.2 Simulation study and experimental evaluation

With the emergence of big data and data mining in business-related issues, new machine learning algorithms need to be thoroughly tested in simulated contexts before being released for public or commercial use. The fields of data science and advanced statistics have been revolutionized by the development of algorithmic and "learning from data" methods [2]. The experimentation methodology is a crucial procedure to validate data-driven research [134, 135]. The methodology followed in this study included a series of experiments conducted on simulated data, with the goal of providing evidence of the reliability of our STEM approach, compared to other existing approaches.

A major obstacle when measuring STEM models involves what is called "the fundamental problem of causal inference" [41], which states that we can only observe one outcome for an individual after being subjected to a specific treatment, and we cannot observe other treatments' outcomes at the same time. In other words, we do not have an individual's true treatment effect. Hence, we do not have a subpopulation's true treatment effect. Simulated experimentation plays an important role in STEM development. This is because it is the only way to obtain the true STE value, which is required for model evaluation.

To evaluate the STEM approaches, we used a simulation framework developed by [11] and then extended by [17]. We improved the simulation to cover over sixteen different scenarios that vary based on three factors, the impact of the main treatment, the magnitude of the noise, and the variable correlation coefficient (check Table 3.1). For the first eight scenarios, the main effect was twice the impact of the STE effect, while in the last eight scenarios, the main effect was four times the STE effect. The magnitude of the noise varied from *medium* to *high* ($\sqrt{2}$ to $2\sqrt{2}$). The correlation among covariates ranged from 0 to 0.7.

For each scenario, we generated a training dataset and a validation dataset. The training datasets contained 200 rows, while the validation datasets contained 1000 rows. Each dataset contained one binary treatment variable, one binary response variable, and twenty continuous features. To minimize the effect of any dubious or biased results, we conducted one hundred simulations for each scenario.

We used two measures to evaluate the performance of the STEM approaches, Spearman's rank correlation coefficient, and the *Qini* coefficient. Spearman's rank correlation coefficient is a statistical procedure that is designed to measure the relationship between the ranks of two variables [136]. We measured Spearman's rank correlation coefficient between the estimated personal STE based on each approach and the real personal STE provided by the simulation. The *Qini* coefficient, which

Table 3.1 – Simulations scenarios

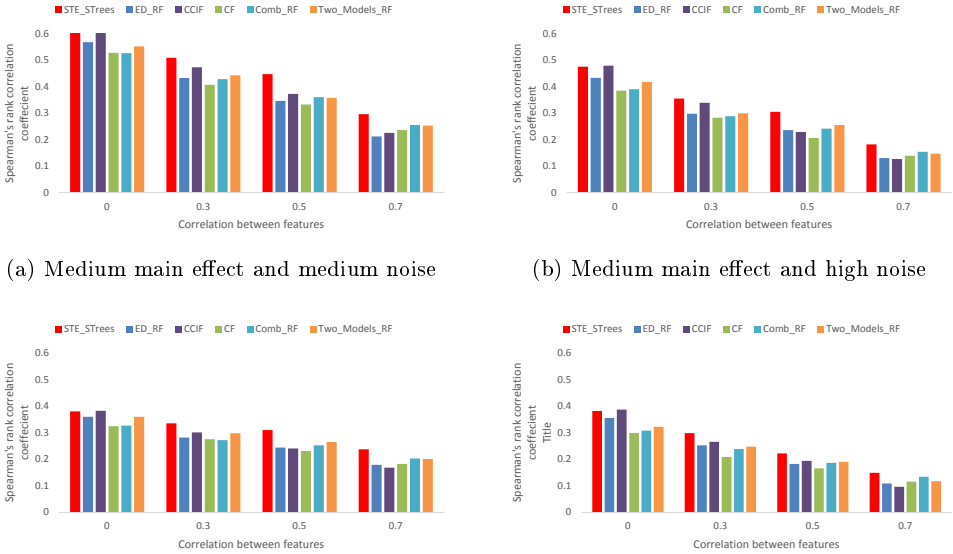
Scenario Number	Impact of the Main Effect	Noise Magnitude	Correlation Coefficient among Covariates
1	*Medium*	*Medium*	0
2	*Medium*	*Medium*	0.3
3	*Medium*	*Medium*	0.5
4	*Medium*	*Medium*	0.7
5	*Medium*	*High*	0
6	*Medium*	*High*	0.3
7	*Medium*	*High*	0.5
8	*Medium*	*High*	0.7
9	*High*	*Medium*	0
10	*High*	*Medium*	0.3
11	*High*	*Medium*	0.5
12	*High*	*Medium*	0.7
13	*High*	*High*	0
14	*High*	*High*	0.3
15	*High*	*High*	0.5
16	*High*	*High*	0.7

was described in the A.2.3 *section*, is a generalized form of the *Gini* coefficient that was developed for STEM evaluation, it is calculated as the area under the cumulative gains curve (*Qini curve*) and above the random line (representing random treatment assignment) [21]. Lastly, for each scenario, we calculated the average of the one hundred simulations' outcomes.

We compared our approach (STE-STree) with the top STEM approaches from the literature. Namely, we compared our algorithm with the Euclidean distance-based uplift trees method developed by [64, 14], the causal conditional inference forest method [17], the causal forest method [91], combined STEM, and two models STEM method. We called the algorithms ED_RF, CCIF, CF, Comb_RF, and Two_Models_RF, respectively.

We built the STEM models using the *random forest* technique [63] with 500 trees as a configuration for the random forest. The combined STEM approach was based on the work of [56, 115]. The *combined STEM* approach was formulated by combining the *treatment and response* subgroup with the *control and no-response* subgroup under one positive group, and then combining the *treatment and no-response* subgroup with the *control and response* subgroup under one negative group. Then, we simply created one binary classification model. On the other hand, the *two models STEM* approach was formulated by building two random forest models, one for the treatment group and the other for the control group.

Figure 3.13 shows the Spearman's rank correlation coefficients for the STEM techniques for the sixteen simulation scenarios. Every four scenarios have the same



(a) Medium main effect and medium noise

(b) Medium main effect and high noise

(c) High main effect and medium noise

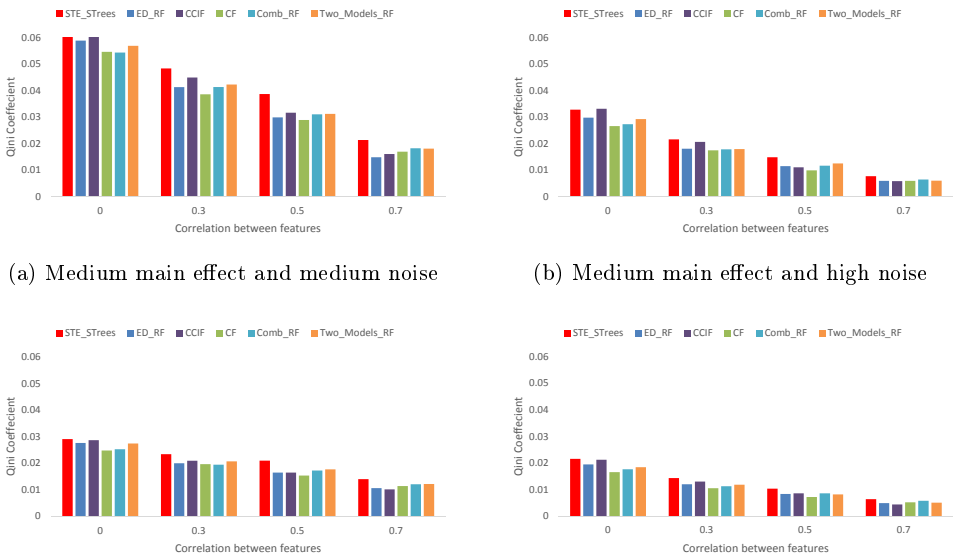
(d) High main effect and high noise

Figure 3.13 – Experiments results for simulated data scenarios. Plots demonstrate results of sixteen simulated data scenarios. Plots contain bar-chart plots of Spearman's rank correlation coefficient between estimated STE and true STE for STE_STrees, ED_RF, CCIF, CF, Comb_RF, and Two_Models_RF methods. Each plot shows the impact of the main effect and noise magnitude, while the correlation among covariates varies between 0 and 0.7.

impact from the main effect and noise magnitude, but they differ in the correlation coefficients between the features (see Table 3.1). Notice the impact of increasing the correlation coefficient on the predictions by the STEM approaches. We can see that among all the scenarios, the STE-STrees approach shows a better correlation with the true value, which verifies the importance of using neighborhood based binning to maximize the prediction accuracy.

Qini is another measure that is used to evaluate the predictability of STEM approaches. The motivation behind developing *Qini* measurements was to assess a STEM model's predictability without the need for a true STE value. *Qini* is a generalization of the *Gini* measure for an STE problem [70, 21]. The *Qini* coefficient is defined as the area between the actual incremental gains curve from the fitted model and the area under the diagonal corresponding to random targeting 3.6.

In Figure 3.14, we present the *Qini* coefficients of the STEM techniques for the sixteen simulation scenarios. Similar to Figure 3.13, every four scenarios share the same impact from the main effect and noise magnitude, but they differ in the correlation coefficients between the features. A high *Qini* coefficient means that the model successfully ranks the members in the right order, i.e., from a high STE to low STE. We can see that in all the scenarios, the STE-STrees approach shows better a *Qini* coefficient, which again proves the significance of using neighborhood based



(a) Medium main effect and medium noise

(b) Medium main effect and high noise

(c) High main effect and medium noise

(d) High main effect and high noise

Figure 3.14 – Experiments results for simulated data scenarios. Plots demonstrate results of sixteen simulated data scenarios. Plots contain bar-chart plots of Qini coefficient for STE_STrees, ED_RF, CCIF, CF, Comb_RF, and Two_Models_RF methods. Each plot shows the impact of the main effect and noise magnitude, while the correlation among covariates varies between 0 and 0.7.

binning for STEM.

In addition to the simulation experimentation, we applied different STEM models to an email marketing campaign provided by an online travel booking website. This email marketing campaign was conducted to determine the effect of a new type of email design on customers: personalized versus standard. The new design comprised a personalized promotional message based on historical data related to the previous activity of each client, whereas the standard design did not vary between clients. The company decided to conduct A\B testing to measure the influence of the new design. A\B testing, or controlled experimentation, is widely used by online businesses (e.g., Facebook, Amazon, and Groupon) during the product development process or to test the effect of a new product [137].

A standard promotional email (which recommends clicking on an online promotional advertisement available on the website) was sent to 52.45% of their clients, who were thus subjected to treatment A (receiving the standard email). The remaining subpopulation (47.55%) received the personalized promotional email (treatment B). The effects of these treatments on the population's behavior (in terms of opening the email, clicking on the advertisement, and purchasing a product) were observed and registered during a monitoring period of thirty days.

We utilized a database of 6,479,601 rows (clients) who were subjected to either

treatment A or treatment B. We used the purchase event variable as a response variable. This means if a client made a purchase after the advertisement, it was coded as a positive response. Otherwise, the client had a negative response. We used the difference (in seconds) between the email's opening timestamp and the clicking on the advertisement timestamp as one of the features. We used five other continuous variables for the modeling procedure. We used STEM approaches on the dataset to detect the subpopulations that were more influenced by one treatment than the other.

We compared our approach to other STEM techniques that can process continuous features. Namely, we compared our approach (STE_STrees) to the uplift tree (uplift_tree), causal conditional inference forest (CCIF), causal forest (CF), and combined models random forest (Comb_RF) methods. We evaluated the performance of these approaches using the *Qini* coefficient measurement. Figure 3.15 presents the *Qini* curves of the STE modeling approaches. This figure shows how our approach outperforms other STEM techniques. In numbers, the Qini coefficients (QC) of the STE methods in descending order are STE_STrees QC = 0.064, Comb_RF QC = 0.052, CCIF QC = 0.048, Uplift_tree QC = 0.031, and CF QC = 0.02.

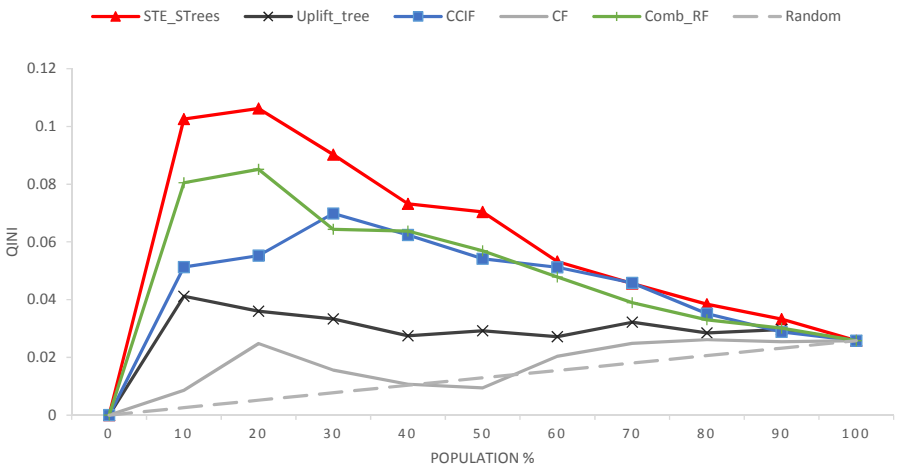


Figure 3.15 – Qini curves results

The good performance of our approach compared to other STE approaches reflects the importance of using sliding trees as a tool to deal with continuous data for STE modeling. In particular, using the STE-NBB technique made it possible to reserve the low variance inside subpopulations, which helped in keeping the information value of each member of the subpopulation.

The good Qini results indicated that our approach successfully differentiated between consumers who favored the personalized email and those who favored the standard email. We calculated the importance of each variable for our STE model. Figure 3.16 shows a pie chart that represents the percentage of importance of each

variable. The importance of a variable means that this variable was informative, and it helped to separate the subgroups. These results show that the age variable and time between opening the email and clicking on the advertisement variable were mostly used to detect the STE between subgroups. Finally, the results for the variable importance require further investigation to be used later in future STEM approaches.

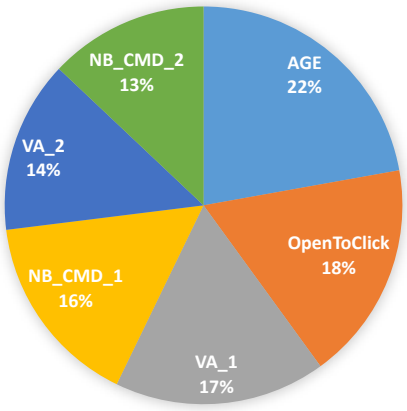


Figure 3.16 – Variable importance

The results of the simulated and real experiments indicated how our STEM approach's performance exceeded those of other STEM approaches. STE-STrees can be used for a marketing campaign as a targeting classifier. It could be used to classify future consumers that are most likely to respond to an email advertisement, and avoid those who are most likely to not react to the email advertisement. Using our approach will make it possible to maximize the response rate for the marketing advertisement and minimize the possibility of any waste.

3.5 Subpopulation treatment effect neighborhood random forests

3.5.1 Proposed apporach

The fundamental motivation behind STE-NRF is to take advantage of neighborhood based binning to increase STE model accuracy. The proposed STE-NRF technique combines neighborhood based binning with classical regression trees [138] in a random forests framework [63].

First, members are sorted based on their feature values. Members with similar feature values are joined as a single member, and the target variable is the average of their targeted variable values. The target variable for a binary treatment can be expressed as

$$\widehat{STE_i} = \frac{\text{sum(Response given treatment)}}{\text{sum(Treatment)}} - \frac{\text{sum(Response given control)}}{\text{sum(Control)}}. \quad (3.4)$$

Window size is then set based on variable variance, low variance requires short window size and vice versa. Step size is set based on population size and computational capability.

To create STE neighborhood random forests (STE-NRF), we build small regression decision trees within each window as a prediction model for $\widehat{STE}$. We randomly select 50% of the features as predictors and 50% of the population to build each decision tree. These configurations minimize overfitting and improve the final model generalization. Finally, collective decision trees (random forests) are used for prediction, where the predicted $\widehat{STE}$ of each member is the average prediction score from all trees that include that member, as discussed above for neighborhood binning. Algorithm 5 shows the STE-NRF algorithm.

Algorithm 5: Building subpopulation treatment effect neighborhood random forest (STE-NRF)

Input: Feature table, F; binary treatment/control variable, T; binary response/no-response variable, R; window size, w; step size, s; feature sample size, q; population sample size, p; and iterations, I.
Output: STE-NRF list.
set member treatment variable, t;
set member response variable, r;
set model target variable,

$$Y = \frac{\sum_{i=1}^{i=n}[r_i = r \wedge t_i = 1]}{\sum_{i=1}^{i=n}[t_i = 1]} - \frac{\sum_{i=1}^{i=n}[r_i = r \wedge t_i = 0]}{\sum_{i=1}^{i=n}[t_i = 0]},$$

repeat
  for *each feature* $f \in F$ **do**
    get population sample (population $\times$ p);
    rank f;
    set pointer = 1;
    repeat
      extract elements in $[pointer, pointer + w]$ from population sample;
      extract feature sample (features list $\times$ q);
      build regression tree for target value Y using feature sample;
      save regression tree in STE-NRF list;
      pointer = pointer + s
    until $(pointer + w) > population\ sample\ size$;
  end
  i ++;
until $i > I$;

We estimate member $\widehat{STE}$ as $\widehat{STE_{tree}}$ average over all trees,

$$STE - NRF_i = \frac{\sum_{k=1}^{n} E_{k_i}(STE)}{n}, \tag{3.5}$$

Table 3.2 – Simulation scenario settings

Scenario Number	Strength of the Main Effect	Correlation among Covariates	Noise Magnitude
1	$(-1)^{j+1}\, I(3 \le j \le 10)/2$	0	$\sqrt{2}$
2	$(-1)^{j+1}\, I(3 \le j \le 10)/2$	0	$2\sqrt{2}$
3	$(-1)^{j+1}\, I(3 \le j \le 10)/2$	0.5	$\sqrt{2}$
4	$(-1)^{j+1}\, I(3 \le j \le 10)/2$	0.5	$2\sqrt{2}$
5	$(-1)^{j+1}\, I(3 \le j \le 10)$	0	$\sqrt{2}$
6	$(-1)^{j+1}\, I(3 \le j \le 10)$	0	$2\sqrt{2}$
7	$(-1)^{j+1}\, I(3 \le j \le 10)$	0.5	$\sqrt{2}$
8	$(-1)^{j+1}\, I(3 \le j \le 10)$	0.5	$2\sqrt{2}$

where n is the number of trees, and $E_{k_i}(STE)$ is the tree k estimation for member i.

We evaluates the proposed STE-NRF performance using simulated and real datasets. Simulated experiments enable comparing proposed STE-NRF performance with PTE for each member compared to current approaches.

3.5.2 Simulation study and experimental evaluation

Simulations settings

We used the experimentation protocol developed by [11] in the simulations to compare the proposed STE-NRF and current STE method performances. For reproducibility, we adopted simulation settings proposed by [17], varying the main effect strength, correlation among covariates (features), and noise magnitude in the dataset as shown in Table 3.2.

Scenarios 1–4 have main effect magnitude twice that of the STE effect, whereas scenarios 5–8 have main effect magnitude four times that of the STE effect. Covariate correlation varies 0–0.5, and noise magnitude = $\sqrt{2}$–$2\sqrt{2}$.

The training dataset contained 200 rows with 1000 rows in the validation dataset. Each dataset contained binary treatment and binary response columns. Simulation datasets also contained twenty features, $X1, X2, \cdots, X20$, but only $X1 - X4$ are affected by STE, with $X3$–$X10$ affected by the main treatment effect. Variables $X10$–$X20$ are not affected by the main or STE effects; their purpose is to add noise to the dataset and harden the modeling process by correlating with other variables. More details are provided in [17].

We built STE-NRF using 50% population sample size, 50% feature sample size, and 50 iterations, with step size = 10 and window size = 30 members. We compared the proposed STE-NRF with the following STE modeling techniques.

— Two random forests models (Two-Models-RF) [1]. We created two random forests models using 500 trees, where each model predicted the response for a specific

1. Random forests models were built using the randomForest R package [63]

treatment. We then estimated STE by calculating score differences between each prediction model.

— Combination random forests (Comb-RF)[1]. This technique simplifies the STE multi-classification problem by combining different labels into two main classes, convert it to a binary classification problem. More information regarding this approach is available at [56, 9]. For these experiments, we used random forests composed of 500 trees.

— Causal conditional inference forests (CCIF)[2]. This is an improved uplift random forests method uses better pruning and stopping criteria to assure statistical significance for each tree in the model. For these experiments, we used a CCIF model composed of 500 trees, and interaction based split criteria.

— Euclidean distance random forests (ED-RF)[2] [115, 62]. This approach modified uplift random forests by changing the splitting criterion during tree construction to calculate squared Euclidean distance between the split and original nodes. For these experiments we used ED-RF models composed of 500 trees.

— Causal forests (CF)[3]. This approach used causal trees developed by [22] in a regression forests framework. For the experiments we used CF model composed of 500 trees with 50% sample fraction.

Simulations results

To guarantee reliability, we repeated each experiment 100 times for each simulation scenario. STE model performance was assessed using two the Qini [21] and Spearman's rank correlation coefficients between the estimated and actual treatment effects from each model. The uplift R package [14] was used to produce the simulated datasets and calculate the Qini coefficient.

Figure 3.17 shows average Qini coefficient from the 100 repeated simulations for each scenario, where higher Qini coefficient reflects better STE detection among the members. The proposed STE-NRF method outperformed all other STE methods in all scenarios.

Figure 3.18 shows average Spearman's rank correlation coefficient between estimated subpopulation treatment effects, $\widehat{STE}$, and real personal treatment effect, PTE over the 100 repeated simulations for each scenario. The proposed STE-NRF method outperformed all other STE methods in all scenarios.

Comparing scenario pairs (1,2), (3,4), (5,6), and (7,8) in Figs. 3.17 and 3.18, shows how noise affects STE scores. The proposed STE-NRF method has superior Qini coefficient regardless of noise magnitude than all other STE models without noise (e.g. Figs. 3.17e and 3.17f; and 3.17g). For the most challenging environment (scenario 8), STE-NRF exhibits 40% average correlation with the true STE, 10% more accurate than the next best model (see Fig. 3.18h).

Correlation between features affects STE model scores (compare scenario pairs (1,3), (2,4), (5,7), and (6,8) in Figs. 3.17 and 3.18), with the proposed STE-NRF

2. CCIF and ED-RF models were built using the uplift R package [17]
3. Causal forests models were constructed using the grf R package [139].

model again outperforming all other STE models for both Qini score and rank correlation. A significant challenge for STE modeling is to differentiate between the main and STE effect. The proposed approach shows better distinguishability between these effects than all other techniques even for the noisiest scenario (scenario 8) (compare scenario pairs (1,5), (2,6), (3,7), and (4,8) in Figs. 3.17 and 3.18).

Experiments using real datasets settings

We conducted four experiments using three real datasets. Each dataset contained binary treatment and binary response variables and were chosen from healthcare and marketing sectors.

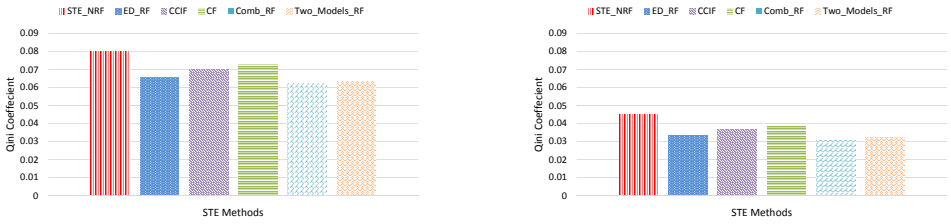
— **Hillstrom visit dataset**: The Hillstrom visit dataset [140] contained data from an email marketing campaign, comprising 64,000 unique customers split into two groups: the treatment group received an email, whereas the control group did not. The dataset considered two email types, based on the advertised product. We were only interested in determining the overall effect of receiving an email, so combined we both email types into one group and compared with the control group. The dataset also included 9 features describing client historical and demographical information.

— **Right heart catheterization dataset**: The right heart catheterization (RHC) dataset comprised information regarding 5735 patients admitted to a hospital [141], and included two groups. Group 1 (2184 patients) were subjected to right heart catheterization, based on their condition, and group 2 (3551 patients) did not receive right heart catheterization. We used `Right Heart Catheterization` to represent the treatment variable, and `Death` the class variable. The monitoring period was 180 days, and patient survival for 180 days after admission was considered a positive outcome. All features related to patient death were removed, and the remaining 58 features were used to building the STE models. The dataset was separated into training (80%) and testing (20%) sets.

— **Bone marrow transplant dataset**: The bone marrow transplant (BMT) dataset comprised information regarding one hundred patients that received a bone marrow transplant [142] extracted from either pelvic bone or peripheral blood. The dataset was binned into treatment and control samples based on the bone marrow source, where pelvic bone was considered the control. The dataset contained seven features that were employed to construct the STE models. We used the BMT dataset once to predict the occurrence of acute graft versus host disease (agvh) and again to predict the occurrence of chronic graft versus host disease (cgvh).

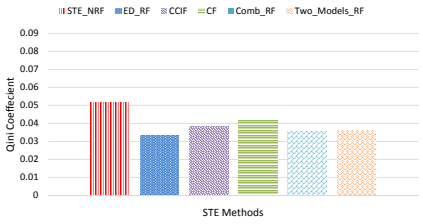
Results of real datasets experimentation

Figures 3.19, 3.20, 3.21, and 3.22 show the Qini coefficients for the Hillstrom visit, RHC, BMT cgvh, and BMT agvh, respectively. The proposed STE-NRF algorithm

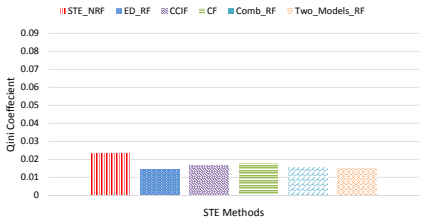


(a) Scenario 1

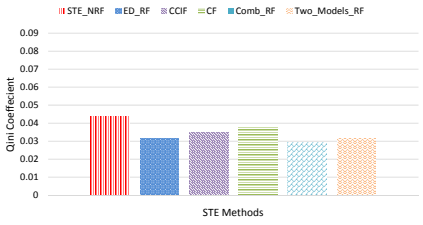
(b) Scenario 2



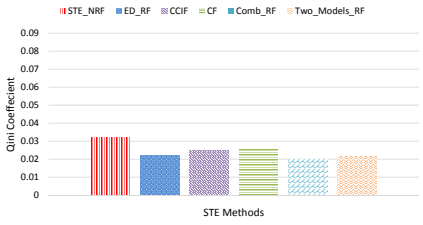
(c) Scenario 3



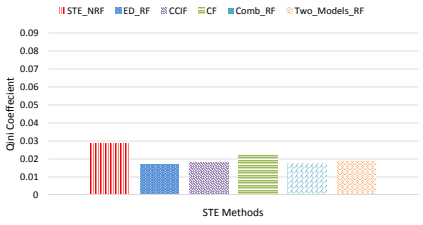
(d) Scenario 4



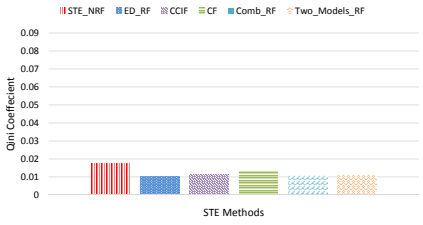
(e) Scenario 5



(f) Scenario 6

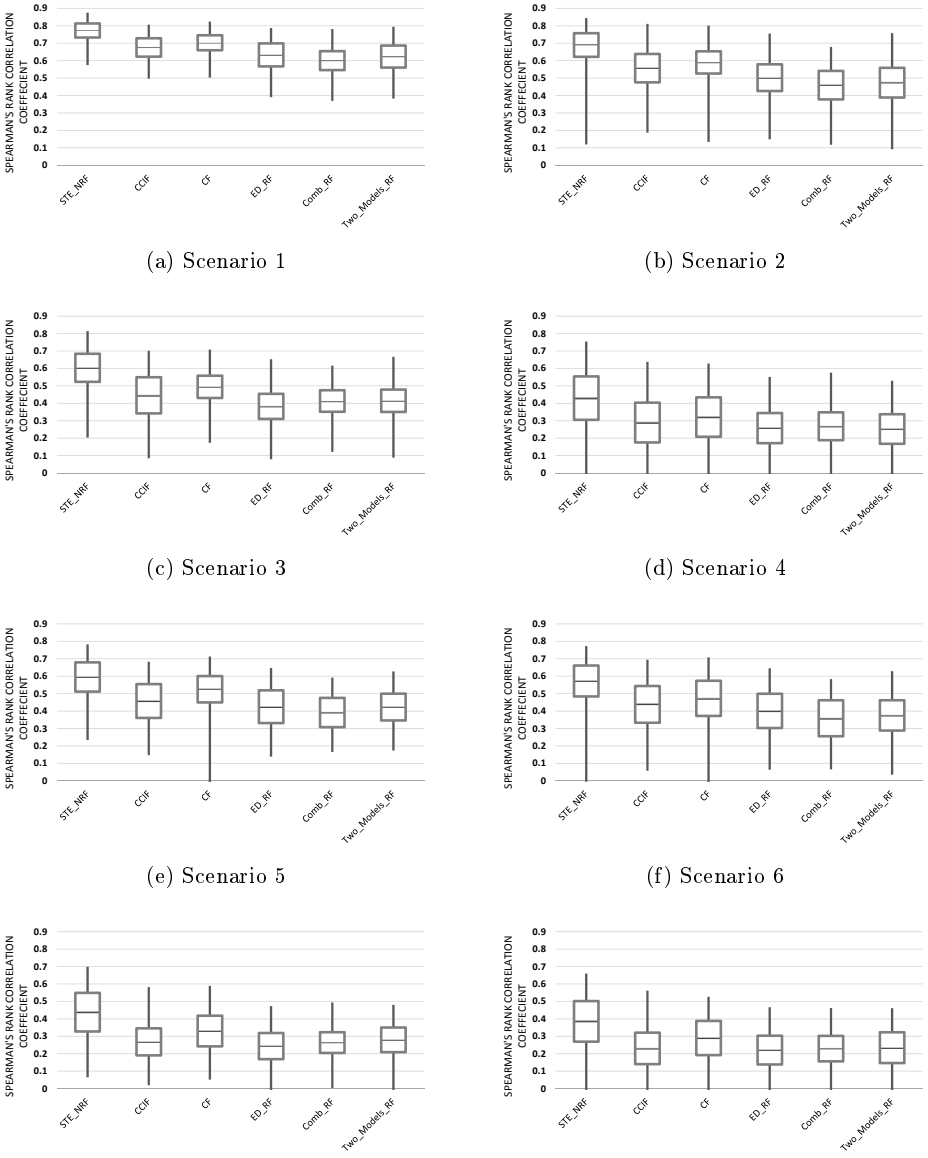


(g) Scenario 7



(h) Scenario 8

Figure 3.17 – Average Qini coefficients over 100 repeated simulations for the various scenarios; STE-NRF = proposed subpopulation treatment effect neighborhood random forests, ED-RF = Euclidean distance based uplift random forests, CCIF = causal conditional inference forests, CF = causal forests, Comb-RF = combined uplift random forests, and Two-Models-RF = two models random forests.



(a) Scenario 1 (b) Scenario 2

(c) Scenario 3 (d) Scenario 4

(e) Scenario 5 (f) Scenario 6

(g) Sceanrio 7 (h) Scenario 8

Figure 3.18 – Average Spearman's rank correlation coefficients over 100 repeated simulations for the various scenarios; STE-NRF = proposed subpopulation treatment effect neighborhood random forests, ED-RF = Euclidean distance based uplift random forests, CCIF = causal conditional inference forests, CF = causal forests, Comb-RF = combined uplift random forests, and Two-Models-RF = two models random forests.

successfully predicting STE for the real datasets, particularly for the first 5% of the population. Table A.1 shows that the proposed STE-NRF model outperformed all other STE models in terms of the Qini coefficient.

We also evaluated the STE methods based on the 15% Qini coefficient [21], to show the effect of minimizing the population size. For example, Fig. 3.19 shows that the advertisement campaign costs could be minimized by targeting only the highest 15%. We can achieve 10.52% more STE gain using the proposed STE-NRF approach.

The proposed STE-NRF model exhibits better capability dealing with continuous features, reflected by the STE evaluation scores, and the sliding window technique provides better matching to minimize binning noise and maximize average predictability.

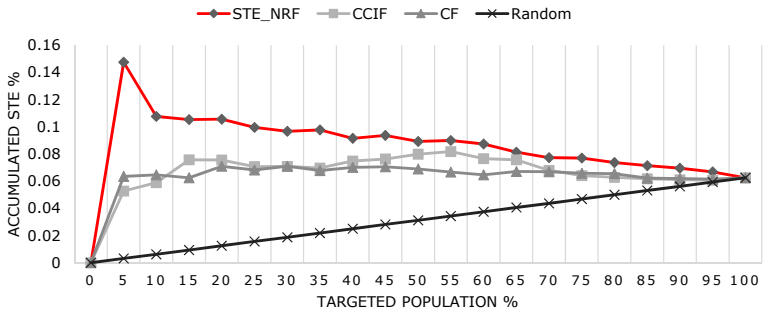


Figure 3.19 – Qini coefficients for experiments on the Hillstrom visit dataset.

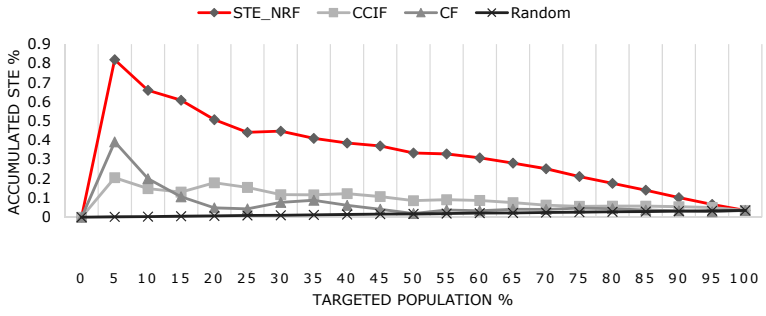


Figure 3.20 – Qini coefficients for experiments on the RHC dataset.

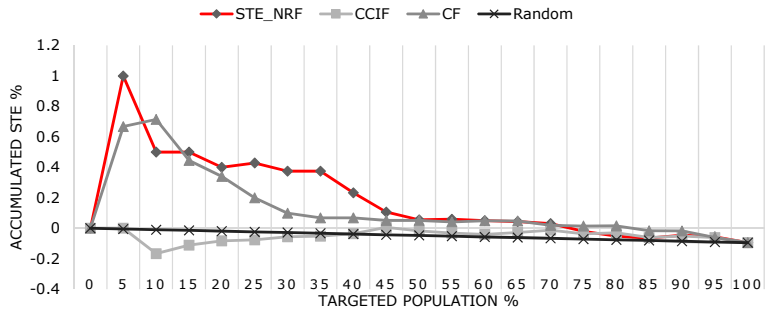


Figure 3.21 – Qini coefficients for experiments on the BMT-cgvh dataset.

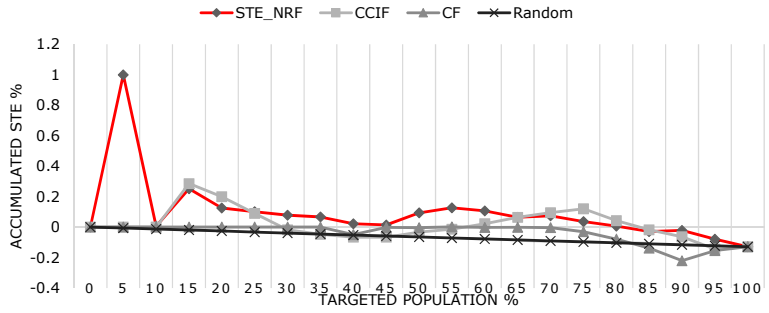


Figure 3.22 – Qini coefficients for experiments on the BMT-agvht dataset.

Table 3.3 – Experimentation outcomes for real datasets.

Dataset Name	Qini Coefficient (%)			15% Qini Coefficient (%)		
	STE-NRF	CCIF	CF	STE-NRF	CCIF	CF
Hillstrom-Visit	**10.20**	6.50	6.00	**10.52**	7.56	6.24
RHC	**51.20**	12.80	8.60	**60.89**	13.10	10.64
BMT-cgvh	**40.90**	0.40	31.40	**50.00**	-11.10	44.44
BMT-agvh	**25.30**	13.00	4.20	25.18	**28.57**	0.00

3.6 Conclusion

This chapter discussed the effect of uncertainty on the performance of subpopulation treatment effect modeling (STEM). We focus on Binning as a technique to minimize the uncertainly effect. We described response based binning, and its limitations for STEM. Theresponse based binning technique works perfectly for the categorical type of features. However, for continuous variables, it may lead to false results and more sensitivity to noise in the data. For this reason, we proposed a new technique called neighborhood based binning (STE-NBB) that solves this problem by taking into account the similarity between members. The proposed neighborhood based binning approach was fully described and performance compared to response based binning using simulated data. The technique reduced estimation error rate by 24% compared to current binning techniques.

After that, we proposed a second contribution, which was (STE-STrees). STE-STrees is an improvement on STE-NBB that uses decision trees as a sliding function. We explained the steps for building an STE-STrees model and the benefits of using it.

We compared our approach to other STEM approaches using synthetic and real-world datasets. For the experimentation, we compared the STEM techniques in sixteen different synthetic simulation scenarios. The simulation scenarios differed in relation to the noise magnitude, feature correlation, and impact of the main effect. The experimental results showed that our approach outperformed the other STEM approaches, as measured by the *Qini* coefficient and Spearman's rank correlation coefficient. We also evaluated our approach using a real-world dataset; we used a database gathered by an online travel booking website. We showed that our approach outperformed other STEM techniques using the *Qini* coefficient measurement.

Finally, we showed the importance of our contribution to real-life marketing applications. We explained how our approach could be beneficial for digital marketing and how it could be applied to maximize the value of marketing actions. In particular, it could be used as a segmentation technique, which is the primary motivation behind STEM. STE-STrees guarantees more homogeneous segments, which will later have a customized solution that suits them. By utilizing STE-STrees, we can enhance the quality of the decisions and reveal valuable insights that will minimize the risks and optimize the customer retention strategies.

We then proposed STE neighborhood random forests (STE-NRF) to take advantage of neighborhood based binning in an STE random forest framework, and provided the pseudo-code algorithm for this approach. Simulation experiments were conducted covering eight scenarios with varying noise magnitude, comparing feature correlation, and main effect strength to conventional STE approaches, including Two-Models-RF, Comb-RF, CCIF, ED-RF, and CF. We used the Qini coefficient to investigate how well the STE models detect treatment effect heterogeneity and Spearman's rank correlation coefficient to investigate rank correlations real personal treatment effects and predicted STE scores for each model. Simulation results verified that the proposed STE-NRF approach provided a model robust to noise that outperformed all other considered STE approaches for all tested scenarios.

We also compared the proposed STE-NRF approach with current approaches using real datasets from marketing and healthcare sectors. Experiments results validated that the proposed STE-NRF approach provided superior Qini coefficients and 15% Qini coefficient compared to all other STE approaches considered for all scenarios. The results also confirmed the sliding window technique importance for the STE modeling framework to provide better predictability and minimize noise effects.

The success of the proposed approach compared with current STE models for Qini and Spearman's rank correlation coefficients is due to neighborhood based binning minimizing model overfitting, due to smoothing from the sliding window, and the sliding window providing better matching, particularly for continuous variables. The random forests framework used in STE-NRF also helped mitigate overfitting. Neighborhood based binning also minimizes noise effects by using the definite member neighborhood, ensuring that any erroneous or biased values are devaluated due to the local average. Thus, the proposed STE-NRF approach significantly outperformed current STE models.

Chapter 4

Disturbance and Subpopulation Treatment Effect Modeling

> In science, read, by preference, the newest works; in literature, the oldest.
>
> Edward G. Bulwer-Lytton,
> *Caxtoniana: A Series of Essays on Life, Literature, and Manners*

Disturbance in data can be defined as any factor that leads to less accurate estimation results. STE data disturbance could be due to biased data, noise, high correlation between the features, or because of low subpopulation treatment effect (STE) compared to the average treatment effect (ATE). The studies of how disturbance affect STE models showed that STE/ATE ratio has the major effect on results, followed by the correlation between the features, then by the noise and bias in data [36, 11, 21].

STE disturbances are categorized as treatment disturbances or response disturbances. Treatment disturbances are the factors that affect the variation of the treatment and control cases, leading to a disparity in treatment assignment (e.g., bias). Response disturbances, on the other hand, are all factors, that change the variation of the respondents and non-respondents (e.g., noise). For example, in a binary response and treatment experiment, we have the four quadrants (see Figure A.1). The perfect state for treatment assignment in the experiment is to have the treatment cases equal to the control cases (TR+TNR) = (CR+CNR). However, treatment disturbances would disturb this perfect state and affects the results negatively.

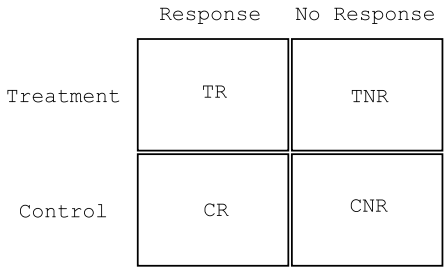


Figure 4.1 – Four quadrants of subpopulation treatment effect experiment.

In any STEM, we are interested in distinguishing the subpopulation treatment effect (STE) from the main treatment effect (ATE). The classical STEM approach models the variation of response rate between treatment and control group.

$$STE = \frac{\text{number of cases in TR}}{\text{number of cases in T}} - \frac{\text{number of cases in CR}}{\text{number of cases in C}}$$

The problem with the classical STEM approach is that it models the response disturbance while trying to maximize TR and minimize CR, Figure 4.2 shows the optimization models of a classical STEM.

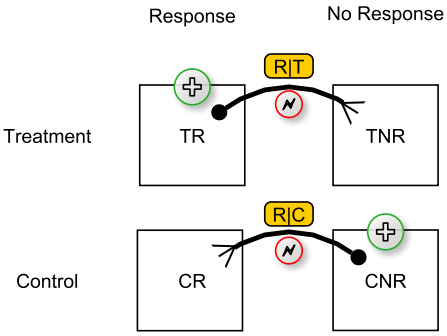


Figure 4.2 – Classical subpopulation treatment effect modeling approach

In the figure, a forked shape headed arrow represents the minimization process, black circle headed arrow represents maximization process. Response disturbance effect represented in the figure by a flash, it is modeled with each optimization models.

4.1 Proposed approach the balanced reflective uplift modeling

We first build what we call a reflective uplift model in order to reduce disturbance sensitivity. It is aimed to handle the variation of the treatment effect rather than the response effect. Reflective uplift model estimates the probability for a specific case to be in the treatment area, given the fact that it has already responded. By doing so, we minimize the effect of response disturbances. In the Figure 4.3, a forked shape headed arrow represents the minimization process.

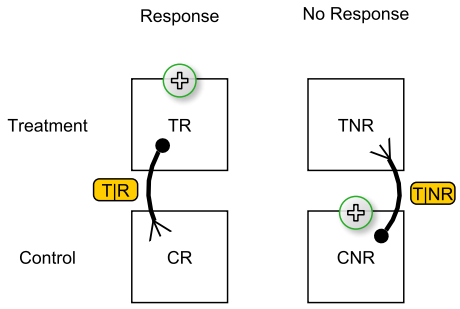


Figure 4.3 – Reflective uplift model

$$
\begin{aligned}
STE &= \frac{TR}{TR+TNR} - \frac{CR}{CR+CNR} \\
&= \frac{(TR*CR)+(TR*CNR)-(CR*TR)-(CR*TNR)}{(TR*CR)+(TR*CNR)+(TNR*CR)+(TNR*CNR)} \\
&= \frac{(TR*CNR)-(CR*TNR)}{(TR*CR)+(TR*CNR)+(TNR*CR)+(TNR*CNR)}
\end{aligned} \tag{4.1}
$$

$$
\begin{aligned}
RU &= \frac{TR}{TR+CR} - \frac{TNR}{TNR+CNR} \\
&= \frac{(TR*TNR)+(TR*CNR)-(TNR*TR)-(TNR*CR)}{(TR*TNR)+(TR*CNR)+(CR*TNR)+(CR*CNR)} \\
&= \frac{(TR*CNR)-(TNR*CR)}{(TR*TNR)+(TR*CNR)+(CR*TNR)+(CR*CNR)}
\end{aligned} \tag{4.2}
$$

By comparing 4.1 and A.4.1, we can see that the STE and RU have the same numerator, but they differ in the denominator. The difference in the denominator $\Delta Deno$ is the following:

$$\begin{aligned}\Delta Deno &= ((TR*CR)+(TR*CNR)+(TNR*CR)+(TNR*CNR)) \\ &\quad -((TR*TNR)+(TR*CNR)+(CR*TNR)+(CR*CNR)) \\ &= (TR*CR)+(TNR*CNR)-(TR*TNR)-(CR*CNR) \\ &= (TR)*(CR-TNR)+(CNR)*(TNR-CR) \\ &= (TR-CNR)*(CR-TNR) \end{aligned} \tag{4.3}$$

Given that we have a perfect experiment environment ($T = C$), we can distinguish based on 4.3 the following three cases:

- $RU > STE$:
  - If $TR > TNR$
  - If $CNR > CR$
- $RU < STE$:
  - If $TR < TNR$
  - If $CNR < CR$
- $RU = STE$:
  - If $TR = CNR$

We can see in the Figure (Fig.4.4), how the RU and STE has the same sign. RU is more responsive to small changes around zero STE, which means it is beneficial for low STE/ATE ratio environment. This makes RU score a good candidate to support STE score.

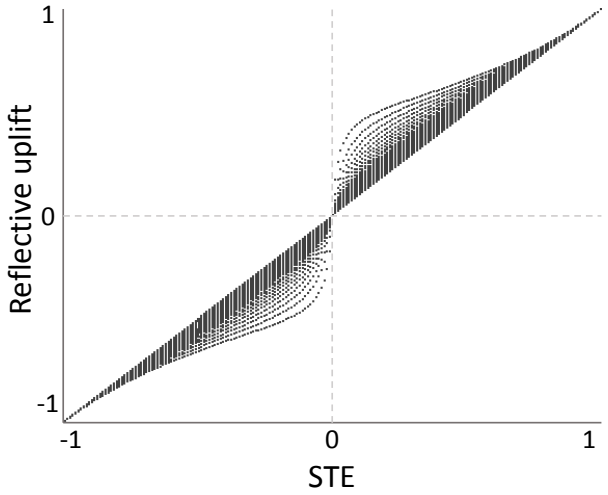


Figure 4.4 – Reflective uplift as a function of subpopulation treatment effect (STE) in a perfect experiment environment

The negative side effect of using reflective uplift is the increase of the sensitivity for treatment disturbances, but it could be held under an acceptable threshold using appropriate sampling, such as stratified random sampling [14].

To build a reflective uplift model, we propose the following procedure:

Algorithm 6: Reflective uplift algorithm

Partition the Responders into two classes:
Treatment Class: TR => $Class = 1$;
Control Class: CR => $Class = 0$;
Build binary classification model with $Class$ as a target variable
$M_i^{TR} = P(Class_{treatment}|R; x_i)$;
Partition the Non Responders into two classes:;
Treatment Class: TNR => $Class = 1$;
Control Class: CNR => $Class = 0$;
Build binary classification model with $Class$ as a target variable
$M_i^{TNR} = P(Class_{treatment}|NR; x_i)$;

$$Reflective\ Uplift_i = M_i^{TR} - M_i^{TNR} \tag{4.4}$$

where x_i is a vector of features for individual i;

We use the ensemble decision trees method as defined by [62] to build each model in order to minimize overfitting and misclassification error rates. For the STE model, we used the simple transform then model approach proposed by [56] (see A.2.2).

By utilizing the RU alongside the regular STE, we reached a balanced score (BRUM) that would be more reliable and robust.

$$BRUM = 1/2 * (STE_i + Reflective\ Uplift_i) \tag{4.5}$$

4.2 Simulation study and experimental evaluation

The experimentation methodology is a crucial procedure to validate data-driven research [134, 135]. With the emergence of business big data related issues, Internet of things and machine learning tools, new digital devices and protocols need to be tested in simulated and real-life contexts before being released for public or commercial use. The fields of data science and advanced statistics have been revolutionized by the development of algorithmic and 'learning from data' [2]. The methodology followed in this paper belongs to this trend. It includes a series of experimentations conducted on simulated data then on real dataset, and is aimed at providing evidence of the reliability of our BRUM approach, compared to other existing approaches.

4.2.1 Simulated Data Set

We use the experiment protocol developed by Tian et al. [11] and compare the performance of our method with existing methods; namely Two-models (Method A), Causal Conditional Inference Forests (CCIF) [17] and Lai's method [56]. We consider eight different scenarios (see Table 4.1).

Scenario	Response Strength	Correlation Strength	Noise Magnitude
1	Low	Low	Low
2	Low	Low	High
3	Low	High	Low
4	Low	High	High
5	High	Low	Low
6	High	Low	High
7	High	High	Low
8	High	High	High

Table 4.1 – Simulation scenarios

The training dataset contains 200 rows while 1000 rows for the validation dataset. Each dataset contains a Treatment column and a Response column. Both of them are binary variables. Datasets further contain twenty features named X_1 to X_{20}, only X_1, X_2, X_3 and X_4 are concerned by treatment heterogeneity effect. We build our models using Ensemble Regression Trees (one hundred ensemble trees are built). Each tree of the ensemble contains randomly selected 80% of the training data. We repeat each experiment one hundred times for each scenario. The performance of the analytical models is assessed using the Spearman's rank correlation coefficient between the estimated STE from each model and the 'real' STE. Uplift R package, developed by [36] is used to produce the simulated data and to apply CCIF method. BRUM is compared with the Two Models method ($Uplift = P(TR/T) - P(CR/C)$) [4], with the generalized Lai's method (as proposed by [6] and finally with CCIF [17] (the default configuration). Ensemble modeling methods are used to build all models.

For the first four scenarios, the main effect is twice the impact of the STE, while in the last four scenarios, the main effect is four times bigger than the STE. The correlation among features varies between 0 and 0.5. The magnitude of noise ranges from $\sqrt{2}$ to $2\sqrt{2}$. As shown in Figure 4.5, our results show that our model scores slightly the same as CCIF in the first, second, fifth and sixth scenarios, and outperforms all the other methods in the seventh and eighth scenarios (i.e. those presenting the highest noise). The Two-models approach (method A) seems the most sensitive to noise. The validity of our approach is therefore confirmed in the simulated dataset.

4.2.2 Using balanced reflective uplift modeling to improve conversion rate of an email marketing campaign (SNCF)

In this section, we move on to real datasets concerning also email marketing campaigns. We apply the BRUM model to an email marketing campaign implemented by an online travel booking website owned by a large railway French company. Besides

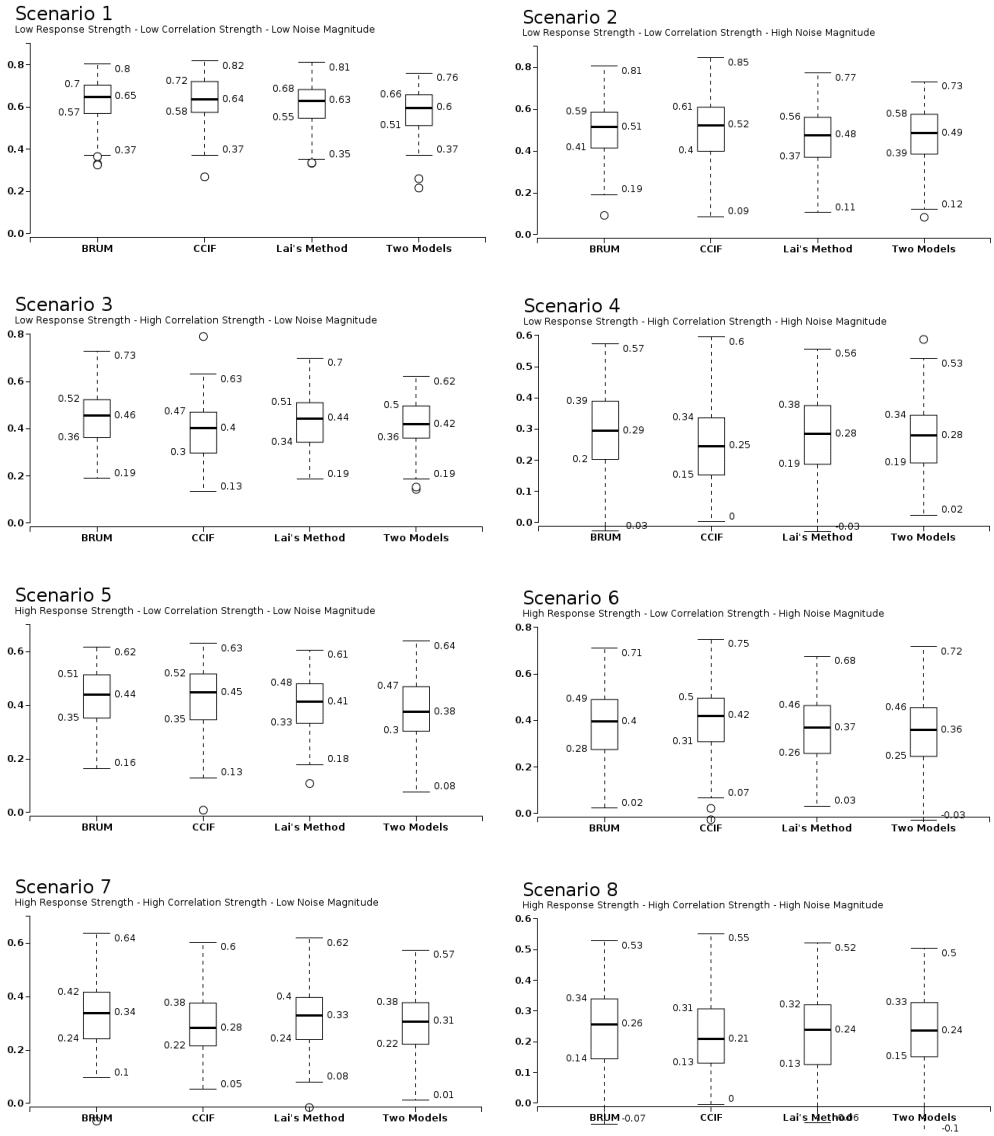


Figure 4.5 – Experiments results of the simulated data scenarios. Plots demonstrate results of eight simulated data scenarios. Plots contain box plots of Spearman's rank correlation coefficient between estimated subpopulation treatment effect and true subpopulation treatment effect for BRUM, CCIF, Lai's and Two Models methods.

selling train tickets online, this website offers the possibility to make hotel reservations, rent cars and other activities related to organizing travels and journeys, all over Europe and in other parts of the world.

An email marketing campaign has been conducted to determine the effect of a new type of email design on customers: personalized versus standard. The new de-

sign comprises a personalized promotional message based on historical data related to the previous activity of each client, whereas standard design does not vary between clients. The company decided to do an A/B testing to measure the influence of the new design. A/B testing, or controlled experimentation, is widely used by online businesses (e.g. Facebook, Amazon, Groupon ...) during product development process or to test the effect of a new product [143]. The company decided to define treatment A as sending the standard promotional email, and Treatment B as sending the personalized promotional email.

Within the prospect population (6,479,601), a standard promotional email (which recommends clicking on an online promotional advertisement available on the website) has been sent to 3,398,669 clients (representing 52.45% of the population) who have so been subject to treatment A (receiving the standard email). The remaining subpopulation (47.55%) received the personalized promotional email (treatment B). The effects of these treatments on the population behavior (in terms of opening the email and clicking on the advertisement) have been observed and registered during a monitoring period of the website activity lasting from 07/09/2015 00:00:00 to 23/09/2015 23:59:59. During this period, 458,068 orders (purchases) have been registered from clients that have been subject to either treatment A or treatment B.

Data set Description and Preparation

We have a database of 6,479,601 rows (clients) who have been subject to either treatment A or treatment B. The Information provided about each client is defined by 18 different variables represented in Table 4.2.

Among those variables, some suffer from missing data, like `Client_Age` and `ZipCode`, which have 52% and 82% missing data respectively. 18.14% of the population (1,221,607 cases) opened the e-mail during the monitoring period, 46.68% of them were in Group A, whereas 53.31% were in Group B.

STEM requires an explicit response variable in addition to the treatment variable. Therefore, it is important, along with the standard data preparation procedures, to determine the specific description of the event that should be flagged as a response. The company monitored three main client-related events (Opening the email, Clicking on the Advertisement, and Purchasing). Each event is recorded in the format of yyyy-MM-dd HH:mm:ss (please check Figure 4.6 for a quantitative aspect of the three main events).

Many scenarios are possible, depending on what happens when the client receives the email. We consequently decided to conduct three different experiments, depending upon how we identify the response event. The three different STEM experiments are:

Variable Name	TYPE	Description
Client ID	String	Unique hash code of the client
Client AGE	Integer	The age of the client
ZipCode	String	The zip code of client's city
Origin	String	The regular origin city of the client
Destination	String	Client's regular destination city
Orders Number 2014/15	Ordinal (Integer) [binned from 1 to 10]	Represents the number of orders the client had in the specific year
Orders Value 2014/15	Ordinal (Integer) [binned from 1 to 10]	Represents the value of orders the client had in the specific year
Segment RFM	Ordinal (String) [11 bins]	Client's RFM segment based on company's algorithm
Segment Reactive	Nominal (String) [6 bins]	Client's Reactive segment based on company's algorithm
Segment Anticipation	Ordinal (String) [7 bins]	Client's Anticipation segment based on company's algorithm
Segment Behavioral	Nominal (String) [9 bins]	Client's Behavioral segment based on company's algorithm
Treatment	Binary [A=0,B=1]	Represents the type of treatment the client has been subjected to
Open/ Click/ Purchase Timestamp	Date Time	Variables include date and time of the three main response events : Opening, Clicking on Advertisement and Purchasing
Order_Type	Nominal (String) [31 Type]	Represents the type of product the client ordered

Table 4.2 – Dataset variables description. 18 variables (features) provided after experimentation. Variables that share similar description are joined in one row.

1. Click vs. No-Click
2. Click & Purchase vs. Click & No Purchase
3. Click & Purchase within 40 Days vs. Click & No Purchase

We generated six new variables to support the analysis. Starting with three binary variables: Opening (the email), clicking (on the advertisement) and purchasing (order). We then have measured the occurrence of each action (0,1). A variable, named Open_To_Click_Hour, is also created to show the lapse of time (in hours) between Opening the email and Clicking on the advertisement. Finally, two variables are added to measure the time gap (in days) between Opening and Purchasing; and

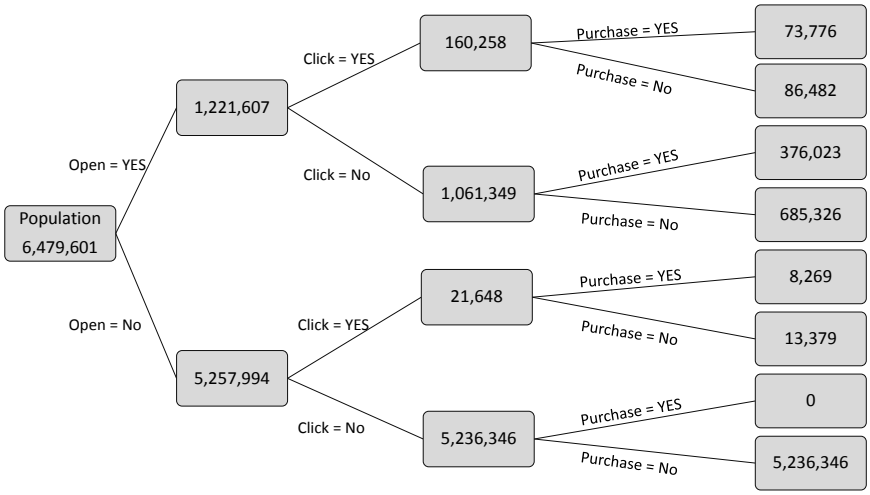


Figure 4.6 – Breakdown tree visualization of population based on main events. Quantitative breakdown tree of population and consecutive recorded three main events (open the email, click on the advertisement, and purchase a product).

the time gap between Clicking and Purchasing actions.

To clean up the data, we replaced the missing values of the `Client_Age` variable with the number (-1) and the missing values of `ZipCode`, `Product_Type`, `Origin` and `Destination` variables with the String "N/A." Finally, we binned the `Orders_Number` and `Orders_Value` columns into ten bins.

For each experiment, we compare our STEM technique (BRUM) with the Causal Conditional Inference Forest (CCIF), CCIF is deemed one of the most performant STEM technique used so far [17]. In addition, we compared BRUM with an improved version of Lai's method (GB-Lai's) and an improved version of the standard Two-models technique (GB-Two models). GB is an abbreviation of Gradient Boosted classification trees method. GB trees technique is used for models construction [144].

To do the experiments we partitioned the data into 80/20, learning/validation set respectively. Learning set is used to train each model and validation set is used just for validation. A major obstacle regarding evaluating STEM approaches is due to what is called 'the fundamental problem of causal inference' [41], which states that we can only observe one outcome of an individual after being subjected to specific treatment, and we cannot observe other treatments' outcomes at the same time. In other words, we do not have individual's other true responses (if he had been exposed to the other treatments) to calculate the true STE score (True STE score = Observed Response after Treatment one - Observed Response after Treatment two). Following the literature [17, 6, 93], we partitioned the population into subgroups (bins) to solve that problem. By doing so, we measure the aggregated STE of a specific subgroup and not the individual treatment effect itself. After sorting the population descendingly based on their *predicted STE* score, We divided the population into ten bins (deciles) with equal rows frequencies, where Bin 1 has the highest average

predicted STE and Bin 10 has the lowest average predicted STE.

We used three measurement criteria to compare between different techniques. First, we used *Qini Coefficient*, which is the ratio between the area under model's Qini Curve and the area under Optimal Qini Curve. A higher Qini coefficient value reflects a better ability to sort the cases from higher STE toward lower STE [21]. In addition, we used a new measurement introduced by [108], which is the χ^2_{net}; this measurement is better when coupled with the Qini coefficient score, to assure that our results are not encountered by accident. χ^2_{net} measures the significant difference between the STE of the top and the STE of the last decile (bin) of each validation set, a higher χ^2_{net} with lower $p-value$ reflects better separation and thereby better STE model. As well, we compared between different STEM methods by computing the Expected STE (ESTE), the ESTE is the STE that we can get if we only target the first decile with treatment B (treatment = 1) and the last decile with treatment A (treatment = 0). Specifically, the Expected STE measures the maximum personalized targeting strategy (remember that the first decile contains cases that are predicted to favor treatment B over A and vice versa for the last decile). The ESTE is useful for having an approximate estimation of the maximum increase in the STE that we will get if we have we used the specified method.

We hereafter present our three experiments and the findings obtained for each of them.

Experiment 1

We started by excluding customers who did not open the email. Then, we posited "*click*" event as a positive response to Treatment. After that, we trained the models using the training set (80% of clients who open the email). Table 4.3 displays the cross tabulation of *Treatment* and *Click* variables for the validation set (20% of clients who opened the email).

Validation set of Experiment 1		**Click=0**	**Click=1**	**Total**
Treatment = A	Count	97,621	18,419	116,040
	Row% = Cell count / Total row count	84.13%	18,419 / 116,040 = **15.87%**	100%
Treatment = B	Count	115,358	17,253	132,611
	Row% = Cell count / Total row count	86.76%	17,253 / 132,611 = **13.01%**	100%
Total		212,979	35,672	248,651

Table 4.3 – Cross tabulation of validation set of experiment 1

The random STE value of the validation set = Response rate for Treatment B - Response rate for Treatment A = 13.01% - 15.87% = -2.86% = -0.0286. This negative

STE implies that Treatment A has generated higher clicking rates than Treatment B. As we can see in from Table 4.4, The Qini coefficient scores for all methods are negative. A negative Qini coefficient means that the methods failed to correctly sort the population, it failed to separate between clients who clicked because of Treatment A and clients who clicked because of Treatment B. Notice that Qini curves are under the *Random STE Line* (Figure 4.7).

Method Name	**Qini Coefficient**	χ^2_{net}	χ^2_{net} $p-value$	**Random STE**	**Expected STE**
BRUM	-0.8%	26.831	0	2.86%	1.726%
CCIF	-0.4%	73.958	0		2.758%
GB Lai's	-1.9%	14.228	0		1.241%
GB Two models	-2.1%	7.085	0.008		1.006%

Table 4.4 – Results table of experiment 1

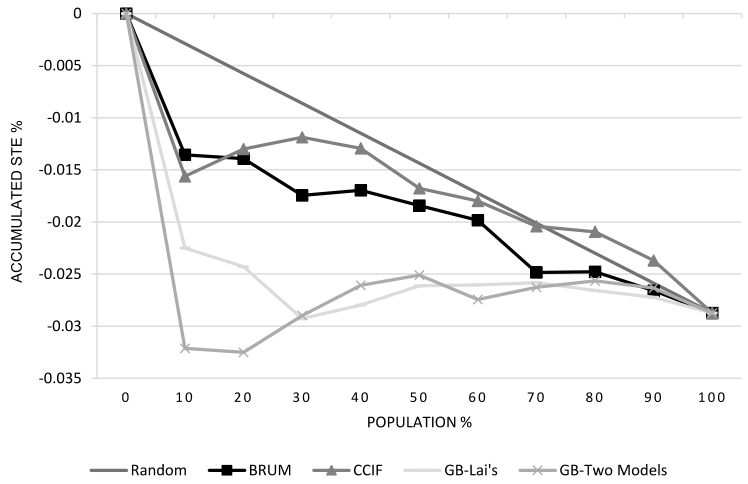


Figure 4.7 – Qini curves for experiment 1: Qini curves for (BRUM, CCIF, GB-Lai's and GB-Two Models). Random line (diagonal line) represents random targeting Qini curve.

These disappointing Qini results could be due to, first to the fact that there is no real STE in the data, which means that the detected random STE might be due to chance. Second, it is also possible that there is in fact excessively random noise compared to the real STE in the data [62], preventing all models available from correctly detecting the real STE. It seems, however, as shown in Table 4.4, that all models have low χ^2_{net} $p-value$, this means that the difference between the STEs of the highest predicted STE and the lowest predicted STE subgroups is significant. The low χ^2_{net} $p-value$ increases the likelihood that the negative Qini coefficient is

due to sampling bias, and not due to models inefficiency. In conclusion, the Expected STE for all methods is less than the Random STE. Therefore, it is better for this case, based on the results, to target customers randomly than to use STEM for targeting.

Experiment 2

We first exclude from our sample the customers who did not open the email, neither click on the advertisement. Second, we posit "Purchase" event as a positive response to the treatment (A or B). Table A.2 shows the cross-tabulation of Treatment and Click & Purchase variables in the validation set.

Validation set of Experiment 2		**Purchase = 0**	**Purchase = 1**	**Total**
Treatment = A	Count	10,364	8,304	18,668
	Row %= Cell count / Total row count	55.52%	**44.48%**	100%
Treatment = B	Count	9,424	8,290	17,714
	Row %= Cell count / Total row count	53.21%	**46.79%**	100%
Total		19,788	16,594	36,382

Table 4.5 – Cross tabulation of validation set of experiment 2

The random STE of the validation set of the second experiment is 46.79% - 44.48% = 2.31% = + 0.0231, that implies that treatment B is more effective than treatment A.

Method Name	**Qini Coefficient**	χ^2_{net}	χ^2_{net} $p-value$	**Random STE**	**Expected STE**
BRUM	**6.3%**	**56.653**	0	2.31%	**8.635%**
CCIF	4.2%	14.065	0		4.389%
GB Lai's	2.2%	2.308	0.129		1.309%
GB Two models	1.6%	0.014	0.906		0.050%

Table 4.6 – Results table of experiment 2

We can see on Table A.3 that BRUM model succeeded in predicting the STE of subpopulations. In addition, we notice, from χ^2_{net} and χ^2_{net} $p-value$ that BRUM model was able to differentiate between customers who purchased after treatment A and customers who purchased after treatment B. The Expected STE for using BRUM model is three times the random targeting STE score; this means that by

using BRUM approach we could increase the random STE (purchase rate) from 2.3% to a maximum of 8.6%.

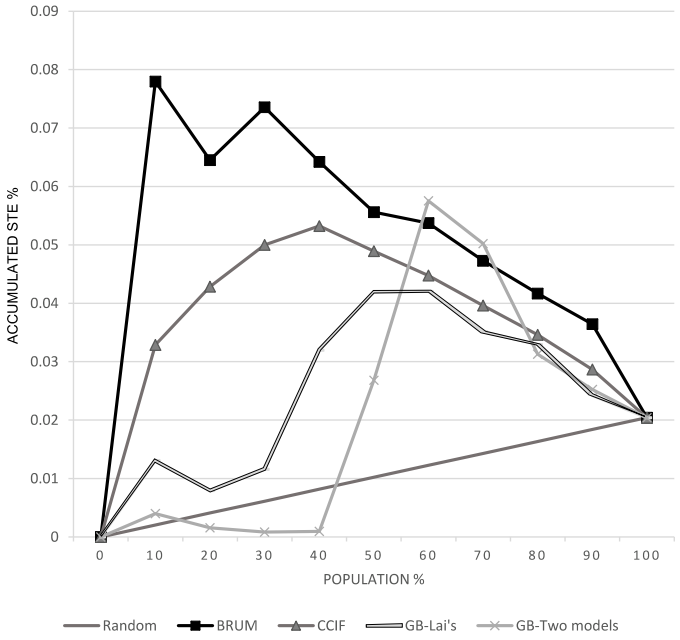


Figure 4.8 – Qini Curve for experiment 2: Qini curves for (BRUM, CCIF, GB-Lai's and GB-Two Models). Random line (diagonal line) represents random targeting Qini curve.

Experiment 3

In the third experiment, we assume that a positive response to Treatment is to purchase (order a travel ticket) within a given period of time (which could equal to the validity time of the special offer proposed in the advertisement). This period of time has been limited to 40 days between clicking and purchasing events (which is the duration on average of the promotional offers proposed by the website). We excluded from the subpopulation under study, the customers who did not open the email, neither clicked on the advertisement. We have also excluded those who have purchased after 40 days from clicking. Table 4.7 contains the cross-tabulation of the variables Treatment and Purchasing within 40 days in the validation set.

Validation set's random STE is 21.61% - 20.26% = 1.35% = 0.0135, which still implies the stronger effect of Treatment B. Based on the Qini coefficient in Table 4.8, BRUM scored the second highest Qini coefficient. (BRUM Qini = 0.035, GB_Two models = 0.042). Nevertheless, BRUM scored the highest χ^2_{net} and lowest χ^2_{net} $p-value$. In addition, BRUM recorded the highest Expected STE. We can see from Figure 4.9 that GB_Two models method failed to sort the subgroups from higher

Validation set of Experiment 3		Purchase = 0	Purchase within 40 days = 1	Total
Treatment = A	Count	10,413	2,647	13,060
	Row %= Cell count / Total row count	79.74%	**20.26%**	100%
Treatment = B	Count	9,543	2,632	12,175
	Row %= Cell count / Total row count	78.39%	**21.61%**	100%
Total		19,956	5,279	25,235

Table 4.7 – Cross tabulation of validation set of experiment 3

to lower STE. This explains the higher Qini coefficient and the low Expected STE. However, by using BRUM method instead of Random targeting, the company could reach 5.8% of incremental purchase rate (STE) instead of 1.35%.

Method Name	Qini Coefficient	χ^2_{net}	χ^2_{net} $p-value$	Random STE	Expected STE
BRUM	3.5%	**19.561**	**0**	1.35%	**5.824%**
CCIF	1.3%	0.3	0.584		0.746%
GB Lai's	1.2%	1.064	0.302		1.005%
GB Two models	**4.2%**	0.029	0.865		-0.101%

Table 4.8 – Results table of experiment 3

4.2.3 Breast Cancer dataset

Breast cancer is a major cause of death for millions of women worldwide. While medical treatments exist and become extensively efficient, the questions of how early the diagnosis is made and how to predict recurrence remain critical and affect the statistics of recovery considerably. STEM approaches have been used to predict the personal recurrence events of breast cancer [17, 19, 65]. To apply BRUM method and compare its performance to other existing STEM methods, we have used real data related to this disease. The dataset is available from the UCI repository [1]; and is provided by Matjaz Zwitter and Milan Soklic at University Medical Centre, Institute of Oncology in Ljubljana, Yugoslavia. It is mainly used as a training set for classification algorithms. The dataset represents breast cancer recurrence events. It holds

1. The dataset can be found at the following URL:(https://archive.ics.uci.edu/ml/datasets/Breast+Cancer).

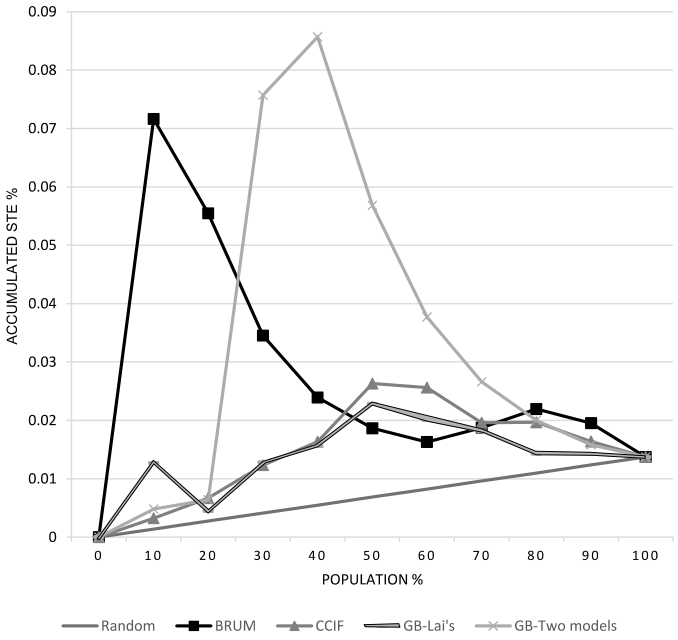


Figure 4.9 – Qini curves for experiment 3: Qini curves for (BRUM, CCIF, GB-Lai's and GB-Two Models). Random line (diagonal line) represents random targeting Qini curve.

286 instances of real patients data divided into two classes, recurrence events, and no recurrence events. Each case is described by nine different attributes sequentially; each attribute has been converted into a nominal attribute [145].

The attributes are: `Age`: in years at last birthday of the patient at the time of diagnosis; `Menopause`: whether the patient is pre- or post-menopausal, at time of diagnosis; `Tumor size`: the greatest diameter (in mm) of the existed tumor; `Inv-nodes`: the number (range 0 - 39) of axillary lymph nodes that contain metastatic breast cancer visible on histological examination; `Node caps`: if the cancer is at a metastatic phase to a lymph node; `Malignancy`: the historical grade (range $1-3$) of the tumor; `Breast`: cancer may occur in either breast; `Breast quadrant`: the breast may be divided into four quadrants; and `Irradiation`: medical protocol using high-energy X-rays to destroy cancer cells.

Following [9, 65], we converted the menopause attribute into treatment attribute and removed all attributes that are correlated with treatment binning or those having missing values. We have partitioned cancer breast dataset into 80/20 percent to build training, and validation sets respectively. We end up with 277 rows after eliminating entries with missing values. Attributes `Age` and `Breast Quadrant` have been removed. We used ensemble decision trees to build the BRUM (Balanced Reflective Uplift Modeling) and CCIF. One of the concerns about STEM is the lack of reliability of measurement tools, mainly because of the absence of the real

treatment effect value [17].

To bypass this obstacle, the segment-based comparison will be applied to calculate the error rate. More particularly, Qini Coefficient and 15% Qini Coefficient are used as scores to compare models performance, defined as the area between the actual incremental gains curve from the fitted model and the area under the diagonal corresponding to random targeting [21]. As shown, in table A.4 and figure 4.10, the results show that BRUM algorithm has better Qini and 15% Qini scores than CCIF. The results mean that the 'treated and positively responded' cases have higher levels compared to 'treated but did not responded' cases. We, therefore, provide evidence, using a real medical dataset, that BRUM outperforms the other STE methods.

	Qini	**15% Qini**
BRUM	0.27	0.31
CCIF	0.12	-0.71

Table 4.9 – Breast cancer experiment findings

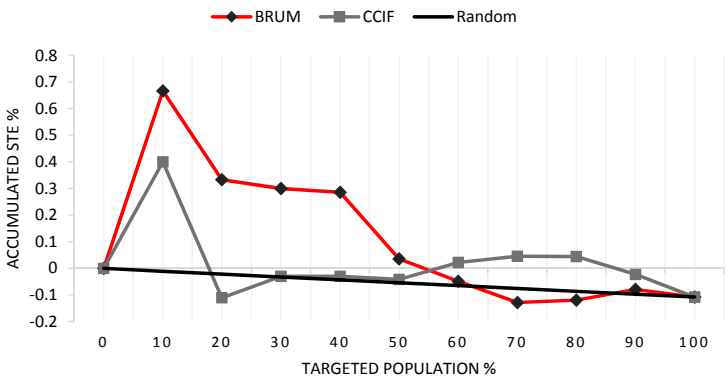


Figure 4.10 – Breast cancer experiment Qini curves

4.3 Conclusion

There are many approaches to building STEM models, most of them struggle to produce reliable scores in noisy datasets. Most of the real world datasets contain high noise and disturbances, especially for STEM, as STE tend to be smaller than the real treatment effect. This led us to develop a new approach: the balanced reflective uplift modeling (BRUM) which is based on building a stabilizing model that would help to avoid results that are based on noise.

We compared our approach with other STEM methods, using simulated and real datasets. Our approach outperformed other STEM approaches using real data set from the marketing and healthcare domains. We used *Qini* coefficient, Spearman's rank correlation coefficient, χ^2_{net} and *Expected STE* to measure and compare BRUM, Two models approach, CCIF and Lai's model. According to our results, BRUM seems more stable and more reliable for decision-making, especially in the case of A/B treatment, for which we can predict, depending on the behavior adopted by different customers, which treatment A or B is more effective for the individual.

Chapter 5

Conclusions

> When I examine myself and my methods of thought, I come to the conclusion that the gift of fantasy has meant more to me than any talent for abstract, positive thinking.
>
> Albert Einstein

The dramatic increase of accessible data and the high demand for techniques to aid decision support systems encouraged the emerging of learning from data phenomenon to provide the essential tools to maximize decision confidence, minimize side effects, and facilitate decision making processes. In many circumstances, customized decisions must be made based on each case's characteristics. Subpopulation treatment effect modeling (STEM) helps the decision making process and maximizes particular action effects.

STEM is a branch of machine learning involved in designing models, using potential outcome inferential framework [146], to map treatment effect fluctuations across the population. STEM aims to understand how particular event effects vary between subpopulations and model the variance for future estimations. The STEM is an effective tool that is used to determine the importance of a treatment. From the statistical learning perspective, the goal of STEM is extremely challenging because the optimal treatment is a priori unknown on a given training dataset [17].

The rest of this chapter is organized as follows. Section A.5.1 reviews and summarizes the contributions of this thesis. Section A.5.3 discusses some limitations of our current work and proposes alternative solutions. To conclude Section A.5.4 surveys future research directions that could originate from the work presented in this thesis.

5.1 Synthesis of the contributions

As stated in the introduction, the contributions of this thesis are given by exploring the problem of STEM, proposing new taxonomy for STEM approaches, introducing STE neighborhood binning technique (STE-NBB), new STEM approach

(STE-Strees), STE Neigborgood random forests (STE-NRF), and the balanced reflective uplift modeling approach (BRUM).

5.1.1 Subpopulation treatment effect modeling survey

Throughout the thesis, we encounter some papers that explain and utilize STEM. We find out that STEM is utilized in many domains under different names. Also, there were variate definitions for STEM. Those were the motivations for us to revisit STEM problem. And because causality is the base of STEM research, we introduce the idea of STEM from causal point of view. Also, we introduced a new taxonomy for the STEM approaches based on three main action (split, model, transform). In addition, we gathered all the terms that are used for STEM in the literature under one section, that would help future STEM researcher. Finally, we list all the measurements that evaluate STEM models performance .

5.1.2 Subpopulation treatment effect neighborhood binning technique

Data uncertainty always exists due to fundamental causal inference [41], i.e., only a subset of treatments can be directly observed. Traditionally, STEM uses response based binning techniques to use the data for inference. We propose new binning technique, called subpopulation treatment effect neighborhood binning technique (STE-NBB), that is not based on the response variable. Our techniques weight and partition cases based on their neighborhood cases. This leads to better partition, less bias and create better subgroups for better prediction score.

5.1.3 Subpopulation treatment effect sliding trees

To use the proposed binning technique (STE-NBB) in a STEM framework, we used decision trees as a base technique and created the Subpopulation treatment effect sliding trees (STE-STrees). Our new approach can be categorized as TTM STEM A.2.2, because we transform the response variable into another form before modeling.

5.1.4 Subpopulation treatment effect neighborhood random forests

We improved the STE-Strees to handle bias and noise using ensemble methods and create subpopulation treatment effect neighborhood random forests (STE-NRF). STE-NRF approach select a sample of the population and a sample of features for building each tree model in the forests. The improvement of STE-NRF are shown in our experiments, where STE-NRF outperforms other STEM approaches in all scenarios.

5.1.5 Balanced reflective uplift modeling

Another major problem that faces STEM are the noise, correlated features and error data (i.e., disturbances). We propose a new measure for STE, we name it the Reflective uplift (RU). Reflective uplift can be used besides the traditional STE to balance the score of in cases of high disturbances. We propose the balanced reflective uplift modeling (BRUM) as a STEM approach which is less sensitive to noise and disturbance in the data.

5.1.6 Experiment to improve marketing decision-making (SNCF dataset)

We conduct an experiment with the marketing team in the French national railway company (SNCF) to improve the decision-making for advertisement campaign. Using data gather through an E-mail A\B testing, we applied the STEM models in three different cases, and we showed that our approach can be a better targeting strategy to maximize response rate and minimize marketing spending.

5.2 Potential applications

Many application across various domains can harness the power of this thesis contributions. Our STEM techniques are beneficial when the data includes continuous variables. The main forms of these potential applications are the same, either to maximize the outcome of a specific intervention, or to minimize the negative side effect of a specific intervention, or both.

In the marketing domain, to maximize the return on investment of an email advertisement, or to maximize the click-through-rate of a specific online promotion. In the medicine domain, to maximize the patient's survival rate of a specific treatment, or to minimize the negative side-effects of a risky medical treatment (e.g., chemo therapy). In the politics domain, to maximize the votes for a specific candidate by focusing on the most persuasive group of voters.

5.3 Limitations and improvements

With respect to the above synthesis of the contributions of this thesis, we discuss here their limitations and propose solutions to improve our work.

Regarding the STE-STrees and STE-NRF approaches, their benefits focus exclusively on improving the processing of continuous and ordinal variable with too many levels. However, if the dataset does not contain those types of variables, our proposed STEM approaches will use a binning technique based on Euclidean distance rather than STE-NBB, similar to the STE random forests technique used by [17, 62].

In addition, our proposed approaches do not take into account the distribution of the continuous variable, which leaves them vulnerable to error and biased data,

such as outliers (extreme cases) in the data. These problems will be investigated in future studies using a dynamic window step sizes.

Another limitation for our approach is the computational load that is required for modeling process, this can be managed by changing the step size based on the total number of rows in the dataset.

Our BRUM approach is subject to other limitations related to the field under study, namely, for example, the exclusive consideration of binary treatments.

The CCIF method, provided in the uplift R package, does not handle missing data. We coded all missing data as values or categories to solve this problem, but then, we could have added more noise to the data by using this naive solution for handling missing data. In addition, by binning numerical variables, we minimized the modeling time, but we may have also introduced noise and bias to the data.

It is worth noticing that the *Qini Coefficient* alone is not a reliable measurement of STEM method effectiveness. Researchers are encouraged to use other methods like χ^2_{net}, *Expected STE* or the *Expected Return on Investment* to compare between different STEM methods.

5.4 Future works and perspectives

In future work, it would be interesting to examine the distribution of continuous treatment. This can be done by adding a step before each binning to detect the distribution of the selected feature. Afterwards, the window size and the step size can be adjusted to match the required distribution. It is interesting to compare the performance of the dynamic step and window size against fixed ones.

In addition, there is a potential for improving the proposed neighborhood based binning by combining it with the causal tree technique proposed by [22]. The idea of separating the binning process and the evaluation process can minimize overfitting and guarantees better prediction.

Our BRUM approach focuses exclusively on binary treatments. It would be interesting to examine our approaches in a multi-category or a continuous treatment settings. Indeed, the case of adaptive treatment strategies, deemed effective in the medical sector, and according to which the treatment typed could be changed according to the ongoing response of the patient [147, 148, 149] needs also to be handled with optimized STEM techniques. More research is certainly needed to handle this case and many others within what we can refer to as perspective analytics using machine-learning techniques.

Finally, we began exploring an application for STEM approach to detect the psychological behavior of online clients. This leads to another interesting area of exploration, which is STEM for stream data. The ultimate solution will be using STEM streaming approach to detect the change of behavior for online clients, and to get future insights based on it.

Appendix A

Résumé de la thèse

> Le signe premier de la certitude scientifique, c'est qu'elle peut être revécue aussi bien dans son analyse que dans sa synthèse.
>
> Gaston L.P. Bachelard,
> *La formation de l'esprit scientifique*

A.1 Introduction

Dans de nombreuses applications, il est essentiel de mesurer les effets d'un traitement spécifique sur la population, par exemple pour étudier l'impact d'une publicité sur le comportement d'achat des consommateurs, pour mesurer l'effet d'un médicament spécifique sur la santé des patients ou pour en calculer le coût. impact de la conception du nouveau site Web sur le taux de clic des visiteurs. Afin d'estimer l'effet de chacun des traitements précédents, une expérience randomisée doit être menée, puis l'effet causal moyen du traitement (ATE) peut être calculé. Cependant, outre l'ATE, les données résultantes de l'expérience contiennent des informations supplémentaires sur la variation de l'effet de causalité dans la population. Ces informations peuvent être utilisées pour tirer des conclusions sur l'effet du traitement sur un groupe spécifique de la population, c'est-à-dire pour mesurer l'effet du traitement de la sous-population (STE).

Avec la taille et la complexité croissantes des structures de données à l'ère du Big Data et les progrès rapides de la technologie de collecte et de stockage de données, le concept d'apprentissage à partir de données à l'aide de statistiques et d'algorithmes d'apprentissage automatique est apparu dans divers domaines [2]. Des algorithmes d'apprentissage automatique ont été développés pour comprendre les modèles de données et découvrir les relations qui sont cachées à la perception humaine. En outre, les vastes développements en matière de traitement informatique nous permettent de modéliser des systèmes complexes qui permettront un meilleur système de recommandation, un logiciel de diagnostic de maladie précis, des traductions linguistiques rapides, une détection proactive de la fraude et des stratégies de marketing

efficaces. Cependant, dans la recherche en science des données, une petite attention a été portée aux problèmes qui concernent l'utilisation de données expérimentales pour découvrir des inférences causales.

Nous nommons le domaine scientifique qui concerne l'application de l'apprentissage automatique pour modéliser l'effet d'un traitement sur une modélisation d'effet de traitement de sous-population (STEM). STEM est l'une des approches émergentes dans les domaines de la science des données et de l'apprentissage automatique. Il est également connu sous le nom de modélisation de soulèvement, de modélisation de résilience nette, de modélisation de réponse différentielle et de modélisation de valeur incrémentale [3, 4, 5, 6]. STEM n'est pas appliqué pour prédire la variable de classe, mais pour prédire la variation de la variable de classe [7]. Dans la plupart des cas, un traitement est rarement bénéfique pour l'ensemble de la population. STEM est donc utilisé pour modéliser l'effet du traitement en fonction des caractéristiques des sous-groupes. STEM est donc devenu un excellent outil pour personnaliser le traitement.

Les techniques STEM ont été appliquées dans de nombreux domaines. Dans le secteur de la santé, les STEM ont été utilisés pour optimiser les avantages du traitement pour chaque patient [8, 9, 10]. En outre, il a été utilisé pour le traitement du cancer, en estimant les avantages de la chimiothérapie sur la base du stade du cancer [11, 12]. En outre, en utilisant les données de survie, STEM a été utilisé pour prédire la probabilité de récupération des patients [13]. Les STEM ont également été utilisés pour explorer le rôle de la consommation de drogues dans le risque sexuel [10].

En marketing, l'utilisation d'une publicité personnalisée était la principale raison d'adopter STEM [14, 6]. Une autre application du marketing est la détection de l'impact négatif des publicités, qui facilite la gestion des dépenses de marketing [15, 16, 17]. De plus, les sciences économiques ont utilisé STEM pour tester l'effet du programme de formation professionnelle sur les gains post-intervention; il a également été utilisé pour sélectionner les stratégies de mobilisation des électeurs les plus efficaces [18, 19]. En finance, STEM a aidé les compagnies d'assurances à maximiser leurs profits en améliorant la modification personnalisée des taux pour les assurés [14].

La littérature a divisé les approches STEM en deux catégories de base [6]. La première approche, appelée STEM indirecte, combine les scores de plusieurs modèles pour générer le score STEM final. La deuxième approche, appelée directe STEM, utilise un seul modèle pour générer le score final STE [14, 20]. Les modèles STEM indirects sont plus simples et faciles à mettre en œuvre. Cependant, les approches STEM indirectes montrent plus de sensibilité au bruit que les méthodes STEM directes [21]. Les différences entre les STEM directes et indirectes ont été abordées dans [21, 14, 22, 20].

Cependant, la littérature sur les STEM est dispersée en raison des multiples termes utilisés (par exemple, modélisation du soulèvement, modélisation de la portance nette, prédiction différentielle ...). En outre, il existe plusieurs définitions de STEM qui diffèrent en fonction du domaine et de la technique utilisée. Cette situation crée

une littérature STEM basée sur un domaine, ce qui pose un problème de recherches répétitives et ralentit le développement scientifique de STEM. En outre, il n'existe pas de mesure d'évaluation cohérente pour les modèles STEM, ce qui crée un autre défi pour l'avancement de la recherche en STEM.

Malgré son importance, STEM montre un manque de fiabilité pour les problèmes réels rencontrés. Principalement en raison du cadre d'inférence causale de STEM, nous ne connaissons pas la valeur réelle de toute action (par personne). Les informations manquantes dans les données STE sont la principale raison du problème d'incertitude. Habituellement, les scientifiques mènent des expériences combinées à des techniques de regroupement pour contourner ce problème. Dans le domaine de l'apprentissage automatique, et surtout lorsqu'il s'agit de données d'observation, des techniques de regroupement spécifiques doivent être appliquées pour formuler les groupes nécessaires permettant d'utiliser les données à des fins d'inférence.

Les techniques de binning actuelles sont basées sur le binning de seuil, à savoir la recherche du point de partitionnement optimal. Threshold-binning essaiera de construire un nouveau noeud enfant (sous-groupe) de manière à maximiser la différence STE entre le noeud enfant et le noeud d'origine. De cette manière, le seuil de classement garantit davantage de sous-groupes homogènes STE après chaque traitement. Pour le type nominal de caractéristiques, le filtrage par seuils fonctionne parfaitement. Néanmoins, pour le type d'attribut continu, le tri de seuil souffre d'un grossissement d'incertitude et d'une sensibilité au bruit plus élevée en raison de trois points:

— Le problème fondamental de l'inférence causale, nous ne connaissons qu'une partie de la valeur vraie, ce qui aboutit à des classes vraies incertaines.

— La procédure de regroupement rigide, qui dépend de ces vraies classes incertaines, conduira à des points de partition incertains, augmentant ainsi le surajustement.

— Les conséquences d'un point de partition d'erreur affecteront tous les autres nœuds enfants, ce qui conduira ultérieurement à des résultats incorrects.

Pour cette raison, il est essentiel de créer une solution personnalisée pour STEM capable de gérer les attributs en continu et de minimiser l'incertitude et la suradaptation des modèles STE.

Dans cette thèse, nous réintroduisons d'abord la problématique STEM du point de vue causal. De plus, nous présentons tous les termes utilisés pour STEM. Nous proposons une nouvelle taxonomie pour les approches STEM. Et nous répertorions toutes les mesures qui évaluent les performances des modèles STEM. Deuxièmement, nous expliquons les problèmes d'incertitude et de perturbations auxquels sont confrontés les STEM. Nous analysons les techniques de tri actuellement utilisées par les approches STEM. Ensuite, nous montrons les limites des techniques de binning actuelles. Ensuite, nous introduisons une nouvelle approche de binning appelée STE binning basée sur le voisinage (STE-NBP). Ensuite, nous comparons différentes techniques de binning en prenant en compte tous les résultats potentiels, et nous présentons les avantages de l'utilisation de notre nouvelle technique de binning.

Troisièmement, nous proposons une nouvelle approche pour modéliser les STE en utilisant des arbres de décision et la technique de binning STE-NBB appelée arbres

à effet de traitement de sous-population (STE-STrees). En utilisant l'approche STE-STrees, les chercheurs seront en mesure de prendre de meilleures décisions à temps en utilisant des informations actualisées, ce qui conduira à des prévisions plus précises et à des résultats utilisables dans les systèmes d'aide à la décision. En outre, dans le domaine du marketing, STE-STrees peut être appliqué en tant que technique de test A\B dynamique, ce qui aidera à entretenir les relations avec les clients à l'ère du big data.

Quatrièmement, nous proposons les forêts aléatoires de voisinage à effet de traitement des sous-populations (STE-NRF) comme une amélioration de STE-STress afin de mieux gérer les biais et le bruit dans les données. Nous montrons comment notre méthode utilise la technique de binning basée sur le voisinage pour améliorer les performances de STEM.

Cinquièmement, nous introduisons une nouvelle approche STEM appelée modélisation équilibrée par réflexion ascendante (BRUM), qui aide la STE standard dans les environnements à fortes perturbations. Ensuite, nous comparons différentes approches STEM en utilisant une étude de simulation approfondie de divers scénarios.

Et plus tard, en utilisant de véritables ensembles de données des domaines médical et commercial. Nous menons une expérience avec l'équipe marketing de la compagnie nationale des chemins de fer français (SNCF) pour améliorer la prise de décision concernant la campagne de publicité. En utilisant les données collectées via un test A\B par courrier électronique, nous avons appliqué les modèles STEM à trois cas différents et nous avons montré que notre approche pouvait constituer une meilleure stratégie de ciblage pour maximiser le taux de réponse et minimiser les dépenses de marketing. Nous montrons comment notre approche surpasse toutes les autres méthodes STEM en termes de coefficient *Qini* et de coefficient de corrélation de rang de *Spearman*. Enfin, nous concluons en présentant les limites et les perspectives futures de notre recherche.

A.2 Examen de la modélisation de l'effet du traitement des sous-populations

Le terme *effet de traitement* est un terme ambigu en soi, il manque l'objet, celui qui a reçu le traitement. Mais généralement, il fait référence à *l'effet moyen du traitement*. L'effet moyen du traitement (ATE), utilisé le plus souvent pour refléter l'impact d'une action, est le plus utilisé et étudié dans la littérature. Cependant, dans de nombreux cas, nous souhaitons savoir comment cela affecte les changements, quels facteurs l'affectent et comment ces changements pourraient être estimés (c'est-à-dire l'effet du traitement de la sous-population). Pour cette raison, les chercheurs effectuent une analyse effect d'effet de traitement hétérogène [26], une analyse de probabilité conditionnelle [27], une représentation graphique du motif d'effet du traitement de sous-population [28] et de nombreuses autres techniques d'analyse. Cependant, en exploitant la puissance de l'exploration de données et de l'apprentissage automatique, de nouvelles méthodes d'estimation de l'effet du traitement de sous-population (STE) sont en augmentation. Ces méthodes peuvent traiter

des données incertaines, manquantes et indépendantes du domaine. Avec l'aide de nouvelles techniques d'apprentissage automatique, le domaine de la modélisation des effets du traitement des sous-populations (STEM) est né.

A.2.1 Notation et Définitions

La notation adoptée dans nos travaux suivra celle de la littérature sur l'inférence causale. Pour simplifier, nous allons utiliser un exemple d'expérimentation publicitaire à côté de la notation. L'expérience aura une variable de traitement binaire qui reflète l'événement de réception d'une publicité. En outre, l'expérience comportera une variable de résultat binaire qui reflète l'événement de l'achat d'un produit.

Soit T la variable de traitement ($T \in \{0,1\}$). T Fait référence à l'événement de réception d'une publicité. Soit Y la variable de résultat ($Y \in \{0,1\}$). Y Fait référence à l'événement d'achat d'un produit. Soit X set l'ensemble des caractéristiques décrivant les sujets de notre expérience ($X = \{X_1, X_2, X_3..\}$). Par exemple, X_1 peut désigner le vecteur des sexes des sujets et X_2 peut désigner le vecteur des âges des sujets.

Supposons que l'indice i de toute variable V représente la valeur individuelle de cette variable v_i. Pour chaque sujet de l'expérience, nous avons des valeurs individuelles y_i, t_i, x_i, représentant respectivement le résultat de l'individu, le traitement de l'individu et ses caractéristiques. Par exemple, x_1 fait référence à un vecteur de caractéristiques qui définissent le premier sujet et $x1_1$ se rapporte au genre du premier sujet.

Soit en exposant j ers toute variable V, on se réfère à la valeur potentielle de la variable V compte tenu de la variable j. Par exemple, $V^{j=k}$ est la valeur potentielle de la variable V étant donné que la variable j est égale à la valeur k.

Il est essentiel de faire la distinction entre le résultat réel (observé) et le résultat potentiel. Dans l'exemple fourni, chaque sujet a deux résultats possibles. Le résultat potentiel si le sujet avait été traité, c'est-à-dire $y_i^{t=1}$, et le résultat potentiel si le sujet n'avait pas été traité et $y_i^{t=0}$, alors qu'un seul résultat potentiel est observable, qui est le résultat réel du sujet y_i.

Il existe trois moyens principaux de mesurer l'effet (effet causal) de la publicité (traitement) sur le comportement d'achat de la population. Il s'agit de la différence de proportions conditionnelles, du rapport des proportions conditionnelles et du rapport des chances conditionnelles [39]. Tous comparent la proportion de sujets ayant répondu (achetés) en traitement à la proportion de sujets ayant répondu dans le groupe témoin.

— La différence entre les proportions conditionnelles

$$Pr(Y = 1|T = 1) - Pr(Y = 1|T = 0)$$

— le rapport des proportions conditionnelles

$$\frac{Pr(Y = 1|T = 1)}{Pr(Y = 1|T = 0)}$$

— le rapport entre les probabilités conditionnelles

$$\frac{Pr(Y=1|T=1)/Pr(Y=0|T=1)}{Pr(Y=1|T=0)/Pr(Y=0|T=0)}$$

Toutes ces mesures évaluent l'effet de causalité de la publicité publicitaire sur le taux d'achat de la population. Mais chaque mesure se concentre sur une échelle différente. Il est à noter que ces mesures ne sont applicables que dans les conditions d'une expérience randomisée.

Dans des conditions expérimentales randomisées parfaites, nous aurons:

$$Pr(Y=1|T=1) - Pr(Y=1|T=0) = Pr(Y^{t=1}=1) - Pr(Y^{t=0}=1)$$

Cependant, dans les expériences réelles, nous ne pouvons pas garantir les conditions expérimentales parfaites.

Nous définissons *l'effet moyen du traitement* (ATE) comme la différence entre la proportion du résultat observé dans la condition de traitement et la proportion du résultat observé dans la condition de contrôle (pas de traitement).

$$ATE = Pr(Y=1|T=1) - Pr(Y=1|T=0)$$

Si nous voulons voir l'effet du traitement sur une personne donnée, nous utilisons *l'effet de traitement personnel* (PTE), qui ne peut être exprimé qu'en utilisant le résultat personnel potentiel:

$$PTE_i = Y_i^{t=1} - Y_i^{t=0}$$

Cependant, comme les informations sur les résultats potentiels ne seront jamais disponibles dans le cadre d'une expérience réelle, nous devons nous appuyer sur d'autres approches pour estimer l'effet du traitement personnel. Nous pouvons principalement faire correspondre notre sujet du groupe de traitement à un autre sujet du groupe de contrôle. Dans ce cas, nous pouvons comparer le taux de réponse entre les deux sous-populations de sujets partageant des caractéristiques communes. Nous appelons cette variance *l'effet du traitement de la sous-population* et nous pouvons la définir mathématiquement comme suit:

$$STE_i = Pr(Y=1|T=1, X=x_i) - Pr(Y=1|T=0, X=x_i) \qquad \text{(A.1)}$$

Pour trouver une correspondance, il faut d'abord compresser les caractéristiques du sujet, malgré leur volume et leur complexité, en des données pouvant être mises en correspondance avec un autre sujet de l'autre traitement. Et même si nous trouvons une correspondance exacte pour le sujet requis, la correspondance sera plus difficile si nous considérons le facteur temps. Ce problème est connu dans le domaine de l'inférence causale comme "*le problème fondamental de l'inférence causale*"[41].

En supposant que nous puissions identifier un individu par ses caractéristiques infinies et que nous sommes limités par nos connaissances et nos capacités de calcul, nous pouvons utiliser STE comme approximation de PTE. Cependant, lorsque nous aurons accès à plus de données, notre entreprise de convergence convergera vers PTE.

Proposition 2 *STE converge vers PTE lorsque les données empiriques augmentent.*
$\lim_{x_i \to \infty} STE_i = PTE_i$

Définition de l'effet du traitement de sous-population

La définition de STE est obtenue à partir de la définition de modification d'effet, c'est lorsque l'effet d'une variable sur une autre varie d'une couche à l'autre ou sur plusieurs variables[45]. La distinction entre interaction et modification d'effet est ici cruciale. Am L'interaction peut être faite lorsque nous avons deux interventions, alors que la modification de l'effet ne nécessite que le conditionnement d'une troisième variable [46].

Definition 3 *L'effet de traitement de sous-population est la disparité d'effet de traitement qui se produit lorsque l'effet d'une variable sur une autre varie d'une couche à l'autre.*

Définition de la modélisation de l'effet du traitement par sous-population

La modélisation des effets de traitement par sous-population est la science qui s'intéresse à la modélisation des relations entre les variables de traitement et les variables de modification d'effet.

Definition 4 *La modélisation de l'effet de traitement de sous-population est le domaine de l'apprentissage automatique qui concerne la modélisation de l'effet de traitement de sous-population dans les données.*

A.2.2 Taxonomie de modélisation d'effet de traitement de sous-population

Les techniques STEM se sont développées dans un environnement multi-domaines. Cette situation permet aux solutions nouvelles et innovantes d'apparaître en tant que techniques STEM. Nous ferons la taxonomie basée sur trois processus principaux.

- Fractionner: fait référence au processus de division du jeu de données en fonction de chaque traitement.
- Modèle: fait référence au processus d'utilisation des données pour créer un modèle.
- Transformer: qui fait référence aux processus de manipulation des données avant la modélisation.

Split puis modéliser

Les approches de modélisation puis de modélisation des effets de traitement des sous-populations (STM-STEM) divisent la base de données en fonction de chaque

traitement. Ensuite, pour chaque sous-ensemble de traitement, un modèle prédictif est construit. Ensuite, l'estimation finale de STE est calculée sur la base des différents modèles de réponse. En raison de sa simplicité, l'approche STM-STEM est la méthode la plus intuitive pour STEM. Par exemple, dans une expérience de traitement / contrôle, un modèle de réponse sera ajusté en fonction du sous-ensemble de traitement $M_{treat} = E[Y|X, T = 1]$, et un autre modèle de réponse sera ajusté en fonction du sous-ensemble de contrôle $M_{control} = E[Y|X, T = 0]$. Ensuite, une estimation de l'EST sera calculée en tant que différence entre les modèles de réponse estimés.

$$\widehat{STE}_i = M_{treat}(Y|X = x_i) - M_{control}(Y|X = x_i) \tag{A.2}$$

Une approche STM-STEM basée sur la régression a été appliquée par [5, 49, 6, 50]. Pour une expérience de traitement binaire, soit x_i un vecteur de variables prédictives pour le cas i, $Pr(Y = 1|T = 1, x_i)$ est la probabilité de réponse du cas i en cours de traitement et $Pr(Y = 1|T = 0, x_i)$ est la probabilité de réponse du cas i sous contrôle,

$$M_{treat}(Y = 1|X = x_i) = \frac{e^{B^{T=1}_{x_i}}}{1 + e^{B^{T=1}_{x_i}}}$$
$$M_{control}(Y = 1|X = x_i) = \frac{e^{B^{T=0}_{x_i}}}{1 + e^{B^{T=0}_{x_i}}}$$
$$\widehat{STE}_i = M_{treat}(Y = 1|X = x_i) - M_{control}(Y = 1|X = x_i)$$

Où $B^{T=1}$ et $B^{T=0}$ sont des vecteurs de coefficients de régression logistique pour le modèle de traitement et de contrôle respectivement [5].

une approche à base d'arbre de décision a été appliquée par [5, 51, 52, 53, 54, 22]. En outre, une approche STM-STEM basée sur la programmation par logique inductive (ILP) [55] a été proposée par [34]. [34] a proposé de construire des arbres séparés augmentés du modèle de Bayes-net Naive Bayes (TAN) pour chaque traitement, puis de maximiser l'aire sous la courbe de soulèvement. Les approches STM ont montré une sensibilité élevée au bruit, comme expliqué dans [21]. Les auteurs ont fait valoir que les modèles distincts dans STM-STEM se concentrent davantage sur la variable de réponse que sur la variance de la réponse, ce qui augmentera l'effet du bruit et réduira la possibilité de modéliser le STE dans les données. En outre, [7, 22] a montré que les approches STM-STEM étaient plus performantes que les autres méthodes STE. Cependant, [50] a contredit l'argument du passé et a défendu que les approches STM sont meilleures que d'autres approches utilisant une approche de régression linéaire.

Transformer puis modéliser

La transformation, puis modéliser les approches de modélisation d'effet de traitement de sous-population (TTM-STEM) ne divise pas l'ensemble de données. Mais ils manipulent les données de manière à s'adapter à certaines exigences spécifiques. Plus tard, ils modélisent le STE en fonction du jeu de données modifié.

C'est l'approche TTM-STEM la plus utilisée. L'idée sous-jacente est de convertir la variable de classe en un autre formulaire pour qu'il soit compatible avec les algorithmes requis. Par exemple, [56] a remodelé la variable de classe pour les ensembles de données de traitement / contrôle conventionnels en une classe binaire. Les auteurs de [56] combinent les cas étiquetés comme traités et y ayant répondu avec les cas étiquetés comme contrôlés et non répondus dans un sous-groupe étiqueté comme étant un sous-groupe positif. Ensuite, les auteurs combinent les cas étiquetés comme traités et n'ayant pas répondu avec les cas étiquetés comme contrôlés et ayant répondu dans un sous-groupe étiqueté comme étant un sous-groupe négatif. Plus tard, les auteurs [56] ont utilisé un modèle de classification binaire pour estimer la réponse positive et ont utilisé cette estimation comme une approche STEM. La simplicité de cette approche a permis aux chercheurs de l'utiliser avec des techniques de modélisation d'ensemble [6]. Cette approche a été utilisée et améliorée ultérieurement par [9, 6].

En outre, l'utilisation de techniques de machines à vecteurs de support (SVM) pour modéliser STE a été utilisée dans [7, 19, 57]. Par exemple, [19] a proposé une technique de SVMs modifiée, la variable de réponse est transformée de $Y \in \{0, 1\}$ en $Y' \in \{-1, 1\}$. De plus, les caractéristiques sont redimensionnées en fonction de [58] pour appliquer deux contraintes distinctes d'opérateur de retrait et de sélection le moins absolu (LASSO)[58], une pour l'effet de traitement et une autre pour STE. Pour estimer l'EST, nous calculons la différence entre les valeurs de la réponse prédite sous le traitement et les conditions de contrôle.

$$STE(x_i) = \frac{1}{2}[(\widehat{Y_i'}|T = 1, X = x_i) - (\widehat{Y_i'}|T = 0, X = x_i)]$$

Où $\widehat{Y_i'}$ est la valeur prédite de Y' pour le sujet i.

Les auteurs de [57] ont maximisé l'aire sous la courbe de soulèvement pour prévoir les sous-groupes STE les plus élevés. Une autre méthode TTM-STEM a été proposée par [11]. Elle consiste à transformer les entités en plus de la transformation de variable de classe, puis à adapter un modèle de régression logistique aux données. Les techniques SVM ont également été utilisées pour détecter STE par [59]. Les auteurs ont nommé *l'apprentissage du poids de leur* approche (OWL). En outre, [22] a proposé une nouvelle *arborescence causale* basée sur l'approche STEM, ils ont transformé le résultat, puis ajouté des méthodes de validation croisée dans l'échantillon et hors échantillon pour choisir la meilleure pénalité.

Modèle alors divisé

Le modèle divise ensuite les effets de traitement par sous-population en modélisant les approches de MTS-STEM ne divise pas les données, ni ne les modifie, mais modélise directement le STE. Ensuite, dans la dernière étape de la modélisation, l'estimation de l'estimilité estimée est calculée en effectuant une division pour un sous-ensemble de données.

Cette approche est principalement utilisée avec des arbres de décision modifiés. Les auteurs de [1] ont proposé le premier modèle utilisé pour calculer le STE, appelé

arbre à réponse différentielle. Il est basé sur un arbre de panier modifié. L'idée est simple. ils modifient le critère de partage binaire de chaque noeud pour favoriser le partage qui maximise la différence STE entre les noeuds enfant et parent. De manière à maximiser la différence entre les deux STE en descendant dans l'arbre.

Les auteurs de [60] ont décrit ce que l'on appelle la règle de décision du seuil de rentabilité incrémental comme une alternative STEM à la règle de décision du seuil de rentabilité normal. La règle de décision relative au seuil de rentabilité est décrite par [5] comme un outil utilisé en économie comme indicateur permettant de choisir les investissements les plus rentables. L'idée est de sélectionner uniquement les clients dont le bénéfice net incrémentiel attendu est supérieur à zéro.

De plus, [61, 14, 62] a utilisé des techniques d'ensemble avec des modèles STE. Par exemple, [36] a utilisé la technique des forêts aléatoires pour calculer STE ajoute un avantage significatif, notamment en raison du problème d'incertitude des données STE 3.1. Les auteurs de [17] ont proposé une nouvelle approche STEM appelée forêt aléatoire Uplift. Il s'agit d'une méthode arborescente qui utilise la méthode standard forest pour les forêts aléatoires [63], mais en utilisant les critères de fractionnement basés sur les STE proposés par [64].

Les critères de scission utilisés dans les forêts aléatoires de soulèvement (divergence de Kullback-Leibler, distance euclidienne au carré, divergence du chi carré et divergence de la norme L1) sont proposés par [65]. Ils sont basés sur la maximisation de la divergence de la répartition des classes entre le groupe de traitement et le groupe de contrôle.

Les auteurs de [17] ont proposé une autre méthode d'ensemble STEM appelée forêt d'inférence conditionnelle causale (CCIF), qui minimise le sur-ajustement et le biais dans le processus de sélection des variables. L'amélioration est réalisée en séparant le processus de séparation des critères de la sélection de variables et en utilisant un critère d'arrêt efficace. Dans CCIF, à chaque nœud terminal (étape 4), nous testons l'hypothèse nulle d'absence d'interaction entre les caractéristiques sélectionnées et la variable de traitement. Nous nous arrêtons si nous ne pouvons pas rejeter l'hypothèse nulle.

En utilisant ce que l'on appelle «une estimation honnête», ce qui assure un processus de tri moins biaisé. Ils résolvent le problème du binning en utilisant deux échantillons, un pour choisir comment partitionner et l'autre pour estimer le STE pour chaque nœud.

[4] a proposé une approche fictive du traitement en ajustant un modèle pour modéliser l'effet de réponse principal et en ajoutant une variable factice de traitement pour détecter l'interaction entre les caractéristiques et les traitements. Soit $T \in \{0, 1\}$ la variable de traitement et Y la variable de réponse. Nous pouvons adapter un modèle de régression logistique comme suit:

$$\widehat{Y}_i = E[Y|X_i] = \frac{exp(\alpha + \beta' X_i + \theta T_i + \gamma' X_i T_i)}{1 + exp(\alpha + \beta' X_i + \theta T_i + \gamma' X_i T_i)}$$

Où α désigne l'interception, β est un vecteur de paramètres indique l'effet de variables, θ désigne l'effet de traitement principal et γ désigne l'effet STE. L'estimation

finale de STE est calculée en définissant la variable muette de traitement dans le modèle comme traitée (T = 1), puis comme contrôlée (T = 0) comme suit:

$$STE(x_i) = \frac{exp(\alpha + \beta' X_i + \theta T_i + \gamma' X_i T_i)}{1 + exp(\alpha + \beta' X_i + \theta T_i + \gamma' X_i T_i)} - \frac{exp(\alpha + \beta' X_i)}{1 + exp(\alpha + \beta' X_i)}$$

En outre, les auteurs de [8, 38] ont proposé des approches de K-plus proches voisins (KNN) [66] pour modéliser une STE. KNN, a d'abord été utilisé pour calculer STE par [8]. L'idée derrière leur méthode est simple, [8] a nommé son algorithme «Algorithme du patient semblable à moi», en utilisant l'algorithme KNN. En appliquant cette approche, l'algorithme teste séquentiellement les patients proches de notre patient cible (c'est-à-dire qu'ils partagent des caractéristiques similaires) jusqu'à ce qu'une différence significative par rapport à l'effet moyen du traitement (ATE) soit trouvée. L'inconvénient de cette méthode est qu'elle nécessite beaucoup de calcul, en particulier pour les données de grande dimension.

A.2.3 Evaluation de la modélisation de l'effet du traitement de sous-population

Mesurer la performance des modèles STE est l'un des défis majeurs du domaine STEM. Les techniques classiques d'évaluation des méthodes d'apprentissage automatique reposent sur la validation croisée: le jeu de données est divisé en un jeu de données d'apprentissage et un jeu de données de validation, puis une fonction de perte est appliquée sur le jeu de données de validation pour mesurer le taux d'erreur reflétant la capacité de la base de données. modèle pour correspondre aux vraies valeurs. Cependant, en raison du problème fondamental de l'inférence causale, nous ne connaissons que le résultat d'un traitement, nous n'avons donc pas la valeur réelle de l'effet du traitement. Pour l'évaluation STEM, le problème est plus compliqué.

Le seul moyen de mesurer les modèles STE consiste à émuler la valeur réelle via une expérience randomisée respectant les trois conditions (interchangeabilité, positivité, cohérence) [39]. Pour les ensembles de données non expérimentaux, la valeur vraie est émulée par un algorithme de correspondance, où des cas similaires sont regroupés pour calculer la valeur vraie agrégée (STE). Dans cette section, nous passons en revue les méthodes d'évaluation permettant d'évaluer la performance des modèles STE.

Mesures Qini

Les mesures Qini sont des mesures fondées sur les rangs introduites par [67] en tant que généralisation des mesures de Gini (c.-à-d. Gains) [68]. Les mesures de Qini sont basées sur la courbe de Qini (graphique des gains de STE), similaire à une courbe de gains. Pour tracer la courbe de Qini, les cas doivent être triés par ordre décroissant en fonction de leur STE prédit (score du modèle). Ensuite, les cas doivent être regroupés en sous-groupes. Le but du binning est d'obtenir le "STE réel" du sous-groupe afin que nous puissions le comparer au STE estimé. Ensuite,

sur les axes Y du graphique, la courbe de Qini ou la courbe des gains incrémentiels cumulés est tracée.

La courbe Qini représente les gains STE cumulés obtenus si nous trions les cas en fonction de leur estimation STE. La motivation derrière la courbe de Qini est de comparer les gains de chaque modèle STE en fonction de l'efficacité du tri des observations et non en comparant la valeur prédite à la valeur vraie (car nous n'avons pas accès à la valeur vraie). La courbe Qini commence au point zéro et se termine aux gains STE totaux. Toutefois, si nous trions les cas de manière aléatoire, nous obtiendrons des gains STE cumulés égaux aux gains STE totaux. Nous pouvons tracer la ligne STE aléatoire sous la forme d'une ligne diagonale entre le point zéro et le total des points STE cumulés.

En outre, pour un traitement binaire et un exemple de réponse binaire, nous pouvons tracer la courbe optimale de Qini en triant les cas dans l'ordre suivant, les cas traités et traités, suivis des cas contrôlés et des cas non répondus, suivis des cas contrôlés et des cas répondus, et enfin les cas traités et non répondus.

Les mesures Qini se composent de la valeur Qini Q, q_0 et du top qini. La valeur Qini Q est le rapport entre l'aire de la courbe Qini du modèle située au-dessus de la ligne aléatoire et la courbe optimale de Qini située au-dessus de la ligne aléatoire. Et le q_0 est le rapport entre l'aire de la courbe Qini du modèle située au-dessus de la ligne aléatoire et la courbe optimale de Qini située au-dessus de la ligne aléatoire sans effet négatif.

Dans certains domaines, il est avantageux de minimiser la population ciblée. Par exemple, en marketing, la réduction de la population ciblée réduira le coût de la campagne, ce qui maximiserait le retour sur investissement marketing. Dans ces cas, la comparaison des Qini des 5%, 10% et 20% supérieurs pour différents modèles d'EST fournira les informations utiles nécessaires pour minimiser la population ciblée. Par exemple, dans une campagne marketing future, ils peuvent cibler des clients dans les 10% de Qini les plus élevés. On prévoit que ces clients auront un taux de réponse élevé pour la publicité.

coefficient de Gini

Les auteurs ont utilisé le coefficient de Gini [69] comme mesure de STE [70, 6]. Le coefficient de Gini est calculé comme le rapport entre l'aire de la courbe des gains cumulatifs du modèle située au-dessus de la ligne aléatoire et la courbe optimale située au-dessus de la ligne statistique.

Surface sous la courbe de soulèvement (AUUC)

L'aire sous la courbe de soulèvement est similaire à la valeur Qini. Il a été utilisé dans [7, 71, 62, 64, 9]. Premièrement, nous établissons des courbes de portance distinctes sur les données de traitement et de contrôle. Ensuite, nous traçons la courbe de soulèvement qui est égale à la différence entre la courbe de levée sur les données de traitement et la courbe de levée sur les données de contrôle. La courbe

de soulèvement indique le gain net en STE si un pourcentage donné de la population est ciblé ou traité.

Moment de soulèvement des mesures

Les mesures du moment de soulèvement sont également classées par [72] pour mesurer l'exactitude et l'audace des prévisions du modèle STE. Les auteurs dans [72] ont conclu que la forme quadratique du moment de soulèvement peut être un meilleur substitut pour Qini et d'autres mesures STE dans l'évaluation du modèle STE. Les mesures de moment de soulèvement se concentrent sur les qualités de prédiction du modèle STE telles que la monotonie, les erreurs de prédiction, la propagation et le maximum de STE.

A.3 Modélisation des effets du traitement de l'incertitude et des sous-populations

A.3.1 Contexte général

L'augmentation spectaculaire du nombre de données accessibles et la forte demande en techniques d'aide aux systèmes d'aide à la décision ont favorisé l'émergence du phénomène de l'apprentissage des données afin de fournir les outils essentiels pour maximiser la confiance en la décision, minimiser les effets secondaires et faciliter les processus de prise de décision. Dans de nombreuses circonstances, des décisions personnalisées doivent être prises en fonction des caractéristiques de chaque cas. La modélisation des effets du traitement des sous-populations (STEM) aide le processus de prise de décision et optimise les effets d'action particuliers.

Supposons qu'un médecin doive choisir un traitement médical spécifique parmi trois traitements disponibles. La modélisation STE peut alors être utilisée pour choisir le traitement le plus bénéfique pour chaque patient et minimiser simultanément les éventuels effets secondaires. Dans un contexte différent, supposons qu'un responsable marketing doive sélectionner le segment de client auquel recevoir une promotion spécifique en raison du budget limité du produit ou du projet. STEM peut être appliqué pour identifier le sous-groupe de clients le plus rentable, en évitant les clients qui achèteront le produit de toute façon (c'est-à-dire sans nécessiter de promotion).

Un problème critique pour STEM est l'incertitude de l'information. L'incertitude sur les données existe toujours en raison d'une inférence causale fondamentale [41], c'est-à-dire que seul un sous-ensemble de traitements peut être directement observé. Ce problème est généralement traité en utilisant des techniques de regroupement. Dans l'apprentissage automatique avec des données d'observation, des techniques de regroupement spécifiques sont nécessaires pour formuler les groupes appropriés afin de permettre l'utilisation des données pour l'inférence.

Ces techniques de binning font face à des défis plus difficiles lorsqu'elles rencontrent une fonctionnalité continue de bin. Les approches STEM précédentes ne

traitent pas correctement les attributs continus. Principalement, la technique de regroupement en cours pour une variable continue dans les arbres de décision STEM (c'est-à-dire les critères de fractionnement) n'aide pas la découverte de STE dans les sous-populations.

Les techniques de binning actuelles reposent sur le partitionnement de seuil, c'est-à-dire la recherche du point de partitionnement optimal. Le partitionnement de seuil essaiera de construire un nouveau nœud enfant (sous-groupe) de manière à maximiser la différence STE entre le nœud enfant et le nœud principal. Ce faisant, le binning de seuil garantit un sous-groupe homogène plus cible après chaque partition. Pour les caractéristiques nominales, le seuil de partition fonctionne parfaitement. Cependant, pour les attributs continus, ce seuil b souffre d'incertitude et d'erreur en raison de trois points:

— Le problème fondamental de l'inférence causale, nous ne connaissons qu'une partie de la valeur vraie, ce qui aboutit à des classes vraies incertaines.

— La procédure de regroupement rigide, qui dépend de ces vraies classes incertaines, conduira à des points de partition incertains, et donc à un sur-ajustement.

— Les conséquences d'un point de fractionnement d'erreur affecteront tous les autres nœuds enfants, ce qui conduira ultérieurement à des résultats incorrects.

Nous avons besoin d'une solution personnalisée pour STEM capable de gérer des attributs continus et de minimiser l'incertitude et le sur-ajustement des modèles STE.

Nous analysons d'abord les techniques de regroupement en cours généralement utilisées par les approches STEM, puis identifions les limites de ces techniques. Nous proposons ensuite l'approche de binning STE basée sur le voisinage et comparons diverses techniques de binning en tenant compte de tous les résultats potentiels, et identifions les avantages proposés pour le binning STE.

Par la suite, nous proposons l'approche des forêts aléatoires de voisinage STE (STE-NRF), utilisant un binning basé sur le voisinage pour améliorer les performances STEM. Nous comparons différentes approches STEM en simulant huit scénarios différents et des ensembles de données réels issus de domaines médicaux et commerciaux. Nous montrons que l'approche proposée surpasse toutes les autres méthodes STE dans les mesures et les termes utilisés dans la littérature. Enfin, nous discutons des limites et des perspectives futures découlant de cette recherche.

Comme indiqué ci-dessus, le sous-regroupement ou le tri des données est le processus principal de tout algorithme STEM, et une faible certitude en matière de données nécessite une extrême prudence lors du traitement des données. Les problèmes de classement sont exacerbés lorsque les données ne sont pas appropriées pour le regroupement, c'est-à-dire non nominales. Dans ce cas, nous nous appuyons sur d'autres outils pour transformer les données afin qu'elles soient mieux adaptées aux modèles STE, en particulier pour regrouper les données en morceaux qui maximisent les informations obtenues de chaque groupe.

A.3.2 Binning

Le binning simplifie les données, accélère le processus et facilite l'interprétation; et a été largement étudié pour l'exploration et l'exploration de données. Certains algorithmes d'exploration de données ne prennent pas en charge les données continues en entrée. Par conséquent, le binning convertit les données continues en données catégorisées [121]. Cependant, toute catégorisation sur une variable continue entraîne inévitablement une perte d'informations [122, 123, 124], et l'objectif est de minimiser cette perte.

L'exemple de classement le plus courant est l'affectation des groupes d'histogrammes, dans laquelle un algorithme détermine le nombre approprié de déciles (groupes) afin de fournir une visualisation simple. Le même concept est appliqué pour l'exploration de données, utilisant le binning pour regrouper les membres de la manière la plus informative possible.

Supposons que nous ayons un ensemble de données avec N membres et que l'algorithme de binning divise la variable continue, X, en m cases $P = p_1, p_2, p_3, ..p_m$, p_i, représentant des intervalles de valeurs x. Ensuite, la moyenne bin peut être exprimée comme pour $\overline{p_i} < \overline{p_{i+1}}$ for $i = 1, 2, \ldots m$.

La technique la plus largement utilisée pour regrouper la population dans STEM est la différence de la différence de STE, $\Delta\Delta STE$ [21], qui calcule les gains obtenus à partir d'un regroupement spécifique en tant que différence entre l'EST de la population d'origine (STE^{original}) et la différence entre les deux. bacs (c.-à-d. $\Delta STE = |STE^{\text{first bin}} - STE^{\text{second bin}}|$). Cela maximise la différence STE entre les nœuds d'origine et enfants dans chaque arborescence. D'autres approches utilisent des techniques de binning de divergence de distribution, par ex. Divergence de Kullback-Leibler, distance euclidienne au carré et χ^2 divergence entre chaque bin STE [65, 115, 14, 17].

Les critères de fractionnement basés sur l'importance ont également été utilisés comme technique de binning [21, 17]. Cette approche est équivalente au χ^2 test de l'effet d'interaction [37]. Les critères de fractionnement basés sur l'importance montrent de bons résultats par rapport aux autres techniques de binning. De plus, le problème de classement a également été utilisé avec une estimation honnête [125], ce qui garantit un processus de tri moins biaisé et utilise deux échantillons pour résoudre le problème de filtrage en utilisant un échantillon pour choisir le mode de partitionnement et un second échantillon pour estimer l'EST de chaque nœud.

Toutes les techniques de binning liées aux STE tentent de trouver des sous-groupes ayant un effet de traitement hétérogène. Les règles de binning sont basées sur la variable de réponse, l'hétérogénéité du taux de réponse définissant les limites des sous-groupes. Cependant, un binning incorrect pourrait produire des estimations erronées et des modèles biaisés. En particulier, le binning des variables continues offre plus de points de scission que les variables catégorielles. Par conséquent, les variables continues posent d'autres problèmes de binning. Les approches STE devraient trouver des solutions améliorées pour les entités ordinales et continues. Le défi STEM consiste à découvrir des sous-groupes avec des réponses hétérogènes partageant des caractéristiques communes, malgré le bruit et l'incertitude des données.

Les sections suivantes expliquent la technique de tri basée sur la réponse pour STEM et ses points faibles. Nous présentons ensuite l'approche de regroupement par quartier proposée par STE et comparons les techniques à l'aide de données simulées tenant compte de toutes les combinaisons de résultats potentiels.

A.3.3 Traitement par sous-population basé sur la réponse par effet de traitement

Comme indiqué dans la section 3.1, STE utilise le binning pour éviter les problèmes de causalité, généralement basés sur la variable de réponse. Le binning basé sur la réponse STE détermine si un multiset donné d'entiers, S, peut être regroupé en sous-ensembles $S1$ et $S2$ de telle sorte que la différence estimée du sous-ensemble STE soit plus significative que le sous-ensemble d'origine. Les membres ayant une réponse commune sont regroupés dans une sous-population.

Les méthodes de tri basées sur les réponses détectent la variance la plus significative entre les sous-populations en fonction d'une mesure de réponse spécifique. Dans l'exemple actuel, nous utilisons la différence entre les deux STE, mais d'autres méthodes de binning ont également été utilisées pour le binning de STE (section (A.2.2), y compris la distance euclidienne [126], la divergence de Kullback-Leibler [127, 128] et la divergence de χ^2 [128, 129]. Toutes ces méthodes sont des techniques de binning basées sur les réponses.

Après chaque tri, le modèle enregistre les emplacements de tri en tant que règle et ces règles de tri sont utilisées pour classer les cas non étiquetés. Le binning basé sur la réponse est efficace pour les variables nominales, mais problématique pour les variables continues, d'intervalle ou ordinales. Dans ces cas, l'exigence de sous-groupes à réponse homogène risque de perdre des informations précieuses, augmentant efficacement le bruit [130, 131, 132, 121].

Solution proposée

Nous proposons un regroupement basé sur le voisinage (STE-NBB) pour atténuer les limitations de regroupement fondées sur la réponse décrites ci-dessus. Le binning basé sur le voisinage suppose que les cas à proximité ont des effets de traitement similaires, par conséquent leurs scores de modèle ne devraient pas varier. Par conséquent, les informations de voisinage peuvent être utilisées pour définir une corbeille, et les valeurs $\widehat{PTE}$ de membre de corbeille sont les $\widehat{STE}$ de leur corbeille.

Nous utilisons des fenêtres coulissantes (SW)[133] pour appliquer le binning basé sur le voisinage STE. La fenêtre glissante limite le nombre de membres dans une fenêtre, garantissant ainsi les descriptions de membre les plus appropriées. Ainsi, deux paramètres principaux sont requis: la taille de la fenêtre et du pas. La taille de la fenêtre spécifie le nombre de membres que chaque fenêtre peut inclure, et la taille de l'étape spécifie le nombre de membres que chaque fenêtre doit ignorer pour démarrer un nouveau groupe de fenêtres.

Notez que l'algorithme SW convertit la variable continue en une sous-population ordinale (variable catégorielle ordonnée). Par conséquent, il s'agit d'une solution

appropriée au problème des STEM (en raison du problème fondamental de l'inférence causale, les modèles STE nécessitent des données de sous-population et non des données pour des individus uniques).

Partant de l'hypothèse que les cas proches forment un sous-groupe plus homogène, le principal avantage de l'utilisation de la technique de la fenêtre glissante pour STEM est l'accent mis sur l'optimisation de l'homogénéité des sous-populations. Le logiciel définit une zone fixe autour de chaque membre (la taille de la fenêtre), ce qui aidera à assurer la localisation du score estimé. D'autre part, en prenant en compte les valeurs entourant chaque membre, le logiciel maximisera la généralité du modèle STE et minimisera l'effet des valeurs erronées (ou aberrantes). Ces caractéristiques du logiciel SW permettent de lisser les estimations du modèle STE et d'assurer l'homogénéité des sous-populations.

Le processus d'application du logiciel dans STEM pose deux problèmes principaux.

— Le premier problème est l'accumulation de bruit provenant d'autres fonctionnalités (fonctionnalités non utilisées par l'algorithme SW).

— Le deuxième problème est l'évaluation (mesure de l'erreur) par le STEM.

Pour surmonter les obstacles mentionnés ci-dessus, nous utilisons une fonction de reconstruction pour estimer le STE de chaque membre de la population.

Étant donné le $SW = \{w_1, w_2, \dots w_{\frac{m-n}{k}+1}\}$, un ensemble de toutes les fenêtres. La fonction de glissement $f(w_j)$, où $j \in [1 \dots \frac{m-n}{k} + 1]$. On note v_i un vecteur contenant les résultats de la fonction glissante $f(w_j)$ appliquée à chaque fenêtre w_j dans le sous-ensemble de fenêtres incluant le membre i.

Nous définissons la variable de fenêtre glissante d'effet de traitement de sous-population $STE - NBB_i$ pour le membre i comme suit:

$$STE - NBB_i = g(v_i) \tag{A.3}$$

Où $g(v_i)$ est la fonction de reconstruction du membre i.

Nous notons que la fonction $g(v_i)$ peut être définie sous différentes formes, en fonction du problème de recherche et de la mesure requise. Par exemple, $g(v_i)$ pourrait être une somme, un produit ou une fonction moyenne. De plus, notez que la fonction $f(w_i)$ suit la même forme que la fonction de fenêtre coulissante conventionnelle.

Le résultat de STE-NBB nous permet de contourner les problèmes du SW conventionnel. Nous pouvons utiliser le résultat de notre approche pour dessiner la courbe de *Qini*. De plus, le bruit des autres fonctionnalités n'est pas agrégé.

A.3.4 Effet de traitement de sous-population des forêts aléatoires de voisinage

La motivation fondamentale de STE-NRF est de tirer parti du binning basé sur le voisinage pour augmenter la précision du modèle STE. La technique STE-NRF

proposée combine le binning basé sur le voisinage et les arbres de régression classiques [138] dans un cadre de forêts aléatoires [63]. Tout d'abord, les membres sont triés en fonction de leurs valeurs de fonctionnalité. Les membres ayant des valeurs de fonctionnalité similaires sont joints en un seul membre et la variable cible est la moyenne de leurs valeurs de variable ciblée. La variable cible pour un traitement binaire peut être exprimée par

$$\widehat{STE_i} = \frac{\text{sum(Response given treatment)}}{\text{sum(Treatment)}} - \frac{\text{sum(Response given control)}}{\text{sum(Control)}}. \quad \text{(A.4)}$$

La taille de la fenêtre est ensuite définie en fonction de la variance variable; une faible variance nécessite une taille de fenêtre courte et inversement. La taille des étapes est définie en fonction de la taille de la population et de la capacité de calcul.

Pour créer des forêts aléatoires de voisinage STE (STE-NRF), nous construisons de petits arbres de décision de régression dans chaque fenêtre en tant que modèle de prédiction pour $\widehat{STE}$. Nous sélectionnons au hasard 50% des caractéristiques comme prédicteurs et 50% de la population pour créer chaque arbre de décision. Ces configurations minimisent les surajustements et améliorent la généralisation du modèle final. Enfin, des arbres de décision collectifs (forêts aléatoires) sont utilisés pour la prédiction, les $\widehat{STE}$ prédits de chaque membre étant le score de prédiction moyen de tous les arbres incluant ce membre.

Nous estimons le membre $\widehat{STE}$ comme la moyenne de tous les arbres $\widehat{STE_{tree}}$,

$$STE - NRF_i = \frac{\sum_{k=1}^{n} E_{k_i}(STE)}{n}, \quad \text{(A.5)}$$

où n est le nombre d'arbres et $E_{k_i}(STE)$ l'arborescence k estimation du membre i.

Nous évaluons les performances proposées de STE-NRF en utilisant des jeux de données simulés et réels. Les expériences simulées permettent de comparer les performances proposées de STE-NRF avec PTE pour chaque membre par rapport aux approches actuelles.

Etude de simulation et évaluation expérimentale

Nous avons utilisé le protocole d'expérimentation développé par [11] dans les simulations pour comparer les performances proposées de STE-NRF et de la méthode STE actuelle. Pour la reproductibilité, nous avons adopté les paramètres de simulation proposés par [17], en faisant varier l'intensité de l'effet principal, la corrélation entre les covariables (caractéristiques) et la magnitude du bruit dans le jeu de données.

Le jeu de données d'apprentissage contient 200 lignes avec 1 000 lignes dans le jeu de données de validation. Chaque ensemble de données contenait des colonnes de traitement binaire et de réponse binaire. Les jeux de données de simulation contenaient également vingt entités, $X1, X2, \cdots, X20$, mais seuls les $X1 - X4$ sont affectés par le STE, les $X3$–$X10$ étant affectés par l'effet de traitement principal. Les variables $X10$–$X20$ ne sont pas affectées par les effets principaux ou STE; leur

objectif est d'ajouter du bruit à l'ensemble de données et de renforcer le processus de modélisation en établissant une corrélation avec d'autres variables. Plus de détails sont fournis dans [17].

Nous avons construit STE-NRF en utilisant 50% de la taille de l'échantillon de population, 50% de la taille de l'échantillon de caractéristiques et 50 itérations, avec une taille de pas = 10 et une taille de fenêtre = 30 membres. Nous avons comparé le STE-NRF proposé avec les techniques de modélisation STE suivantes.

- Deux modèles de forêts aléatoires (Two-Models-RF). Nous avons créé deux modèles de forêts aléatoires utilisant 500 arbres, où chaque modèle prédit la réponse pour un traitement spécifique. Nous avons ensuite estimé les STE en calculant les différences de score entre chaque modèle de prédiction.
- Combinaison de forêts aléatoires (Comb-RF). Cette technique simplifie le problème de multi-classification STE en combinant différentes étiquettes en deux classes principales et convertit le problème en un problème de classification binaire. Plus d'informations sur cette approche sont disponibles sur [56, 9]. Pour ces expériences, nous avons utilisé des forêts aléatoires composées de 500 arbres.
- Forêts d'inférence conditionnelle causale (CCIF). Il s'agit d'une méthode améliorée de forêts aléatoires à soulèvement utilisant de meilleurs critères d'élagage et d'arrêt pour assurer la signification statistique de chaque arbre du modèle. Pour ces expériences, nous avons utilisé un modèle CCIF composé de 500 arbres et des critères de scission basés sur l'interaction.
- Forêts aléatoires à distance euclidienne (ED-RF). [115, 62] Cette approche a modifié les forêts aléatoires de soulèvement en modifiant le critère de fractionnement lors de la construction de l'arborescence afin de calculer la distance euclidienne au carré entre les nœuds fractionné et d'origine. Pour ces expériences, nous avons utilisé des modèles ED-RF composés de 500 arbres.
- Forêts causales (FC). Cette approche a utilisé des arbres de causalité développés par [22] dans un cadre de régression forestière. Pour les expériences, nous avons utilisé un modèle CF composé de 500 arbres avec une fraction d'échantillon de 50%.

Pour garantir la fiabilité, nous avons répété chaque expérience 100 fois pour chaque scénario de simulation. Les performances du modèle STE ont été évaluées à l'aide des deux coefficients de corrélation de rang de Qini [21] et de Spearman entre les effets de traitement réels et estimés de chaque modèle. Le paquet uplift R [14] a été utilisé pour produire les jeux de données simulés et calculer le coefficient Qini. Les résultats de la simulation ont confirmé que l'approche STE-NRF proposée fournissait un modèle robuste au bruit qui surpassait toutes les autres approches STE envisagées pour tous les scénarios testés.

Expériences utilisant des paramètres de jeux de données réels

Nous avons mené quatre expériences en utilisant trois ensembles de données réels. Chaque ensemble de données contenait des variables de traitement et de réponse binaires binaires et avait été choisi dans les secteurs de la santé et du marketing.

— Jeu de données de visite Hillstrom: le jeu de données de visite Hillstrom [140] contenait les données d'une campagne de marketing par courrier électronique comprenant 64 000 clients uniques répartis en deux groupes: le groupe de traitement a reçu un courrier électronique, contrairement au groupe de contrôle. L'ensemble de données a pris en compte deux types de courrier électronique, basés sur le produit annoncé. Nous voulions uniquement déterminer l'effet global de la réception d'un courrier électronique. Nous avons donc combiné les types de courrier électronique dans un groupe et comparé au groupe de contrôle. L'ensemble de données comprenait également 9 fonctionnalités décrivant les informations historiques et démographiques du client.

— Ensemble de données sur le cathétérisme cardiaque droit: l'ensemble de données sur le cathétérisme cardiaque droit (RHC) comprenait des informations sur 5735 patients admis dans un hôpital [141] et comprenait deux groupes. Le groupe 1 (2184 patients) a été soumis à un cathétérisme cardiaque droit, en fonction de son état, et le groupe 2 (3551 patients) n'a pas subi de cathétérisme cardiaque droit. Nous avions l'habitude `Right Heart Catheterization` de représenter la variable de traitement et `Death` la variable de classe. La période de surveillance était de 180 jours et la survie du patient pendant 180 jours après l'admission était considérée comme un résultat positif. Toutes les fonctionnalités liées au décès du patient ont été supprimées et les 58 fonctionnalités restantes ont été utilisées pour créer les modèles STE. L'ensemble de données a été séparé en formations (80%) et tests (20%).

— Ensemble de données sur la greffe de moelle osseuse: l'ensemble de données sur la greffe de moelle osseuse (BMT) comprenait des informations sur cent patients ayant reçu une greffe de moelle osseuse [142] extraite à partir d'os pelvien ou de sang périphérique. L'ensemble de données a été regroupé dans des échantillons de traitement et de contrôle en fonction de la source de la moelle osseuse, l'os pelvien étant considéré comme le contrôle. L'ensemble de données contenait sept caractéristiques qui ont été utilisées pour construire les modèles STE. Nous avons utilisé le jeu de données BMT une fois pour prédire la survenue d'une maladie aiguë du greffon contre l'hôte (agvh) et à nouveau pour prévoir la survenue d'une maladie chronique du greffon contre l'hôte (cgvh).

L'algorithme STE-NRF proposé permet de prédire avec succès l'EST pour les jeux de données réels, en particulier pour les 5% premiers de la population. le tableau A.1 montre que le modèle STE-NRF proposé a surpassé tous les autres modèles STE en termes de coefficient Qini.

Nous avons également évalué les méthodes STE sur la base du coefficient Qini de 15% [21], afin de montrer l'effet de la minimisation de la taille de la population. Nous pouvons réaliser un gain de STE supérieur de 10,52% en utilisant l'approche STE-NRF proposée.

Le modèle STE-NRF proposé présente une meilleure capacité de traitement des caractéristiques continues, reflétée par les scores d'évaluation STE, et la technique de la fenêtre glissante offre une meilleure correspondance pour minimiser le bruit de stockage et maximiser la prévisibilité moyenne.

Table A.1 – Experimentation outcomes for real datasets.

Dataset Name	Qini Coefficient (%)			15% Qini Coefficient (%)		
	STE-NRF	CCIF	CF	STE-NRF	CCIF	CF
Hillstrom-Visit	**10.20**	6.50	6.00	**10.52**	7.56	6.24
RHC	**51.20**	12.80	8.60	**60.89**	13.10	10.64
BMT-cgvh	**40.90**	0.40	31.40	**50.00**	-11.10	44.44
BMT-agvh	**25.30**	13.00	4.20	25.18	**28.57**	0.00

A.4 Modélisation des effets du traitement des perturbations et des sous-populations

La perturbation des données peut être définie comme tout facteur entraînant des résultats d'estimation moins précis. La perturbation des données STE peut être due à des données biaisées, au bruit, à une forte corrélation entre les caractéristiques, ou à un effet de traitement faible (STE) par rapport à l'effet de traitement moyen (ATE). Les études sur l'effet des perturbations sur les modèles STE ont montré que le rapport STE / ATE avait un effet majeur sur les résultats, suivi de la corrélation entre les caractéristiques, puis du bruit et du biais dans les données [36, 11, 21].

Les perturbations STE sont classées en tant que perturbations de traitement ou perturbations de réponse. Les perturbations du traitement sont les facteurs qui affectent la variation des cas de traitement et des cas témoins, ce qui entraîne une disparité dans l'attribution du traitement (par exemple, un biais). Les perturbations de la réponse, en revanche, sont tous des facteurs qui modifient la variation des répondants et des non-répondants (par exemple, le bruit). Par exemple, dans une expérience de traitement et de réponse binaire, nous avons les quatre quadrants (voir la figure A.1). L'état idéal pour assigner un traitement à l'expérience est que les cas de traitement soient égaux aux cas témoins (TR + TNR) = (CR + CNR). Cependant, des perturbations du traitement perturberaient cet état parfait et affecteraient négativement les résultats.

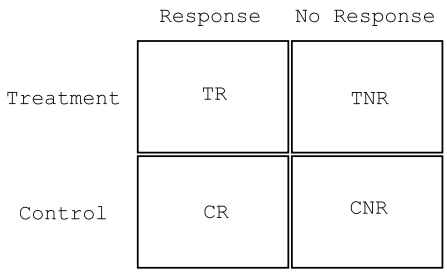


Figure A.1 – Quatre quadrants de l'expérience d'effet de traitement de sous-population.

Dans toute STEM, nous souhaitons distinguer l'effet de traitement de sous-population (STE) de l'effet de traitement principal (ATE). L'approche classique STEM modélise la variation du taux de réponse entre le traitement et le groupe témoin.

$$STE = \frac{\text{number of cases in TR}}{\text{number of cases in T}} - \frac{\text{number of cases in CR}}{\text{number of cases in C}}$$

L'approche STEM classique modélise la perturbation de la réponse en maximisant TR et en minimisant CR.

A.4.1 Approche proposée de la modélisation équilibrée du soulèvement réflexif

Nous construisons d'abord ce que nous appelons un modèle de soulèvement réfléchissant afin de réduire la sensibilité aux perturbations. L'objectif est de gérer la variation de l'effet du traitement plutôt que celle de la réponse. Le modèle de soulèvement réfléchissant estime la probabilité pour un cas spécifique de se trouver dans la zone de traitement, compte tenu du fait qu'il a déjà répondu. Ce faisant, nous minimisons les effets des perturbations de la réponse.

$$RU = \frac{TR}{TR+CR} - \frac{TNR}{TNR+CNR}$$

Étant donné que nous avons un environnement d'expérience parfait ($T = C$), nous pouvons distinguer les trois cas suivants:

- $RU > STE$:
  - Si $TR > TNR$
  - Si $CNR > CR$
- $RU < STE$:
  - Si $TR < TNR$
  - Si $CNR < CR$

— $RU = STE$:

— Si $TR = CNR$

RU est plus sensible aux changements mineurs autour de zéro STE, ce qui signifie qu'il est bénéfique pour un environnement à faible ratio STE / ATE. Cela fait du score RU un bon candidat pour soutenir le score STE.

L'augmentation de la sensibilité aux perturbations dues au traitement est un effet secondaire négatif de l'utilisation du soulèvement par réflexion, mais il pourrait être maintenu sous un seuil acceptable en utilisant un échantillonnage approprié, tel qu'un échantillonnage aléatoire stratifié [14].

Pour construire un modèle de soulèvement réfléchissant, nous utilisons la méthode des arbres de décision d'ensemble telle que définie par [62] pour construire chaque modèle afin de minimiser les taux d'erreur de surajustement et de classement erroné. Pour le modèle STE, nous avons utilisé l'approche de transformation simple puis modèle proposée par [56] (voir A.2.2).

En utilisant l'EF parallèlement à la STE régulière, nous avons atteint un score équilibré (BRUM) qui serait plus fiable et plus robuste.

$$BRUM = 1/2 * (STE_i + Reflective\ Uplift_i) \tag{A.6}$$

A.4.2 Utilisation de la modélisation équilibrée du soulèvement réfléchi pour améliorer le taux de conversion d'une campagne de marketing par courrier électronique (SNCF)

Dans cette section, nous passons à de véritables ensembles de données concernant également les campagnes de marketing par courrier électronique. Nous appliquons le modèle BRUM à une campagne de marketing par courriel mise en œuvre par un site Web de réservation de voyages en ligne appartenant à une grande compagnie de chemin de fer française. Outre la vente en ligne de billets de train, ce site Web offre la possibilité de faire des réservations d'hôtel, de louer des voitures et d'autres activités liées à l'organisation de voyages et de voyages, partout en Europe et dans le monde.

Une campagne de marketing par courrier électronique a été menée pour déterminer l'effet d'un nouveau type de conception de courrier électronique sur les clients: personnalisé par rapport au standard. La nouvelle conception comprend un message promotionnel personnalisé basé sur des données historiques liées à l'activité antérieure de chaque client, tandis que la conception standard ne varie pas entre les clients. La société a décidé de faire un test A / B pour mesurer l'influence du nouveau design. Les tests A / B, ou expérimentations contrôlées, sont largement utilisés par les entreprises en ligne (Facebook, Amazon, Groupon, etc.) au cours du processus de développement d'un produit ou pour tester les effets d'un nouveau produit [143]. La société a décidé de définir le traitement A comme l'envoi du courrier électronique promotionnel standard et le traitement B comme l'envoi du courrier promotionnel personnalisé.

Dans la population cible (6 479 601), un courrier électronique promotionnel standard (qui recommande de cliquer sur une annonce promotionnelle en ligne disponible sur le site Web) a été envoyé à 3 398 669 clients (représentant 52,45% de la population) qui ont ainsi été soumis au traitement A l'email standard). La sous-population restante (47,55%) a reçu le courrier électronique promotionnel personnalisé (traitement B). Les effets de ces traitements sur le comportement de la population (en termes d'ouverture du courrier électronique et de clic sur la publicité) ont été observés et enregistrés au cours d'une période de surveillance de l'activité du site web du 07/09/2015 00:00:00 au 23 / 09/2015 23:59:59. Au cours de cette période, 458 068 commandes (achats) ont été enregistrées auprès de clients ayant fait l'objet d'un traitement A ou d'un traitement B.

Nous avons une base de données de 6 479 601 lignes (clients) qui ont été soumis au traitement A ou au traitement B. Les informations fournies sur chaque client sont définies par 18 variables différentes. Parmi ces variables, certaines souffrent de données manquantes, comme *ClientAge* et *ZipCode*, qui ont respectivement 52% et 82% de données manquantes. 18,14% de la population (1 221 607 personnes) ont ouvert le courrier électronique au cours de la période de surveillance, 46,68% d'entre elles étaient dans le groupe A, contre 53,31% dans le groupe B.

STEM nécessite une variable de réponse explicite en plus de la variable de traitement. Par conséquent, il est important, avec les procédures de préparation de données standard, de déterminer la description spécifique de l'événement qui doit être signalé en tant que réponse. La société a surveillé trois événements principaux liés au client (ouverture du courrier électronique, clic sur la publicité et achat). Chaque événement est enregistré au format aaaa-MM-jj HH: mm: ss.

1. Click vs. No-Click
2. Cliquez sur Achat vs Cliquez sur Aucun achat
3. Cliquez sur Achat dans les 40 jours par rapport à Cliquez sans achat.

Nous avons généré six nouvelles variables pour appuyer l'analyse. Commençant par trois variables binaires: Ouverture (l'email), en cliquant sur (sur la publicité) et sur les achats (ordre). Nous avons ensuite mesuré l'occurrence de chaque action (0,1). Une variable, nommée OpenToClickHour, est également créée pour afficher le laps de temps (en heures) entre l'ouverture du courrier électronique et le clic sur la publicité. Enfin, deux variables sont ajoutées pour mesurer l'écart de temps (en jours) entre l'ouverture et l'achat; et l'intervalle de temps entre les actions Clic et Achat.

Pour nettoyer les données, nous avons remplacé les valeurs manquantes de la variable *ClientAge* par le nombre (-1) et les valeurs manquantes des variables *ZipCode*, *ProductType*, *Origin* et *Destination* par la chaîne "N / A". Enfin, nous avons regroupé les colonnes **OrdersNumber** et *OrdersValue* en dix zones.

Pour chaque expérience, nous comparons notre technique STEM (BRUM) à la forêt d'inférence conditionnelle causale (CCIF). CCIF est considéré comme l'une des techniques STEM les plus performantes utilisées jusqu'à présent [17]. De plus, nous avons comparé BRUM avec une version améliorée de la méthode de Lai (GB-Lai) et

une version améliorée de la technique standard à deux modèles (modèles GB-Two). GB est une abréviation de la méthode des arbres de classification à gradient renforcé. La technique des arbres GB est utilisée pour la construction des modèles [144].

Pour faire les expériences, nous avons partitionné les données en 80/20, ensemble apprentissage / validation. L'ensemble d'apprentissage est utilisé pour former chaque modèle et l'ensemble de validation est utilisé uniquement pour la validation. Un obstacle majeur à l'évaluation des approches STEM est dû à ce que l'on appelle «le problème fondamental de l'inférence causale» [41], qui stipule que nous ne pouvons observer qu'un résultat chez un individu après avoir été soumis à un traitement spécifique, et que nous ne pouvons pas observer les résultats d'autres traitements. en même temps. En d'autres termes, nous ne disposons pas des autres réponses vraies de l'individu (s'il avait été exposé aux autres traitements) pour calculer le véritable score STE (score True STE = Réponse observée après le premier traitement - Réponse observée après le deuxième traitement). Suivant la littérature [17, 6, 93], nous avons divisé la population en sous-groupes (bacs) pour résoudre ce problème. Ce faisant, nous mesurons le STE agrégé d'un sous-groupe spécifique et non l'effet de traitement individuel lui-même. Après avoir trié la population par ordre décroissant en fonction de leur score *STE estimé*, nous avons divisé la population en dix groupes (déciles) avec des fréquences égales, où Bin 1 a la plus haute moyenne prédite STE et Bin 10 a la plus faible moyenne prédite STE.

Nous avons utilisé trois critères de mesure pour comparer différentes techniques. Tout d'abord, nous avons utilisé le *coefficient Qini*, qui correspond au rapport entre l'aire sous la courbe Qini du modèle et l'aire sous la courbe Qini optimale. Une valeur de coefficient Qini plus élevée reflète une meilleure capacité à trier les cas d'une STE supérieure à une STE inférieure [21]. De plus, nous avons utilisé une nouvelle mesure introduite par [108], qui est la χ^2_{net}; cette mesure est meilleure lorsqu'elle est associée au score du coefficient Qini, afin de garantir que nos résultats ne sont pas obtenus par accident. χ^2_{net} mesure la différence significative entre le STE du sommet et le STE du dernier décile (bin) de chaque ensemble de validation, un supérieur χ^2_{net} avec un $p - value$ inférieur reflète une meilleure séparation et donc un meilleur modèle STE. De même, nous avons comparé différentes méthodes STEM en calculant la STE attendue (ESTE). ESTE est la STE que nous pouvons obtenir si nous ne ciblons que le premier décile avec le traitement B (treatment = 1) et le dernier décile avec le traitement A (treatment = 0). En particulier, le STE attendu mesure la stratégie de ciblage personnalisé maximum (rappelez-vous que le premier décile contient des cas susceptibles de favoriser le traitement B par rapport à A et inversement pour le dernier décile). L'ESTE est utile pour obtenir une estimation approximative de l'augmentation maximale de l'EST que nous obtiendrons si nous utilisions la méthode spécifiée.

Les techniques de BRUM montrent de meilleures capacités de prédiction par rapport aux autres méthodes dans toutes les expériences. Dans la deuxième expérience, par exemple, nous excluons d'abord de notre échantillon les clients qui n'ont pas ouvert le courrier électronique, ni cliqué sur la publicité. Deuxièmement, nous posons l'événement "Achat" comme une réponse positive au traitement (A ou B). Le tableau A.2 montre la tabulation croisée des variables Traitement et Clic Achat dans le jeu

de validation.

Jeu de validation de l'expérience 2		**achat = 0**	**achat = 1**	**Total**
Treatment = A	Compter	10,364	8,304	18,668
	Row %= Nombre de cellules / Nombre total de lignes	55.52%	**44.48%**	100%
Treatment = B	Compter	9,424	8,290	17,714
	Row %= Nombre de cellules / Nombre total de lignes	53.21%	**46.79%**	100%
Total		19,788	16,594	36,382

Table A.2 – Tableau croisé du jeu de validation de l'expérience 2

Le STE aléatoire de l'ensemble de validation de la deuxième expérience est 46,79% - 44,48% = 2,31% = + 0,0231, ce qui implique que le traitement B est plus efficace que le traitement A.

Méthode Nom	**Coefficient Qini**	χ^2_{net}	χ^2_{net} $p-value$	**STE Aléatoire**	**STE attendu**
BRUM	**6.3%**	**56.653**	0	2.31%	**8.635%**
CCIF	4.2%	14.065	0		4.389%
GB Lai's	2.2%	2.308	0.129		1.309%
GB Deux modèless	1.6%	0.014	0.906		0.050%

Table A.3 – Table de résultats de l'expérience 2

On peut voir sur le tableau A.3 que le modèle BRUM a permis de prédire le STE des sous-populations. De plus, nous remarquons, d'après χ^2_{net} et χ^2_{net} $p-value$, que le modèle BRUM a été en mesure de faire la différence entre les clients qui ont acheté après le traitement A et les clients qui ont acheté après le traitement B. Le STE attendu pour l'utilisation du modèle BRUM est trois fois le score STE ciblé au hasard; Cela signifie qu'en utilisant l'approche BRUM, nous pourrions augmenter le STE (taux d'achat) aléatoire de 2,3% à un maximum de 8,6%.

Ensemble de données sur le cancer du sein

Le cancer du sein est une cause majeure de décès pour des millions de femmes dans le monde. Bien que les traitements médicaux existent et deviennent extrêmement efficaces, la question de savoir à quel moment le diagnostic est établi et comment prédire les récidives reste critique et affecte considérablement les statistiques de rétablissement. Les approches STEM ont été utilisées pour prédire les événements de récurrence personnelle du cancer du sein [17, 19, 65]. Pour appliquer la méthode BRUM et comparer ses performances à d'autres méthodes STEM existantes, nous avons utilisé des données réelles relatives à cette maladie. Le jeu de données est disponible dans le référentiel UCI. L'ensemble de données peut être trouvé à l'URL suivante: (https://archive.ics.uci.edu/ml/datasets/ Breast + Cancer). ; et est fourni par Matjaz Zwitter et Milan Soklic au Centre médical universitaire de l'Institut d'oncologie de Ljubljana, en Yougoslavie. Il est principalement utilisé en tant qu'ensemble d'apprentissage pour les algorithmes de classification. L'ensemble de données représente les événements de récurrence du cancer du sein. Il contient 286 instances de données de patients réels divisées en deux classes, les événements de récurrence et aucun événement de récurrence. Chaque cas est décrit par neuf attributs différents séquentiellement; chaque attribut a été converti en un attribut nominal [145].

Les attributs sont les suivants: *Âge*: au dernier anniversaire du patient au moment du diagnostic; *Ménopause*: que la patiente soit pré ou post-ménopausée, au moment du diagnostic; *Taille de la tumeur*: le plus grand diamètre (en mm) de la tumeur existante; *Inv-node*: nombre (entre 0 et 39) de ganglions axillaires contenant un cancer du sein métastatique visible à l'examen histologique; *Capsules de ganglions*: si le cancer est à une phase métastatique vers un ganglion lymphatique; *Tumeur maligne*: le grade historique (intervalle 1 - 3) de la tumeur; *Sein*: le cancer peut survenir dans l'un ou l'autre sein; *Quadrant de la poitrine*: la poitrine peut être divisée en quatre quadrants; et *irradiation*: protocole médical utilisant des rayons X à haute énergie pour détruire les cellules cancéreuses.

Après [9, 65], nous avons converti l'attribut ménopause en attribut de traitement et avons supprimé tous les attributs en corrélation avec le traitement en binôme ou ceux pour lesquels des valeurs étaient manquantes. Nous avons divisé le jeu de données sur le cancer du sein en 80/20 pour cent afin de créer des ensembles de formation et de validation, respectivement. Nous nous retrouvons avec 277 lignes après avoir éliminé les entrées avec des valeurs manquantes. Les attributs *âge* et *quadrant du sein* ont été supprimés. Nous avons utilisé des arbres de décision d'ensemble pour construire les modèles BRUM (modélisation équilibrée du soulèvement réfléchissant) et CCIF. L'une des préoccupations concernant les STEM est le manque de fiabilité des outils de mesure, principalement en raison de l'absence de la valeur réelle de l'effet du traitement [17].

Pour contourner cet obstacle, la comparaison segmentée sera appliquée pour calculer le taux d'erreur. Plus particulièrement, le coefficient Qini et le coefficient Qini à 15% sont utilisés comme scores pour comparer les performances des modèles, définies comme étant la surface entre la courbe des gains incrémentaux réels du

modèle ajusté et la surface sous la diagonale correspondant au ciblage aléatoire [21]. Comme indiqué dans le tableau A.4, les résultats montrent que l'algorithme BRUM a de meilleurs scores Qini et 15% Qini que CCIF. Les résultats signifient que les cas "traités et ayant répondu positivement" ont des niveaux plus élevés par rapport aux cas "traités mais n'ayant pas répondu". Par conséquent, nous fournissons la preuve, à l'aide d'un ensemble de données médicales réel, que BRUM est plus performant que les autres méthodes STE.

	Qini	**15% Qini**
BRUM	0.27	0.31
CCIF	0.12	-0.71

Table A.4 – Experiment résultats de l'expérience du cancer du sein

A.5 Conclusions

L'augmentation spectaculaire du nombre de données accessibles et la forte demande en techniques d'aide aux systèmes d'aide à la décision ont favorisé l'émergence du phénomène de l'apprentissage des données afin de fournir les outils essentiels pour maximiser la confiance en la décision, minimiser les effets secondaires et faciliter les processus de prise de décision. Dans de nombreuses circonstances, des décisions personnalisées doivent être prises en fonction des caractéristiques de chaque cas. La modélisation des effets du traitement des sous-populations (STEM) aide le processus de prise de décision et optimise les effets d'action particuliers.

STEM est une branche de l'apprentissage automatique impliquée dans la conception de modèles, utilisant le cadre inférentiel de résultats potentiels outcome [146], pour cartographier les fluctuations des effets du traitement sur l'ensemble de la population. STEM vise à comprendre comment les effets d'événements particuliers varient d'une sous-population à une autre et à modéliser la variance pour des estimations futures. La STEM est un outil efficace permettant de déterminer l'importance d'un traitement. Du point de vue de l'apprentissage statistique, l'objectif de STEM est extrêmement difficile, car le traitement optimal est a priori inconnu pour un ensemble de données d'entraînement donné [17].

La suite de ce chapitre est organisée comme suit. La section A.5.1 examine et résume les contributions de cette thèse. La section A.5.3 traite de certaines limites de notre travail actuel et propose des solutions alternatives. En conclusion, Section A.5.4 examine les axes de recherche futurs pouvant découler des travaux présentés dans cette thèse.

A.5.1 Synthèse des contributions

Comme indiqué dans l'introduction, les contributions de cette thèse proviennent de l'exploration du problème des STEM, proposant une nouvelle taxonomie pour

les approches STEM, introduisant la technique de binning de voisinage STE (STE-NBB), la nouvelle approche STEM (STE-Strees), STE Neigborgood random forêts (STE-NRF) et l'approche de modélisation équilibrée du soulèvement par réflexion (BRUM).

Enquête de modélisation sur les effets du traitement des sous-populations

Tout au long de la thèse, nous rencontrons des articles qui expliquent et utilisent STEM. Nous découvrons que STEM est utilisé dans de nombreux domaines sous des noms différents. En outre, il y avait différentes définitions pour STEM. C'est la raison pour laquelle nous avons réexaminé le problème STEM. Et comme la causalité est la base de la recherche en STEM, nous introduisons l'idée de STEM du point de vue causal. De plus, nous avons introduit une nouvelle taxonomie pour les approches STEM basée sur trois actions principales (split, model, transform). De plus, nous avons rassemblé tous les termes utilisés pour STEM dans la littérature dans une section, ce qui aiderait les futurs chercheurs en STEM. Enfin, nous répertorions toutes les mesures qui évaluent les performances des modèles STEM.

Technique de binning de voisinage par effet de traitement de sous-population

L'incertitude sur les données existe toujours en raison d'une inférence causale fondamentale [41], c'est-à-dire que seul un sous-ensemble de traitements peut être directement observé. Traditionnellement, STEM utilise des techniques de binning basées sur les réponses pour utiliser les données à des fins d'inférence. Nous proposons une nouvelle technique de binning, appelée technique de binning de voisinage par effet de traitement de sous-population (STE-NBB), qui ne repose pas sur la variable de réponse. Nos techniques pondèrent et répartissent les cas en fonction de leur cas de voisinage. Cela conduit à une meilleure partition, moins de biais et crée de meilleurs sous-groupes pour un meilleur score de prédiction.

Effet de traitement de sous-population glissant des arbres

Pour utiliser la technique de binning proposée (STE-NBB) dans un cadre STEM, nous avons utilisé les arbres de décision comme technique de base et avons créé les arbres à effet de glissement de traitement par sous-population (STE-STrees). Notre nouvelle approche peut être classée dans la catégorie TTM STEM A.2.2, car nous transformons la variable de réponse en un autre formulaire avant la modélisation.

Effet de traitement de sous-population des forêts aléatoires de voisinage

Nous avons amélioré STE-Strees pour gérer les biais et le bruit à l'aide de méthodes d'ensemble et créer des forêts aléatoires de voisinage avec effet de traitement de sous-population (STE-NRF). La méthode STE-NRF sélectionne un échantillon de la population et un échantillon de caractéristiques pour la construction de chaque

modèle arborescent dans les forêts. L'amélioration de STE-NRF est illustrée dans nos expériences, où STE-NRF surpasse les autres approches STEM dans tous les scénarios.

Modélisation équilibrée du soulèvement réfléchissant

Un autre problème majeur auquel STEM est confrontée est le bruit, les caractéristiques corrélées et les données d'erreur (à savoir, les perturbations). Nous proposons une nouvelle mesure pour STE, nous l'appelons le soulèvement réfléchissant (RU). Le soulèvement réfléchissant peut être utilisé en plus du STE traditionnel pour équilibrer le score en cas de perturbations importantes. Nous proposons la modélisation équilibrée par remontée réfléchie (BRUM) comme une approche STEM moins sensible au bruit et aux perturbations dans les données.

Expérimenter pour améliorer la prise de décision marketing (jeu de données SNCF)

Nous menons une expérience avec l'équipe marketing de la compagnie nationale des chemins de fer français (SNCF) pour améliorer la prise de décision concernant la campagne de publicité. En utilisant les données collectées via un test A\B par courrier électronique, nous avons appliqué les modèles STEM à trois cas différents, et nous avons montré que notre approche peut constituer une meilleure stratégie de ciblage pour maximiser le taux de réponse et minimiser les dépenses de marketing.

A.5.2 Applications potentielles

De nombreuses applications dans divers domaines peuvent exploiter la puissance de cette thèse. Nos techniques STEM sont bénéfiques lorsque les données incluent des variables continues. Les principales formes de ces applications potentielles sont les mêmes, soit pour maximiser le résultat d'une intervention spécifique, soit pour réduire au minimum les effets secondaires négatifs d'une intervention spécifique, ou les deux.

Dans le domaine marketing, pour maximiser le retour sur investissement d'une publicité par courrier électronique ou pour maximiser le taux de clics d'une promotion en ligne spécifique. Dans le domaine de la médecine, maximiser le taux de survie du patient d'un traitement spécifique ou minimiser les effets secondaires d'un traitement médical à risque (chimiothérapie, par exemple). Dans le domaine politique, maximiser les votes pour un candidat spécifique en se concentrant sur le groupe d'électeurs le plus persuasif.

A.5.3 Limites et améliorations

En ce qui concerne la synthèse ci-dessus des contributions de cette thèse, nous discutons ici de leurs limites et proposons des solutions pour améliorer notre travail.

En ce qui concerne les approches STE-STrees et STE-NRF, leurs avantages sont exclusivement axés sur l'amélioration du traitement des variables continues et ordinales avec trop de niveaux. Cependant, si l'ensemble de données ne contient pas ces types de variables, nos approches STEM proposées utiliseront une technique de binning basée sur la distance euclidienne plutôt que sur STE-NBB, similaire à la technique de forêts aléatoires STE utilisée par [17, 62].

De plus, nos approches proposées ne prennent pas en compte la distribution de la variable continue, ce qui les rend vulnérables aux erreurs et aux données biaisées, telles que les valeurs aberrantes (cas extrêmes) dans les données. Ces problèmes seront examinés dans des études futures utilisant une taille de fenêtre dynamique.

Une autre limitation de notre approche est la charge de calcul requise pour le processus de modélisation. Elle peut être gérée en modifiant la taille de l'étape en fonction du nombre total de lignes dans le jeu de données.

Notre approche BRUM est soumise à d'autres limitations liées au domaine à l'étude, à savoir, par exemple, la prise en compte exclusive des traitements binaires.

La méthode CCIF, fournie dans le package uplift R, ne traite pas les données manquantes. Nous avons codé toutes les données manquantes sous forme de valeurs ou de catégories pour résoudre ce problème, mais nous aurions pu ajouter plus de bruit aux données en utilisant cette solution naïve pour gérer les données manquantes. De plus, en regroupant les variables numériques, nous avons minimisé le temps de modélisation, mais nous avons peut-être également introduit du bruit et des biais dans les données.

Il convient de noter que le *coefficient Qini* ne constitue pas à lui seul une mesure fiable de l'efficacité de la méthode STEM. Les chercheurs sont encouragés à utiliser d'autres méthodes telles que χ^2_{net}, *STE attendu* ou le *retour sur investissement attendu* pour comparer différentes méthodes STEM.

A.5.4 Travaux futurs et perspectives

Dans les travaux futurs, il serait intéressant d'examiner la répartition du traitement en continu. Cela peut être fait en ajoutant une étape avant chaque binning pour détecter la distribution de la fonctionnalité sélectionnée. Ensuite, la taille de la fenêtre et du pas peut être ajustée pour correspondre à la distribution requise. Il est intéressant de comparer les performances du pas dynamique et de la taille de la fenêtre avec celles fixes.

De plus, il est possible d'améliorer le binning basé sur le voisinage proposé en le combinant avec la technique d'arborescence causale proposée par [22]. L'idée de séparer le processus de tri et le processus d'évaluation peut minimiser le surajustement et garantir de meilleures prévisions.

Notre approche BRUM se concentre exclusivement sur les traitements binaires. Il serait intéressant d'examiner nos approches dans un contexte de traitement multi-catégories ou continu. En effet, le cas des stratégies de traitement adaptatif, jugées efficaces dans le secteur médical et selon lesquelles le traitement typé pourrait être

modifié en fonction de la réponse en cours du patient [147, 148, 149] doit également être traité avec des techniques STEM optimisées. Davantage de recherche est certainement nécessaire pour traiter ce cas et de nombreux autres au sein de ce que nous pouvons appeler une analyse de perspective utilisant des techniques d'apprentissage automatique.

Enfin, nous avons commencé à explorer une application pour l'approche STEM afin de détecter le comportement psychologique des clients en ligne. Cela conduit à un autre domaine d'exploration intéressant, à savoir STEM pour les données de flux. La solution ultime utilisera l'approche de la diffusion en continu STEM pour détecter le changement de comportement des clients en ligne et pour obtenir des informations futures sur cette base.

Bibliography

[1] NJ Radcliffe and PD Surry. Differential response analysis: Modeling true response by isolating the effect of a single action. *Credit Scoring and Credit Control VI. Edinburgh, Scotland*, 1999. xviii, 35, 37, 44, 109

[2] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, NY, USA:, 2012. 1, 63, 83, 101

[3] Nicholas J Radclifte and Rob Simpson. Identifying who can be saved and who will be driven away by retention activity. *Journal of Telecommunications Management*, 1(2), 2008. 1, 102

[4] Victor SY Lo. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86, 2002. 1, 39, 46, 84, 102, 110

[5] Behram Hansotia and Brad Rukstales. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35–46, 2002. 1, 33, 34, 35, 45, 102, 108, 110

[6] Kathleen Kane, Victor SY Lo, and Jane Zheng. Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, 2(4):218–238, 2014. 1, 2, 33, 34, 42, 46, 84, 88, 102, 108, 109, 112, 125

[7] Lukasz Zaniewicz and Szymon Jaroszewicz. Support vector machines for uplift modeling. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 131–138. IEEE, 2013. 1, 34, 42, 47, 102, 108, 109, 112

[8] Farrokh Alemi, Harold Erdman, Igor Griva, and Charles H Evans. Improved statistical methods are needed to advance personalized medicine. *The open translational medicine journal*, 1:16, 2009. 2, 11, 39, 46, 102, 111

[9] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, 2012. 2, 34, 42, 47, 71, 94, 102, 109, 112, 119, 127

[10] Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. *arXiv preprint arXiv:1701.05306*, 2017. 2, 45, 102

[11] Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates 2. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014. 2, 35, 45, 63, 70, 79, 83, 102, 109, 118, 121

[12] Finn C Kuusisto. *Machine Learning for Medical Decision Support and Individualized Treatment Assignment*. phdthesis, The University of Wisconsin-Madison, 2015. 2, 11, 12, 45, 102

[13] Szymon Jaroszewicz and Piotr Rzepakowski. Uplift modeling with survival data. 2014. 2, 11, 47, 102

[14] Leo Guelman, Montserrat Guillen, Pérez-Marín , Ana M., et al. Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study. Technical report, 2014. 2, 35, 45, 46, 53, 64, 71, 82, 102, 110, 115, 119, 123

[15] Behram J Hansotia and Bradley Rukstales. Direct marketing for multichannel retailers: Issues, challenges and solutions. *Journal of Database Marketing & Customer Strategy Management*, 9(3):259–266, 2002. 2, 102

[16] David Jingjun Xu, Stephen Shaoyi Liao, and Qiudan Li. Combining empirical experimentation and modeling techniques: A design research approach for personalized mobile advertising applications. *Decision Support Systems*, 44(3):710–724, February 2008. 2, 102

[17] Leo Guelman, Montserrat Guillén, and Pérez-Marín , Ana M. A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems*, 72:24–32, 2015. 2, 13, 35, 38, 45, 53, 63, 64, 70, 71, 83, 84, 88, 93, 95, 97, 99, 102, 110, 115, 118, 119, 124, 125, 127, 128, 131

[18] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999. 2, 13, 102

[19] Kosuke Imai, Marc Ratkovic, and others. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013. 2, 12, 13, 34, 44, 45, 46, 93, 102, 109, 127

[20] Pierre Gutierrez and Jean-Yves Gérardy. Causal Inference and Uplift Modelling: A Review of the Literature. In *International Conference on Predictive Applications and APIs*, pages 1–13, 2017. 2, 102

[21] Nicholas J Radcliffe and Patrick D Surry. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, 2011. 2, 34, 38, 47, 53, 57, 64, 65, 71, 75, 79, 89, 95, 102, 108, 115, 119, 120, 121, 125, 128

[22] Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050:5, 2015. 2, 34, 35, 44, 45, 71, 100, 102, 108, 109, 119, 131

[23] Francis MacDonald Cornford. *Plato's cosmology: the Timaeus of Plato*. Routledge, 2014. 8

[24] Terence Irwin and Gail Fine. Aristotle: selections, translated with introduction, notes, and glossary. *Hackett, Indianapolis Google Scholar*, 1995. 8

[25] Menno Hulswit. Causality and causation: The inadequacy of the received view. *SEED*, 2:3–23, 2004. 8

[26] Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012. 8, 104

[27] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015. 8, 44, 104

[28] Ann A Lazar, Bernard F Cole, Marco Bonetti, and Richard D Gelber. Evaluation of treatment-effect heterogeneity using biomarkers measured on a continuous scale: subpopulation treatment effect pattern plot. *Journal of Clinical Oncology*, 28(29):4539, 2010. 8, 45, 104

[29] CH Liu and Benjamin Paul Chamberlain. Online controlled experiments for personalised e-commerce strategies: Design, challenges, and pitfalls. *arXiv preprint arXiv:1803.06258*, 2018. 9

[30] Suresh Vittal, Chritine Spivey Overby, and Emily Bowen. Optimizing Customer Retention Programs. *Available online at http : //www.pbinsight.com/assets_microsite/resources/files/telenor-cs.pdf(accessed July*1, 2018). 10

[31] Case Study - U.S. Bank, 2012. Available online at http://www.pb.com/docs/US/Products-Services/Software/Analytics/MarketingAnalytics/PortraitUplift/PDFs/CAI-PortraitCaseStudyUSBank.pdf (accessed July 1, 2018). 10

[32] Case Study - T-Mobile Austria uses Portrait Customer Analytics to reduce customer churn and make significant savings in software and consultancy costs, 2011. Available online at http://www.pbinsight.com/assets_microsite/resources/files/tmobilecs.pdf (accessed July 1, 2018). 10

[33] Jyotirmay Nag. A Generic Uplift Modeling Framework to Calculate ROI - Application in Promotion Effectiveness - Predictive Analytics World. 2013. 10

[34] Houssam Nassif, Finn Kuusisto, Elizabeth S Burnside, David Page, Jude Shavlik, and Vitor Santos Costa. Score as you lift (SAYL): A statistical relational learning approach to uplift modeling. In *Machine Learning and Knowledge Discovery in Databases*, pages 595–611. Springer, 2013. 11, 34, 47, 108

[35] Swiss town set for universal basic income experiment, June 2018. https://www.thelocal.ch/20180606/swisstownsetforuniversal-basicincome-experimentrheinau. 12

[36] Leo Guelman and others. *Thesis - Optimal personalized treatment learning models with insurance applications*. phdthesis, Universitat de Barcelona, 2014. 12, 13, 35, 79, 84, 110, 121

[37] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, 10:141–158, 2009. 13, 46, 53, 115

[38] Xiaogang Su, Joseph Kang, Juanjuan Fan, Richard A Levine, and Xin Yan. Facilitating score and causal inference trees for large observational studies. *The*

Journal of Machine Learning Research, 13(1):2955–2994, 2012. 13, 39, 44, 49, 111

[39] Hernán MA and Robins JM. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming., 2018. 14, 17, 39, 105, 111

[40] Lawrence V Fulton, Francis A Mendez, Nathaniel D Bastian, and R Muzaffer Musal. Confusion between odds and probability, a pandemic? *Journal of Statistics Education*, 20(3), 2012. 15

[41] Paul W. Holland, Clark Glymour, and Clive Granger. STATISTICS AND CAUSAL INFERENCE*. *ETS Research Report Series*, 1985(2):i–72, 1985. 18, 50, 63, 88, 98, 106, 113, 125, 129

[42] K Thulasiraman and MNS Swamy. 5.7 acyclic directed graphs. *Graphs: Theory and Algorithms*, 118, 1992. 19

[43] Felix Elwert. Graphical causal models. In *Handbook of causal analysis for social research*, pages 245–273. Springer, 2013. 19, 45

[44] Tyler J VanderWeele, Miguel A Hernán, and James M Robins. Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology (Cambridge, Mass.)*, 19(5):720, 2008. 20

[45] Tyler J VanderWeele and James M Robins. Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, pages 561–568, 2007. 21, 107

[46] Tyler J VanderWeele. On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871, 2009. 21, 107

[47] Judea Pearl. Simpson's paradox, confounding, and collapibility. *Causality: models, reasoning and inference*, pages 173–200, 2000. 28

[48] Heather Shaw, David Ellis, Libby-Rae Kendrick, Richard Wiseman, et al. Individual differences between iphone and android smartphone users. British Psychological Society, 2016. 29

[49] Kim Larsen. Net lift models. Slides of a talk given at the M2009 – 12thAnnual SAS Data Mining Conference, October 26–27, Las Vegas, NV. 33, 46, 108

[50] Krzysztof Rudaś and Szymon Jaroszewicz. Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, pages 1–31, 2018. 33, 34, 108

[51] David Maxwell Chickering and David Heckerman. A decision theoretic approach to targeted advertising. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 82–88. Morgan Kaufmann Publishers Inc., 2000. 34, 45, 108

[52] James M Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412, 1994. 34, 108

[53] Stijn Vansteelandt and Els Goetghebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):817–835, 2003. 34, 108

[54] James Robins and Andrea Rotnitzky. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, 91(4):763–783, 2004. 34, 108

[55] Nada Lavrac and Saso Dzeroski. Inductive logic programming. In *WLP*, pages 146–160. Springer, 1994. 34, 108

[56] Lily Yi-Ting Lai. *Influential marketing: a new direct marketing strategy addressing the existence of voluntary buyers.* phdthesis, Citeseer, 2006. 34, 46, 64, 71, 83, 109, 119, 123

[57] Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. Support Vector Machines for Differential Prediction. In *Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer, 2014. 34, 35, 44, 46, 109

[58] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 34, 109

[59] Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012. 35, 46, 109

[60] Behram Hansotia and Brad Rukstales. *Research Council Journal 2001 (Incremental Value Modeling)*. 2001. 35, 45, 110

[61] Glacy Elizabeth Jacob and M Sunitha. Evaluation of Ensemble methods for uplift modeling. *International Journal of Research*, 2(11):351–356, 2015. 35, 110

[62] Michal Soltys, Szymon Jaroszewicz, and Piotr Rzepakowski. Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, pages 1–29, 2014. 35, 42, 47, 71, 83, 90, 99, 110, 112, 119, 123, 131

[63] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 35, 64, 68, 70, 110, 118

[64] Piotr Rzepakowski and Szymon Jaroszewicz. Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology*, pages 43–50, 2012. 35, 42, 64, 110, 112

[65] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 441–450. IEEE, 2010. 37, 47, 53, 93, 94, 110, 115, 127

[66] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967. 39, 111

[67] NJ Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Market J Direct Market Assoc Anal Council*, 1:14–21, 2007. 40, 111

[68] John Banslaben. Predictive modeling. *The Direct Marketing Handbook, McGraw-Hill, New York*, pages 620–636, 1992. 40, 111

[69] David J Hand. *Construction and assessment of classification rules*, volume 15. Wiley Chichester, 1997. 42, 112

[70] NJ Radcliffe. Generating Incremental Sales: Maximizing the Incremental Impact of Cross-Selling, Up-Selling and Deep-Selling Through Uplift Modelling. *Stochastic Solutions Limited*, 2007. 42, 47, 65, 112

[71] Houssam Nassif, Finn Kuusisto, Elizabeth S Burnside, and Jude Shavlik. Uplift Modeling with ROC: An SRL Case Study. *ILP 2013 Late Breaking Papers*, page 40, 2013. 42, 112

[72] Oscar Mesalles Naranjo. Testing a New Metric for Uplift Models. Master's thesis, The University of Edinburgh School of Mathematics, 2012. 42, 47, 113

[73] Arthur Lewbel. Selection model and conditional treatment effects, including endogenous regressors. Technical report, mimeo, Boston College, 2002. 44

[74] Avi Feller and Chris C Holmes. Beyond toplines: Heterogeneous treatment effects in randomized experiments. *Unpublished manuscript, Oxford University*, 2009. 44

[75] Justin Grimmer, Solomon Messing, and Sean J Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017. 44

[76] Sokbae Lee, Ryo Okui, and Yoon-Jae Whang. Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, 32(7):1207–1225, 2017. 44

[77] Houssam Nassif, Vıtor Santos Costa, Elizabeth S Burnside, and David Page. Relational differential prediction. In *Machine Learning and Knowledge Discovery in Databases*, pages 617–632. Springer, 2012. 44

[78] Herbert I Weisberg and Victor P Pontes. CAUSALYTICS, LLC. *Cadit modeling for estimating individual causal effects*, 2012. 45

[79] HI Weisberg and VP Pontes. *Cadit modeling for estimating individual causal effects*. 2012. 45

[80] Lu Tian, Ash Alizadeh, Andrew Gentles, and Robert Tibshirani. A simple method for detecting interactions between a treatment and a large number of covariates. *arXiv preprint arXiv:1212.2995*, 2012. 45

[81] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177 – 188, 1986. 45

[82] A. C. Davison. Treatment effect heterogeneity in paired data. *Biometrika*, 79(3):463–474, 1992. 45

[83] Joshua D. Angrist. Treatment effect heterogeneity in theory and practice*. *The Economic Journal*, 114(494):C52–C83. 45

[84] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008. 45

[85] Felix Elwert and Christopher Winship. Effect heterogeneity and bias in main-effects-only regression models. *Heuristics, probability and causality: A tribute to Judea Pearl*, pages 327–36, 2010. 45

[86] Justin Grimmer. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(1):80–83, 2015. 45

[87] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials revisited. *Contemporary clinical trials*, 45:139–145, 2015. 45

[88] Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *arXiv preprint arXiv:1510.04342*, 2015. 45

[89] Susan Athey, Julie Tibshirani, and Stefan Wager. Solving Heterogeneous Estimating Equations with Gradient Forests. *arXiv preprint arXiv:1610.01271*, 2016. 45

[90] Ann A Lazar, Marco Bonetti, Bernard F Cole, Wai-ki Yip, and Richard D Gelber. Identifying treatment effect heterogeneity in clinical trials using subpopulations of events: STEPP. *Clinical Trials*, 13(2):169–179, 2016. 45

[91] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017. 45, 64

[92] Doug Freud and Robert Cooley. *Approaches to Incremental Response Modeling.* 2009. 45

[93] Taiyeong Lee, Ruiwen Zhang, Xiangxiang Meng, and Laura Ryan. Incremental Response Modeling Using SAS® Enterprise Miner™. Technical report, SAS Institute Inc., 2013. 45, 88, 125

[94] Ryan Burton. *Analyzing Collection Effectiveness using Incremental Response Modeling.* 2014. 45

[95] Sankara Prasad Kondareddy, Shruti Agrawal, and Shishir Shekhar. Incremental Response Modeling Based on Segmentation Approach Using Uplift Decision Trees. In *Industrial Conference on Data Mining*, pages 54–63. Springer, 2016. 45

[96] Kevin Rosamont Prombo. Modélisation incrémentale par méthode bayésienne. 2016. 45

[97] Johannes AN Dorresteijn, Frank LJ Visseren, Paul M Ridker, Annemarie MJ Wassink, Nina P Paynter, Ewout W Steyerberg, Yolanda van der Graaf, and Nancy R Cook. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *Bmj*, 343:d5888, 2011. 45

[98] Graham Dunn, Richard Emsley, Hanhua Liu, and Sabine Landau. Integrating biomarker information within trials to evaluate treatment mechanisms and efficacy for personalised medicine. *Clinical Trials*, 10(5):709–719, 2013. 45, 46

[99] Florence Hiu-Ling Yong. *Thesis - Quantitative methods for stratified medicine.* phdthesis, 2015. 45, 46

[100] Laura Forastiere, Edoardo M Airoldi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *arXiv preprint arXiv:1609.06245*, 2016. 45

[101] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint arXiv:1606.03976*, 2016. 45

[102] Boriska Toth. *Targeted learning of individual effects and individualized treatments using an instrumental variable*. PhD thesis, UC Berkeley, 2016. 45

[103] Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011. 46

[104] Kim Larsen. Net lift models. Slides of a talk given at SAS A2010 analytics conference. 46

[105] Bruce Lund. Direct marketing profit model. In *Proceeding in midwest sas user group conference 2012*, 2012. 46

[106] R Kubiak. Net Lift Model for Effective Direct Marketing Campaigns at 1800flowers. com. In *SAS Global Forum*, 2012. 46

[107] René Michel, Igor Schnakenburg, and Tobias von Martens. Methods of variable pre-selection for net score modeling. *Journal of Research in Interactive Marketing*, 7(4):257–268, 2013. 46

[108] Rene Michel, Igor Schnakenburg, and Tobias von Martens. A modified χ^2-test for uplift models with applications in marketing performance measurement. *arXiv preprint arXiv:1401.7001*, 2014. 46, 47, 89, 125

[109] René Michel, René Michel, Igor Schnakenburg, Igor Schnakenburg, Tobias von Martens, and Tobias von Martens. Effective customer selection for marketing campaigns based on net scores. *Journal of Research in Interactive Marketing*, 11(1):2–15, 2017. 46

[110] Kathleen Kane, Victor SY Lo, Jane Zheng, and Alex Arias-Vargas. (PPT)True-Lift Modeling: Mining for the Most Truly Responsive Customers and Prospects. page 20, NewYork City, 2011. 46

[111] Portrait Software^{TM}. *Optimal Targeting through Uplift Modeling: Generating higher demand and increasing customer retention while reducing marketing costs.* 2006. 47

[112] Andy Littleton. Driving business decision for maximal impact through uplift modeling, 2011. SAS Forum Moscow. 47

[113] David P Hofmeyr. An Application of Genetic Algorithms to Uplift Modelling. Master's thesis, University of Edinburgh, 2011. 47

[114] E Siegel. Uplift Modeling: Predictive Analytics Can't Optimize Marketing Decisions Without It. *Prediction Impact white paper sponsored by Pitney Bowes Business Insight*, 2011. 47

[115] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012. 47, 53, 64, 71, 115, 119

[116] Victor SY Lo and Dessislava A Pachamanova. From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk. *Journal of Marketing Analytics*, 3(2):79–95, 2015. 47

[117] Szymon Jaroszewicz and Lukasz Zaniewicz. Székely Regularization for Uplift Modeling. In *Challenges in Computational Statistics and Data Mining*, pages 135–154. Springer, 2016. 47

[118] Szymon Jaroszewicz. *Uplift Modeling*. Springer, 2017. 47

[119] Yan Zhao, Xiao Fang, and David Simchi-Levi. Uplift Modeling with Multiple Treatments and General Response Types. *arXiv preprint arXiv:1705.08492*, 2017. 47

[120] Michał Sołtys and Szymon Jaroszewicz. Boosting algorithms for uplift modeling. *arXiv preprint arXiv:1807.07909*, 2018. 49

[121] Salvador Garcia, Julian Luengo, José Antonio Sáez, Victoria Lopez, and Francisco Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750, 2013. 53, 54, 115, 116

[122] Robert C MacCallum, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. On the practice of dichotomization of quantitative variables. *Psychological methods*, 7(1):19, 2002. 53, 115

[123] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25(1):127–141, 2006. 53, 115

[124] Oliver Kuss. The danger of dichotomizing continuous variables: A visualization. *Teaching Statistics*, 35(2):78–79, 2013. 53, 115

[125] Susan Athey, Raj Chetty, Guido Imbens, and Hyunseung Kang. Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index. *arXiv preprint arXiv:1603.09326*, 2016. 53, 115

[126] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 1999. 54, 116

[127] Kingo Kobayashi. *Mathematics of information and coding*, volume 203. American Mathematical Soc., 2007. 54, 116

[128] Imre Csiszár and Paul C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004. 54, 116

[129] Szymon Jaroszewicz and Dan A. Simovici. A general measure of rule interestingness. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 253–265. Springer, 2001. 54, 116

[130] Julie R. Irwin and Gary H. McClelland. Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, 40(3):366–371, 2003. 54, 116

[131] Gavan J. Fitzsimons. *Death to dichotomizing*. The University of Chicago Press, 2008. 54, 116

[132] O. Naggara, J. Raymond, F. Guilbert, D. Roy, A. Weill, and Douglas G. Altman. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, 32(3):437–440, 2011. 54, 116

[133] Vladimir Braverman. *Sliding Window Algorithms*, pages 1–6. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. 55, 116

[134] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, 2011. 63, 83

[135] Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013. 63, 83

[136] Gregory W. Corder and Dale I. Foreman. *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons, 2014. 63

[137] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1168–1176. ACM, 2013. 66

[138] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees. Monterey, Calif., USA: Wadsworth*. Inc, 1984. 68, 118

[139] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized Random Forests. *arXiv:1610.01271 [econ, stat]*, October 2016. arXiv: 1610.01271. 71

[140] K Hillstrom. The minethatdata e-mail analytics and data mining challenge. *MineThatData blog*, 2008. 72, 120

[141] Connors AF, Jr, Speroff T, Dawson NV, and et al. The effectiveness of right heart catheterization in the initial care of critically iii patients. *JAMA*, 276(11):889–897, 1996. 72, 120

[142] Melania Pintilie. *Competing risks: a practical perspective*, volume 58. John Wiley & Sons, 2006. 72, 120

[143] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1168–1176. ACM, 2013. 86, 123

[144] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. 88, 125

[145] José G Dias. Breast cancer diagnostic typologies by grades of membership fuzzy modeling. In *Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering*, 2009. 94, 127

[146] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 97, 128

[147] Daniel Almirall, Inbal Nahum-Shani, Nancy E Sherwood, and Susan A Murphy. Introduction to smart designs for the development of adaptive interventions: with application to weight loss research. *Translational behavioral medicine*, 4(3):260–274, 2014. 100, 132

[148] Susan A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005. 100, 132

[149] Susan A Murphy and Linda M Collins. Customizing treatment to the patient: Adaptive treatment strategies. *Drug and alcohol dependence*, 88(Suppl 2):S1, 2007. 100, 132

Subpopulation Treatment Effect Modeling : machine learning approaches to model treatment effect heterogeneity

Atef SHAAR

RESUME : La modélisation des effets de traitement de sous-population (STEM) est une technique d'apprentissage automatique utilisée pour choisir le traitement optimal (c'est-à-dire un stimulus) pour chaque sous-groupe. L'incertitude de l'information est un problème critique pour le STEM. L'incertitude sur les données existe en raison du problème fondamental de l'inférence causale, c'est-à-dire que seul un sous-ensemble des réponses des traitements est observé. Dans le domaine de l'apprentissage automatique, des techniques de tri spécifiques sont appliquées pour contourner le problème de l'incertitude. Cependant, l'un des inconvénients des méthodes de tri STEM actuelles est le traitement médiocre des variables de données continues, ordonnées et chronologiques, ce qui conduit à des résultats peu fiables et non interprétables.

Dans cette thèse, nous avons d'abord comblé les lacunes de la littérature et proposé une étude détaillée des techniques actuelles. Deuxièmement, nous résolvons les insuffisances en STEM concernant l'incertitude dans les données en proposant des arbres à effet de traitement de sous-population glissant. Troisièmement, nous proposons les forêts aléatoires de voisinage avec effet de traitement des sous-populations afin de minimiser l'effet du bruit dans les données. Quatrièmement, nous abordons le problème de la perturbation dans les données en proposant la technique de modélisation équilibrée du soulèvement par réflexion. Nous évaluons la performance des solutions proposées en utilisant des jeux de données simulés et réels, et nous montrons comment nos approches surpassent les autres méthodes en termes de coefficient de corrélation de rang de Qini et Spearman.

MOTS-CLEFS : Modélisation de l'effet du traitement de sous-population, effet du traitement hétérogène, modélisation Uplift, traitement personnalisé, effet causal

ABSTRACT : Subpopulation treatment effect modeling (STEM) is a machine learning technique that is used to choose the optimal treatment (i.e., stimulus) for each subgroup. A critical challenge facing the STEM is information uncertainty. Data uncertainty exists due to the fundamental problem of causal inference, i.e., only a subset of treatments' responses are observed. In machine learning domain, specific binning techniques are applied to bypass the problem of uncertainty. However, one drawback of current STEM binning approaches is the poor handling of continuous, ordered, and time-series data variables, leading to unreliable and non-interpretable results.

In this thesis, first, we fill the gaps in the literature and propose a detailed study of current techniques. Second, we solve STEM shortcomings regarding uncertainty in the data by proposing subpopulation treatment effect sliding trees. Third, we propose the subpopulation treatment effect neighborhood random forests to minimize the effect of noise in data. Fourth, we address the problem of disturbance in data by proposing the balanced reflective uplift modeling technique. We evaluate the performance of the proposed solutions using simulated and real datasets, and we show how our approaches outperform other methods in terms of Qini and Spearman's rank correlated coefficient.

KEY-WORDS : Subpopulation treatment effect modeling, heterogeneous treatment effect, uplift modeling, personalized treatment, causal effect

TELECOM ParisTech

EDiE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY