A unified survey of treatment effect heterogeneity modelling and uplift modelling

Weijia Zhang* Jiuyong Li[1] Lin Liu[1]

* weijia.zhang.xh@gmail.com
[1]University of South Australia
April 27, 2021

Abstract

A central question in many fields of scientific research is to determine how an outcome is affected by an action, i.e., to estimate the causal effect or treatment effect of an action. In recent years, in areas such as personalised healthcare, sociology, and online marketing, a need has emerged to estimate heterogeneous treatment effects with respect to individuals of different characteristics. To meet this need, two major approaches have been taken: treatment effect heterogeneity modelling and uplifting modelling. Researchers and practitioners in different communities have developed algorithms based on these approaches to estimate the effect of heterogeneous treatment. In this paper, we present a unified view of these two seemingly disconnected yet closely related approaches under the potential outcome framework. We provide a structured survey of existing methods following either of the two approaches, emphasising their inherent connections and using unified notation to facilitate comparisons. We also review the main applications of the surveyed methods in personalised marketing, personalised medicine, and sociology. Finally, we summarise and discuss the available software packages and source codes in terms of their coverage of different methods and applicability to different datasets, and we provide general guidelines for method selection.

1 Introduction

A fundamental question for scientific research and applications in many disciplines is to determine whether and to what extent changing the value of one variable (i.e., a treatment) affects the value of another variable (i.e., an outcome), which is one of the main tasks in causal inference (Imbens and Rubin, 2015; Morgan and Winship, 2015). To answer this question, we need to estimate the causal effect of the treatment on the outcome. For example, an oncologist might estimate the average causal effect that a cancer therapy has on prognosis outcomes, such as the expected survival time after treatment. An employment assistance project might study the average causal effect of a job training program on employment

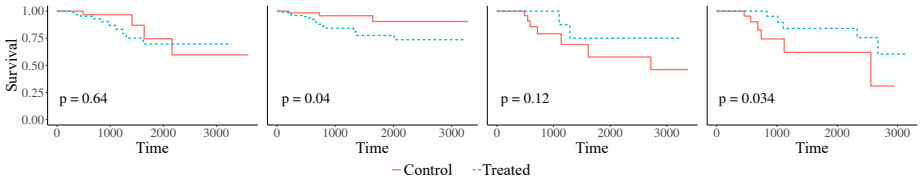


Figure 1: An example of the heterogeneous treatment effects of radiotherapy on the survival of breast cancer patients (Zhang et al., 2017). Patients in four different subgroups are distinguished by their gene expressions and have different survival times among treated patients (dashed lines) and control patients (solid lines). This means that patients in different subgroups may respond differently to radiotherapy.

prospects, i.e., whether the program reduces unemployment. An online retailer might want to model the average causal effect that an advertisement will have on sales.

However, it is often insufficient merely to determine the average causal effect. For example, a cancer patient may be more interested in individual-level causal effects, asking "Would this treatment be effective for a patient like me with a specific gene mutation?", since many cancer treatments are known to be effective for those with certain gene expression patterns (Bellon, 2015). From the perspective of policymakers, it is more reasonable to offer a job training program to those who will benefit from it, since the program's effectiveness may depend on the education background, experience, and employment history of the participants. For an online retailer, it is also preferable to target only persuadable customers in order to reduce advertisement costs, and to avoid disturbing customers who do not wish to receive unsolicited advertisements (Rzepakowski and Jaroszewicz, 2012b).

Motivated by the different application scenarios, researchers from two closely related yet surprisingly isolated research communities—the treatment effect heterogeneity modelling and the uplift modelling communities—have contributed significantly to solving this causal inference problem regarding conditional causal effects, i.e., causal effects in different sub-populations. Many methods have been developed by the treatment effect heterogeneity modelling community (Su et al., 2009; Hill, 2011; Su et al., 2012; Imai and Ratkovic, 2013; Athey and Imbens, 2015; Louizos et al., 2017; Atan et al., 2018; Wager and Athey, 2018; Zhang et al., 2017; Hassanpour and Greiner, 2018; Künzel et al., 2019) and the uplift modelling community (Hansotia and Rukstales, 2002; Eustache et al., 2018; Radcliffe, 2007; Rzepakowski and Jaroszewicz, 2010; Guelman et al., 2012; Zaniewicz and Jaroszewicz, 2013; Sołtys et al., 2015; Guelman et al., 2015a; Gutierrez and Gérardy, 2017; Yamane et al., 2018).

An exemplar application of treatment effect heterogeneity modelling is personalised medical treatment, where the goal is to determine whether a treatment is effective for a patient with certain characteristics, such as a certain gene

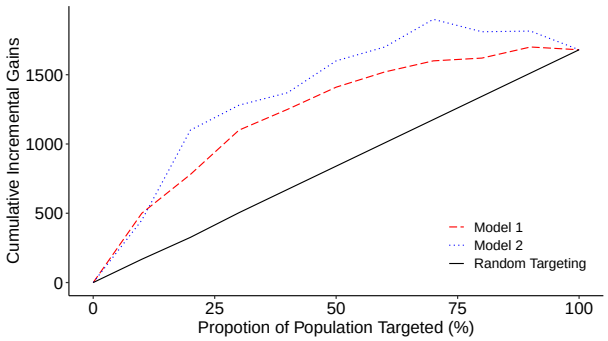


Figure 2: Incremental gain curves showing the increase in sales from targeting different proportions of customers with two uplift modelling models versus random targeting using data from an online retailer (Gubela et al., 2019).

expression profile. A causal tree-based method was developed to model the heterogeneous effects of radiotherapy on the survival time of breast cancer patients (Zhang et al., 2017). Four subgroups of patients characterised by their gene expressions responded differently to the radiotherapy treatment. The Kaplan–Meier survival curves (Kaplan and Meier, 1958) of the subgroups are shown in Figure 1. From the curves, we can see that the effects of the treatment are positive for patients in the third and fourth subgroups, negative for those in the second subgroup, and marginal for patients in the first subgroup. Correctly identifying treatment effect heterogeneity is beneficial for both patients and healthcare systems.

A typical application of uplift modelling is targeted advertising in online marketing, where the goal is to predict whether a promotion will be effective for customers with a certain purchase history and certain preferences. Uplift is a term in the marketing application and refers to the differences in the purchasing behaviour between customers who are offered the promotion (treated) and those who are not (control). As we will discuss in Section 2, with some assumptions, uplift is the conditional average causal effect of a treatment (promotion) on the outcome (purchase behaviour) for a given sub-population of customers with certain characteristics (e.g. purchase history and preferences). Figure 2 shows the incremental gain curves of smartphone sales by promotional emails from an e-commerce company with two uplift modelling methods (Gubela et al., 2019). We can see that when targeting the same proportion (10% to 100%) of customers, both uplift models achieved higher sales increases over random targeting. Model 2 performs better than Model 1 except when targeting only the first 10% of the customers.

Treatment effect heterogeneity modelling and uplift modelling share a common objective: to estimate the change in the outcome caused by the change of the treatment for some given subjects—e.g., changes in the survival time that result from radiotherapy treatment, or changes in purchasing behaviour that result

from promotional emails. However, a distinct difference between the algorithms from the two communities is that those from the uplift modelling community are implicitly designed for data from randomised experiments, and no assumptions for using the methods are explicitly discussed. By contrast, methods developed in the treatment effect heterogeneity modelling community are explicitly specified with assumptions and they can be used for experimental and observational data that satisfy these assumptions.

Over the last few years, several surveys have been published in the uplift modelling community (Gutierrez and Gérardy, 2017; Devriendt et al., 2018; Gubela et al., 2019). Gutierrez and Gérardy (Gutierrez and Gérardy, 2017) were the first to briefly discuss the link between uplift modelling and treatment effect heterogeneity modelling. However, to the best of our knowledge, no literature explicitly discusses the necessary assumptions for unifying the methods from the two communities. Furthermore, none of the existing work has discussed the connections between the methods developed by the two communities using unified notation. Researchers from the two communities have been working in parallel, and the progress in one community is not contributing to the progress of the other. This article contributes to bridging the gap between the two communities and accelerating the research progress by showing that the methods, benchmarks, evaluation metrics, and applications in both communities can be used to cross-fertilize each other.

Causal inference is a diverse research area, where the potential outcome framework (Rubin, 1974) and the structural causal models (Pearl, 2009) are two of the most influential paradigms. This work follows the potential outcome framework for estimating the causal effects of a treatment on an outcome, which is a central task in causal inference. Treatment effects can be estimated from either experimental or observational data (Imbens and Rubin, 2015; Morgan and Winship, 2015). We focus on estimating the conditional average treatment effect (CATE) within different sub-populations, whereas most treatment effect estimation methods estimate the average treatment effect in a general population. This article presents a unified survey of the CATE estimation methods developed by the treatment effect heterogeneity modelling community and the uplift modelling community. For more general topics on data-driven causal inference, including learning structural causal models and estimating average causal effects, please refer to the surveys of Knaus et al. (2020) and Guo et al. (2020).

This article serves as a timely survey that connects the literature from the two seemingly different yet closely related communities, and enables the cross-fertilisation of the methods for solving essentially the same causal inference problem. The survey contributes to the literature in the following ways. Firstly, we explicitly present and discuss the fundamental assumptions needed for unifying the methods proposed by the two communities under the potential outcome framework (Rubin, 1974). Secondly, we survey and discuss the methods proposed by both communities using unified notation, which allows us to clarify the connections and distinctions between the methods developed by the two communities. Thirdly, we discuss the feasibility of the methods by evaluating software packages developed by both communities, and we outline the challenges

for both treatment effect heterogeneity modelling and uplift modelling.

2 Unifying Treatment Effect Heterogeneity Modelling and Uplift Modelling

In this section, we discuss and unify the definitions, assumptions, and objectives for treatment effect heterogeneity modelling and uplift modelling under the potential outcome framework (Rubin, 1974).

2.1 Objectives of Treatment Effect Heterogeneity Modelling and Uplift Modelling

We use $T \in \{0, 1\}$ to denote a binary treatment variable, $T = 0$ for having no treatment (control), and $T = 1$ for having treatment (treated). Under the potential outcome framework, every subject i has two *potential outcomes*: $Y_i(0)$, the potential outcome if the subject had received no treatment; and $Y_i(1)$, the potential outcome if the subject had received the treatment. Potential outcomes can be either continuous or discrete. For example, in cancer treatment, $Y_i(0)$ corresponds to a continuous variable indicating the number of years the patient would have survived without treatment, and $Y_i(1)$ corresponds to the number of years the patient would have survived with the treatment. In marketing, binary potential outcomes are used to indicate whether a customer would purchase a product if that customer were offered a promotion ($Y_i(1)$) or not offered the promotion ($Y_i(0)$).

For subject i, the individual treatment effect (ITE) of a treatment T, denoted as τ_i, is defined as the difference between the two potential outcomes:

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

In an ideal world, if we could determine the ITE τ_i as defined in Equation (1), then the ultimate goal of both treatment effect heterogeneity modelling and uplift modelling, i.e., predicting individual treatment effect, would be fulfilled, and we would know exactly whether an individual should be prescribed a treatment.

Unfortunately, only one of the two potential outcomes can be observed for any subject. For example, if we observe the potential outcome of a cancer patient who receives a radiotherapy treatment, it will be impossible for us to observe the potential outcome of the same patient receiving no treatment. The unobserved potential outcome is often referred as the *counterfactual* of the observed outcome. We use Y_i to denote the observed outcome of subject i, which can be expressed using the interaction between the treatment and the two potential outcomes:

$$Y_i = TY_i(1) + (1 - T)Y_i(0). \tag{2}$$

Hereafter, when the context is clear, we drop the subscript i for simplicity of notation.

Since we never observe both potential outcomes for any subject, ITE τ_i is not identifiable. With some adequate assumptions, however, it is possible to estimate the conditional average treatment effect (CATE), which is the average treatment effect conditioning on a set of covariates that describe the subjects. Specifically, let $X = \boldsymbol{x} \in \mathbb{R}^k$ be a k-dimensional covariate vector describing the pre-treatment characteristics of a group of subjects. The CATE, denoted $\tau(\boldsymbol{x})$, is defined as:

$$\tau(\boldsymbol{x}) := \mathbb{E}[\tau \mid X = \boldsymbol{x}] = \mathbb{E}[Y(1) - Y(0) \mid X = \boldsymbol{x}]. \tag{3}$$

Equation (3) is frequently used as the objective of treatment effect heterogeneity modelling methods (Athey and Imbens, 2016; Zhang et al., 2017; Künzel et al., 2019). Although the CATE is not the same as the ITE, it has been shown that the CATE is the best estimator for the ITE in terms of the mean squared error (Künzel et al., 2019).

Uplift modelling techniques assume that data are obtained from experiments with randomised treatment assignment, and have the following objective:

$$\mathrm{Upl}(\boldsymbol{x}) = \mathbb{E}(Y \mid T = 1, X = \boldsymbol{x}) - \mathbb{E}(Y \mid T = 0, X = \boldsymbol{x}). \tag{4}$$

From Equation (3), the objective of treatment effect heterogeneity modelling involves the conditional expectations of two potential outcomes, and thus cannot be directly estimated from data. By contrast, the objective of uplift modelling (Equation (4)) involves two conditional expectations of the observed outcomes and thus can be estimated from data, but assuming the use of randomised experiment data for unbiased estimation. In the next section, we discuss the link between the objectives and the assumptions of treatment effect heterogeneity modelling and uplift modelling.

2.2 Linking the Objectives

We can re-arrange the objective of treatment effect heterogeneity modelling, i.e., Equation (3), using rules of conditional probability and conditional expectations, as follows:

$$\begin{aligned}
\tau(\boldsymbol{x}) =& \mathbb{E}[Y(1) - Y(0) \mid X = \boldsymbol{x}] \\
=& \mathbb{E}[Y(1) \mid T = 1, X = \boldsymbol{x}]P(T = 1) + \mathbb{E}[Y(1) \mid T = 0, X = \boldsymbol{x}]P(T = 0) \\
& - \mathbb{E}(Y(0) \mid T = 1, X = \boldsymbol{x})P(T = 1) - \mathbb{E}[Y(0) \mid T = 0, X = \boldsymbol{x}]P(T = 0) \\
=& \mathbb{E}[Y(1) \mid T = 1, X = \boldsymbol{x}]P(T = 1) + \mathbb{E}[Y(1) \mid T = 0, X = \boldsymbol{x}]P(T = 0) \\
& - \mathbb{E}(Y(0) \mid T = 1, X = \boldsymbol{x})P(T = 1) - \mathbb{E}[Y(0) \mid T = 0, X = \boldsymbol{x}]P(T = 0) \\
& + \mathbb{E}[Y(1) \mid T = 1, X = \boldsymbol{x}]P(T = 0) - \mathbb{E}[Y(1) \mid T = 1, X = \boldsymbol{x}]P(T = 0) \\
& + \mathbb{E}[Y(0) \mid T = 0, X = \boldsymbol{x}]P(T = 1) - \mathbb{E}[Y(0) \mid T = 0, X = \boldsymbol{x}]P(T = 1) \\
=& \mathbb{E}[Y(1) \mid T = 1, X = \boldsymbol{x}]P(T = 1) + \mathbb{E}[Y(1) \mid T = 1, X = \boldsymbol{x}]P(T = 0) \\
& - \mathbb{E}[Y(0) \mid T = 0, X = \boldsymbol{x}]P(T = 0) - \mathbb{E}[Y(0) \mid T = 0, X = \boldsymbol{x}]P(T = 1) \\
& + \mathbb{E}[Y(0) \mid T = 0, X = \boldsymbol{x}]P(T = 1) - \mathbb{E}(Y(0) \mid T = 1, X = \boldsymbol{x})P(T = 1) \\
& + \mathbb{E}[Y(1) \mid T = 0, X = \boldsymbol{x}]P(T = 0) - \mathbb{E}[Y(1) \mid T = 1, X = \boldsymbol{x}]P(T = 0),
\end{aligned}$$

which gives us:

$$\begin{aligned}\tau(\boldsymbol{x}) =& \underbrace{\mathbb{E}[Y(1) \mid T=1, X=\boldsymbol{x}] - \mathbb{E}[Y(0) \mid T=0, X=\boldsymbol{x}]}_{\text{observed}} \\ &+ P(T=1)\{\underbrace{\mathbb{E}[Y(0) \mid T=0, X=\boldsymbol{x}]}_{\text{observed}} - \underbrace{\mathbb{E}[Y(0) \mid T=1, X=\boldsymbol{x}]}_{\text{unobserved}}\} \\ &+ P(T=0)\{\underbrace{\mathbb{E}[Y(1) \mid T=0, X=\boldsymbol{x}]}_{\text{unobserved}} - \underbrace{\mathbb{E}[Y(1) \mid T=1, X=\boldsymbol{x}]}_{\text{observed}}\}. \end{aligned} \tag{5}$$

The above process decomposes the objective into three components. For the two conditional expectations in the first line of Equation (5), observing that the potential outcome of treatment (or control) equals the observed outcome when conditioning on $T = 1$ (or $T = 0$), we have:

$$\mathbb{E}[Y(1) \mid T=1, X=\boldsymbol{x}] - \mathbb{E}[Y(0) \mid T=0, X=\boldsymbol{x}] = \mathbb{E}[Y \mid T=1, X=\boldsymbol{x}] - \mathbb{E}[Y \mid T=0, X=\boldsymbol{x}]. \tag{6}$$

It is worth noting that the expectations on the right-hand side of Equation (6) are the same as the uplift modelling objective in Equation (4). Furthermore, both expectations only involve observed outcomes, without any counterfactuals. Their values can be estimated without bias using the overlap, the stable unit treatment value (SUTV), and the unconfoundedness assumption described below.

Assumption 1. ***(Overlap)*** *Any subject has a non-zero probability of receiving the treatment and the control. In other words, for all $\boldsymbol{x}$ in the support of X, we have:*

$$0 < P(T=1 \mid X=\boldsymbol{x}) < 1. \tag{7}$$

The overlap assumption states that the probability of any subject with covariates $\boldsymbol{x}$ being treated is bounded away from 0 and 1. This ensures that all types of individuals have been observed in both treatment and control groups. This is necessary because if subjects with some covariate value $\boldsymbol{x}$ always receive treatment (or control) in the data, the expectations cannot be estimated.

Assumption 2. ***(SUTV)*** *The stable unit treatment value (SUTV) assumption states that the individuals do not interfere with each other. In other words, a treatment applied to one subject does not affect the outcome of other subjects.*

The SUTV assumption is usually satisfied in health and bioinformatics applications since it is reasonable to assume that giving radiotherapy to a patient will not affect the life expectancy of other patients. However, the assumption should be considered carefully in marketing where treatment to one subject may affect the outcomes of other subjects. For example, in online advertising, there is no guarantee that sending an email promotion (treatment) to an individual

will not affect other individuals' knowledge of the promotion. For example, a treated individual could forward the email to friends and thus their decision to purchase the product might change.

At this point, we have seen that the differences between the objective of treatment effect heterogeneity modelling and the objective of uplift modelling lie in the last two components of Equation (5). These components are not estimable from data, since each of them involves an unobservable potential outcome (counterfactual), $\mathbb{E}[Y(0) \mid T = 1, X = \boldsymbol{x}]$, the conditional average of potential outcome $Y(0)$ in the treatment group, and $\mathbb{E}[Y(1) \mid T = 0, X = \boldsymbol{x}]$, the conditional average of potential outcome $Y(1)$ in the control group. To overcome the counterfactual problem, methods in the treatment effect heterogeneity modelling literature have introduced an important assumption, namely, that potential outcomes are independent of treatment assignment when conditioning on a set of covariates. This is known as the unconfoundedness assumption:

Assumption 3. ***(Unconfoundedness)*** *The distribution of treatment is independent of the distribution of potential outcomes when conditioning on a set of observed variables. Formally, we have:*

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X = \boldsymbol{x}. \tag{8}$$

The unconfoundedness assumption is also often referred as the strong ignorability assumption in the literature. By applying the unconfoundedness assumption to the last two components of Equation (5), we get the following results:

$$\mathbb{E}[Y(0) \mid T = 0, \boldsymbol{x}] - \mathbb{E}[Y(0) \mid T = 1, \boldsymbol{x}] = \mathbb{E}[Y(0) \mid \boldsymbol{x}] - \mathbb{E}[Y(0) \mid \boldsymbol{x}] = 0, \tag{9}$$
$$\mathbb{E}[Y(1) \mid T = 0, \boldsymbol{x}] - \mathbb{E}[Y(1) \mid T = 1, \boldsymbol{x}] = \mathbb{E}[Y(1) \mid \boldsymbol{x}] - \mathbb{E}[Y(1) \mid \boldsymbol{x}] = 0, \tag{10}$$

and the objective of treatment effect heterogeneity can be written as:

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y \mid T = 1, X = \boldsymbol{x}] - \mathbb{E}[Y \mid T = 0, X = \boldsymbol{x}], \tag{11}$$

which is the same as the objective of uplift modelling given in Equation (4).

Therefore, when the assumptions of overlap, SUTV, and unconfoundedness are all satisfied, the objective of treatment effect heterogeneity modelling (Equation (3)) and the objective of uplift modelling (Equation (4)) are the same. However, it is worth mentioning that the above three assumptions are not explicitly stated or discussed in the majority of uplift modelling literature. If the data do not satisfy these assumptions, the estimated uplift will be biased, since Equation (4) does not correspond to the true causal effect of an action on the outcome.

In practice, it is challenging to determine whether the unconfoundedness assumption is satisfied, as it is untestable from data. In other words, it is hard to know if the covariate set used is correct when exploring treatment heterogeneity or building uplift models. This is an important related topic, but one that is outside the scope of our survey. A simple criterion for covariate selection is to

have all direct causes of outcome Y as covariates and exclude all effect variables of Y from the covariate set, as shown in (Li et al., 2021). For an in-depth discussion of the topic, we refer the reader to VanderWeele and Shpitser (2011); de Luna et al. (2011); Entner et al. (2013); Maathuis et al. (2015).

3 Methods for CATE Estimation

In this section, assuming that the data satisfy the overlap, SUTV and unconfoundedness assumptions (i.e., Assumptions 1, 2, and 3), we provide an extensive survey of the existing treatment effect heterogeneity modelling and uplift modelling algorithms using unified notation. We categorise the methods for treatment effect heterogeneity modelling and uplift modelling into two major categories. The first category consists of methods that extend existing supervised learning methods for CATE estimation, and the second category consists of tailored methods for treatment effect heterogeneity modelling or uplift modelling.

3.1 Methods Extending Existing Supervised Learning Models

3.1.1 The Single-model Approach

Estimating the CATE in Equation (11) can be achieved by estimating conditional expectations (or probabilities) from data. The problem reduces to a regression or classification problem. Specifically, given a dataset D for (Y, X, T) that satisfies Assumptions 1–3, the single-model approach uses the concatenation of treatment and covariates $[T, X]$ as the features, and Y as the target to train a supervised model, where $Y = \hat{\mu}(T, X)$ from D, e.g., a regression model if Y is a continuous variable, or a classification model if Y is a categorical variable. Then, the trained model is used to predict the CATE for a subject described by covariates $\boldsymbol{x}$ as follows:

$$\hat{\tau}(\boldsymbol{x}) = \hat{\mu}([T = 1, \boldsymbol{x}]) - \hat{\mu}([T = 0, \boldsymbol{x}]), \tag{12}$$

where $\boldsymbol{x}$ is the short-hand form for $X = \boldsymbol{x}$.

Lo (Lo, 2002) used two estimators, a logistic regression model and a neural network with one hidden layer, to estimate the uplift in a market experimental dataset. Lo (Lo, 2002) also proposed that standard supervised methods such as linear regression, regression tree, and spline regression could be used as the estimators. Athey and Imbens (Athey and Imbens, 2015) implemented a single-model method using a regression tree for treatment effect heterogeneity modelling.

The single-model approach is simple, easy to implement, and has the flexibility of being able to use any off-the-shelf supervised learning algorithm. However, a major drawback to this approach is that a single model may not model both potential outcomes well, and hence the estimation of CATE may be biased. Another problem with the single-model approach is that T may not be selected

by a model that only uses a subset of the features for prediction (such as a tree model), and thus the CATE will be estimated as zero for all subjects.

3.1.2 The Two-Model Approach

An improvement to the single-model approach is to model the two potential outcomes using two separate models. The two-model approach trains two models on the control and treated subjects separately, and then uses the difference of the two predictions as the estimated conditional average causal effect or uplift.

Specifically, given a dataset D for (Y, X, T) that satisfies Assumptions 1–3, the two-model approach estimates the CATE as:

$$\begin{aligned}\hat{\tau}(\boldsymbol{x}) &= \mathbb{E}[Y|T=1, X=\boldsymbol{x}] - \mathbb{E}[Y \mid T=0, X=\boldsymbol{x}] \\ &= \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x}),\end{aligned}$$

where $\hat{\mu}_1$ is trained using the sub-dataset of D containing the samples of treated subjects only, and $\hat{\mu}_0$ is trained using the sub-dataset of D containing the samples of control subjects only.

Any off-the-shelf estimator can be used to estimate $\hat{\mu}_1(\boldsymbol{x})$ and $\hat{\mu}_0(\boldsymbol{x})$. Popular choices include linear regression, as in (Hansotia and Rukstales, 2002; Cai et al., 2011); regression trees (Breiman et al., 1984), as in the two-tree method in (Athey and Imbens, 2015); decision trees (Quinlan, 1993), as in (Sołtys et al., 2015); rule-based methods, as in (Nassif et al., 2012a, 2013); gradient boosting trees, as in (Kane et al., 2014); and Bayesian additive regression trees (BART) (Chipman et al., 2010), as in (Hill, 2011).

The two-model approach is simple and flexible. The models for the control and treated subjects can be built straightforwardly using a wide range of off-the-shelf estimators. The freedom of choice in the estimators also provides flexibility for modelling various treatment and outcome relationships. However, as the two models are built separately, they do not utilise the information shared by the control and treated subjects. Furthermore, they cannot mitigate the impact of the disparity in covariate distributions between the treatment and control groups on CATE estimation (Athey and Imbens, 2015).

3.1.3 X-Learner

Künzel et al. (2019) proposed an improvement to the two-model approach called the X-Learner. A motivation for the X-Learner approach is that the sample size in the treatment group is usually very small, and thus the estimator $\hat{\mu}_1(\boldsymbol{x})$ may not be modelled accurately. The X-Learner addresses this problem by information crossover between treatment and control groups.

Specifically, the X-Learner consists of three steps. First, two separate estimators, $\hat{\mu}_1(\boldsymbol{x})$ and $\hat{\mu}_0(\boldsymbol{x})$, are built using the subjects from the treatment and control groups, respectively, similar to the model building process in the two-model approach. Then, the treatment effect for a subject in the treatment group, denoted as $\hat{\tau}_{1i}$, is inputted using the observed outcome and the estimator $\hat{\mu}_0(\boldsymbol{x})$ from the control group. The treatment effects for a subject in the control

group, denoted as $\hat{\tau}_{0i}$, is inputted using the observed outcome and the estimator $\hat{\mu}_1(\boldsymbol{x})$ from the treatment group. That is,

$$\hat{\tau}_{1i} = Y_i - \hat{\mu}_0(\boldsymbol{x}_i), \text{for subject } i \text{ belonging to the treated group, and}$$
$$\hat{\tau}_{0i} = \hat{\mu}_1(\boldsymbol{x}_i) - Y_i, \text{for subject } i \text{ belonging to the control group.}$$

Now we have two sets of inputted CATE estimations: $\hat{\tau}_1$, corresponding to the CATE estimations for subjects in the treatment group; and $\hat{\tau}_0$, corresponding to the CATE estimations for subjects in the control group. Using these two sets of inputted CATE estimations, the X-Learner builds two estimators, $\hat{\tau}_1(\boldsymbol{x})$ and $\hat{\tau}_0(\boldsymbol{x})$, using covariates $\boldsymbol{x}$, with $\hat{\tau}_1$ and $\hat{\tau}_0$, respectively. Finally, the CATE is estimated using a weighted average of the two estimators:

$$\hat{\tau}(\boldsymbol{x}) = e(\boldsymbol{x})\hat{\tau}_0(\boldsymbol{x}) + (1 - e(\boldsymbol{x}))\hat{\tau}_1(\boldsymbol{x}), \tag{13}$$

where $e(\boldsymbol{x}) \in [0, 1]$ is a weight function defined as the estimated probability of a subject receiving the treatment, i.e., $\hat{e}(\boldsymbol{x}) = P(T = 1|\boldsymbol{x})$, as suggested in Künzel et al. (2019). This probability is referred to as the propensity score Rubin (1997) and in practice is often estimated using logistic regression, decision trees, neural networks (Setoguchi et al., 2008), or boosting algorithms (McCaffrey et al., 2004). For a comprehensive discussion of propensity score, see Austin (Austin, 2011).

An advantage of the X-Learner is that it cross-references the data in the treatment and control groups, and thus can perform better than the two-model approach when the number of subjects in the treatment group is significantly smaller than that in the control group. However, the X-Learner requires building four estimators, twice as many as the two-model approach. This increases the risk of overfitting and the difficulty tuning parameters.

3.1.4 R-Learner

Recently, Nie and Wager (2020) proposed the R-Learner, which is a class of two-step algorithms designed for treatment effect heterogeneity modelling. R-Learner first estimates the treatment outcome and control outcomes $\hat{\mu}(\boldsymbol{x})$ along with the propensity score $\hat{e}(\boldsymbol{x})$ using a base learner trained by k-fold cross-validation, e.g., penalised regression, neural networks, or boosting. Then, it estimates the CATE by minimising the following loss function:

$$L[\hat{\tau}(\boldsymbol{x})] = \frac{1}{n}\sum_{i=1}^{N}\{[(Y_i - \hat{\mu}^{(-i)}(\boldsymbol{x}_i)] - [T_i - \hat{e}^{(-i)}(\boldsymbol{x}_i)]\hat{\tau}(\boldsymbol{x}_i)\}^2, \tag{14}$$

where $\hat{\mu}^{(-i)}(\boldsymbol{x}_i)$ and $\hat{e}^{(-i)}(\boldsymbol{x}_i)$ denote the predictions made from the samples excluding the cross-validation fold that the i-th training sample belongs.

An advantage of the R-Learner is that it has guaranteed error bounds when using penalised kernel regression as the base learner. Nie and Wager (2020) showed that the asymptotic error bound of the R-Learner is the same as the bound of an oracle learner that has access to how the outcomes and treatments are generated, but not to the ground-truth treatment effects.

3.1.5 Transformed outcome Approach

The transformed outcome approach transforms the observed outcome Y to Y^* such that the CATE equals the conditional expectation of the transformed outcome Y^*. After the transformation, an off-the-shelf estimator can be applied to the dataset containing the original covariates and the transformed outcomes for estimating the CATE.

For example, the following transformed outcome for continuous outcomes is used in (Athey and Imbens, 2015):

$$Y^* = \frac{Y}{e(\boldsymbol{x})} \cdot T - \frac{Y}{(1 - e(\boldsymbol{x}))} \cdot (1 - T), \tag{15}$$

where Y^* is the transformed outcome, and $e(\boldsymbol{x})$ is the probability of a subject with covariates $\boldsymbol{x}$ receiving the treatment, i.e., $e(\boldsymbol{x}) = P(T = 1|\boldsymbol{x})$.

Using the transformed outcome described above, it is straightforward to derive that the conditional expectation of the transformed outcome equals the CATE, i.e., $\mathbb{E}[Y^*|X = \boldsymbol{x}] = \tau(\boldsymbol{x})$. Therefore, an off-the-shelf regression algorithm can be used to estimate the CATE by using Y^* as the target and X as the covariates. In the treatment effect heterogeneity modelling community, an instantiation of this approach using regression trees is discussed by Athey and Imbens (2015).

For a binary outcome, Jaskowski and Jaroszewicz (2012) proposed the following transformation:

$$Y^* = YT + (1 - Y)(1 - T), \tag{16}$$

where the transformed outcome Y^* corresponds to one of the following cases: $Y^* = 1$ when ($T = 1$ and $Y = 1$) or ($T = 0$ and $Y = 0$); and $Y^* = 0$ otherwise. Under the assumption that $e(\boldsymbol{x}) = 0.5$ for all $\boldsymbol{x}$, i.e., that a subject has an equal chance to be in the treatment or the control group, Jaskowski and Jaroszewicz (2012) proved that the CATE can be estimated as:

$$\tau(\boldsymbol{x}) = 2P(Y_i^* = 1|\boldsymbol{x}) - 1. \tag{17}$$

The transformed outcome in Equation (17) can be viewed as a special case of the transformation in Equation (15), where Y is binary and $e(\boldsymbol{x}) = 0.5$ for all $\boldsymbol{x}$. Jaskowski and Jaroszewicz (2012) built a logistic regression model using the transformed outcomes. Weisberg and Pontes (2015) and Pechyony et al. (2013) also used the same transformation and logistic regression as the base learner.

A main advantage of the transformed outcome approach is that, after transformation, the CATE can be modelled directly. Furthermore, it provides the flexibility for choosing any existing off-the-shelf supervised methods for CATE estimation. However, the transformed outcome approach relies heavily on the accurate estimation of $e(\boldsymbol{x})$, the propensity score. In Equation (15) the estimated $e(\boldsymbol{x})$ appears in the denominator, and thus a small variation in the estimation of $e(\boldsymbol{x})$ will lead to a large variation in the transformed outcomes.

3.1.6 Deep Learning-Based Methods

Recently, several deep learning-based treatment effect heterogeneity modelling algorithms have been proposed (Shalit et al., 2017; Louizos et al., 2017; Yao et al., 2018; Hassanpour and Greiner, 2018). Here, we introduce three types of deep learning-based methods that extend the single-model approach, the two-model approach, and the X-Learner approach. Some other deep learning-based algorithms that do not fall into the above three categories, i.e., those specifically designed as neural network-based methods, are discussed later in Section 3.2.3.

The main advantages of deep learning-based methods are that they can model complex non-linear relationships between the treatment, covariates, and the outcome, and can handle high-dimensional and large data. However, deep learning-based methods are difficult to interpret and they offer no convergence guarantees or error bounds. Furthermore, these methods are sensitive to the selection of parameters and their parameter tuning procedures are difficult.

Deep-Treat (Atan et al., 2018) Deep-Treat is a deep learning-based single-model approach that consists of two stages. The first stage takes the covariates and the treatment as input, and utilises a debiasing auto-encoder (Vincent et al., 2010) to learn a representation $\Phi(\boldsymbol{x})$ of the original covariates $\boldsymbol{x}$ such that the treatment and control groups are balanced. Balancing of the learned representation $\Phi(\boldsymbol{x})$ is measured by the cross-entropy loss between the marginal treatment distribution $P(T)$ and the conditional treatment distribution, given the learned representation $P(T|\Phi(\boldsymbol{x}))$. In other words, the learned representations are considered balanced if the cross-entropy loss between the marginal distribution and the conditional distribution is minimised. In the second stage, Deep-Treat uses the learned representation $\Phi(\boldsymbol{x})$, the treatment T, and the outcome Y as inputs, and trains a single neural network to predict the outcome Y using the concatenated features $[T, \Phi(\boldsymbol{x})]$. The main advantage of Deep-Treat over the single-model is that Deep-Treat learns a balanced encoding for the control and treated subjects. However, it also inherits the disadvantages of the single-model approach.

Counterfactual Regression (CFR) (Johansson et al., 2016; Shalit et al., 2017) CFR is a deep learning-based method that extends the two-model approach. CFR estimates $\tau(\boldsymbol{x})$ by learning two functions parameterised by two neural networks (similar to the two-model approach). Before learning the two functions, CFR utilises representation learning to minimise the discrepancy between the two distributions, $P(\boldsymbol{x} \mid T = 0)$ and $P(\boldsymbol{x}|T = 1)$, measured by either the maximum mean discrepancy or the Wasserstein distance.

Several works have been proposed to improve upon CFR. In particular, CFR with importance sampling weights (CFR-ISW) (Hassanpour and Greiner, 2018) is based on the problem that the representation learned by CFR cannot completely eliminate bias. Thus, CFR-ISW adds a propensity network to alleviate this problem. The similarity-preserved individual treatment effect (SITE) estimation algorithm (Yao et al., 2018) improves the learning of the common representation

by adding a position-dependent deep metric (PPDM) and middle-point distance minimisation (MPDM) constraints.

The main advantage of CFR and its improvements over the two-model approach is that they are designed to learn a shared and balanced feature representation across the treated and control subjects, to reduce bias in the CATE estimation.

3.2 Tailored Methods for Treatment Effect Heterogeneity Modelling or Uplift Modelling

In this section, we survey methods that are specifically designed for treatment effect heterogeneity modelling or uplift modelling. We categorise the methods into four categories: tree-based methods, which build binary trees using designed splitting criteria; support vector machine (SVM)-based methods, which reformulate uplift modelling within the SVM framework; generative deep learning methods, which utilise a variational autoencoder or generative adversarial network to estimate the potential outcomes; and ensemble-based methods, which build upon tree-based methods.

3.2.1 Tree-Based Methods

Tree-based methods build binary tree models for estimating the CATE. Both treatment effect heterogeneity modelling and uplift modelling communities have developed tree-based methods separately.

The procedure of building a tree-based CATE estimation model is similar to that of building a normal regression/decision tree (Breiman et al., 1984), in the sense that they all build tree models using recursive partitioning, which, starting from the root node, recursively splits the node into two child nodes using a splitting criterion. The major difference between existing decision/regression trees and tree-based CATE estimation algorithms lies in how they define their splitting criteria. In the same fashion, the main difference among different tree-based CATE estimation algorithms also lies in their splitting criteria. Therefore, we focus our discussion on the difference in the splitting criteria of tree-based CATE estimation algorithms.

The main advantage of tree-based methods lies in the interpretability, which is very important for many applications of CATE estimation. Furthermore, tree-based methods naturally provide groups of subjects with heterogeneous CATEs as defined by the paths from the root to leaf nodes of a tree model. However, a major drawback to all tree-based methods is that the tree construction process is fundamentally greedy and does not return the "optimal" tree. One tree and an alternative tree (by slightly perturbing data) from the same algorithm can differ significantly.

Uplift Incremental Value Modelling (UpliftIVM) Hansotia and Rukstales (2002) UpliftIVM is one of the earliest tree-based methods proposed

in the uplift modelling community. UpliftIVM searches for a splitting point that maximises the following criterion:

$$\mathcal{C}^{\text{UpliftIVM}} := |\hat{\tau}_L - \hat{\tau}_R|, \tag{18}$$

where $\hat{\tau}_L$ and $\hat{\tau}_R$ are the estimated conditional average treatment effect within the left and right child nodes, respectively. In other words, UpliftIVM aims to find the split that maximises the difference between the estimated CATEs of the two child nodes.

Specifically, in UpliftIVM the within-node CATEs $\hat{\tau}_L$ and $\hat{\tau}_R$ are estimated as the average outcome difference between the treatment and control groups using training data within the node, i.e., $\hat{\tau}_L = \frac{\sum_{i=1}^{n_L} T_i Y_i}{\sum_{i=1}^{n_L} T_i} - \frac{\sum_{i=1}^{n_L} (1-T_i) Y_i}{\sum_{i=1}^{n_L} (1-T_i)}$ and $\hat{\tau}_R = \frac{\sum_{i=1}^{n_R} T_i Y_i}{\sum_{i=1}^{n_R} T_i} - \frac{\sum_{i=1}^{n_R} (1-T_i) Y_i}{\sum_{i=1}^{n_R} (1-T_i)}$, where n_L and n_R denote the number of subjects in left and right child nodes, respectively.

An advantage of UpliftIVM is its simplicity, and it performs well when the magnitude of CATE heterogeneity is large. However, it is prone to outliers and spurious treatment effect heterogeneities.

Squared t-Statistics Tree (t-stats) (Su et al., 2009) t-stats is an early tree-based algorithm for modelling treatment effect heterogeneity. It builds a tree model by seeking the split with the largest value of the squared t-statistic for testing the null hypothesis that the average treatment effect is the same in the two potential child nodes. The t-stats tree maximises the following splitting criterion:

$$\mathcal{C}^{tstats} := \frac{(\hat{\tau}_L - \hat{\tau}_R)^2}{\hat{\sigma}^2 (1/n_{1L} + 1/n_{0L} + 1/n_{1R} + 1/n_{0R})}, \tag{19}$$

where $\hat{\sigma}^2 = \sum_{i \in \{0,1\}} \sum_{j \in \{L,R\}} \frac{n_{ij}-1}{n-4} \sigma_{ij}^2$ and $n = n_{0L} + n_{0R} + n_{1L} + n_{1R}$. Here, n_{1L} (n_{1R}) and n_{0L} (n_{0R}) denote the number of treated and control subjects, respectively, in the left (right) child node; and σ_{1L}^2 (σ_{1R}^2) and σ_{0L}^2 (σ_{0R}^2) are the sample variances of the treated and control subjects, respectively, in the left (right) child node.

The within-node CATEs $\hat{\tau}_L$ and $\hat{\tau}_R$ are estimated in the training data in the same way as in UpliftIVM discussed above. t-stats differs from UpliftIVM in that t-stats (Equation (19)) uses a pooled variance estimator (for estimating the common sample variances of various populations with different means) to normalise the difference of within-node CATEs.

Uplift Decision Tree (UpliftDT) (Rzepakowski and Jaroszewicz, 2010) UpliftDT is developed in the uplift modelling community for binary outcomes, and its motivation is different from the previously described tree-based methods. The previous tree-based methods aim to find the split that maximises the difference between the estimated CATEs of the left and right child nodes, whereas UpliftDT aims to maximise the estimated CATE within each child node. Specifically,

UpliftDT maximises the following splitting criterion:

$$\mathcal{C}^{Eu} = \frac{n_L}{n}\hat{\tau}_L^2 + \frac{n_R}{n}\hat{\tau}_R^2, \tag{20}$$

where $\hat{\tau}_L$ and $\hat{\tau}_R$ are estimated as the within-node CATEs in the training data as before.

The above splitting criterion is referred to as the Euclidean criterion by Rzepakowski and Jaroszewicz (2010). To see this, note that for the binary outcome Y, the treatment effect within a node can be written as $\hat{\tau} = P(Y|T = 1) - P(Y \mid T = 0)$, and thus $\hat{\tau}^2$ can be viewed as the Euclidean distance between the treated and control subjects within the node.

Rzepakowski and Jaroszewicz (2010) proposed two splitting criteria based on KL divergence and χ^2 divergence. However, they argued that the Euclidean criterion is superior because it is more stable than the other criteria and has the important property of being symmetric. It is worth noting that the χ^2 splitting criterion has also been investigated by others in the uplift modelling community. In the work of Michel et al. (2017), a similar tree-based method was proposed utilising the χ^2 divergence.

A benefit of the splitting criterion in UpliftDT is that it can be extended to handle categorical outcomes and multiple branch splitting for a decision tree. To see this, for a p-way split, we can rewrite the splitting criterion in Equation (20) as $\mathcal{C}^{Eu} = \sum_{i=1}^{p} \frac{n_p}{n}\hat{\tau}_p^2$. UpliftDT has also been extended to handle multiple treatments by Rzepakowski and Jaroszewicz (2012a).

Balance-Based or Significance-Based Uplift Tree (Radcliffe and Surry, 2011) These two types of trees are designed in the uplift modelling community for binary outcomes.

The balance-based uplift tree aims to maximise the uplift difference in two splitting nodes while minimising the difference in size between the nodes. The following splitting criterion is used:

$$\mathcal{C}^{BL} := |\hat{\tau}_L - \hat{\tau}_R \mid (1 - \mid \frac{n_L - n_R}{n_L + n_R} \mid^{\alpha}), \tag{21}$$

where n_L and n_R are the number of subjects in the left and right nodes, respectively, $0 \leq \alpha \leq 1$ is a hyperparameter, and $\hat{\tau}_L$ and $\hat{\tau}_R$ are the within-node CATEs in the training data as before.

The significance-based uplift tree uses the significance of the interaction between the treatment variable and a candidate splitting variable as a measure for the splitting quality. In each partition, the data in the current node are fitted with a linear model where each candidate split variable and the treatment form an interaction term. The significance of the interaction is tested by a t- statistic:

$$\mathcal{C}^{SIG} := \frac{(n-4)(\tau_L - \tau_R)^2}{\text{SSE} \cdot (1/n_{1L} + 1/n_{0L} + 1/n_{1R} + 1/n_{0R})}, \tag{22}$$

where $\text{SSE} = \sum_{i\in\{1,0\}} \sum_{j\in\{L,R\}} n_{ij} P_{ij}(Y = 1)(1 - P_{ij}(Y = 1))$, and $\hat{\tau}_L$ and $\hat{\tau}_R$ are estimated as the within-node CATEs in the training data as before.

The splitting criterion of the balance-based uplift tree is closely related to the criterion of the UpliftIVM. Furthermore, for the significance-based uplift tree, it can be seen that the SSE in Equation (22) is the weighted sum of the population variances. Contrasting this criterion with the one from t-stats (Equation (19)), it can be seen that the two criteria are equivalent except that t-stats uses sample variances instead of population variances in the denominator.

Causal Inference Tree (CIT) (Su et al., 2012) A CIT is a tree-based method for treatment effect heterogeneity modelling, and it was proposed by the same author as the t-stats tree. The CIT assumes that the potential outcomes, $Y(0)$ and $Y(1)$, come from two Gaussian distributions with the same variance. In other words, $Y(T) \sim \mathcal{N}(T\mu_1 + (1-T)\mu_0, \sigma^2)$, where σ^2 is the variance, and μ_1 and μ_0 are the means of the treatment and control outcomes, respectively. Furthermore, the CIT assumes that the treatment follows a Bernoulli distribution $T \sim \text{Bernoulli}(\pi)$. Using these assumptions, the CIT then proposes to find the split that maximises the following log-likelihood within the node:

$$\mathcal{C}^{CIT} := -\frac{n_L}{2} \cdot \ln(n_L \text{SSE}_L) - \frac{n_R}{2} \cdot \ln(n_R \text{SSE}_R) + n_{1L} \ln n_{1L} + n_{0L} \ln n_{0L} + n_{1R} \ln n_{1R} + n_{0R} \ln n_{0R}, \tag{23}$$

where SSE_L and SSE_R are the sum of squared errors for the left and right child nodes, respectively. SSE_R is defined as $\text{SSE}_R = \sum_{i=1}^{N_{1R}} (Y_i - \hat{Y}_1)^2 + \sum_{i=1}^{N_{0R}} (Y_i - \hat{Y}_0)^2$, where $\hat{Y}_1 = \sum_{i=1}^{N_R} Y_i \cdot T_i / \sum_{i=1}^{N_R} T_i$ and $\hat{Y}_0 = \sum_{i=1}^{N_R} Y_i \cdot (1-T_i) / \sum_{i=1}^{N_R} (1-T_i)$ are the means of the treated and the control outcomes, respectively, within the right child node. SSE_L is defined similarly. The authors showed that the CIT consistently outperforms the t-stats tree.

Causal Tree (CT) (Athey and Imbens, 2016) A CT is a more recent tree-based algorithm developed specifically to estimate the CATE. A main difference between the CT and the previously mentioned tree-based methods is that the authors designed the CT to be an "honest" approach in the sense that, instead of using the training dataset as a whole, the CT divides the training dataset (of size n) into two parts, the splitting set (of size n_s) and the estimation set (of size n_e), and then uses the subjects in the splitting set to determine the split and the subjects in the estimation set to estimate the CATE within the node. The splitting criterion of the CT can be represented as follows:

$$\mathcal{C}^{CT} := (\frac{n_L}{n} \hat{\tau}_L^2 + \frac{n_R}{n} \hat{\tau}_R^2) - (\frac{1}{n} + \frac{1}{n_s})(\frac{S_{1L}^2}{p} + \frac{S_{0L}^2}{1-p} + \frac{S_{1R}^2}{p} + \frac{S_{0R}^2}{1-p}), \tag{24}$$

where $S_{1L}, S_{0L}, S_{1R}, S_{0R}$ denote sample variances of treated and control subjects in left and right nodes, respectively. $(\frac{1}{n} + \frac{1}{n_s})$ is the weight to penalise small-sized leafs, and p is the marginal treatment probability in data samples including the training dataset.

Unlike other tree methods, in the CT, the CATEs are estimated using inverse propensity score weighting (Rosenbaum and Rubin, 1983) as $\hat{\tau}(\boldsymbol{x}) =$

$\sum_{\boldsymbol{x}_i \in \mathcal{N}} \frac{T_i \cdot Y_i}{e(\boldsymbol{x}_i)} / \sum_{\boldsymbol{x}_i \in \mathcal{N}} \frac{T_i}{e(\boldsymbol{x}_i)} T - \sum_{\boldsymbol{x}_i \in \mathcal{N}} \frac{(1-T_i) \cdot Y_i}{(1-e(\boldsymbol{x}_i))} / \sum_{\boldsymbol{x}_i \in \mathcal{N}} \frac{1-T_i}{(1-e(\boldsymbol{x}_i))}$, where $\mathcal{N}$ is a node, and $e(\boldsymbol{x}) = P(T = 1|\boldsymbol{x})$ denotes the propensity score for subject $\boldsymbol{x}$.

Athey and Imbens (2016) also discussed an "adaptive" CT where the splitting criterion does not have the variance part (the second term in Equation (24)) and the algorithm does not divide the data into two sets. Recalling the splitting criterion of UpliftDT (Equation (20)), it can be clearly seen that the splitting criteria of the adaptive version of the CT and UpliftDT are equivalent. In their evaluation, the honest CT performed consistently better than the adaptive CT. However, a drawback to the honest CT is that it effectively only utilises a portion of the dataset, because it requires dividing the subjects into the splitting and the estimation sets.

Bayesian Score Tree (Chickering and Heckerman, 2000) This is an early method designed for targeted advertising and is different from the previously surveyed methods. The Bayesian score tree does not directly maximise the uplift difference in each split. Instead, it models the uplift as a new split by the treatment variable on a child node. Each child node is forced to be split by the treatment variable if its path does not contain the treatment variable. A child node is removed if the overall tree quality score does not increase when it is split by the treatment variable. The uplift is calculated by the difference between the outcome probabilities in the left and right child nodes. A Bayesian score, such as the one used by Buntine (1993), is employed to calculate the score of the candidate trees.

3.2.2 Support Vector Machine-Based Uplifting Modelling Methods

In this section, we discuss methods that utilise a support vector machine (SVM) (Cortes and Vapnik, 1995) for CATE estimation proposed in the uplift modelling community. The main benefit of these methods is that by reformulating the uplift modelling problem within the SVM framework, they enjoy the benefit of an SVM, which has been proven to be effective in many supervised learning applications. However, a problem with these methods is that none of them has implementation available online for potential users.

L_1 and L_p Uplift SVMs (L_1-USVM and L_p-USVM) (Zaniewicz and Jaroszewicz, 2013, 2017) These two SVM-based methods were proposed by the same authors, Zaniewicz and Jaroszewicz, who recast the uplift modelling problem as a three-class classification problem. Specifically, the two methods aim to predict whether the treatment has a positive treatment effect, no treatment effect, or a negative treatment effect. They achieve this goal by using the following two parallel hyperplanes:

$$H_1 : \langle \boldsymbol{w}, \boldsymbol{x} \rangle - b_1 = 0, \qquad H_2 : \langle \boldsymbol{w}, \boldsymbol{x} \rangle - b_2 = 0, \tag{25}$$

where $b_1, b_2 \in \mathbb{R}$ are the intercepts, and $\boldsymbol{x}$ denotes the coefficients of the decision boundary. The predicted treatment effect is then specified as:

$$\hat{\tau}(\boldsymbol{x}) = \begin{cases} +1 & \text{if } \langle \boldsymbol{w}, \boldsymbol{x} \rangle > b_1 \text{ and } \langle \boldsymbol{w}, \boldsymbol{x} \rangle > b_2, \\ 0 & \text{if } \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq b_1 \text{ and } \langle \boldsymbol{w}, \boldsymbol{x} \rangle > b_2, \\ -1 & \text{if} \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq b_1 \text{ and } \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq b_2. \end{cases} \tag{26}$$

L_1-USVM (Zaniewicz and Jaroszewicz, 2013) utilises the L_1-norm as the regularisation for $\boldsymbol{w}$, but it is sensitive to the parameter setting because of the discontinuity problem with the L_1-norm. Then, L_p-USVM (Zaniewicz and Jaroszewicz, 2017) was proposed to utilise the L_p-norm in the optimisation to replace L_1-norm. L_p-USVM not only resolves the discontinuity problem, but also improves the convergence and efficiency and provides more stable results. However, it also introduces an additional hyperparameter p that needs to be tuned.

The above two methods are designed for data from randomised controlled trials. The authors also proposed extended methods to observational data by adding one more regularisation term (Jaroszewicz and Zaniewicz, 2015). However, adding the additional regularisation term makes the objective function difficult to be optimised, because the new objective is not differentiable.

Lift Curve SVMs Another contribution from the uplift modelling community is the SVM for differential prediction (SVM^{upl}) (Kuusisto et al., 2014). In SVM^{upl}, the authors proposed to directly find the decision boundary that maximises the area under the uplift curve (AUUC). The uplift curve is an evaluation metric used by the uplift modelling community for comparing the performance of uplift models. We give the precise definition of the uplift curve (Equation (33)) when introducing the evaluation metrics (Section 5.3). Intuitively, the idea of SVM^{upl} is similar to the SVM-based methods for supervised learning that maximise the area under the ROC curve (AUC), i.e., SVM^{perf} (Joachims, 2005). However, instead of maximising the AUC, SVM^{upl} maximises the AUUC for uplift modelling.

The authors showed that maximising the AUUC is equivalent to maximising a weighted difference between the AUC for the treatment group and the AUC for the control group. Specifically,

$$\max(AUUC) \equiv \max(AUC_{T=1} - \lambda AUC_{T=0}), \tag{27}$$

where $\lambda = \frac{\sum_{i=1}^{n} Y_i(1-T_i) \sum_{i=1}^{n} (1-Y_i)(1-T_i) \sum_{i=1}^{n} T_i}{\sum_{i=1}^{n} Y_i T_i \sum_{i=1}^{n} (1-Y_i) T_i \sum_{i=1}^{n} (1-T_i)}$. The difference in Equation (27) is equivalent to the sum of the AUC for the treatment group and the control group by flipping the control group outcomes:

$$\max(AUUC) = \max(AUC_{T=1} - \lambda(1 - AUC^{-}_{T=0})) = \max(AUC_{T=1} + \lambda AUC^{-}_{T=0}) \tag{28}$$

where $AUC^{-}_{T=0}$ indicates the AUC of the control group with flipped outcome labels.

By showing that maximising the AUUC is equivalent to maximising the sum of two AUCs, the authors then solved the SVM^{Upl} optimisation problem by utilising the SVM^{perf} algorithm (Joachims, 2005), which is designed to optimise the AUC directly. Since none of the methods proposed by the treatment effect heterogeneity modelling community directly maximises the AUUC, it could be beneficial to apply this approach to sociology and medical problems.

3.2.3 Deep Learning-Based Methods

Recently, several generative learning methods for treatment effect heterogeneity modelling have been proposed. We survey them separately from the deep learning methods discussed above, since the generative approach of these methods does not fall into the single-model, two-model, X-Learner, or R-Learner approaches discussed above. It is worth noting that the uplift modelling community seldom discusses deep learning-based methods. Since deep learning-based methods excel at dealing with large-scale datasets, these methods may contribute to online marketing where the number of subjects is usually abundant.

Causal Effect Variational Autoencoders The causal effect variational autoencoder (CEVAE) Louizos et al. (2017) is a variational autoencoder (VAE)-based treatment effect heterogeneity modelling approach. This method uses a VAE to learn a latent confounding set $\boldsymbol{z}$ from the observed covariates $\boldsymbol{x}$, and then uses $\boldsymbol{z}$ to estimate the CATE. A VAE (Kingma and Welling, 2014) is a new type of autoencoder based on variational inference that is able to approximately infer the intractable posteriors of the latent variables. The CEVAE assumes that the observed covariates ($\boldsymbol{x}$) are independent of both treatment and outcome conditioning on the latent confounders $\boldsymbol{z}$. It adopts a VAE to infer the posterior of the latent confounders $p(\boldsymbol{z}|\boldsymbol{x})$ and estimate the CATE as:

$$\hat{\tau}(\boldsymbol{x}) = \int_{\boldsymbol{z}} p(Y|\boldsymbol{z}, T=1)p(\boldsymbol{z} \mid \boldsymbol{x})d\boldsymbol{z} - \int_{\boldsymbol{z}} p(Y \mid \boldsymbol{z}, T=0)p(\boldsymbol{z} \mid \boldsymbol{x})d\boldsymbol{z} \tag{29}$$

CEVAE can infer unobserved confounders that are difficult to measure. For example, the income of a patient is rarely available from electronic medical records but can be inferred from the patient's postcode and occupation. However, a drawback to the CEVAE is that there is no guarantee that the inferred latent posterior $p(\boldsymbol{z}|\boldsymbol{x})$ will converge to the true posterior, because the CEVAE relies on variational approximations. Another drawback to the CEVAE is that it assumes that all the factors in $\boldsymbol{z}$ are confounders, which is often not satisfied by data.

A recently proposed improvement to the CEVAE is the disentangled variational autoencoder (TEDVAE) (Zhang et al., 2021). The TEDVAE learns three disentangled sets of latent factors: the instrumental factors $\boldsymbol{z}_t$, which affect only the treatment; the confounding factors $\boldsymbol{z}_c$, which affect both the treatment and the outcome; and the risk factors $\boldsymbol{z}_y$, which affect only the outcome. Disentangling the latent factors facilitates accurate CATE estimations, and alleviates the

user's burden of choosing the appropriate set of covariates, since users can safely include all observed variables without the implication that including variables unrelated to the outcome may increase the bias and the variance of the CATE estimations.

Generative Adversarial Network for Individualised Treatment Effects (GANITE) (Yoon et al., 2018) GANITE utilises a generative adversarial network (GAN) to model treatment effect heterogeneity. GANITE consists of two components, a counterfactual block that generates the counterfactual outcome Y^{cf} with input $(\boldsymbol{x}, T, Y)$, and an ITE block that generates the CATE $\hat{\tau}(\boldsymbol{x})$ for the subjects.

Specifically, the counterfactual block contains a generator $\boldsymbol{G}$ paired with a discriminator $\boldsymbol{D_G}$. $\boldsymbol{G}$ takes input $(\boldsymbol{x}, T, Y)$ and generates a counterfactual Y^{cf} for the treatment $1-T$, while $\boldsymbol{D_G}$ takes $(\boldsymbol{x}, Y, Y^{cf})$ as input and outputs whether the outcome is generated by $\boldsymbol{G}$. During training, $\boldsymbol{G}$ is trained to maximise the probability of $\boldsymbol{D_G}$ incorrectly identifying whether Y^{cf} is factual or counterfactual, while $\boldsymbol{D_G}$ is trained to maximise the probability of correctly distinguishing Y^{cf} from Y. After training the counterfactual block, the counterfactual outcome Y^{cf} generated by $\boldsymbol{G}$ along with $(\boldsymbol{x}, T, Y)$ are fed into the ITE block, which consists of a generator $\boldsymbol{I}$ paired with an discriminator $\boldsymbol{D_I}$. The generator $\boldsymbol{I}$ takes the covariates $\boldsymbol{x}$ as input and generates the potential outcomes Y and Y^{cf}. The discriminator $\boldsymbol{D_I}$ aims to discriminate whether the outcomes generated by $\boldsymbol{I}$ are the inputs from the counterfactual block (generated by $\boldsymbol{G}$). After training, only the generator $\boldsymbol{I}$ in the ITE block is used for predicting the CATE for new subjects.

3.2.4 Ensemble-Based Methods

Ensemble-based CATE estimation methods have been proposed to address the high variance problem of tree-based methods. Most ensemble methods use tree-based methods as base learners. Generally speaking, ensemble-based methods perform better than a single tree-based model; however, ensemble-based methods lose the interpretability possessed by the tree-based methods and have higher time complexity than tree-based methods.

Uplift Bagging Bagging (Breiman, 1996) is a simple and popular ensemble method in supervised learning. When using bagging for uplift modelling, a set of bootstrap training datasets is randomly sampled from the original training dataset with replacement. A bootstrap training dataset has the same size as the original training dataset. An uplift model is built on each bootstrap training dataset. The final prediction for a test subject is the average of predicted uplifts of all models in the example.

For CATE estimation, Radcliffe and Surry (2011) stated that they used bagging in real-world applications, but they did not provide any experimental results. Sołtys et al. (2015) implemented and compared two bagging methods. The base learners used were UpliftDT (using the Euclidean distance (Rzepakowski

and Jaroszewicz, 2010) and the two-model decision tree (using C4.5 (Quinlan, 1993), as discussed in Section 3.1.2). Based on their evaluations (Sołtys et al., 2015), the uplift bagging methods performed significantly better than the uplift decision trees, and were competitive with the uplift random forest method discussed below.

Uplift Random Forest (UpliftRF) (Guelman et al., 2012, 2015a) UpliftRF is an uplift ensemble-based on the idea of random forests (Breiman, 2001). It consists of four steps. Firstly, a set of bootstrap training datasets is randomly sampled from the original training dataset with replacement. Secondly, each bootstrap training dataset is projected to a fixed number of k randomly selected covariate spaces. Thirdly, UpliftDT (Rzepakowski and Jaroszewicz, 2010) is built on every training dataset from the above two steps. Fourthly, the set of uplift trees is used to predict the uplift for a new subject by using the average predictions of all trees.

Sołtys et al. (2015) implemented an ensemble-based algorithm called double uplift random forests (DURF). DURF is a bagged ensemble of the two-model approach using randomised trees from Weka (Hall et al., 2009). In their evaluation, both UpliftRF (using Euclidean distance) and DURF performed better than the two-model decision trees, but were not significantly different from the bagged UpliftDT (using the Euclidean distance) (Rzepakowski and Jaroszewicz, 2010) or the bagged two-model decision tree (using C4.5 (Quinlan, 1993)).

Causal Conditional Inference Forest (UpliftCCIF) (Guelman et al., 2015b) UpliftCCIF is based on tree models and uses a similar strategy as UpliftRF to construct the ensemble of trees, i.e., by randomly sampling the training subjects and covariates with replacement. The major difference between UpliftCCIF and UpliftRF lies in the tree splitting procedure. Whether to split the node in a base learner tree model of UpliftCCIF is determined by testing the null hypothesis that there are no interactions between the treatment T and any of the covariates in X. Specifically, the null hypothesis is formulated as $H_0 = \cap_{j=1}^{k} H_0^j$ with $H_0^j : \mathbb{E}[Y^*|x_j] = \mathbb{E}[Y^*]$, where Y^* is the transformed outcome, as discussed in Section 3.1.5, and x_j is a covariate value in $\boldsymbol{x}$. The authors used Bonferroni-adjusted p-values for handling the multiplicity in the statistical tests. If the null hypothesis is rejected, the splitting covariate is selected as the one with the smallest p-value. After the trees are built, the uplift of a subject is estimated as the average of the uplifts predicted by the individual trees.

Causal Forests (CF) (Wager and Athey, 2018) CF is a random forest-like algorithm for treatment effect heterogeneity modelling. CF uses the Causal Tree (CT) algorithm (discussed in Section 3.2.1) as its base learner, and constructs the forest from an ensemble of k causal trees where each tree provides a CATE estimation $\hat{\tau}_b(\boldsymbol{x})$ for a subject. The forest then uses the average of the predicted CATEs from k trees as its prediction, i.e., $\hat{\tau}(\boldsymbol{x}) = \frac{1}{k}\sum_{b=1}^{k} \hat{\tau}_b(\boldsymbol{x})$.

An important advantage of CF over the other surveyed ensemble methods is that the estimations of CF are asymptotically Gaussian and unbiased for the true CATE $\tau(\boldsymbol{x})$. In other words, $(\hat{\tau}(\boldsymbol{x}) - \tau(\boldsymbol{x}))/\sqrt{Var(\tau(\boldsymbol{x}))} \rightarrow \mathcal{N}(0,1)$. Furthermore, the authors provides a way to estimate the asymptotic variances. CF is a general framework in the sense that its theoretical properties are valid as long as the trees used as base learners are "honest", i.e., that the outcome of any sample is not used for both selecting the split and estimating the within-node CATE $\hat{\tau}$. Based on this property, the authors proposed a CF instantiated using propensity tree, which completely ignores the outcome Y when choosing the splits and builds the tree using the Gini criterion (Breiman et al., 1984) of the treatment T.

4 Applications

In this section, we discuss the main applications of targeted advertising from the uplift modelling community, and applications in personalised medical treatment and social sciences from the treatment effect heterogeneity modelling community. While applications in the uplift modelling community mostly utilise data from randomised experiments, those in the treatment effect heterogeneity community often make use of observational data, since ethical and cost concerns often prohibit controlled trials in many medical and sociology studies.

4.1 Applications in Marketing

Driven by the need by companies for increased sales and minimal advertisement costs, targeted advertising has long been the focus of the uplift modelling community. Most of the work in this area is based on real-world datasets obtained by randomised experiments, and some of the results have been monetised by companies according to the literature. The results in these applications confirm that using uplift modelling to target customers is more effective than random targeting or using standard supervised learning methods.

Chickering and Heckerman (2000) applied the Bayesian score tree (discussed in Section 3.2.1) to an MSN advertising experimental dataset. Registrants of Windows 95 were randomly divided into treatment and control groups, where the treated subjects were mailed advertisements and the control subjects received nothing. The outcome was an MSN sign-up within a time period. 110,000 subjects were involved in the experiment, where the treatment subjects accounted for 90% of the total. 70% of the data were used for training and the rest were used for evaluation. Their results showed that the uplift model achieved more revenue when compared to the mail-to-all strategy.

Hansotia and Rukstales (2002) applied the UpliftIVM algorithm (discussed in Section 3.2.1) to analyse customer responses to a promotion for $10 off of a $100 purchase. The data were obtained from a holiday promotion of a major national retailer in the United States. Promotional mail (treatment) was randomly sent to 50% of the total 282,277 customers. They compared a two-model approach

using logistic regression with UpliftIVM. The evaluation criterion was the uplift in the 50% reserved evaluation dataset (and the other 50% was used for building models). Their results showed that UpliftIVM performed better than the two-model approach when targeting the top 10% of customers, while their overall performance was similar.

Radcliffe (2007) discussed three real-world applications of uplift modelling in marketing: "deep-selling", where the goal is to use a promotion to increase the frequency or size of customer transactions; customer retention, which aims to mitigate customer attrition; and cross-selling, which involves selling new products to existing customers. The deep-selling experiment included 100,000 subjects, and the treatment and control groups were split by 50:50. The sample sizes in two other examples were not reported. They used a balance-based uplift tree (Radcliffe and Surry, 2011) (discussed in Section 3.2.1). With deep-selling, the authors showed that the uplift modelling approach was better at increasing revenue than standard supervised methods. For customer retention, they considered a problem where a mobile service provider was experiencing an annual churn rate of 9%. Originally, the company targeted its entire customer base with a retention offer and the churn rate increased to 10%. Using uplift modelling, they targeted 30% of customers, and were able to reduce the customer churn from 9% to 7.8%. Using an estimated average revenue per user of \$400/year, the increase in revenue was around \$8.8 million. For cross-selling, they tackled a banking problem where banks wanted to sell new banking products to their existing customers and found that by using uplift modelling, banks could achieve between 80% and 110% increases in sales while reducing the volume of mailing offers by 30% to 80% when compared to a standard supervised learning approach.

Guelman et al. (2012) applied uplift modelling methods to a customer retention dataset from a Canadian insurance company. A randomised experiment was conducted with a treatment group of 8249 subjects and a control group of 3719 subjects whose policies were due for renewal. The customers in the treatment group received a letter explaining that there would be an increase in their insurance premiums and a phone call from an insurance advisor. The customers in the control group received no retention effort. Four methods were applied to the dataset: UpliftRF (discussed in Section 3.2.4), the two-model approach with logistic regression, the single-model approach using logistic regression with interaction terms between the treatment and all covariates (Lo, 2002), and the uplift decision tree (Rzepakowski and Jaroszewicz, 2010) (discussed in Section 3.2.1). Based on the uplift curves obtained from 10-fold cross validation, all methods performed better than the baseline retention rate by random targeting. Furthermore, UpliftRF performed better than the others, especially in top-ranked customer subgroups, but did not dominate other methods in the other customer subgroups.

Hillstrom's Email Advertisement dataset is an open-access uplift modelling dataset (Hillstrom, 2008) that contains 64,000 samples collected from a randomised experiment of an email advertisement campaign. The customers were evenly distributed in two treatment groups and a control group, where the first treatment group was sent a "Men's Advertisement Email" and the second treat-

ment was sent a "Women's Advertisement Email". The control group received no email. The outcome variables were the visit and conversion status of the customers. Radcliffe (Radcliffe, 2008) conducted analysis using the Hillstrom dataset, in which it was shown that the "Men's Email" was more effective than the "Women's Email" and the best customer subgroup to target was the subgroup of those who had visited the store using both a phone and web browser, and had spent more than $160. The analysis identified that some customers were negatively affected by the "Women's Email".

Another recent dataset from the uplift modelling community is the Criteo Uplift modelling dataset (Eustache et al., 2018). It contains 25,309,483 subjects where each row represents the behaviour of a customer. There are 11 anonymised pre-treatment covariates for each customer, a treatment indicator representing whether the customer received a promotional email, and two outcomes indicating the visiting and conversion status of the customer. Recently, several uplift modelling methods (Heliou et al., 2020; Betlei et al., 2020) have been proposed to maximise the AUUC metric (as discussed in Equation 27) for the dataset.

4.2 Applications in Social Science

Social science is one of the main applications areas of the treatment effect heterogeneity modelling community. In the following, we present some examples of the applications.

Imai and Ratkovic (2013) analysed the get-out-the-vote (GOTV) (Gerber and Green, 2000) randomised experiment where 69 different voting mobilisation methods, including canvassing (by in-person direct contact to encourage voting), phone calls and mailing were randomly given to registered voters in New Haven and Connecticut during the 1998 U.S. presidential election. The goal of their analysis was to select the best voting mobilisation strategy for individuals. To avoid the interference between voters within the same household (which violates the SUTV assumption), they focused on a subset of 14,774 voters in single-voter households, of which 5269 voters belonged to the control group. They used a single-model approach (discussed in Section 3.1.1) with a modified SVM algorithm (Cortes and Vapnik, 1995) as the base learner. Specifically, they used two separate L1-regularisation terms for covariates that only affected the outcome and covariates that interacted with both the outcome and the treatment (the type of covariates were manually identified by the authors). Their analysis showed that canvassing was the most effective treatment. When canvassing was used, any additional treatment such as phone calling and mailing in combination with canvassing reduced the effectiveness of canvassing. In addition, when canvassing was absent, the most effective treatment was contact by mail with a civic duty appeal. Any other treatment was found to be less effective or have negative effects.

Imai and Ratkovic (2013) analysed a dataset for a national supported work (NSW) program (LaLonde, 1986; Dehejia and Wahba, 1999) to determine whether a job training program increased the income of workers. The dataset contained 297 and 425 workers randomly assigned to the treatment and control groups,

respectively, and the 1978 panel study of income dynamics workers (PSID) from low-income subjects. In other words, the dataset consisted of two components, where the first component was obtained from a randomised controlled trial (the NSW sample), and the second component was obtained from an observational study (the PSID sample). The method used was the same as the one described in the GOTV application. They built the model on the randomised NSW samples and applied it to the PSID samples. Their analysis showed that, overall, the training program was beneficial. However, it benefited educated Hispanics and low-income non-Hispanics the most, and it did not help employed white workers with high-school degrees.

Künzel et al. (2019) analysed a field experiment for the effect of canvassing (an in-person direct contact by conversation) on reducing transphobia (i.e., prejudice against transsexual and transgender people). The dataset was originally used by Broockman and Kalla (2016) where the analysis showed that brief but high-quality canvassing significantly reduced prejudice against transgender individuals for at least three months. The dataset consisted of three groups: a treatment group of 913 individuals who were canvassed on the topic of reducing transphobia, a placebo group of 912 individuals who were canvassed by an irrelevant conversation (i.e., about recycling), and a control group of 68,278 individuals who had not been canvassed at all. The outcomes of interest were the results of an online survey that measured the subjects' attitudes towards transphobia at 3 days after treatment. Using covariates such as religion, ideology, and demographics, the authors applied the single-model approach, the two-model approach, and the X-Learner (as discussed in Section 3.1) using random forests as the base learner. The results showed that the treatment effects estimated by the single-model approach were mostly 0 or almost 0, despite the fact that the average effect of the treatment was 0.22 (Broockman and Kalla, 2016). The two-model approach and the X-Learner produced similar results, but the estimates of the two-model approach had increased variance.

4.3 Applications in Personalised Medical Treatments

Personalised medical treatment is another major application area of the treatment effect heterogeneity community, although some uplift modelling literature has also considered it (Nassif et al., 2012b; Jaskowski and Jaroszewicz, 2012). However, research in the treatment effect heterogeneity modelling community is more focused on the medical implications of the results, whereas research from the uplift modelling community uses medical datasets to compare different models.

Cai et al. (2011) studied the personalised treatment problem with the two-model approach. They used a clinical trial dataset from the AIDS Clinical Trials Group (ACTG), which studied the effect of a protease inhibitor for treating human immunodeficiency virus (Hammer et al., 1997). A total of 1156 patients were included in the trial and were randomly divided into treatment and control groups. The control group was given a two-drug combination, while the treatment group was given a three-drug combination. Previous studies showed that the treatment was significantly more effective than the control. However, it was

also noted that some patients did not respond to the treatment and instead suffered from toxic side effects. The authors considered three covariates and used the change of CD4 count at week 24 from the baseline level as the continuous outcome. They used linear regression as the base learner for the two-model approach. The results showed that the treatment effects of patient groups defined by different covariates were significantly heterogeneous.

Weisberg and Pontes (2015) applied CATE estimation algorithms to clinical trial data from the Randomized Aldactone Evaluation Study. Aldactone is medicine for treating fluid build-up due to heart failure, liver scarring, or kidney disease. The trial was designed to test whether the medicine could reduce the mortality of patients who had suffered from severe heart failure. With Aldactone as the treatment, the study used 63 variables describing the demographics, history, and concomitant medications as covariates, and the maximum potassium level (continuous) within the first 12 weeks of treatment as the outcome. The authors used a transformed outcome approach (discussed in Section 3.1.5). The model was built on 80% of the total 1632 subjects and tested on the remaining 20% of the subjects. It showed that the treatment, although highly significant in the original study (Pitt et al., 1999), was no longer significant when the characteristics of the subjects were considered.

Some works extend CATE estimation to censored survival outcomes. Zhang et al. (2017) extended the causal tree (Athey and Imbens, 2016) (discussed in Section 3.2.1) to censored outcomes and used it to study the heterogeneous treatment effects of radiotherapy on the survival outcomes of breast cancer patients and glioma cancer patients using patients' gene profiles as covariates. The breast cancer dataset contained 964 subjects and the expression profile of 11,535 genes, and the glioma cancer dataset contained 632 subjects and 11,543 genes. For each dataset, about 50% of the subjects were treated with radiotherapy while the rest were not. The datasets can be accessed at TCGA (The Cancer Genome Atlas). The proposed method was compared with popular cancer subtype clustering methods including semi-supervised clustering and the L1-regularised COX proportional hazard model. Results on individual test sets showed that the CATE estimation method was better at finding subgroups with treatment effect heterogeneity. Recently, Tabib and Larocque (2020) proposed a random forest-based method to censored outcomes and showed the effectiveness of their algorithm on a breast cancer dataset and a colon cancer dataset. It is also possible to extend other surveyed methods to censored survival outcomes.

5 Software, Metrics, Demonstration, and Discussion

In this section, we summarise the available software packages and source codes for CATE estimation. We also illustrate the methods on synthetic, semi-synthetic, and real-world datasets, and discuss their modelling behaviour, usability, interpretability, and scalability. We did not aim to find the best performing method,

Table 1: Software packages for CATE estimation and uplift modelling. B, C, and N denote binary, categorical, and numerical variables, respectively.

Package	Covariates B C N	Outcome B N	Methods	Language	URL
causalTree	✓ ✓ ✓	✓ ✓	Causal Tree t-stats Tree TO# Tree	R	https://github.com/susanathey/causalTree
causalToolbox	✓ ✓ ✓	✓ ✓	Single-model Two-Model X-Learner TO#	R	https://github.com/soerenkuenzel/causalToolbox
grf	✓ × ✓	✓ ✓	Causal Forest	R	https://CRAN.R-project.org/package=grf
Uplift	✓ ✓ ✓	✓ ×	UpliftDT UpliftRF UpliftCCIF	R	https://cran.r-project.org/package=uplift
CausalML	✓ ✓ ✓	✓ ✓*	UpliftDT UpliftRF UpliftCCIF Single-model Two-Model X-Learner	Python	https://github.com/uber/causalml
pylift	✓ ✓ ✓	✓ ×	TO#	Python	https://github.com/wayfair/pylift

*: the uplift modelling methods in CausalML are implemented for binary outcomes only. #: TO refers to the transformed outcome approach.

Table 2: Source codes for deep learning-based CATE estimation algorithms. B, C, and N denote binary, categorical, and numerical variables, respectively.

	Covariates B C N	Outcome B N	Method	Language	URL
CFR	✓ × ✓	× ✓	CFR	Python	https://github.com/clinicalml/cfrnet
SITE	✓ × ✓	× ✓	SITE	Python	https://github.com/Osier-Yi/SITE
CEVAE	✓ × ✓	× ✓	CEVAE	Python	https://github.com/AMLab-Amsterdam/CEVAE
TEDVAE	✓ × ✓	× ✓	TEDVAE	Python	https://github.com/WeijiaZhang24/TEDVAE

since this is beyond the scope of this survey. Rather, we provide discussion based on empirical evaluations of the methods in the available packages and recently proposed methods from both the causal effect heterogeneity modelling and uplifting modelling communities. For an empirical evaluation focused on comparing the accuracy of some of the algorithms for treatment effect heterogeneity modelling, we refer the reader to Dorie et. al. (Dorie et al., 2019). For empirical evaluations focused on comparing the performances of some uplift modelling algorithms, see Devriendt et al. (Devriendt et al., 2018) and Gubela et al. (Gubela et al., 2019).

5.1 Software Packages and Source Codes

Table 1 presents a summary of the available software packages for uplifting modelling and CATE estimation. We include those which are well documented

and are easy for end-users to use.

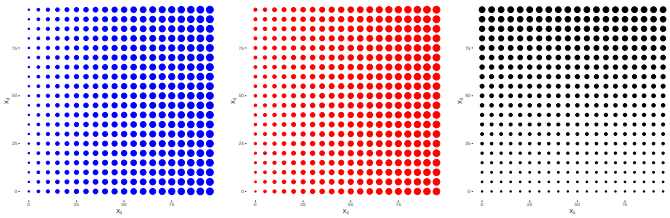
The causalTree package implements the causal tree (Athey and Imbens, 2016) (discussed in Section 3.2.1), the t-stats tree (Su et al., 2009) (discussed in Section 3.2.1), and the transformed outcome approach (discussed in Section 3.1.5) using a regression tree (RT) as the base learner. The causalToolbox package implements the single-model, two-model, X-Learner, and transformed outcome approaches (discussed in Section 3.1.1, 3.1.2, 3.1.3, and 3.1.5) using random forests (RF) and BART as base learners. The grf package implements the causal forest algorithm, as discussed in Section 3.2.4. The Uplift package was developed by the uplift modelling community and provides implementations of the uplift decision trees (UpliftDT) (discussed in 3.2.1), uplift random forests (UpliftRF), and uplift causal conditional inference forest (UpliftCCIF) algorithms (discussed in Section 3.2.4). The CausalML package implements the same uplift modelling methods as the Uplift package and the same CATE approaches as the causalToolbox package. However, the base learners used in CausalML are linear regression (LR), XGBoost, and multi-layer perceptron, which are different from those provided in causalToolbox. Finally, the pylift is a uplift modelling package that implements the transformed outcome approach using XGBoost and random forests as base learners.

Table 2 summarises the source codes for some recently proposed deep learning-based algorithms, including CFR (Shalit et al., 2017), SITE (Yao et al., 2018), CEVAE (Louizos et al., 2017), and TEDVAE (Zhang et al., 2021), as discussed in Sections 3.1.6 and 3.2.3. These methods have shown potential for CATE estimation tasks. However, the main difference between these source codes and the software packages is that the former do not provide ready-to-use interfaces for users to apply the methods to their own datasets or to tune parameters.

Recently, a few packages have been proposed specifically for generating synthetic datasets in order to evaluate treatment effect estimation algorithms: the OPOSSUM package Winkel and Krebs (2017) provides a Python interface for generating synthetic datasets with ground-truth CATEs, and aciccomp Dorie (2017) is an R package which contains the datasets used in the Atlantic Causal Inference Competitions.

5.2 Illustration

Firstly, we use a synthetic dataset to provide a visual illustration for the behaviour in the CATE estimation of some representative methods. The dataset we used was first introduced in (Radcliffe and Surry, 2011) to provide an illustration of uplift modelling methods. It consists of 100,000 subjects of two covariates, a binary treatment and a continuous outcome. The data generating procedure is described as follows. First, two covariates, X_1 and X_2, are uniformly sampled from integers ranging from 0 to 99. Then, the two potential outcomes are generated based on the covariates and the treatment. The outcomes in the control group are uniformly sampled from the interval defined by $[0, x_1)$, and the outcomes in treatment groups are uniformly sampled from the interval defined by $[0, x_1) + [0, x_2)/10 + 3$. In other words, the ground truth of the CATE has a



(a) Treatment (b) Control (c) Theoretical CATE

Figure 3: Disc plots of the treatment, control, and ground-truth CATEs of the synthetic dataset. Values in covariates X_1 and X_2 are grouped in bins of size 5. The disc areas are proportional to the average outcome values / CATEs in the squares. Note that the areas of discs in Subfigure (c) are in a different scale from Subfigures (a) and (b). The scale is around seven times smaller, since the magnitude of CATEs is small. Disc plots in Figures 4 and 5 use the same scale as in Subfigure (c).

linear relationship with the covariates as $\tau(\boldsymbol{x}) = x_2/20 + 3$. We generate three different sets of data with different treatment assignment ratios. Specifically, the treatment assignments are sampled from a Bernoulli distribution with $p = 0.5$, $p = 0.1$, and $p = 0.9$ for assigning a subject to the treatment. Visualisations of the outcomes in both the treatment and control groups, along with the ground truth CATEs, are illustrated in Figure 3.

We use disc plots for visualisation. In the disc plots, the horizontal and vertical axes represent the covariates X_1 and X_2, respectively. The size of a disc is proportional to the average value of the outcomes (Y in Figure 3(a) and Figure 3(b))) or CATEs (Figure 3(c)) for the subjects binned within the two-dimensional space with a width of 5. For example, the size of the lower-left-most disc in Figure 3(a) represents the average treated outcomes for subjects in the space of $x_1 \in [0, 5)$ and $x_2 \in [0, 5)$. The lower-left disc in Figure 3(c) indicates the average CATE for subjects in that square. We keep the scale of the discs the same for Figure 3(a) and Figure 3(b), but use a different scale for Figure 3(c), since the average CATEs are much smaller than the outcome values. The estimated CATEs of some representative methods are shown in Figure 4 on the balanced dataset with $p = 0.5$. Furthermore, we also illustrate the behaviour of the selected methods (the best from the previous illustration) on datasets with an imbalanced treatment ratio $p = 0.1$ and $p = 0.9$ in Figure 5.

When the models are specified in the same way as the data generating procedure, i.e., using linear regression (LR) as the base learner, the CATE modelling behaviour of the algorithms is similar to the ground-truth CATEs. This can be seen from the disc plots of the two-model LR (Figure 4(a)), X-Learner LR (Figure 4(d)), and transformed outcome LR (Figure 4(g)), where the trends of the disc plots are similar to the trends in Figure 3(c). However, it is worth noting that in most applications the ground-truth relationships between the CATE and the covariates are unknown to the users, and thus specifying the

parametric form of the base learner is difficult.

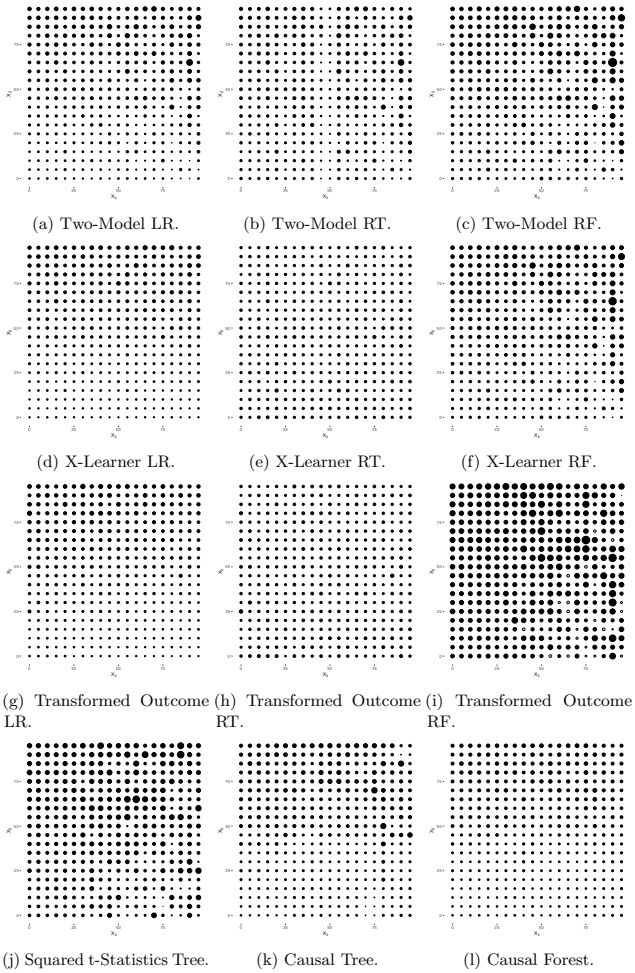
When the models are specified differently from the data generating procedure, i.e., using regression trees (RT) or random forests (RF) as base learners, we observe that the modelling behaviour of different methods is significantly different. In the second and the third columns of Figure 4, we can hardly see any two methods which produce the same CATE estimations. Furthermore, the disc plots of the two tailored tree-based methods—the squared t-statistics tree (Figure 4(j)) and causal tree (Figure 4(k))—are also different from each other. Even with the same method, the performance changes significantly when datasets change with different treatment and control sample ratios, as shown in Figure 5. We can see that X-Learner RF performs well when the data are balanced and when the number of treated subjects is smaller than the number of control subjects ($p = 0.1$ in Figure 5(h)). However, its performance significantly worsens when the treated subjects are dominant in the data ($p = 0.9$ in Figure 5(i)). An explanation for this can be seen from Equation (13). When $p = 0.9$, the dominating component of the equation is $\hat{\tau}_0(\boldsymbol{x})$, which is estimated from the control subjects and may underestimate the treatment effects. We observe that when the weight for $\hat{\tau}_0(\boldsymbol{x})$ is reduced, X-Learner RF performs better. However, in practice, the weight can be difficult to determine when the ground truth is unavailable.

CATEs are difficult to model. The synthetic dataset has only two covariates with a sufficiently large number of subjects, and the ground-truth CATEs follow a linear relationship with one covariate (X_2). It is reasonable to expect that most methods should present nearly perfect estimations of the datasets. However, the results are far from perfect. When a method is not correctly specified, i.e., when a non-liner modelling method is used for the data generated by a linear relationship, the CATE estimations change significantly when the proportion of treated samples changes, as shown in Figure 5. Even when a method is correctly specified (two-model LR, X-Learner LR, and transformed outcome LR), there is still a significant degree of underestimation of the CATEs.

A major reason for the difficulties with CATE estimation is the counterfactual problem where we can only observe one treatment/control outcome for any subject. Another reason is that the average values of the CATEs are weak in the data, since they are only at a magnitude of around 1/20 of an outcome value. Unfortunately, such weak signals are common in many real-world applications. In marketing promotion and personalised medicine, the scale of treatment effects is commonly quite small. Therefore, treatment effect heterogeneity modelling and uplift modelling are challenging and will require a significant amount of research in the future.

5.3 Evaluation Metrics

Before advancing to further discussions, we formally introduce the metrics commonly used when evaluating causal effect heterogeneity modelling or uplift modelling methods. The evaluation of CATE estimation methods on real-world datasets is a difficult task. Several metrics have been proposed from both



(a) Two-Model LR. (b) Two-Model RT. (c) Two-Model RF.

(d) X-Learner LR. (e) X-Learner RT. (f) X-Learner RF.

(g) Transformed Outcome LR. (h) Transformed Outcome RT. (i) Transformed Outcome RF.

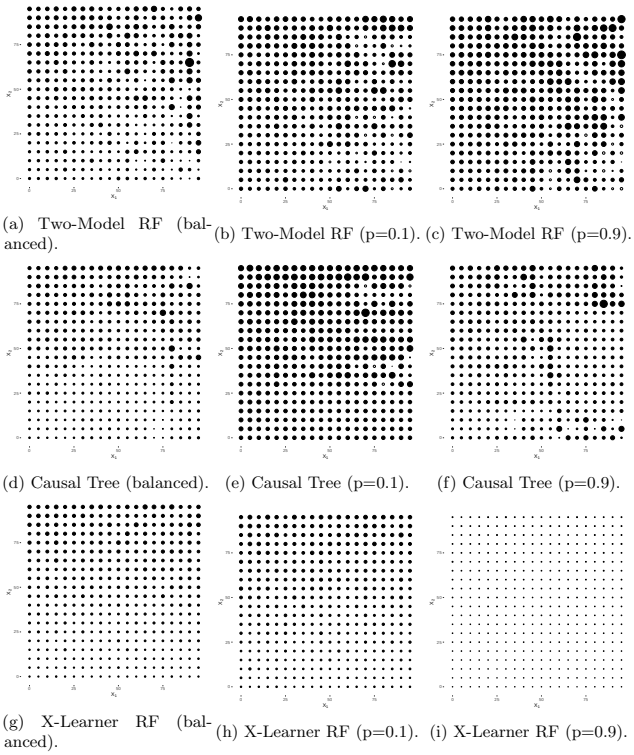
(j) Squared t-Statistics Tree. (k) Causal Tree. (l) Causal Forest.

Figure 4: Estimated CATEs of representative methods on the balanced synthetic dataset. The disc plots use the same scale as in Subfigure 3(c).

communities depending on whether ground-truth CATEs are available.

5.3.1 Metrics With Known Ground-Truth Treatment Effects

When the ground-truth CATEs are known, i.e., in synthetic datasets or semi-synthetics where the potential outcomes are simulated based on real-world covariates, the precision in the estimation of heterogeneous effects (PEHE) (Hill, 2011) is a straightforward metric used for datasets with ground-truth individual



(a) Two-Model RF (balanced). (b) Two-Model RF (p=0.1). (c) Two-Model RF (p=0.9).

(d) Causal Tree (balanced). (e) Causal Tree (p=0.1). (f) Causal Tree (p=0.9).

(g) X-Learner RF (balanced). (h) X-Learner RF (p=0.1). (i) X-Learner RF (p=0.9).

Figure 5: Illustration of X-Learner linear regression, transformed outcome linear regression, and causal forest in datasets of different treatment and control sample ratios. Hollow discs indicate that the estimated CATE is negative. The disc plots use the same scale as in Subfigure 3(c).

treatment effects. Given the ground-truth CATE, the PEHE is defined as:

$$\text{PEHE} = \frac{1}{n}\sum_{i}^{n}(\hat{\tau}(\boldsymbol{x}_i) - \tau(\boldsymbol{x}_i))^2. \tag{30}$$

In other words, the PEHE measures the mean squared error between the estimated treatment effects and the ground-truth treatment effects.

Another performance metric is the mean absolute percentage error (MAPE) of the CATE estimation, which is defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i}^{n}|\frac{\hat{\tau}(\boldsymbol{x}_i) - \tau(\boldsymbol{x}_i)}{\tau(\boldsymbol{x}_i)}| \times 100\% \tag{31}$$

5.3.2 Metrics Without Ground-Truth Treatment Effects

The uplift curve and the Qini curve can be used for the evaluation when the outcome variable is binary without ground-truth CATEs.

The uplift and Qini curves are two closely related metrics proposed in the uplift modelling literature (Radcliffe, 2007; Gutierrez and Gérardy, 2017; Eustache et al., 2018). The intuition behind these metrics is that when the subjects are ranked in descending order by their estimated CATEs, with an accurate CATE estimation, subjects with positive outcomes in the treatment group should be ranked higher than those with negative outcomes in the treatment group. Likewise, subjects with negative outcomes in the control group should be ranked higher than those with positive outcomes in the control group.

To formerly define these metrics, we introduce some notation. For a given CATE estimator $\hat{\tau}$ and subjects $\boldsymbol{x_i}$, let π be a descending ordering of subjects according to their estimated treatment effects, i.e., $\hat{\tau}^{\pi}(\boldsymbol{x}_i) \geq \hat{\tau}^{\pi}(\boldsymbol{x}_j), \forall i < j$. We use $\pi(k)$ to denote the first k subjects. Let $R_{\pi(k)}$ be the count of positive outcomes in $\pi(k)$, i.e., $R_{\pi(k)} = \sum_{i \in \pi(k)} \mathbb{1}[Y_i = 1]$, where $\mathbb{1}$ denotes the indicator function. Furthermore, let $R^{T=1}_{\pi(k)}$ and $R^{T=0}_{\pi(k)}$ be the number of positive outcomes in the treatment and control groups, respectively, from $\pi(k)$. Finally, let $N^{T=1}_{\pi(k)}$ and $N^{T=0}_{\pi(k)}$ be the number of subjects in the treatment and control groups from $\pi(k)$. Now we can define the values of the two curves as:

$$\text{uplift}(k) = \left(\frac{R^{T=1}_{\pi(k)}}{N^{T=1}_{\pi(k)}} - \frac{R^{T=0}_{\pi(k)}}{N^{T=0}_{\pi(k)}}\right) \cdot \left(N^{T=1}_{\pi(k)} + N^{T=0}_{\pi(k)}\right) \tag{32}$$

$$\text{Qini}(k) = R^{T=1}_{\pi(k)} - R^{T=0}_{\pi(k)} \frac{N^{T=1}_{\pi(k)}}{N^{T=0}_{\pi(k)}}. \tag{33}$$

The uplift and Qini curves can be drawn by varying k in the above equations. The uplift and Qini curves are similar in terms of their shape (Gutierrez and Gérardy, 2017).

Despite the prevalence of the uplift and Qini curves in the uplift modelling community, the treatment effect heterogeneity modelling community has not adopted these metrics. Since the ground truth CATEs are often not known in many sociology and medical applications, the uplift and Qini curves could be utilised to alleviate some of the evaluation difficulties encountered in the treatment effect heterogeneity modelling community.

Furthermore, due to ethical and cost concerns, evaluations of CATE estimation methods have been difficult. The current literature often relies on synthetically generated data to obtain ground-truth CATEs. Due to different data generating procedures, such evaluations may produce biased conclusions. Therefore, a standardised and comprehensive set of benchmark datasets would be instrumental in providing fair comparisons of methods and advancements in the field. Recently, both communities have made efforts in this direction. For example, the treatment effect heterogeneity modelling community has proposed several benchmarks (Winkel and Krebs, 2017; Dorie et al., 2019), and the uplift modelling community has contributed a large-scale online advertising benchmark Eustache et al. (2018).

5.4 Demonstration of Semi-Synthetic and Real-World Datasets

In this section, we show how the methods work using three datasets: two semi-synthetic datasets frequently used in CATE estimation literature, and a real-world advertisement campaign dataset frequently used in the uplift modelling literature. Our purpose here is to show how the methods are used in real-world problems, how they are evaluated, and their limitations, instead of assessing which method is better. Since not all methods are applicable to all types of datasets—e.g., most uplift modelling methods cannot be applied to a dataset with continuous outcomes—we only apply the applicable methods to each of the datasets.

Infant Health Development Program (IHDP) The IHDP dataset comes from a randomised study designed to evaluate the effects of home visits from doctors on the cognitive scores of premature infants (Brooks-Gunn et al., 1992). The program began in 1985, and the subjects of the program were low-birth-weight, premature infants. Subjects in the treatment group were provided with intensive, high-quality childcare and home visits from a trained health-care provider. The program was effective at improving the cognitive function of the treated subjects when compared with the control subjects.

A version of this dataset was first used as a semi-synthetic dataset for evaluating CATE estimation in (Hill, 2011), where the outcomes were synthetically generated according to the original covariates, and selection bias was introduced by removing all subjects with non-Caucasian mothers. The resulting dataset contained 747 subjects (608 control and 139 treated) with 25 covariates (6 continuous and 19 binary covariates) that described both the characteristics of the infants and the characteristics of their mothers. The methods for generating the synthetic outcomes are described below.

We followed the same procedure as that described in (Hill, 2011; Johansson et al., 2016; Louizos et al., 2017) to replicate two settings of this semi-synthetic dataset, in which the counterfactual outcomes were simulated using the non-parametric causal inference (NPCI) package (Dorie, 2016). Since the covariates are the same as in the original IHDP dataset, the difference between the two settings lies in how the outcomes are simulated. That is, "Setting A" simulates a linear relationship between the outcome and the covariates, whereas "Setting B" simulates an exponential relationship. The reported performance was calculated by averaging over 100 replications with a training/validation/test split proportion of 60%/30%/10%.

For the parameters of neural network-based methods, we used a grid search to search for an optimal parameter set that achieved the minimum loss on the validation dataset, which consisted of 30% of the whole dataset. We then trained the network using the selected parameters on the entire set. For CFR, the parameters we grid-searched included combinations of representation layers ($\{3, 4, 5\}$), regression layers ($\{3, 4, 5\}$), pre-representation layer dimensions ($\{100, 200, 300, 400, 500\}$), post-representation layer dimensions ($\{100, 200, 300, 400, 500\}$), and imbalance regularisation parameters ($\{0, 0.001, 0.01, 0.1, 0.316, 1, 3.16, 10\}$).

Table 3: Means and standard errors of PEHE and MAPE (smaller is better) across 100 replications for training and test sub-datasets of IHDP. The first group of methods are tree-based. "T-" stands for instantiations of the two-model approach, "X-" for instantiations of X-Learner, and "TO-" for instantiations of the transformed outcome approach. The last group of methods are deep learning-based. Tailored uplift modelling methods are not reported, since their implementations are restricted to datasets with binary outcomes.

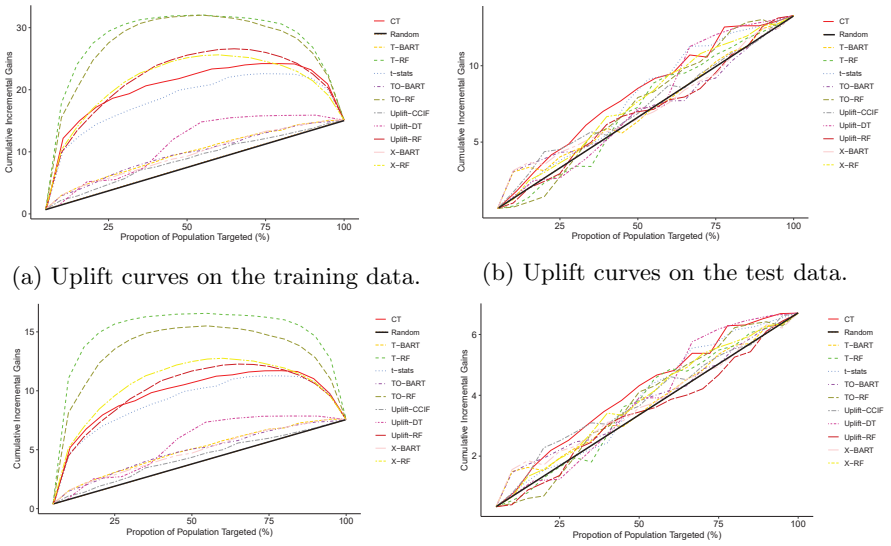
	Setting A				Setting B			
	PEHE^{tr}	PEHE^{te}	MAPE^{tr} (%)	MAPE^{te} (%)	PEHE^{tr}	PEHE^{te}	MAPE^{tr} (%)	MAPE^{te} (%)
t-stats	1.48 ± 0.12	1.56 ± 0.13	48.3 ± 2.5	113.8 ± 25.3	6.92 ± 0.10	5.68 ± 0.09	771.3 ± 193.4	867.7 ± 127.6
CT	1.48 ± 0.12	1.56 ± 0.13	56.5 ± 4.8	148.5 ± 54.8	6.92 ± 0.10	5.70 ± 0.10	631.6 ± 81.4	841.4 ± 113.5
CF	1.01 ± 0.08	1.09 ± 0.16	34.5 ± 4.1	65.6 ± 16.9	2.77 ± 0.03	3.02 ± 0.03	331.9 ± 87.8	436.6± 103.0
T-RF	0.86 ± 0.69	0.99 ± 0.09	30.5 ± 3.4	56.5 ± 18.1	2.89 ± 0.03	3.15 ± 0.04	426.7 ± 94.5	516.3 ± 214.9
T-BART	0.60 ± 0.02	0.68 ± 0.04	19.1 ± 1.7	34.1 ± 10.7	2.30 ± 0.03	2.51 ± 0.04	302.9 ± 50.7	320.5 ± 110.8
X-RF	0.98 ± 0.08	1.09 ± 0.15	42.6 ± 5.1	109.5 ± 40.3	3.50 ± 0.04	3.59 ± 0.06	421.3 ± 92.3	516.3 ± 216.0
X-BART	0.58 ± 0.02	0.66 ± 0.04	18.7 ± 1.9	32.5 ± 9.6	2.29 ± 0.03	2.49 ± 0.04	301.2 ± 52.3	313.5 ± 114.8
TO-RF	1.01 ± 0.09	1.07 ± 0.10	32.8 ± 3.8	80.2 ± 29.0	2.93± 0.03	3.15 ± 0.05	378.5 ± 95.3	450.3 ± 98.6
TO-BART	0.86 ± 0.02	0.91± 0.04	25.0 ± 2.7	45.8 ± 16.0	2.40 ± 0.03	2.60 ± 0.04	338.2 ± 68.9	336.6 ± 138.4
CFR	0.67 ± 0.02	0.73 ± 0.04	23.4 ± 2.5	40.6 ± 13.8	2.60 ± 0.04	2.76 ± 0.04	338.2 ± 68.9	385.6 ± 161.9
SITE	0.65 ± 0.07	0.67 ± 0.06	22.3 ± 2.4	38.5 ± 11.2	2.65 ± 0.04	2.87 ± 0.05	338.2 ± 68.9	390.7 ± 168.0
CEVAE	1.13 ± 0.07	1.37 ± 0.19	46.6 ± 4.2	86.7 ± 20.8	3.06 ± 0.03	3.42 ± 0.05	400.5 ± 110.8	490.8 ± 135.8

The grid consisted of 1800 parameter combinations. For CEVAE, we grid-searched combinations of hidden layers ($\{3, 4, 5\}$), hidden layer dimensions ($\{100, 200, 300, 400, 500\}$), latent factor dimensions ($\{5, 10, 15, 20, 25, 30\}$), and learning rates ($\{0.0001, 0.005, 0.001, 0.05, 0.01\}$). This grid consisted of 450 parameter combinations. For all methods based on an ensemble of trees (i.e., random forests and BART-based algorithms) the best number of trees was selected from 100 to 1000 in intervals of 100. For tree-based methods, pruning was conducted by cross-validation implemented within the software packages.

The results of the methods on the two settings of the IHDP datasets are shown in Table 3. For Setting A, we note that at least one method in each group performs well, and we cannot categorically say which group of methods is better than others. From the application viewpoint, about half of the methods produce some useful models, since their MAPEs are around 50% or less, and a quarter of methods do not produce useful models, since their MAPEs are more than 100%. This further shows that CATE estimation is very challenging, considering that the underlying relationships between CATEs and the covariates are linear. For Setting B, we can see that the performance of all the methods is not satisfactory, since the smallest MAPE is larger than 300%. This can be explained by the fact that the sample size of the IHDP semi-synthetic datasets is rather small (747 samples), and thus it would be difficult for the non-parametric recovery of the non-linear exponential relationships between the CATEs and the covariates.

Hillstrom's Email Advertisement Dataset Here we use the Hillstrom's Email Advertisement dataset Hillstrom (2008) (as discussed in Section 4.1) to illustrate an uplift modelling example. We used the "men's advertisement email" as the treatment and the visit status as the outcome. The algorithms showed similar performance when using the "women's advertisement" as the treatment.

On average, the "men's advertisement" treatment increased the "visit" out-



(a) Uplift curves on the training data.

(b) Uplift curves on the test data.

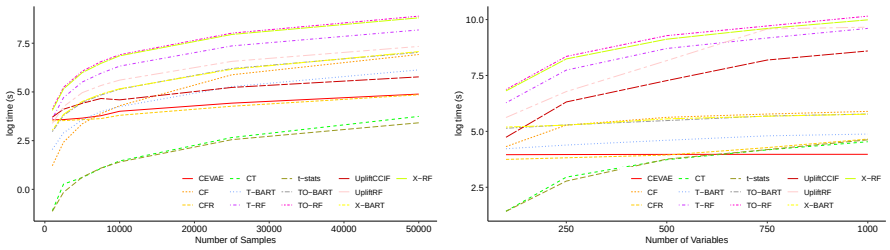
(c) Qini curves on the training data.

(d) Qini curves on the test data.

Figure 6: Uplift and Qini curves for the compared methods on the training and test sets of the Hillstrom's Email Advertisement dataset. The black lines indicate the curves by random predictions. "T-" stands for instantiations of the two-model approach, "X-" for instantiations of X-Learner, and "TO-" for instantiations of the transformed outcome approach. Figures are best viewed in colour.

come by 7.6%. We ran the algorithms using a 70%/30% split of training and test sets without repeats, and we used the same parameter selection procedure as that described in the last section. Note that the source codes for CF, CFR, and CEVAE are not designed to handle categorical covariates. We tried transforming the categorical covariates into binary codes with one-hot encoding. However, the results were poor, and thus we have not included them here.

The uplift curves and the Qini curves of the evaluated methods on the Hillstrom Email dataset are illustrated in Figure 6. Firstly, the majority of models do predict uplifts in the test dataset, but the uplifts in the test dataset are much smaller than the uplifts in the training dataset. The rankings of the methods based on the uplifts between the training dataset and the test dataset are also inconsistent. The above two observations indicate difficulty in evaluating the performance of uplift modelling when the ground truth is unknown (which is common in practice). Cross-validation is often used in evaluating an uplift model, but it is an open question whether cross-validation is a valid means, since the uplifts are unobserved in the test datasets. In contrast, for evaluating a supervised method, the outcomes are observed in the test dataset. Secondly, the trends of uplift curves and Qini curves differ slightly, although they are largely consistent. For example, X-RF is better than Uplift-RF in the uplift curves, but



(a) Number of subjects.

(b) Number of covariates.

Figure 7: Running time of the compared methods. The curves for CFR and SITE are similar. We only show CFR, since one is hidden behind the other. The Y axis is in logarithmic form. "T-" stands for instantiations of the two-model approach, "X-" for instantiations of X-Learner, and "TO-" for instantiations of the transformed outcome approach. Figures are best viewed in colour.

worse than Uplift-RF in the Qini curves.

The illustrations of the IHDP and Hillstrom's Email Advertisement datasets serve as a good example of cross-fertilisation between the two communities. Almost all methods developed in the uplift modelling community are designed for binary outcomes, whereas most methods from the treatment effect heterogeneity modelling community are designed for continuous outcomes. On the one hand, suppose the uplift needs to be modelled for the amount of sales increase instead of purchasing or not purchasing, methods in the treatment effect heterogeneity modelling community can be directly applied to solve this extended uplift modelling problem. On the other hand, uplift modelling methods can also be applied to sociology and medical problems when the outcomes of interest are binary, e.g., whether a medicine is effective for patients with specific gene mutations.

5.5 Scalability

As datasets get larger, it becomes imperative to understand how the CATE estimation and uplift modelling methods scale. We compared the running time of different methods by varying the numbers of subjects and covariates. The comparison was conducted using an AMD Ryzen 3700x CPU and 32 GB of RAM. For deep learning-based methods, the running time was obtained using a single Nvidia GeForce GTX 1080 Ti GPU with 11 GB of RAM. For GPU training, the batch sizes were set to one-fifth the number of subjects, and a total of 200 epochs of training were performed. Each experiment was repeated 10 times and the average running time was reported.

We utilised a synthetic dataset proposed in (Häggström, 2017). In this dataset, the causal structure between the 10 covariates, the treatment T, and the outcome Y is depicted in Häggström (2017). To evaluate the scalability with regard to the number of variables, additional variables (which are not related to the treatment

Table 4: Summary of the usability, interpretability, and scalability of the uplift modelling and CATE estimation packages and algorithms.

Methods	Usability	Interpretability	Scalability
Single-model **Two-Model** **X-Learner** **Transformed outcome**	Depends on the base method, and can be very easy to use.	Treatment effect is not given and needs to be derived. Results are interpretable.	Scalability to the #covariates and #subjects can be very good.
Causal Tree **t-stats Tree** **Uplift Tree**	Easy to use with few parameters to set.	Treatment effect is directly given and interpretable.	Scalability to the #covariates* and #subjects is very good.
Causal RF **Uplift RF** **Uplift CCIF**	Easy to use with few parameters to set.	Treatment effect is directly given. Results are not interpretable.	The time for building a model can be long, and the scalability to #covariates and #subjects is not good.
CFR **CEVAE** **SITE**	The network structure and parameters are difficult to set.	Treatment effect is directly given. Results are not interpretable.	Good for big data. It takes a long time to train a model, but the training time is constant and unaffected by #covariates or #subjects.

*When using propensity score matching, the scalability with #covariates is not good.

or the outcome) were randomly sampled from Gaussian distributions and added to the covariates. As a result, the number of variables varied from 10 to 1000 with the number of subjects fixed at 10,000. For evaluating scalability with regard to the number of subjects, we fixed the number of variables at 100, and varied the number of subjects from 1000 to 50,000. Multiple CPU parallelisation was not used for the ensemble-based algorithms. In other words, the time efficiency for ensemble-based methods can be improved by utilising multiple CPUs.

The running time comparison for different numbers of subjects is shown in Figure 7(a). The two tree-based methods, causal tree (CT) and t-statistics tree (t-stats), are the fastest among all the compared algorithms. The representative deep learning-based algorithms CEVAE and CFR use almost constant time regardless of the number of subjects. This means that they are efficient when the number of subjects is large. The slowest running methods are those based on BART, including X-BART, T-BART, and TO-BART. Methods based on random forests, i.e., causal forest (CF), X-RF, T-RF, and TO-RF, are faster than the deep learning-based algorithms when sample sizes are small, but become slower when the numbers of subjects increase.

The comparison of different numbers of covariates is shown in Figure 7(b). The two tree-based methods, CT and t-stats, are the fastest. Deep learning-based algorithms CEVAE and CFR also use almost constant time regardless of the number of variables. They can handle datasets with a large number of variables very well. The slowest running methods are those based on random forests, including X-RF, T-RF, TO-RF, and UpliftRF. The BART-based algorithms take similar time, and they are faster than RF-based algorithms but slower than deep learning-based algorithms.

5.6 Discussions

Table 4 summarises the usability, interpretability, and scalability of the methods implemented in the packages and codes listed in Tables 1 and 2. Usability is about the ease of use of the implementation of a method in the packages or codes, i.e., parameter setting and tuning. Interpretability refers to the level of explanation provided for a prediction, for example, why a predicted outcome may be linked to some specific covariate value. Scalability is about the speed of a method in relation to the dataset size and number of covariates. Based on our experience with various uplift modelling and treatment effect heterogeneity modelling methods, we have the following observations:

- The accuracy of the methods is data-specific. For example, X-Learner BART performed well with the IHDP dataset, but only performed marginally better than the baseline for Hillstrom's Email Advertisement dataset. Currently, there is no good understanding of which method suits the best type of data.

- The accuracy of ensemble methods is stable across different datasets, although they may not be the best all the time. This is expected, as discussed in (Sołtys et al., 2015) where the authors observed that ensemble methods frequently outperformed methods that build a single model. However, ensemble methods suffer from a lack of interpretability and their training time is long.

- The accuracy of the single-model approach is low, while the accuracies for the two-model approach and the X-Learner are competitive. Based on their design, the two-model approach is good for datasets with balanced treatment and control subjects, and the X-Learner is good for unbalanced data. A strength of both the X-Learner and the two-model approaches is that they provide good flexibility for using a rich set of existing supervised methods.

- The performance of deep learning-based methods is determined by dedicated parameter tuning for each specific dataset. Neural networks are known to be prone to data variability and parameter selection (Novak et al., 2018). This is more problematic in CATE estimation than in supervised learning, since there are no ground-truth treatment effects available in most cases.

In this survey, we focus our discussion on a single binary treatment and exclude the discussion of multiple treatments and non-binary treatments (i.e., ordinal or continuous treatments). We focus on a single treatment, since almost all of the surveyed methods can be extended to multiple treatments, in a manner similar to extending binary classification to multiple classes. There are some methods specifically designed for handling multiple treatments. We refer the reader to Olaya, Coussement, and Verbeke (Olaya et al., 2020) for a recent survey on uplift modelling with multiple treatments. We exclude a discussion

of non-binary treatments, since most existing methods are designed for binary treatments. Recently, several methods for non-binary treatments have been proposed from the treatment effect heterogeneity modelling community, and we refer the reader to Zhao, Dyk, and Imai (Zhao et al., 2020) for an overview of these methods.

6 Conclusions

Estimating the heterogeneous effects of an action on the outcomes of individual subjects is an important problem with a wide range of applications. Motivated by different applications, researchers from the treatment effect heterogeneity modelling and the uplift modelling communities have both contributed to the problem. In this article, we provided a unified survey of the methods proposed by the two communities. Using the potential outcome framework, we showed that, with the overlap, SUTV, and unconfoundedness assumptions, the objectives of treatment effect heterogeneity modelling and uplift modelling are the same: they both aim to estimate the conditional average treatment effect (CATE) from data. With a unified objective and notation, we systematically reviewed the methods developed by the two communities, focusing on the inherent connections between them. Although most methods are strongly linked to supervised machine learning methods, we should stress here that CATE estimation is not supervised learning, and that it is crucial to understand the assumptions to ensure the correct use of the methods. We discussed applications of the methods in targeted marketing, personalised medicine, and social studies. Finally, we summarised existing open-source software packages and source codes, and demonstrated their implementation with synthetic, semi-synthetic, and real-world datasets. We showed that CATE estimation and evaluation are challenging tasks, and we offered some general guidelines for method and software-tool selection based on our experience.

An important direction for future work is to find balance between the interpretability, ease of use, and accuracy of CATE estimation algorithms. On the one hand, black-box methods such as deep learning-based methods are accurate at CATE estimation with carefully tuned parameters. However, their parameter tuning procedures are difficult, and the models are generally not interpretable. On the other hand, tree-based methods are easy to interpret and require minimal parameter tuning, although their performance is often limited. An emerging direction for bridging this gap is to construct tree-based models on top of the results from other CATE estimators (Lee et al., 2020).

acknowledgement

The work has been partially supported by ARC Discovery Projects grant DP170101306.

References

Onur Atan, James Jordon, and Mihaela van der Schaar. 2018. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2071–2078.

Susan Athey and Guido W. Imbens. 2015. *Machine learning for estimating heterogeneous causal effects*. Technical Report. Stanford University. `https://www.gsb.stanford.edu/gsb-cmis/gsb-cmis-download-auth/406621`

Susan Athey and Guido W. Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

Peter C. Austin. 2011. An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46, 3 (2011), 399–424.

Jennifer R. Bellon. 2015. Personalized radiation oncology for creast cancer: The new frontier. *Journal of Clinical Oncology* 33, 18 (2015), 1998–2000.

Artem Betlei, Eustache Diemert, and Massih-Reza Amini. 2020. Optimization of treatment assignment with generalization guarantees. In *Causal Learning for Decision Making: ICLR 2020 Workshop*.

Leo Breiman. 1996. Bagging Predictors. *Machine Learning* 24, 2 (1996), 123–140.

Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. 1984. *Classification and Regression Trees*. Wadsworth and Brooks.

David Broockman and Joshua Kalla. 2016. Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* 352, 6282 (2016), 220–224.

Jeanne Brooks-Gunn, Fong ruey Liaw, and Pamela K. Klebanov. 1992. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of Pediatrics* 120, 3 (1992), 350–359.

Wray Buntine. 1993. Learning classification trees. *Statistics and Computing* 2 (1993), 63–73.

Tianxi Cai, Lu Tian, Peggy H. Wong, and L. J. Wei. 2011. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12, 2 (2011), 270–82.

David M. Chickering and David E. Heckerman. 2000. A decision theoretic approach to targeted advertising. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. 82–88.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (1995), 273–297.

Xavier de Luna, Ingeborg Waernbaum, and Thomas S. Richardson. 2011. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* 98, 4 (2011), 861–875.

Rajeev H. Dehejia and Sadek Wahba. 1999. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 448 (1999), 1053–1062.

Floris Devriendt, Darie Moldovan, and Wouter Verbeke. 2018. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: a stepping stone toward the development of prescriptive analytics. *Big Data* 6, 1 (2018), 13–41.

Vincent Dorie. 2016. NPCI: Non-parametrics for causal inference. (2016). `https://github.com/vdorie/npci`

Vincent Dorie. 2017. Tools and data for the Atlantic causal Inference conference competition. (2017). `https://github.com/vdorie/aciccomp`

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. 2019. Automated versus do-It-yourself methods for causal inference: lessons learned from a data analysis competition. *Stastical Science* 34, 1 (2019), 43–68.

Doris Entner, Patrik Hoyer, and Peter Spirtes. 2013. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*. 256–264.

Diemert Eustache, Betlei Artem, Renaudin Christophe, and Massih-Reza Amini. 2018. A large scale benchmark for uplift modeling. In *Proceedings of the 2018 AdKDD and TargetAd Workshop*.

Alan S. Gerber and Donald P. Green. 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review* 94, 3 (2000), 653–663.

Robin Gubela, Artem Bequ é, Stefan Lessmann, and Fabian Gebert. 2019. Conversion uplift in e-commerce: A systematic benchmark of modeling strategies. *International Journal of Information Technology and Decision Making* 18, 03 (2019), 747–791.

Leo Guelman, Montserrat Guillén, and Ana M. Pérez-Marín. 2012. Random forests for uplift modeling: An insurance customer retention case. In *Proceedings of the International Conference on Modeling and Simulation in Engineering, Economics and Management*. 123–133.

Leo Guelman, Montserrat Guillén, and Ana M. Pérez-Marín. 2015a. Uplift random forests. *Cybernetics and Systems* 46, 3-4 (2015), 230–248.

Leo Guelman, Montserrat Guillén, and Ana M. Pérez-Marín. 2015b. A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems* 72, C (2015), 24–32.

Ruocheng Guo, Lu Cheng, Jundong Li, P.Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys* 53, 4 (2020), Article 75.

Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal inference and uplift modelling: A review of the literature. In *Proceedings of The 3rd International Conference on Predictive Applications and APIs*. 1–13.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

Scott M. Hammer, Kathleen E. Squires, Michael D. Hughes, Janet M. Grimes, et al. 1997. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine* 337, 11 (1997), 725–733.

Behram Hansotia and Brad Rukstales. 2002. Incremental value modeling. *Journal of Interactive Marketing* 16, 3 (2002), 35–46.

Negar Hassanpour and Russel Greiner. 2018. Counterfactual regression with importance sampling weights. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 5880–5887.

Amelie Heliou, Matthieu Martin, Christophe Renaudin, and Eustache Diemert. 2020. Individual treatment effect in presence of observable interference. In *Causal Learning for Decision Making: ICLR 2020 Workshop*.

Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.

Kevin Hillstrom. 2008. The MineThatData e-mail analytics and data mining challenge. (2008). https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html.

Jenny Häggström. 2017. Data-driven confounder selection via Markov and Bayesian networks. *Biometrics* 74, 2 (2017), 389–398.

Kosuke Imai and Marc Ratkovic. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7, 1 (2013), 443–470.

Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Szymon Jaroszewicz and Łukasz Zaniewicz. 2015. Székely regularization for uplift modeling. In *Challenges in Computational Statistics and Data Mining*. Studies in Computational Intelligence, Vol. 605. Springer, 135–154.

Maciej Jaskowski and Szymon Jaroszewicz. 2012. Uplift modeling for clinical trial data. In *Machine Learning for Clinical Data Analysis: ICML 2012 Workshop*. `http://people.cs.pitt.edu/~milos/icml_clinicaldata_2012/Papers/Oral_Jaroszewitz_ICML_Clinical_2012.pdf`

Thorsten Joachims. 2005. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine learning*. 377–384.

Fredrik D. Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning*. 3020–3029.

Kathleen Kane, Victor S.Y. Lo, and Jane Zheng. 2014. Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics* 2, 4 (2014), 218–238.

Edward L. Kaplan and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 282 (1958), 457–481.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Michael C Knaus, Michael Lechner, and Anthony Strittmatter. 2020. Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal* 24, 1 (2020), 134–161.

Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. 2014. Support vector machines for differential prediction. In *Lecture Notes in Computer Science*. ECML PKDD, Vol. 8725. 50–65.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165.

Robert J. LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* 76, 4 (1986), 604–620.

Hyun-Suk Lee, Yao Zhang, William Zame, Cong Shen, Jang-Won Lee, and Mihaela van der Schaar. 2020. Robust recursive partitioning for heterogeneous treatment effects with uncertainty quantification. In *Advances in Neural Information Processing Systems*, Vol. 33. 2282–2292.

Jiuyong Li, Weijia Zhang, Lin Liu, Kui Yu, Thuc Duy Le, and Jixue Liu. 2021. A general framework for causal classification. *International Journal of Data Science and Analytics* (2021), Advance online publication.

Victor S.Y. Lo. 2002. The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter* 4, 2 (2002), 78–86.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, Vol. 30. 6449–6459.

Marloes H Maathuis, Diego Colombo, et al. 2015. A generalized back-door criterion. *The Annals of Statistics* 43, 3 (2015), 1060–1088.

Daniel F. McCaffrey, Greg Ridgeway, and Andrew R. Morral. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 9, 4 (2004), 403–425.

René Michel, Igor Schnakenburg, and Tobias von Martens. 2017. Effective customer selection for marketing campaigns based on net scores. *Journal of Research in Interactive Marketing* 11, 1 (2017), 2–15.

Stephen L. Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference: methods and principles for social research.* Cambridge University Press.

Houssam Nassif, Finn Kuusisto, Elizabeth S. Burnside, David Page, Jude Shavlik, and Vítor Santos Costa. 2013. Score As You Lift (SAYL): A statistical relational learning approach to uplift modeling. In *Lecture Notes in Computer Science*. ECML PKDD, Vol. 8190. 595–611.

Houssam Nassif, Vítor Santos Costa, Elizabeth S. Burnside, and David Page. 2012a. Relational differential prediction. In *Lecture Notes in Computer Science*. ECML PKDD, Vol. 7523. 617–632.

Houssam Nassif, Yirong Wu, David Page, and Elizabeth Burnside. 2012b. Logical Differential Prediction Bayes Net, improving breast cancer diagnosis for older women. In *American Medical Informatics Association Annual Symposium Proceedings*. 1330–1339.

Xinkun Nie and Stefan Wager. 2020. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* (2020), Advance online publication.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. 2018. Sensitivity and generalization in neural networks: An empirical study. In *Preceedings of the 6th International Conference on Learning Representations*.

Diego Olaya, Kristof Coussement, and Wouter Verbeke. 2020. A survey and benchmarking study of multitreatment uplift modeling. *Data Mining and Knowledge Discovery* 34, 2 (2020), 273–308.

Judea Pearl. 2009. *Causality: Models, reasoning and inference*. Cambridge University Press.

Dmitry Pechyony, Rosie Jones, and Xiaojing Li. 2013. A joint optimization of incrementality and revenue to satisfy both advertiser and publisher. In *Proceedings of the 22nd International Conference on World Wide Web*. 123–124.

Bertram Pitt, Faiez Zannad, Willem J. Remme, Robert Cody, Alain Castaigne, Alfonso Perez, Jolie Palensky, and Janet Wittes. 1999. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. *New England Journal of Medicine* 341, 10 (1999), 709–717.

John R. Quinlan. 1993. *C4. 5: Programs for Machine Learning*. Morgan Kaufmann.

Nicholas J. Radcliffe. 2007. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal* 3 (2007), 14–21.

Nicholas J. Radcliffe. 2008. *Hillstrom's MineThatData email analytics challenge: An approach using uplift modelling*. Technical Report. Stochastic Solutions. `https://www.stochasticsolutions.com/pdf/HillstromChallenge.pdf`

Nicholas J. Radcliffe and Patrick Surry. 2011. *Real-world uplift modelling with significance-based uplift trees*. Technical Report. Stochastic Solutions.

Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688–701.

Donald B. Rubin. 1997. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine* 127, 8 (1997), 757–763.

Piotr Rzepakowski and Szymon Jaroszewicz. 2010. Decision trees for uplift modeling. In *Proceedings of the 10th IEEE International Conference on Data Mining*. 441–450.

Piotr Rzepakowski and Szymon Jaroszewicz. 2012a. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems* 32, 2 (2012), 303–327.

Piotr Rzepakowski and Szymon Jaroszewicz. 2012b. Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology* (2012), 43–50.

Soko Setoguchi, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn, and E. Francis Cook. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety* 17, 6 (2008), 546–555.

Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*. 3076–3085.

Michał Sołtys, Szymon Jaroszewicz, and Piotr Rzepakowski. 2015. Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1531–1559.

Xiaogang Su, Joseph Kang, Juanjuan Fan, Richard A. Levine, and Xin Yan. 2012. Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research* 95, 13 (2012), 2955–2994.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nkckerson, and Bogong Li. 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10 (2009), 141–158.

Sami Tabib and Denis Larocque. 2020. Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics* 36, 2 (2020), 629–636.

Tyler J. VanderWeele and Ilya Shpitser. 2011. A new criterion for confounder selection. *Biometrics* 67(4) (2011), 1406–1413.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: learning useful pepresentations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11 (2010), 3371–3408.

Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113, 523 (2018), 1228–1242.

Herbert I. Weisberg and Victor P. Pontes. 2015. Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials* 12, 4 (2015), 357–364.

Julian Winkel and Tobias Krebs. 2017. Data generating process simulation: the OPOSSUM package. (2017). `https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/data_generating_process_blogpost/#Package-application`

Ikko Yamane, Florian Yger, Jamal Atif, and Masashi Sugiyama. 2018. Uplift modeling from separate labels. In *Advances in Neural Information Processing Systems*, Vol. 31. 9949–9959.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, Vol. 31. 2638–2648.

Jinsung Yoon, James Jordan, and Mihaela van der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *Proceedings of the 6th International Conference on Learning Representations*.

Łukasz Zaniewicz and Szymon Jaroszewicz. 2013. Support vector machines for uplift modeling. In *Proceedings of the 13th International Conference on Data Mining Workshops*. 131–138.

Łukasz Zaniewicz and Szymon Jaroszewicz. 2017. L_p-Support vector machines for uplift modeling. *Knowledge and Information Systems* 53, 1 (2017), 269–296.

Weijia Zhang, Thuc Duy Le, Lin Liu, Zhi-Hua Zhou, and Jiuyong Li. 2017. Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics* 33, 15 (2017), 2372–2378.

Weijia Zhang, Lin Liu, and Jiuyong Li. 2021. Treatment effect estimation with disentangled latent factors. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. Advance online publication.

Shandong Zhao, David A van Dyk, and Kosuke Imai. 2020. Propensity score-based methods for causal inference in observational studies with non-binary treatments. *Statistical Methods in Medical Research* 29, 3 (2020), 709–727.