Chapter 4
Causal Machine Learning in Social Impact Assessment

Nuno Castro Lopes
https://orcid.org/0000-0003-2448-9798
Universidade Aberta, Portugal

Luís Cavique
https://orcid.org/0000-0002-5590-1493
Universidade Aberta, Portugal

ABSTRACT

Social impact assessment is a fundamental process to verify the achievement of the objectives of interventions and, consequently, to validate investments in the social area. Generally, this process is based on the analysis of the average effects of the intervention, which does not allow a detailed understanding of the individualization of these effects. Causal machine learning methods mark an evolution in causal inference, as they allow for a more heterogeneous assessment of the effects of interventions. Applying these methods to evaluate the impact of social projects and programs offers the advantage of improving the selection of target audiences and optimizing and personalizing future interventions. In this chapter, in a non-technical way, the authors explore classical causal inference methods to estimate average effects and new causal machine learning methods to evaluate heterogeneous effects. They address adapting the Uplift Modeling method to assess social interventions. They also address the advantages, limitations, and research needs for using these new techniques in social intervention.

1. INTRODUCTION

Recently, due to the emergence of ChatGPT and similar technologies, there has been much talk about Artificial Intelligence (AI) and Machine Learning (ML). As we explore the potential of AI/ML, we are seeing significant advances in natural language processing, neural networks, generative networks, computer vision, and others. However, despite these advances, it is only recently that the scientific community has begun to look more deeply into the relationship between Causality and AI/ML. This work

DOI: 10.4018/978-1-6684-9591-9.ch004

focuses precisely on this interconnection. The text is divided into two distinct parts: a brief introduction to the classical approach to causal inference, with an emphasis on econometrics, followed by a second, more specific part on how to apply AI/ML to causal inference. We will explore the advantages of these new approaches for a more complete understanding of causality, focusing on the practical application of causal inference in social impact assessment.

According to a report by the Organization for Economic Cooperation and Development (OECD) published in 2021, the need for better management of funding for social programs and projects has led to an increasing number of funders requesting impact evaluation reports from third-sector organizations. Assessing the impact of social intervention is becoming increasingly essential for these organizations and public entities that develop similar intervention actions. This approach is because funders of social programs and projects, whether public entities, foundations, companies, or individual patrons, increasingly demand transparency and accountability.

According to the same organization (OECD, 2002, p.4), the impact can be defined as "Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended.". Impact evaluation demonstrates how projects, programs, or policies achieve their objectives, showing possible changes in participants' behaviors, skills, and living conditions (Rogers, 2014a).

The outcome of a social intervention's impact may condition its continuation or replication. On the one hand, funders will find it more challenging to re-fund interventions that do not seem to show results. On the other hand, less effective intervention strategies and practices may be internally modified or even abandoned. Promoters and funders can also, based on the impact evaluation results, take advantage of the successful cases and be presented as good practices to be replicated in other contexts and by other organizations. For these reasons, impact assessment is fundamental in seeking greater transparency in social investment and greater effectiveness in social and community interventions.

The need to evaluate social programs and projects promoted by non-profit organizations and public bodies has driven the creation of consulting firms and research centers dedicated to this purpose. Despite several methodologies used to conduct an impact evaluation, starting the process by building a solid "Theory of Change" has become a frequent procedure among evaluators (Rogers, 2014a).

The "Theory of Change" is a model that describes how the activities of programs, projects, policies, or even the mission of specific organizations are expected to produce a series of results that contribute to achieving the intended final impacts. This method can be an asset in understanding the intended mechanisms and outcomes of these interventions to show which indicators should be considered to assess potential changes in people (Rogers, 2014b). A key element to consider in impact evaluation is that its purpose is not only to measure and describe the changes that occur but also to seek to understand the role of that program or project in producing the changes. This process is often referred to as causal inference. Several methods for analyzing causal inference benefit from being based on a solid "Theory of Change" (Rogers, 2014a).

However, it is crucial to note that, in general, the approaches adopted in these evaluations provide only the average effects on the group receiving the intervention, failing to consider the individual effects. However, looking only at average effects can provide misguidance to decision-makers (White et al., 2014a).

1.1. Problem

In the classic context of evaluating the impact of programs and social projects, the approach focuses on analyzing the average effects of this intervention (ATE- Average treatment Effect). However, this approach is limited in not capturing the heterogeneity of observed effects (HTE- Heterogeneity of Treatment Effect); that is, it does not analyze how an intervention can affect groups or individuals in a targeted way. In many situations, the intervention may have harmful average effects while benefiting specific subgroups within the same intervention group. In addition, it is possible to observe positive overall effects that may be negative for particular subgroups. This outcome variation highlights the importance of understanding the heterogeneity of the impact across different individuals or segments of the target population.

1.2. Objectives

Using stratification techniques to analyze HTE, namely trying to identify subgroups demonstrating divergent effects about the average, may lead to biases or ambiguous conclusions. Even when pre-experimental analysis plans are prepared, they often need a more comprehensive exploration of all potential heterogeneous effects of the intervention (Athey & Imbens, 2016; Wager & Athey, 2018).

Thus, this work aims to demonstrate how current and future evaluators of social projects can benefit from using casual Machine Learning methods to infer the so-called CATE (Conditional Average Treatment Effect). Enabling a deeper understanding of the effects allows decision-makers to obtain more accurate information for developing future interventions that are more personalized and adapted to the target audience (Lecher, 2023)

1.3. Contributions

This chapter presents how Causal Machine Learning methods can be a valuable resource for those evaluating social programs and projects. As a contribution, we demonstrate the adaptation of the Uplift Modeling method, a tradition used in marketing, for its application in social intervention. Through the Transformed Outcome Approach, it is possible to expand decision-makers ability to assess the heterogeneity of effects and, at the same time, contribute to improving decision-making related to future social programs and projects.

1.4. Organization

The chapter is divided as follows: Section 2: Causal Inference will begin by addressing the frequently used traditional causal inference methods, which focus on evaluating the Average Treatment Effect (ATE). We will explore techniques such as “Propensity Score Matching”, “Regression Discontinuity”, “Difference in Differences”, and “Synthetic Control”. Section 3: Machine Learning for Causal Inference, we will discuss the application of Machine Learning in causal inference, especially the evaluation of the Conditional Average Treatment Effect (CATE). For ease of understanding, we will briefly explain how Supervised Machine Learning techniques are generally used to make variations. We will also discuss some strategies and algorithms that use Machine Learning to perform causal inference, allowing the estimation of the heterogeneous effects of an intervention. These techniques include the “Metalearners”: “S-learner”, “T-learner” and “X-learner”, as well as the “Causal Trees and Causal Forests” algorithms.

Section 4: Application - Uplift Modeling: this section will demonstrate how to adopt the Uplift Modeling method, traditionally used in marketing. We will explain how this method is typically applied in marketing and introduce the Transformed Results Approach technique. Finally, we will show how this adaptation can be used to evaluate the impact of programs and social projects. Finally, in section 5, there are some considerations. It is important to note that this chapter addresses techniques and methodologies that still need to be widely disseminated in social impact assessment. We aim to make this text accessible to anyone, even with little prior knowledge about Machine Learning algorithms, avoiding the excessive use of formulas or precise terminology. However, we will provide references to some of the main works in the area, allowing those interested in deepening the topic to consult relevant bibliographies for a deeper understanding.

2. CAUSAL INFERENCE

Developing a "Theory of Change" helps impact evaluators to define the relevant indicators better, usually referred to as dependent variables and designated with the letter "Y," as well as the variables that may be associated, traditionally referred to as covariates or independent variables and defined with the letter "X." In addition to the " Theory of Change," which is a more general and comprehensive representation, in causal inference, we can use the so-called "Causal Directed Acyclic Graphs" (Pearl, 1995; Pearl & Mackenzie, 2018) to represent the cause-effect relationships between the different variables that can affect the impact of these programs or projects. These diagrams make it possible to visualize the causal connections between variables, facilitating the analysis of the effects of interest. They are represented by nodes and arrows, where each node represents a variable, and the arrows indicate the direction of causality between them. In this way, it is possible to examine how different variables influence each other, allowing us to understand better and determine which variables we wish to control and which we wish to incorporate into our causal inference process effectively.

In the context of causal inference, Rubin's (1974) causal model is often used as a basis. In this context, it is usual to use the letter "T" to denote the intervention (or treatment). One of the critical concepts of this model is the potential outcomes, which correspond to the outcomes that would be observed if a person[1] were subjected to a particular intervention. According to this model, each person will have a potential outcome for each possible level of intervention, and only one of them can be observed. These potential outcomes are calculated by the difference between the value of the dependent variable at a time before the intervention starts (time=1 or t1) and the value of that variable at a later stage (time=2 or t2), which usually corresponds to the time of the evaluation. They may occur either after the end of the intervention or at an intermediate stage. Thus, the causal effect is defined as the difference between the potential outcome that would be observed if the person received the intervention (Y^1) and the potential outcome that would be observed if the same person did not receive the intervention (Y^0) or even if they received another type of intervention. This notion of causal effect can be summarized with the formula Y^1-Y^0.

However, in practice, we face the fundamental problem of causal inference. What we want to know is what happened to the people who participated in the intervention and what would have happened if they had not participated in the intervention. However, it is logical that we cannot simultaneously and directly observe the effects of people exposed to the intervention and those not. Creating counterfactual strategies and alternative ways of simulating what would have happened in situations different from reality is necessary (Imbens & Rubin, 2015; Rubin, 1974).

The gold standard strategy for conducting causal inference is experimental studies with randomly selected groups. This study randomly divides participants into the treatment or intervention (T=1) and the control group (T=0). This random selection ensures that each participant has an equal chance of being assigned to either group, making them comparable concerning the dependent variable (Y) and eliminating potential bias. Following this approach, when the sample is significant, and the random selection is performed correctly, the groups become similar at the level of their independent variables (X_1, X_i, ..., X_n).

In experimental studies, the assignment of a person to one of the groups must be individual, probabilistic, and unbiased, and this random assignment can be carried out in different ways (Imbens & Rubin, 2015; Kang et al., 2008; Stanley, 2007). According to Imbens & Rubin (2015), this assignment can be carried out by randomly dividing the sample into two groups; Bernoulli's method, where for each person in the sample, a 50% random choice is made (such as flipping a coin to determine whether they will belong to one group or the other); stratification, where first the sample is divided into blocks based on a certain covariate that may influence the outcome (such as gender, age, etc.) and within each block, the selection is carried out entirely randomly; or paired comparison, where within each stratified block the sample is divided into pairs and, for each pair, the selection is carried out completely randomly and within each block, the selection is carried out entirely randomly; or paired comparison, where within each stratified block, that block is divided into pairs and, for each pair, a random choice is made to determine which of the people in the pair will go to each of the groups.

Despite recognition by the scientific community, conducting experimental studies in social intervention is not always feasible or appropriate. The need to intervene with individuals with specific characteristics makes it impossible to randomly select intervention and control groups. Usually, the intervention group is chosen before the procedures inherent to the evaluation are carried out. In many situations, it would be unethical to determine who would and would not receive the intervention randomly. Also, in many circumstances, the start of the evaluation occurs after the intervention has begun, making it impractical to conduct experimental studies, leading to the use of observational data.

For these reasons, social impact assessment often resorts to quasi-experimental methods to conduct the evaluation. In several of these situations, there are recorded data from other people with similar characteristics (X_1, X_i, ..., X_n) who have not received an intervention, which can help to create the counterfactuals.

For example, when a project is implemented in a particular class in a school or a specific group of inmates in prison, if there is relevant information regarding the same covariates (X_1, X_i, ..., X_n) and dependent variables (Y) for other people in the same organization, it is possible to use some techniques to create the control groups. We will discuss some of these techniques below. It is important to note that these techniques only determine the ATE.

2.1. Propensity Score Matching

One of the techniques used to address bias in the selection of a control group, known as control group selection bias, is Propensity Score Matching (PSM). This approach allows for matching between each person who received the intervention and all the other people who were not intervened but with similar characteristics (covariates) and belong to the same context.

PSM aims to create a control group comparable to the intervention group's characteristics using various statistical techniques or supervised machine learning algorithms (which will be explained later); PSM seeks to create a control group comparable to the intervention group's characteristics. The larger the sample of possible counterfactuals compared to the sample of intervened persons, the greater the

possibility of performing the matching properly. In addition, a significant number of covariates also requires a higher proportion of counterfactual candidates compared to the number of intervention subjects.

For example, in an intervention in a school context, PSM assigns a propensity score to each student in the same school, indicating their similarity to the student who received the intervention. Intervened students with the highest propensity scores and similar characteristics are selected for the control group. This procedure will be done for all other subjects who received the intervention. In the end, the sample obtained through the PSM will serve as the control group, just as in an experimental study, since the groups will have similar characteristics.

The most commonly used technique to perform this propensity score is logistic regression, but discriminant analysis, probit models, classification or regression trees, neural networks, support vector machines, or meta-classifiers can also be applied (Cunningham, 2021; Fan & Nowell, 2011; Fougère & Jacquemet, 2019; Huntington-Klein, 2022; Imbens, 2000; Imbens & Rubin, 2015; Rosenbaum & Rubin, 1983; Tu, 2019).

2.2. Regression Discontinuity

Unlike PSM, regression discontinuity (RD) is a methodology that cannot be applied to all situations. RD is suitable in specific settings where the intervention is defined based on a cut-off index, value, or percentage that unambiguously determines which individuals can and cannot participate.

For example, let us consider a school support program that offers intervention only to students who are beneficiaries of scholarships. These scholarships are awarded only to students whose household income per capita is below a specific cut-off value. In this context, the RD is applied to distinguish the intervention group (i.e., the scholarship beneficiaries) from the control group (i.e., the non-beneficiaries) based on this specific cut-off criterion. Moreover, to ensure the precision of the analysis, the RD methodology requires the definition of a bandwidth, which delimits the upper and lower limits related to the cut-off.

In this way, RD allows the analysis of the causal effect of the intervention, focusing on subjects close to the cut-offline, both in the intervention group and the control group. The comparison between these groups makes it possible to investigate the intervention's impact on those on the borderline of eligibility, contributing to more robust causal inferences.

Importantly, RD is a methodology that presents its effectiveness and applicability in specific settings where the intervention is clearly defined through a cut-off point. However, it is critical to recognize that this approach is only suitable for evaluating a relatively small number of real-world social projects and programs (Cook, 2008; Fougère & Jacquemet, 2019; Huntington-Klein, 2022; Gopalan, Rosinger & Ahn, 2020; Thistlewaite & Campbell, 2017).

2.3. Difference-in-Differences and Synthetic Control

In Social Impact Assessment, it is only sometimes feasible to find individual observations that can serve as a counterfactual to assess the impact of a specific intervention. In such situations, the strategy used is to compare aggregate data, i.e., to analyze the average effect of the intervention on the dependent variable (Y), based on observational data from other contexts in which that intervention has not been applied but where we have access to data regarding the value of the indicator of that dependent variable. The Difference in Differences and the Synthetic Control are two strategies that allow us to perform this type of analysis.

The Difference-in-Differences (DiD) is one of the most classical and widely used methods in social impact assessment. The DiD method starts by looking for a population group similar to the intervention group, which has yet to be exposed to the intervention, acting as a control group. These groups should be selected in contexts identical to the intervention, such as in another school, prison, municipality, etc.

For both groups, the intervention and the control, we need information regarding two distinct temporal moments: before and after the implementation of the intervention. In addition, it is desirable to have other intermediate time points that allow analyzing the trajectory of the values of the dependent variable over time.

For the method to be applied correctly, the trajectory of the two groups must be similar during the period before the intervention, even if the values themselves are different. This prior similarity of trajectory helps ensure that any differences observed after the intervention are caused by the intervention itself and not by pre-existing factors.

The DiD consists of two differences. First, the difference between the periods before and after the intervention is calculated for both groups. Then, these two differences are subtracted, resulting in the intervention effect.

To estimate the effect, many studies use a regression model, which allows estimating what would have happened to the intervention group if it had not been exposed to the intervention, based on the trajectory observed in the control group after that intervention (Cunningham, 2021; Crato & Paruolo, 2019; Fougère & Jacquemet, 2019; Huntington-Klein, 2022; Gopalan, Rosinger & Ahn, 2020; Lechner, 2011).

For its part, Synthetic Control (SC) is a methodology that addresses the difficulty of finding control groups with similar trajectories in the pre-intervention phase. Instead of looking for an authentic context identical to the intervention group, SC uses an algorithm to create a synthetic group that is as similar as possible to that intervened group.

CS is an approach that offers an alternative to directly comparing the intervention group to a specific school, prison, or municipality. Instead, SC creates a synthetic entity, such as a synthetic school, synthetic prison, or synthetic city, through an algorithm. Creating the synthetic group is based on other schools, prisons, or municipalities with similar characteristics to the intervention group. To carry out the process, it is necessary to have observational data related to the dependent variable and the covariates of these similar entities. The SC algorithm combines these similar entities' characteristics (covariates) to create a synthetic entity virtually identical to the intervened group. In this way, the synthetic group acts as a control group with similar characteristics and behaviors similar to what would have happened to the intervention group in the absence of the intervention (Abadie & Gardeazabal, 2003; Abadie, Diamond, & Hainmueller, 2010; Fougère & Jacquemet, 2019; Cunningham, 2021; Huntington-Klein, 2022).

3. CAUSAL MACHINE LEARNING

Using these traditional causal inference methods makes it possible to evaluate the practical impact of interventions, demonstrating the viability of programs and social projects before funders. By determining the ATE of social interventions, decision-makers can better understand actions that should be considered best practices for future interventions and those that should not be continued due to their reduced average effectiveness.

But, in social intervention, especially with vulnerable or at-risk populations, there are many cases of failure, even in projects that obtain good results overall. In several situations, the average effects of the intervention may be adverse despite being beneficial for specific groups within the same intervention population. Additionally, it is possible to identify positive results that may harm particular subgroups. This variation in results highlights the importance of understanding diversity in effects among different individuals or segments of the target population. The conventional approach to social impact assessment may not be sufficient to fully capture these nuances and individual differences, which limits a comprehensive understanding of the effects of social interventions.

In this logic, new approaches that combine Machine Learning techniques and algorithms with causal inference have emerged in recent years, known as "Causal Machine Learning". These new approaches allow a more in-depth analysis of the heterogeneity of effects; that is, they will enable us to understand the impact of a project or program on individuals with specific characteristics based on the interventions already carried out. Despite their advantages, they are not widely applied in social impact assessment processes. However, their implementation can boost the prescription of future projects and interventions by improving the selection of target audiences, optimizing the interventional process, and personalizing future interventions.

"Causal Machine Learning" is a broad term encompassing Machine Learning methodologies for inferring causality. According to Kaddour, Lynch, Liu, Kusner, & Silva (2022), the area can be subdivided into five subgroups: (1) causal supervised learning, (2) causal generative modeling, (3) causal explanations, (4) causal fairness, and (5) causal reinforcement learning.

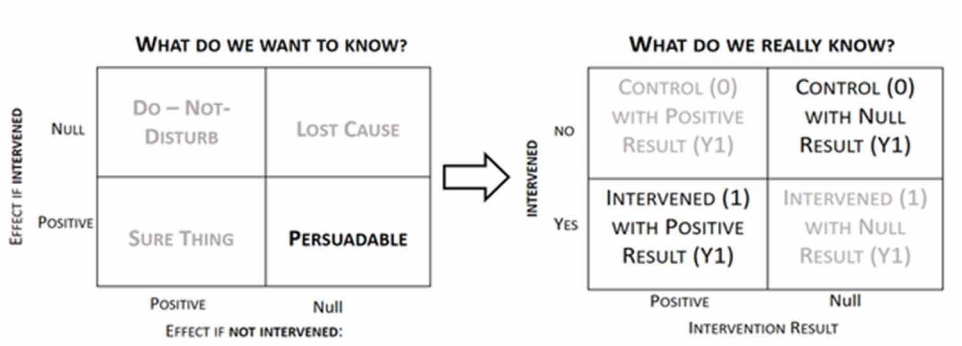
In this paper, we only focus on Causal Supervised Machine Learning and its application to infer the heterogeneous effect of interventions (determine the CATE). Within the various strategies to combine Supervised Machine Learning algorithms or new specific Causal Machine Learning algorithms, we present the Metalearners (S-Learner, T-Learner, and X-Learner), the Transformed Outcome Approach, and Causal Trees and Forests. To facilitate the understanding of these methods, we will only present how to use them in simple situations where there is a single intervention, where an individual can be intervened (T=1) or not (T=0), and the outcome is binary (Y^1 or Y^0).

Before exploring these methodologies and algorithms, for better understanding, we will briefly explain how Supervised Machine Learning algorithms are generally used to make predictions.

3.1. Supervised Machine Learning for Forecasting

According to Gama, Lorena, Faceli, Oliveira & Carvalho (2017), machine learning can be defined as the process of inducing a hypothesis from past experiences. In machine learning, algorithms learn from experience, causing general conclusions from specific examples. Generally, machine learning tasks are categorized into three types: supervised, unsupervised, and reinforcement learning. In this context, we focus on supervised learning. As noted by Lechner (2023), the task of supervised learning involves identifying structures in the space of covariates (X_1, X_i, ..., X_n) that result in accurate predictions of the dependent variable (Y). Within this category are classification methods for predicting specific values and regression methods for predicting discrete or continuous values. Supervised Machine Learning involves training a model using relationships of the type X_1, X_i, ..., X_n -> Y and employing that model to predict Y based on new datasets X_1, X_i, ..., X_n (Figure 1).

Figure 1. Supervised machine learning model for prediction



Translating this example to a practical application in the social area, we consider a scenario of a school that wants to use a supervised machine learning model to predict whether students in a given school year will transition or be retained based on the co-variables of students from previous years (such as sociodemographic characteristics, academic performance, behavior, among others). In this context, these co-variables would be the values X, Y would be 1 in the case of transition, and 0 in the retention situation.

3.2. Metalearners

As we have seen, the use of Supervised Machine Learning requires two types of attributes: the independent variables (X_1, X_i, ..., X_n) and a dependent variable (Y) aims to train the model to predict Y based on the new variables X. To infer causality, a new variable referring to the treatment/intervention (T) is added. In this latest action context, the data sets must include variables of type X, T, and Y for applying Causal Machine Learning methods.

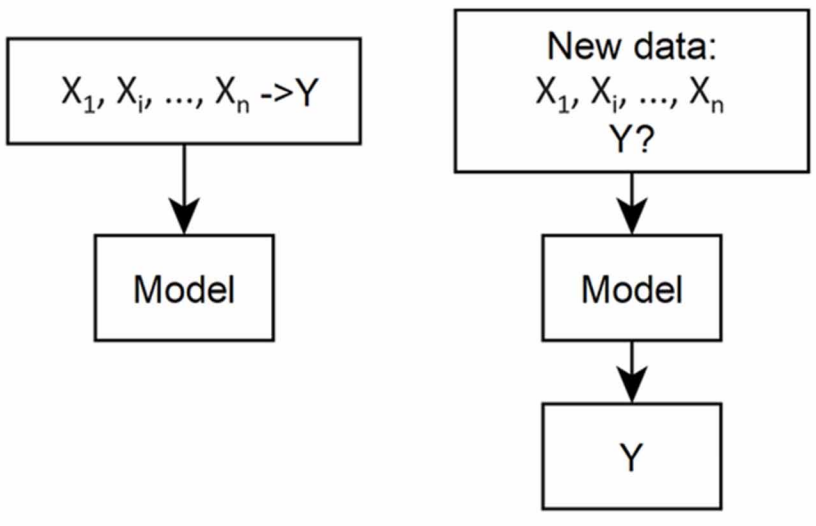
Metalearners enable the use of traditional Supervised Machine Learning algorithms or statistical regression methods to estimate the heterogeneous effects of an intervention.

3.2.1. S-Learner

One of the simplest metalearner is the S-Learner (Figure 2), in which the letter "S" stands for using a single estimation model. The S-Learner model is trained to estimate the target variable "Y" based on variables X and T. In this method, the intervention-related variable (T) and the other independent variables (X) are another input attribute. Thus, the algorithm is trained and tested according to the following logic: X_1, X_i, ..., X_n, T -> Y. Because we only consider binary interventions and outcomes, T and Y can only have the values 0 or 1.

After training and testing the model, causal inference is performed using two estimates obtained by the same model. In the first estimate, it is assumed that the person underwent the intervention (T=1), and in the second estimate, it is assumed that this person was not subject to the intervention (T=0). The individual treatment effect can be obtained by subtracting the estimated value when T=1 from the estimated value when T=0 (Alves, 2022 & Künzel, Sekhon, Bickel, & Yu, 2019).

Figure 2. S-Learner model



Returning to the example of school retention, the focus here would not be to predict retention but to understand the effect an intervention (e.g., a tutoring program) would have on this retention. To do this, we would consider the history of results obtained in that school in previous years, both in students who participated in tutoring and in those who did not. Thus, the model could be trained with all students in the school, regardless of whether they participated in the mentoring programs. The model would be trained with the specific attributes of each student (X), plus a feature indicating whether they had participated (T=1) or not (T=0) in tutoring, with the target variable for training being whether these students had been retained (Y^0) or not (Y^1). After training, the model could be used to estimate the individual effect of future pupils based on their characteristics (X). Thus, for each new student, an estimate would be made with all their features X and a T=1, and another would be made with the same characteristics but with a T=0. The subtraction of these two scenarios would provide the estimated individual effect of this mentoring program.

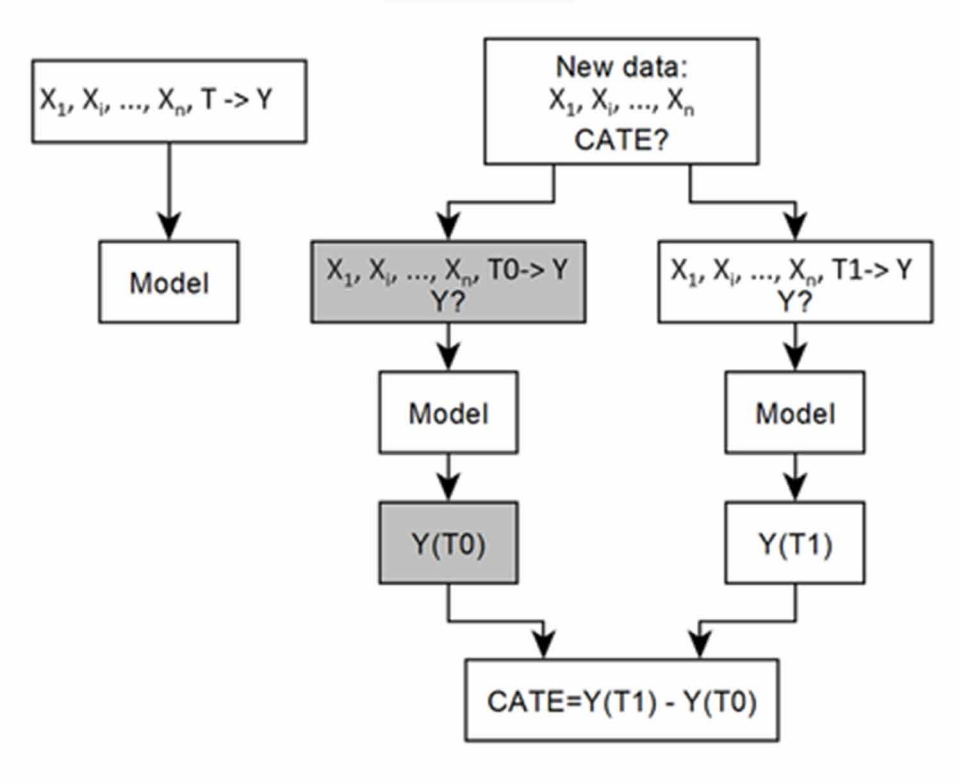
3.2.2. T-Learner

Another metalearner is the T-Learner model (Figure 3), which is also relatively simple to apply. It is so named because it involves the use of two separate models. The process starts by splitting the initial dataset into two subsets: one containing only those cases where people were intervened (T=1) and one containing those that were not (T=0).

Subsequently, we train these subsets separately using supervised machine learning algorithms or regression models, following the logic: Model (T1) = X_1, X_i, ..., X_n -> Y and Model (T0) = X_1, X_i, ..., X_n -> Y.

To estimate the heterogeneous effect of a given individual based on their specific attributes (X), we apply the model created with the subset of those intervened (Model T1) and then use these same attributes to the model developed with those not intervened (Model T0). The estimated individual treatment effect for that subject would be the effect estimated by Model T1minus, estimated by Model T0 (Alves, 2022; Künzel, Sekhon, Bickel, & Yu, 2019).

Figure 3. T-Learner model



In the mentoring scenario, the process would involve dividing the former students into two groups: those who participated in the mentoring (T=1) and those who did not (T=0). An estimation model based on a machine learning or regression algorithm would be developed for each group to estimate retention (Y) based on student characteristics (X). To estimate the individual effect on new students, it would be sufficient to estimate the potential outcome in each model considering the characteristics of these new students and subtract the potential outcome obtained in Model T1 from that obtained in Model T0.

3.2.3. X-Learner

Another method is the X-Learner, which is a more complex approach than the previous ones, as it involves the interaction of models and datasets. Due to this complexity, we recommend following the explanation while analyzing the process in Figure 4.

The first stage of the X-Learner is like the T-Learner, where we build separately Model(T1) for the intervened (T=1) and Model(T0) for the non-intervened (T=0). Still, we will create a third propensity model (PS) with machine learning or regression algorithms in this first stage. This other propensity model aims to estimate the probability of a given person with specific characteristics belonging to the intervention group. To this end, this model is developed based on the factors of all people in the initial data set, where the target variable is the fact of having been intervened (X_1, X_i, ..., X_n ->T).

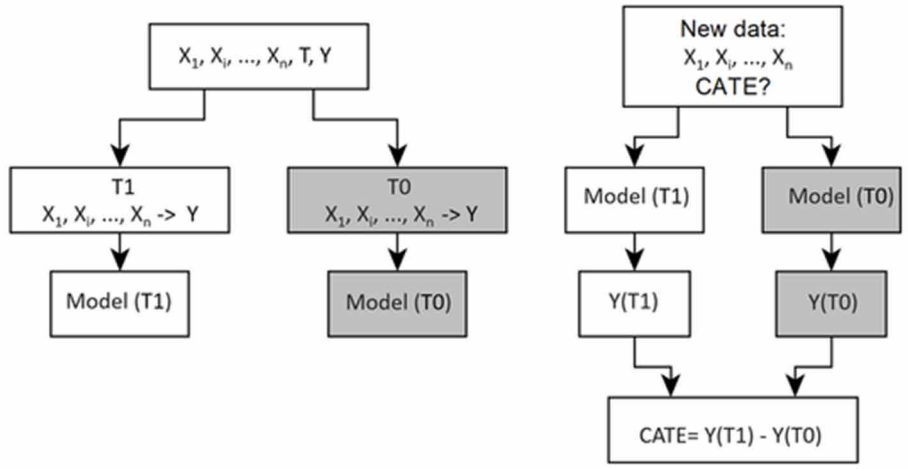
In the second stage, a significant difference appears in Model(T1) and Model(T0), compared to T-Learner: a cross between data and models is carried out. Thus, we apply the Model(T1) model to the initial set of non-intervened people (Model(T1)-> T0), and we use the Model(T0) to the set of intervened people (Model(T0)-> T1). This approach allows us to obtain inverse estimates in both groups: in the intervention group, we get an estimate of the potential outcome that each person would have had if they had not been intervened, Y[Model(T0)], while in the non-intervention group, we obtain an estimate of the potential outcome that each individual would have been if intervened, Y[Model(T1)].

In this way, we generated two sets of data. In the first set CATE(1), we subtract from the actual result Y(T1) the potential result obtained with the estimate Y[Model(T0)]. In the case of the other subgroup, CATE(0), we use the estimate of the potential result Y[Model(T1)] and subtract the actual value from Y(T1). The CATE(1) and CATE(0) are designated as the imputed treatment effect. We then developed two more machine learning models, each to estimate the value of the effects attributed to the treatment, based on the characteristics of each of the groups, since in the case of the intervention group, we obtain the Model[CATE(1)], and there is no control of Model[CATE(0)].

Thus, to estimate the effect of the intervention on new data (X1, Xi, ..., Xn), we apply three models to this data set: Model[CATE(1)], Model[CATE(0)] and PS and we obtain the values Y(T1), Y(T0) and P. Thus, to estimate the effect of the intervention, to the estimated value Y(T1) we multiply the propensity value of the individual belonging to the intervention group Y(T1). P and Y(T0), we multiply the propensity value of the individual who does not belong to the intervention group Y(T0). (1-P), in the end, we subtract one value from the other to know the effect: P.[Y(T1)] - (P-1).[Y(T0)].

In addition to the methods mentioned, there are others, most of them more complex, such as the Doubly Robust (DR-learner) method (Kennedy, Ma, McHugh, & Small, 2017), R-learner (Nie & Wager, 2021), M-Learner (Acharki, Lugo, Bertoncello, & Garnier, 2023) among others. Some of these methods were created for multiple interventions, and each has its advantages and disadvantages depending on the dataset, namely the sample size, number of covariates, and the ratio of the proportion between intervention and control groups (Acharki, Lugo, Bertoncello, & Garnier, 2023).

Figure 4. X- Learner model



3.3. Causal Trees and Causal Forests

In addition to methods that employ conventional Machine Learning algorithms to simulate the creation of counterfactuals and estimate heterogeneous intervention effects, specific algorithms have been developed for this purpose. One example is the Causal Tree (Athey & Imbens, 2016), which derives from traditional Decision Tree algorithms.

Decision Trees are Machine Learning algorithms that divide data into branches, forming a tree-like structure in which each node represents a feature of the dataset, and each branch denotes a decision based on that feature. These decisions are made until the data is separated into smaller, more homogeneous groups, allowing the prediction of an outcome for each group.

The Causal Tree follows a detached approach. It also segments the data into branches to identify heterogeneity in treatment effects. Rather than aiming only for accuracy in predicting the outcome, the Causal Tree focuses on finding groups of individuals who react differently to the intervention, manifesting distinct responses. This algorithm divides the data into two sets. The first set is used to build the tree, establishing rules that segment the data into terminal nodes. At the same time, the second set is used to calculate the effects of the intervention at each terminal node by calculating the mean difference between the outcomes observed in the intervention and control groups. This split is known as "honest" as it avoids the bias that could arise if the same dataset was used for both tasks. This approach allows the Causal Tree to more accurately identify subgroups of people who may or may not benefit from the intervention.

Like Random Forests, which expedite Decision Trees by constructing multiple trees that work together and are randomly generated using only a subset of the data and features to reduce overfitting and improve prediction accuracy, Causal Random Forests (Wager & Athey, 2018) also represent new causal machine learning algorithms that extend the notion of Causal Trees. The Causal Forests algorithm builds multiple causal trees and subsequently combines the predictions from these trees to obtain a more accurate estimate of the intervention effect.

It should be noted that, in addition to implementing the methods described, it is essential to evaluate the effectiveness of these models since the procedures used for this evaluation differ from those used in predictions made through Supervised Machine Learning (Devriendt, Moldovan, & Verbeke, 2018; and Pinheiro & Cavique, 2022).

4. APPLICATION: UPLIFT MODELLING

In social intervention, especially with vulnerable or at-risk populations, there are many cases of failure, even in projects that obtain good results overall. In many situations, the average effects of the intervention may be harmful, although they may be favorable for specific subgroups within that same intervention group. In addition, it is possible to observe positive overall effects that may be negative for particular subgroups. This outcome variation highlights the importance of understanding the heterogeneity of effects across different individuals or segments of the target population.

The traditional methods are presented in *section 2. Causal Inference* is widely used in social impact assessment. Despite their advantages, these approaches provide only the average effects of the intervention (usually referred to as ATE - Average Treatment Effect), not considering the individual effects that the intervention may have had on each person.

In social intervention, it is critical to understand the heterogeneity of effects across different individuals or segments of the target population. Traditional approaches to social impact assessment can only sometimes fully capture these nuances and individual differences, which limits our complete understanding of the effects of social interventions. Understanding the heterogeneity of the effects of social interventions can be crucial to refining future interventions, including the most appropriate selection of candidates for a given project or program, optimizing the intervention process itself, and tailoring the intervention to individual needs.

To illustrate the practical application of a Causal ML technique to address the heterogeneity of the effects of the impacts of social interventions, we will demonstrate adapting the Uplift Modeling method to evaluate social programs and projects. Although initially developed for marketing, this methodology also applies to other areas. We will initially discuss how this approach is commonly used in marketing campaigns for a more precise understanding. We will then explain how to implement this methodology through the Transformed Outcome Approach. Finally, we will propose a specific practical adaptation of this method to social impact assessment.

4.1. Uplift Modelling in Marketing

This approach is central to the field of prescriptive marketing and customer analytics, as it allows us to analyze what effect the campaign has on customer behavior.

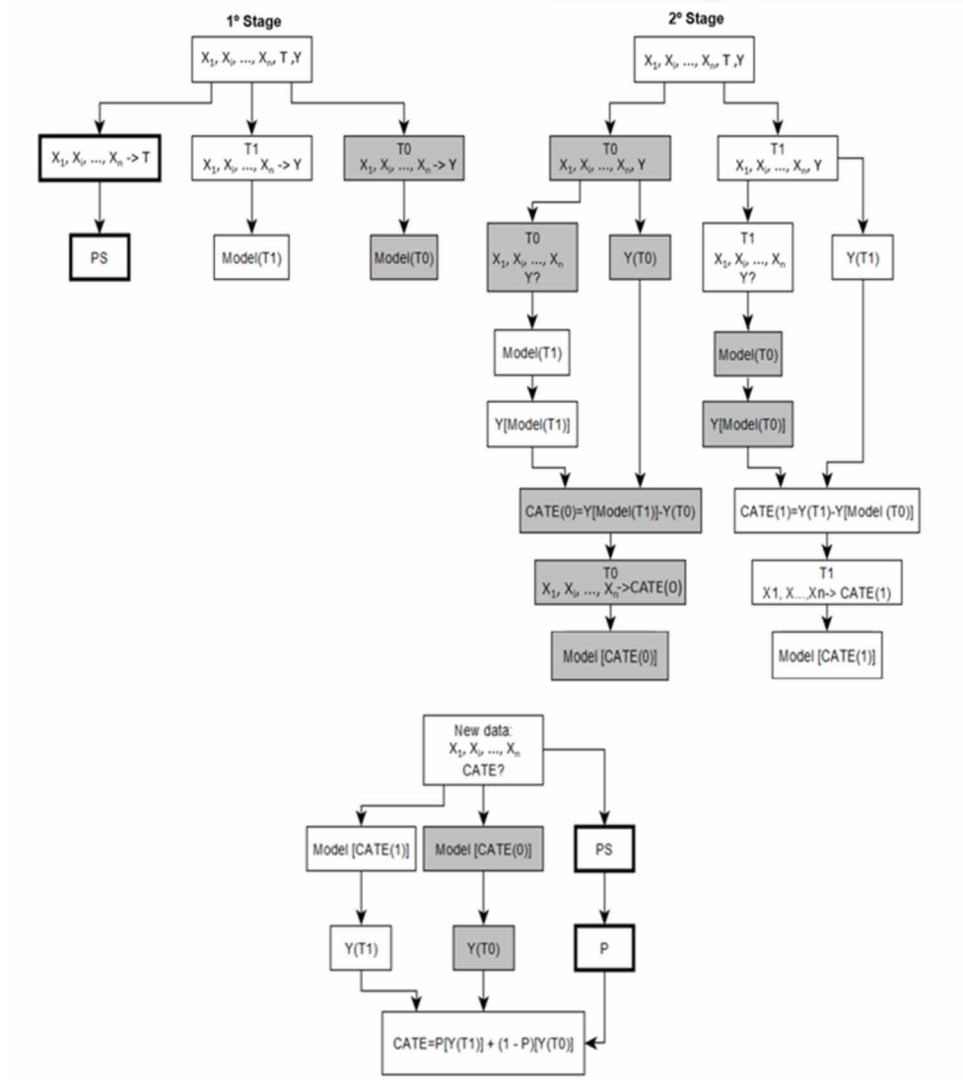
This methodology was developed to reduce the costs of direct marketing campaigns and increase the return on investment. In its original version, based on previous campaigns, the objective is to categorize customers into four quadrants, considering their behavior towards a marketing campaign.

In the first quadrant, the "Persuadable" customers only buy when a marketing campaign targets them. In the second quadrant, we have the "Sure Thing," individuals who buy regardless of whether the campaign targets them. In the third quadrant, we have the "Lost Cause," individuals who never buy, irrespective of whether they are targeted or campaigned. Finally, in the fourth quadrant, we have "Do-Not-Disturb" or "Sleeping Dog", which are individuals who only buy if they are not targeted by the campaign (Pinheiro & Cavique, 2022; and Devriendt, Moldovan, & Verbeke, 2018).

Based on this method of prescription for a future marketing campaign, companies seek to avoid sending this campaign to the so-called "Do-Not-Disturb" and direct their efforts to the "persuadable". In addition, they seek to avoid investing resources in the "Sure Thing" and the "Lost Cause". In this way, they can maximize the effectiveness of their campaigns by focusing resources where there is a greater propensity to achieve positive results.

However, in reality, this model corresponds to what we want to know rather than what we know (see Figure 5); we fall back into the fundamental problem of causal inference. We cannot simultaneously know what would have happened to a given intervened person if they had not been intervened, nor the opposite: inability to observe effects simultaneously with and without exposure.

Figure 5. Uplift modeling



4.2. Transformed Outcome Approach

A common form of Uplift Modeling is through the use of the Transformed Outcome Approach (Figure 6). This approach, similar to Metalearners, is a technique that allows estimating the effect of heterogeneous impact through classical supervised machine learning methods. As the name suggests, the aim is to transform the outcome approach to facilitate causal inference. The symbology that represents this transformation may vary according to the author. In this case, we express it as "Y*".

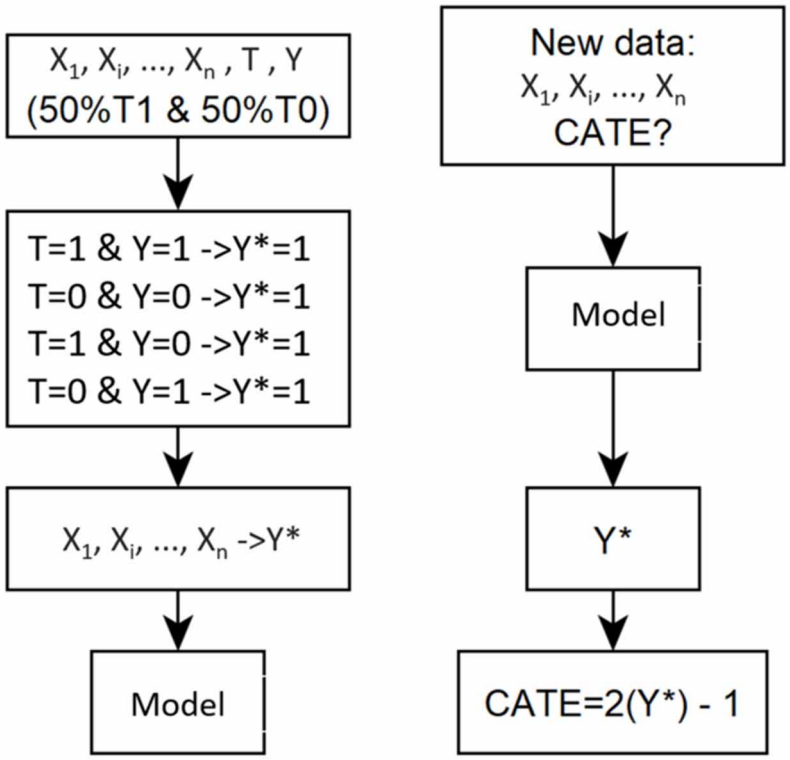
According to Jaskowski and Jaroszewicz (2012), the transformation from the initial values to the value Y* is intuitive in the case of binary outcomes. We consider Y* equal to 1 if the intervention outcome is at least as good as the outcome that would have occurred in the control group had we known

the outcome in both groups. In practice, Y* is equal to 1 if the intervention results in a positive outcome (T=1 & Y=1) or if no intervention results in a null outcome (T=0 & Y=0); in all other situations, Y* will be equal to 0.

By performing this transformation on all the observations (individuals) in our dataset, we can train a machine learning or regression model using the attributes (X_1, X_i, ..., X_n) as independent variables and the value of Y* as the independent variable. Without going into the mathematical particularities inherent in this method, to calculate the individual treatment effect, multiply the value of the estimate by two and subtract 1, i.e. (2xY*)-1. It is important to note that this method is only effective if there is an initial probability of 50% of an individual belonging to the intervention group or the control group (Devriendt, Moldovan, & Verbeke, 2018; Jaskowski & Jaroszewicz, 2012 and Pinheiro & Cavique, 2022).

In the example of tutoring at school, the model would be trained by transforming the value Y*=1 for all students who participated in tutoring and were not retained (T=1&Y=1) or did not participate and were retained (T=0 & Y=0), with all other situations, the value of Y* would be equal to 0. It is important to note that the sample size of those who did not participate in tutoring should be similar to those who did. After the estimation for each student, the individual effect of tutoring would be calculated by applying the formula mentioned above.

Figure 6. Transformed outcome approach model



4.3. Uplift Modelling in Social Impact Assessment

To improve the effectiveness of interventions, just as in Marketing, where we seek to avoid wasting resources on campaigns that aim to "Do Not Disturb" and try to focus our efforts and investments on reaching "Persuadable", the objective can be similar in social intervention. As mentioned previously, funders of educational, social, and community projects and programs are increasingly demanding and looking for more evidence of the effectiveness of interventions through impact evaluation projections. Detecting which quadrant of the uplift modeling model each potential beneficiary of a policy, program, or social project is in can help decision-makers develop more targeted and/or personalized interventions.

In this sense, using Uplift Modeling, through the employment of the Transformed Outcome Approach to estimate the heterogeneous effects of interventions, can lead to the improved success of each intervention and, consequently, a greater monetization of social investment. Given this, promoting these procedures can revolutionize the general approach to impact assessment, moving from a merely evaluative vision to an approach more based on guiding/recommending future actions.

The application of this methodology in the evaluation of social interventions still requires significant studies. In this context, it is relevant to mention two recent studies by Olaya et al. (2020) and Tanai & Ciftci (2023), who used this approach to analyze the heterogeneous effects of specific interventions. One of the studies investigated the impact of participating in tutorials. At the same time, the other explored the effects of taking preparatory courses in English and mathematics, both about the risk of dropping out of university. The main objective is personalizing future interventions based on the "persuadable" segment.

To understand these advantages, it is essential to consider the social impact and the individual perspectives. Under a perspective focused on social impact, future social programmers and projects could achieve more promising results since it would be easier to select the people who would benefit from that specific intervention, optimizing the return on investment made in those actions. It is particularly relevant for interventions with relatively low average success outcomes, such as social projects and programs targeting young people with justice problems or substance use reduction programs.

On the other hand, we have the person-centered perspective, in which we can use this knowledge to estimate the heterogeneity of effects to support the creation of individualized intervention pathways. Social workers can accompany and guide these people in trajectories more suited to their characteristics and needs. An example could be job centers' personalized referral of unemployed people to training or apprenticeships that best meet their characteristics and needs. The same approach could benefit any intervention to develop individualized social support pathways: rehabilitation, social reintegration, and promotion of behaviors or skills, among others. In this way, the Social Impact Assessment and Prescription Approach can be valuable in optimizing social interventions and providing personalized support for each individual, allowing a more effective targeting of resources and a more attentive approach to the needs and potential of the people involved.

Despite the advantages mentioned, it is essential to highlight that, as in any data science project, especially when it involves large volumes of data and machine learning algorithms, it is crucial to consider the quality of data collection and the preservation of the privacy of that data. Aragon et al. (2022) argue that data is rarely neutral, as it reflects the biases and subjectivity of the people and systems that collect, process, and interpret it. Ignoring this fact can perpetuate discriminatory patterns and biases harmful to specific groups.

In addition to concerns about potential biases in the data collected, another significant challenge is the availability of this data. Data related to the characteristics of each subject participating in the evaluation

and prescription processes (both in the intervention and control groups) must be available to achieve these objectives. It may include data on the characteristics of the subjects, such as demographic data, behaviors, needs, and results obtained in this intervention or previous interventions. Another challenge is data collection itself. Data must be collected before and after the intervention to evaluate its effects. Additionally, having many observations (usually subjects) is vital to making the most of these methods because sample size can influence the effectiveness of causal machine-learning methods and algorithms.

In many social projects, outcome indicators are based on standard questionnaires applied before and after interventions. However, it is only sometimes feasible to have control groups for comparison. Considering these difficulties from the project design stage is crucial to maximize future benefits when assessing and predicting varied effects. Finally, it is essential that funding entities recognize the importance of this type of evaluation/prescription, considering the investment in social interventions as a long-term return, as specific projects may present less favorable initial results but can be a source of information for the future improvement and personalization of these interventions.

5. CONCLUSION

The classic approach to social impact assessment focuses on traditional methods of causal inference, which only provide us with the average effects of the intervention without capturing how social programs and projects can affect people or groups differently. Traditional stratification methods require detailed analysis before intervention to identify groups in detail, limiting later discovery of effects as they could be biased. In contrast, new Causal Machine Learning approaches allow for a more heterogeneous analysis of the effects of interventions, allowing us to understand existing effects and estimate possible future effects of new programs and projects. Causal Machine Learning provides decision-makers with more accurate information for developing future personalized interventions adapted to the target audience.

In this chapter, we have explored the potential advantages of applying Machine Learning algorithms for causal inference in the impact assessment processes of social projects and programs. In addition to a non-technical review of the classic causal inference methods used to evaluate the social impact of programs and projects, we present some of the new Causal Machine Learning techniques that can be applied to assess the heterogeneous effects of this type of intervention. To understand how to adapt these methods to social impact assessment, we demonstrate the Uplift Modeling method's adaptation to assess social projects and programs. Although it is usually used in Marketing, this method allows for capturing the heterogeneity of effects, grouping individuals into four categories. With this methodology, impact evaluators can present decision-makers with the ability to identify, based on previous interventions, the people most likely to benefit from this social intervention, that is, those who probably fall into the persuadable group. In this approach, we indicate that, through the Transformed Results Approach, it is possible to increase the ability of decision-makers to evaluate the heterogeneity of effects and, at the same time, contribute to improving decision-making related to future social programs and projects.

Despite remarkable advances in Artificial Intelligence, its use remains predominantly restricted to sectors with substantial economic returns. Contrary to other fields, social intervention often faces investment constraints, as its sustainability derives primarily from public funding or patronage, resulting in a relative neglect of investment in new Artificial Intelligence and Machine Learning methodologies and algorithms. The emerging area of Causal Machine Learning is no exception, as it needs more research aimed at its applicability in assessing and prescribing the social impact of projects and programs. The

number of studies employing these techniques is scarce, except for some research on the impact evaluation of public policies. Thus, it is essential to invest in disseminating these processes among funders and promoters and training new impact evaluators in contemporary data science techniques, namely those related to Causal Machine Learning techniques.

To improve evaluation and prescription for a more effective intervention practice, academia must promote more applied research in this field, aggregating contributions from econometrics, computer science, and social and behavioral sciences. This chapter contributes to developing a new prescriptive vision, promoting the future design of more personalized social, educational, and community intervention projects and programs with a more significant impact on society.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for Comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. doi:10.1198/jasa.2009.ap08746

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *The American Economic Review*, *93*(1), 113–132. doi:10.1257/000282803321455188

Acharki, N., Lugo, R., Bertoncello, A., & Garnier, J. (2023). Comparison of metalearners for estimating multi-valued treatment heterogeneous effects. *ICML2023- Fortieth International Conference on Machine Learning*. Doi:10.48550/arXiv.2205.14714

Alves, M. (2022). Causal inference for the brave and true. *Matheus Facture*. https://matheusfacure.github.io/python-causality-handbook/landing-page.html

Aragon, C., Guha, S., Kogan, M., Muller, M., & Neff, G. (2022). *Human-centered data science: An introduction*. MIT Press.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7353–7360. doi:10.1073/pnas.1510489113 PMID:27382149

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*(2), 636–654. doi:10.1016/j.jeconom.2007.05.002

Crato, N., & Paruolo, P. (2019). *Data-driven policy impact evaluation: How access to microdata is transforming policy design*. Springer Nature. doi:10.1007/978-3-319-78461-8

Cunningham, S. (2021). *Causal inference: The mixtape*. Online Ebook Version. https://mixtape.scunning.com/

Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward developing prescriptive analytics. *Big Data*, *6*(1), 13–41. doi:10.1089/big.2017.0104 PMID:29570415

Fan, X., & Nowell, D. L. (2011). Using propensity score matching in educational Research. *Gifted Child Quarterly*, *55*(1), 74–79. doi:10.1177/0016986210390635

Fougère, D., & Jacquemet, N. (2019). Causal inference and impact evaluation. *Economie et Statistique/ Economics and Statistics*, (510-511-512), 181-200. doi:10.24187/ecostat.2019.510t.1996

Gama, J., Lorena, A., Faceli, K., Oliveira, O., & Carvalho, A. (2017). Extração de Conhecimento de dados: Data Mining. Edições Sílabo- 3ª Edição.

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, *44*(1), 218–243. doi:10.3102/0091732X20903302

Huntington-Klein, N. (2022). *The effect: An introduction to research design and causality*. Chapman and Hall/CRC Online Ebook Version. https://theeffectbook.net/index.html

Imbens, G., & Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. doi:10.1017/CBO9781139025751

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710. doi:10.1093/biomet/87.3.706

Jaskowski, M., & Jaroszewicz, S. (2012). Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis* (Vol. 46, pp. 79-95).

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). (Manuscript submitted for publication). Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475. Work (Reading, Mass.)*. doi:10.48550/arXiv.2206.15475

Kang, M., Ragan, B. G., & Park, J. H. (2008). Issues in outcomes research: An overview of randomization techniques for clinical trials. *Journal of Athletic Training*, *43*(2), 215–221. doi:10.4085/1062-6050-43.2.215 PMID:18345348

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *79*(4), 1229–1245. doi:10.1111/rssb.12212 PMID:28989320

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4156–4165. doi:10.1073/pnas.1804597116 PMID:30770453

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, *4*(3), 165–224. doi:10.1561/0800000014

Lechner, M. (2023). Causal Machine Learning and its use for public policy. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, *159*(1), 1–15. doi:10.118641937-023-00113-y

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. doi:10.1093/biomet/asaa076

OECD. (2002). *Glossary of key terms in evaluation and results-based management. DAC Network on Development Evaluation*. OECD.

OECD. (2021). *Social Impact Measurement for the Social and Solidarity Economy*. Doi:10.1787/20794797

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift modeling for preventing student dropout in higher education. *Decision Support Systems*, *134*, 113320. doi:10.1016/j.dss.2020.113320

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. doi:10.1093/biomet/82.4.669

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Pinheiro, P., & Cavique, L. (2022). Uplift Modeling Using the Transformed Outcome Approach. In G. Marreiros, B. Martins, A. Paiva, B. Ribeiro, & A. Sardinha (Eds.), Lecture Notes in Computer Science: Vol. 13566. *Progress in Artificial Intelligence. EPIA 2022*. Springer., doi:10.1007/978-3-031-16474-3_51

Rogers, P. (2014a). Overview of Impact Evaluation: Methodological Briefs - Impact Evaluation. *Methodological Briefs*. UNICEF Office of Research-Innocenti. https://www.unicef-irc.org/publications/746-overview-of-impact-evaluation-methodological-briefs-impact-evaluation-no-1.html

Rogers, P. (2014b). Theory of Change: Methodological Briefs - Impact Evaluation No. 2, *Methodological Briefs*. UNICEF Office of Research-Innocenti. https://www.unicef-irc.org/publications/747-theory-of-change-methodological-briefs-impact-evaluation-no-2.html

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in Observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi:10.1093/biomet/70.1.41

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

Stanley, K. (2007). Design of randomized controlled trials. *Circulation*, *115*(9), 1164–1169. doi:10.1161/CIRCULATIONAHA.105.594945 PMID:17339574

Tanai, Y., & Ciftci, K. (2023). How to customize an early start preparatory course policy to improve student graduation success: An application of uplift modeling. *Annals of Operations Research*. doi:10.100710479-023-05607-9

Thistlewaite, D. L., & Campbell, D. T. (2017). Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment. *Observational Studies*, *3*(2), 119–128. doi:10.1353/obs.2017.0000

Tu, C. (2019). Comparison of various machine learning algorithms for estimating generalized propensity score. *Journal of Statistical Computation and Simulation*, *89*(4), 708–719. doi:10.1080/00949655.2019.1571059

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. doi:10.1080/01621459.2017.1319839

White, H., Sabarwal, S., & de Hoop, T. (2014). Randomized Controlled Trials (RCTs): Methodological Briefs - Impact Evaluation No. 7, *Methodological Briefs*. UNICEF Office of Research-Innocenti. https://www.unicef-irc.org/publications/752-randomized-controlled-trials-rcts-methodological-briefs-impact-evaluation-no-7.html

ENDNOTE

[1] In this text, we have used the term "person" as a generic reference to facilitate the reading and understanding of impact assessment methods. However, these methods can be applied to other units of interest, such as organizations (schools, companies, hospitals, prisons, health facilities, etc.), localities, and groups of people.