Causal Classification: Treatment Effect vs. Outcome Estimation

Carlos Fernandez
NYU Stern
New York, NY, USA
cfernand@stern.nyu.edu

Foster Provost
NYU Stern
New York, NY, USA
fprovost@stern.nyu.edu

ABSTRACT

The goal of causal classification is to identify (classify) individuals whose outcome would be positively changed by a treatment. Large-scale examples include targeting (online) advertisements and targeting retention incentives to high-value customers or employees who are at high-risk for attrition. Causal classification can be remarkably difficult for multiple reasons, most importantly because data at best show each individual under only one treatment condition, so it is impossible to know for certain which individuals had a positive outcome *due to* the treatment. Therefore, the potential outcomes for each treatment condition are estimated either explicitly or implicitly in causal classification. Then, these estimates are used to decide which individuals to target with a treatment. Curiously, in practice we see causal classification problems being treated simply as outcome prediction rather than as a causal inference task, e.g., will someone purchase if shown the ad? We might write that off as naivete, but perhaps there is a good reason. In this paper, we undertake a theoretical analysis comparing treatment effect estimation vs. simple outcome prediction when addressing causal classification. The analytical results show a bias/variance trade-off: because treatment effect estimation depends on two outcome estimates instead of one, the larger variance may lead to higher misclassification error than the (biased) outcome prediction approach. As the analytical results include approximations, we next introduce a flexible simulation environment for experimenting with causal classification; simulation results support the analytical results. The bottom line is that using outcome prediction sometimes is indeed preferable to treatment effect estimation, even when the best-possible models are used for both approaches and there are no estimation challenges (such as confounding). Specifically, outcome prediction is preferable when positive outcomes are (1) very rare, (2) difficult to predict, and when (3) treatment effects are small.

CCS CONCEPTS

• **Information systems → Data mining**;

KEYWORDS

Bias-variance trade-off, predictive modeling, treatment effects

ACM Reference Format:
Carlos Fernandez and Foster Provost. 2018. Causal Classification: Treatment Effect vs. Outcome Estimation. In *Proceedings of* . ACM, New York, NY, USA, 10 pages.

, ,

1 INTRODUCTION

Predictive modeling is applied to improve increasingly many tasks, such as predicting whether customers will leave after their contracts expire, whether customers will purchase after having seen an ad, whether there is fraud present on an account, and many many more. Often, the fundamental problem in these tasks is one of assessing whether an *intervention* will have an effect. For example, when attempting to stop customers from leaving, we don't really just want to target the customers most likely to leave, but the customers for whom our incentive will cause them to stay when they otherwise would not have. For many advertising settings, the goal is not simply to target ads to people who will purchase after seeing the ad, but to target people for whom the ad will increase their likelihood of purchasing. Generally, instead of simply predicting the (likelihood of the) outcome, for tasks such as these we ideally would like to estimate whether the outcome of a particular individual can be changed with a treatment (i.e., the ad or the retention incentive in our examples). These are *causal classification* tasks.

Curiously, for ad targeting, churn incentive targeting, and other large-scale predictive applications where causal classification seems to be called for, practitioners stubbornly continue simply to target based on models predicting outcomes (e.g., whether someone will buy after being targeted) rather than models predicting treatment effects [1]. For example, in targeted online advertising, although ad campaigns are sometimes evaluated for causal effect (e.g., through A/B testing), seldom (if ever) are causal-effect models used for actually targeting the ads! Researchers and savvy practitioners see this as due either to naivite or pragmatics [1, 23]. However, we also see continued use of outcome prediction models even in cases where the users have looked carefully at the estimation of causal treatment effects (cf., [27] and [21] for a clear example).

This paper examines the question: might targeting based on outcome prediction actually be more effective than targeting based on treatment effect estimation in certain important settings? We analyze the problem theoretically and show that there indeed are settings where targeting based on outcome prediction will be more effective than targeting based on treatment effects. This is the case *even if* one has access to the Bayes-optimal models—so the result is not due to the difficulty of learning causal models in the presence of selection biases and other confounding. Which strategy is theoretically better depends on a combination of aspects of the setting, including the fundamental predictability (i.e., the Bayes error rates), the size of the treatment effect, and the base rate of the outcome. We provide additional support for the theoretical results based on a large-scale simulation allowing us to know all of these parameters with certainty, as well as the true Bayes-optimal models. We return to the contributions of the paper and the differences from prior work in the next sections, after defining the problem more precisely and discussing related work.

Table 1: Causal classification labels

Obs. Type	Outcome (no treatment)	Outcome (treatment)	Causal Label
Never-Taker	Negative	Negative	Negative
Always-Taker	Positive	Postive	Negative
Defier	Positive	Negative	Negative
Complier	**Negative**	**Positive**	**Positive**

2 PROBLEM DEFINITION

For this paper, we will consider problems with binary outcomes (e.g., purchase or not) and binary treatments (e.g., show an ad or not). Moreover, we will consider *outcome classification* models, which predict the probability of a positive outcome for treated and untreated instances. Ideally, in *causal classification* we would like to identify *Compliers*, those individuals whose outcome would change from negative to positive if they were to be treated.[1] We use the potential outcomes framework [25] to define the causal-prediction labels in terms of what the outcomes would be with and without the treatment (Table 1)—although in practice one of the outcomes would be counterfactual. So, in addition to Compliers, we have individuals who would have a positive outcome without treatment, individuals who would have a negative outcome even with treatment, and to complete the picture, individuals for whom treatment would have a negative effect on the outcome.

It is important to be clear that we have two different sorts of classification tasks here: outcome prediction and Complier prediction. Ideally, we would like to target Compliers and not waste resources treating the other three categories. More specifically, causal classification is a classification problem in which the goal is to identify Compliers, usually by assigning scores to each observation and giving higher scores to Compliers to separate them from other types of observations.

The main challenge for causal classification, as for other sorts of causal inference tasks, is that in practice we never observe the counterfactual outcome, and therefore the causal label is not visible either. For each observation we know its outcome and whether there was a treatment, but not its counterfactual outcome. Continuing with the previous example, if we observe that a customer made a purchase (i.e., the outcome) after seeing an ad (i.e., the treatment), we do not know if that customer would have purchased *even without* the ad. This has obvious implications for both training and evaluation. First, we cannot label historical cases as Compliers, and second, if a model's prediction leads to a different treatment decision from the one observed in the data (e.g., not showing the ad), we cannot tell if the outcome would have changed.

There are two ways to deal with this challenge. The first is to make strong assumptions about the counterfactuals, such as assuming that outcomes are always negative when there is no treatment. This assumption reduces causal classification to a regular classification task where the goal is to target the observations most likely to be positive given a treatment [18, 23]. However, as the assumption will be violated in the cases we care about here, these predictions are likely to include Always-Takers and therefore will seem wasteful. A second, seemingly more sensible, approach is to estimate the counterfactual by training two classification models to predict the outcomes for the treated and the untreated, and then to score each unlabeled observation based on the difference between the predictions of the two models.[2] This approach has been referred to as "true lift" or "uplift modeling" [18], and can be used for *causal targeting*: targeting observations based on the estimated treatment effect [23]. Causal targeting can also take into account costs and benefits that may affect treatment decisions [23].

As we mentioned in the introduction, despite the first approach seeming to be naive and the second approach seeming to be well considered, practitioners nonetheless broadly follow the first approach, even when they have the data and resources to follow the second. We therefore should think carefully about reasons why the first might actually be preferable. Targeting based on outcome prediction will clearly be biased when used for Complier prediction. However, treatment effect prediction is based on two estimates rather than one and therefore, ceteris paribus, the variance in the scores will tend to be larger. Thus, outcome prediction might perform better, if its errors due to systematic bias are small compared to the errors due to variance for the treatment effect prediction.

We model these effects theoretically and assess them experimentally. We find that each approach is preferable in different situations depending on (1) how unbalanced are the classes, (2) how large is the treatment effect, and (3) how easy it is to predict the outcome. Considering that the treatment effect approach is frequently more expensive in data acquisition and more difficult to implement from a technical point of view, our findings are particularly relevant because they describe the conditions in which a simple outcome model can defeat a causal model in a task that is inherently causal.

3 RELATED WORK

Aspects of causal classification have been considered from three different perspectives: heterogeneous treatment effect (HTE) estimation (e.g., [28, 31–33]), optimal treatment policies (OTP) (e.g., [3, 4, 7]), and uplift (or true lift) modeling (UM) (e.g., [16, 18, 24]). HTE studies differ from ours because their main object of study is the understanding of treatment effects in heterogeneous subpopulations, not treatment assignment.[3] These studies are concerned with the accurate estimation of the difference between the expectation of potential outcomes, whereas we care about this difference only to the extent that it helps us to discriminate Compliers.

On the other hand, UM and OTP studies are concerned with treatment assignment, which is an analogous problem to causal classification. The following are some of the best-known examples. In the econometrics literature, it has been argued that assigning treatments to maximize social welfare is a distinct problem from the point estimation and hypothesis testing problems usually considered in the treatment effects literature [13]. As a result, several authors have proposed parametric or semi-parametric models for

[1] In practice, we may instead be modeling the change in the likelihood of outcome with the treatment, but when the models are applied, identifying Compliers would still be our ideal goal.

[2] One version of this approach is to train a single model to predict the outcome using the treatment as an additional covariate, such as one often does in regression modeling for treatment effect estimation. For the purposes of this paper, the treated and untreated models would be the model with the appropriate setting for the treatment covariate.

[3] Compare with discussions about differences between explanation and prediction [26].

Table 2: Intended targets by approach

	Most Likely (ML)	Max Treatment (MT)
Scoring function	$\hat{P}(Y=1\|X,T=1)$	$\hat{P}(Y=1\|X,T=1)$ $-\hat{P}(Y=1\|X,T=0)$
Intended Targets	Compliers Always-Takers	Compliers
Unintended Targets	Never-Takers	Always-Takers Never-Takers

optimal policies [4, 5, 19]. In the setting of contextual bandits [2], Beygelzimer and Langford [3] propose an algorithm to learn how to make decisions in situations where the payoff of only one choice is observed rather than for all choices. Finally, Kane et al. [16] explain in detail how uplift modeling can be understood as a methodology that combines predictive modeling and experimental design to enable marketers to identify the characteristics of treatment responders separately from the characteristics of control responders.

Our work is different from these studies in at least four important ways. First, we analyze and evaluate causal classification from a standpoint where the goal is to identify Compliers rather than to maximize a reward while acting (as with contextual bandits), mean social welfare (in econometrics), or the difference between treatment and control subgroups (in uplift modeling). Thinking about causal classification in this way is crucial because there are many cases where the treatment cost and the outcome values are not readily available (e.g., when deciding which employees to train), and because it allows us to focus on the real goal: to target the observations for which the treatment changes the outcome into a positive one. Second, previous work has been evaluated mostly through ad hoc simulations or small experiments; we propose a broad, highly flexible but simple simulation framework that can be easily incorporated to test causal classification methods in future research. Third, the focus of our study is to analyze the (bias-variance) trade-off that exists between ignoring the counterfactual (i.e., outcome prediction) and trying to estimate it (i.e., treatment effect estimation). While the trade-off between different types of errors in classification tasks has received significant research attention since Friedman's work [8], to our knowledge this is the first paper that has attempted to transfer these ideas to causal classification. Finally, we develop a set of guidelines to decide whether to deal with causal classification as an outcome predictive task or as a causal estimation task depending on the specific characteristics of the setting and the quality of the models. Overall, our study has important implications for the way we should think about modeling causal classification tasks, as well as for other phases in the data mining life cycle, such as business understanding, data acquisition, and model evaluation.

There are other aspects of causal inference in which the KDD community has made promising strides. Examples include the automatic identification of natural experiments in observational data [15], the estimation [11] and evaluation [9] of causal graphical models, and the estimation of causal effects through mediator variables rather than instrumental variables [12].

4 CAUSAL TARGETING, BIAS & VARIANCE

Consider three approaches to address causal classification:

(1) **Most Likely (ML)**: Target the observations with the highest probability of a positive outcome when treated.
(2) **Least Likely (LL)**: Target the observations with the lowest probability of a positive outcome when not treated.
(3) **Max Treatment (MT)**: Target the observations with the largest difference between the probability of a positive outcome when treated and the probability of a positive outcome when not treated.

The first two essentially ignore the counterfactual, while the latter explicitly estimates the counterfactual.[4] In this section, we compare both types of approach analytically. Due to limited space, we will focus exclusively on ML and MT, but the results for LL are symmetrical to the ones obtained for ML. Moreover, for this analysis we will assume for simplicity that there are no Defiers, leaving that for future research.

Classification problems generally are modeled as scoring problems, where we want observations with a positive class to have a higher score than observations with a negative class; the scores rank observations and classifications are made using a specified threshold. Estimated probabilities of class membership often are used as scores and a default threshold of 0.5 used to discriminate observations, but more generally we would choose a threshold appropriate for the problem at hand [23].

Table 2 shows how ML and MT would attempt to score and discriminate observations. Both approaches favor Compliers, but ML gives a high score to Always-Takers too, which suggests that this approach will have a higher false positive rate due to bias. We can view MT as correcting for the bias in ML, since in principle it reduces the score of the observations that would have a positive outcome even without the treatment. However, the MT scoring function uses two outcome estimates rather than one. Even if both are unbiased, MT will misclassify Compliers if the variance in the untreated estimate "accidentally" lowers the score (and subsequently the rank) of a Complier. This suggests that MT will have a smaller true positive rate due to variance. Therefore, we compare the predictive power of both approaches in terms of their true positive rate (TPR) and their false positive rate (FPR).

Suppose that we have an outcome classifier for treated observations (the treated classifier) and another for untreated observations (the untreated classifier), and that they provide scores S_T and S_U respectively. Without loss of generality, assume that both scores are between zero and one, as is the case with classifiers that give probabilities. In addition, suppose we want to establish the quality of these classifiers. If we set a threshold τ_T for the treated classifier, it will have some TPR δ_T and some FPR ξ_T. Similarly, if we set a threshold τ_U for the untreated classifier, it will have some TPR δ_U and some FPR ξ_U. Then, if Y_T and Y_U are the (potential) outcomes for the treated and the untreated, the following would define the TPR and the FPR for both classifiers:

$$P(S_T > \tau_T | Y_T = 1) = \delta_T$$

$$P(S_T > \tau_T | Y_T = 0) = \xi_T$$

[4] In the uplift modeling literature, ML and LL would usually be referred to as response modeling, while MT would be referred to as uplift modeling

$$P(S_U > \tau_U | Y_U = 1) = \delta_U$$
$$P(S_U > \tau_U | Y_U = 0) = \xi_U$$

All of the previous definitions relate to *outcome* prediction rather than *causal* prediction, but they will become useful in the following analysis. As shown in Table 2, the scoring functions for ML and MT are S_T and $S_T - S_U$ respectively.[5] Observations can then be classified as either Compliers or non Compliers based on these scoring functions and specified thresholds. The optimal threshold for each approach will ultimately depend on the circumstances of the problem at hand, such as the benefit of finding a Complier and the cost of incorrectly classifying a non Complier [23]. However, when specific values for costs and benefits are not available, the optimal threshold will still represent some TPR and some FPR [22]. Thus, in theory, we can compare MT and ML based on their *Complier* TPR and FPR (different from the *outcome* TPRs and FPRs).

Now, suppose that τ is the optimal threshold for MT. What we seek to answer in this section is whether ML can do better than MT under this condition (and by better, we mean in terms of the Complier TPR and FPR). For simplicity, we will assume the same threshold τ for ML. Notice that this does not imply that τ is optimal for ML. It only implies that if ML performs better than MT using τ, then ML should be preferable. There could be another threshold for ML that provides even better results. The following would be the Complier TPR and FPR for both approaches:

$$TPR_{ML} = P(S_T > \tau | Complier)$$
$$TPR_{MT} = P(S_T > \tau + S_U | Complier)$$
$$FPR_{ML} = P(S_T > \tau | \neg Complier)$$
$$FPR_{MT} = P(S_T > \tau + S_U | \neg Complier)$$

Clearly, $TPR_{ML} > TPR_{MT}$ and $FPR_{ML} > FPR_{MT}$ given that $S_U > 0$. These results highlight the key trade-off we hinted at before: the TPR increases when using ML and the FPR decreases when using MT. Ideally, we would like to quantify and characterize this trade-off. Thus, in order to derive analytical results, we make the additional assumptions that (1) the scores are conditionally independent of each other given the potential outcomes, (2) and each score is conditionally independent of its counterfactual outcome given the outcome that it is estimating. These assumptions imply that:

$$S_T \perp S_U | Y_T, Y_U$$
$$S_T \perp Y_U | Y_T$$
$$S_U \perp Y_T | Y_U$$

While these assumptions might be considered restrictive, they are necessary to avoid further parametric assumptions about the distributions of the scores. We discuss them further in the limitations section. Under these assumptions, the Complier TPR for ML is:

$$\begin{aligned} TPR_{ML} &= P(S_T > \tau | Complier) \\ &= P(S_T > \tau | Y_T = 1, Y_U = 0) \end{aligned}$$

Choose $\tau_T = \tau$. Then:

$$\begin{aligned} &= P(S_T > \tau_T | Y_T = 1) \\ &= \delta_T \end{aligned}$$

Similarly, the TPR for MT can be derived as:[6]

$$\begin{aligned} TPR_{MT} &= P(S_T - S_U > \tau | Complier) \\ &= P(S_T > \tau + S_U | Y_T = 1, Y_U = 0) \\ &= \int_0^1 f_U(u|Y_U = 0) \int_{\tau+u}^1 f_T(t|Y_T = 1) dt du \\ &= \int_0^{1-\tau} f_U(u|Y_U = 0) \int_{\tau+u}^1 f_T(t|Y_T = 1) dt du \\ &\leq \int_0^{1-\tau} f_U(u|Y_U = 0) du \int_{\tau}^1 f_T(t|Y_T = 1) dt \end{aligned}$$

Choose $\tau_T = \tau$ and $\tau_U = 1 - \tau$. Then:

$$\begin{aligned} &= \int_0^{\tau_U} f_U(u|Y_U = 0) du \int_{\tau_T}^1 f_T(t|Y_T = 1) dt \\ &= (1 - \xi_U)\delta_T \end{aligned}$$

Notice that what we derived for TPR_{MT} is an upper bound, and this upper bound is smaller than TPR_{ML}. We repeat the process with the probability of misclassifying an Always-Taker:

$$\begin{aligned} FPR_{Alw.ML} &= P(S_T > \tau | Always) \\ &= P(S_T > \tau | Y_T = 1, Y_U = 1) \\ &= \delta_T \\ FPR_{Alw.MT} &= P(S_T - S_U > \tau | Always) \\ &= P(S_T > \tau + S_U | Y_T = 1, Y_U = 1) \\ &\leq \delta_T(1 - \delta_U) \end{aligned}$$

And with the probability of misclassifying a Never-Taker:

$$\begin{aligned} FPR_{Nev.ML} &= P(S_T > \tau | Never) \\ &= P(S_T > \tau | Y_T = 0, Y_U = 0) \\ &= \xi_T \\ FPR_{Nev.MT} &= P(S_T - S_U > \tau | Never) \\ &= P(S_T > \tau + S_U | Y_T = 0, Y_U = 0) \\ &\leq \xi_T(1 - \xi_U) \end{aligned}$$

These results also imply that the FPR for MT is smaller than the FPR for ML. Now, suppose that the base rate for positive outcomes when untreated is α, and the average treatment effect is β. Then:

$$E[Y_U] = \alpha$$
$$E[Y_T] = \alpha + \beta$$

Assuming there are no Defiers, these parameters describe the proportion of Compliers (β), Always-Takers (α), and Never-Takers ($1-\alpha-\beta$). As a result, we can formulate the FPR for both approaches in terms of these quantities:

$$\begin{aligned} FPR_{ML} &= \frac{Always * FPR_{Alw.ML} + Never * FPR_{Nev.ML}}{100\% - Compliers} \\ &= \frac{\alpha\delta_T + (1 - \alpha - \beta)\xi_T}{1 - \beta} \\ FPR_{MT} &\leq \frac{Always * FPR_{Alw.MT} + Never * FPR_{Nev.MT}}{100\% - Compliers} \\ &= \frac{\alpha\delta_T(1 - \delta_U) + (1 - \alpha - \beta)\xi_T(1 - \xi_U)}{1 - \beta} \end{aligned}$$

[5] Let $S_T = \hat{P}(Y = 1|X, T = 1)$ and $S_U = \hat{P}(Y = 1|X, T = 0)$.

[6] In the following derivation, f_j is the conditional density function of S_j

Now, suppose that we use the upper bounds as the actual FPR_{MT} and TPR_{MT} to choose an approach. Assuming that an increase in the TPR has a value of w_t and that a decrease in the FPR has a value of w_f, we could decide which approach to choose based on the differences between the TPRs and the FPRs. Specifically, we would choose ML if the following inequality is met:

$$0 < w_t(TPR_{ML} - TPR_{MT}) - w_f(FPR_{ML} - FPR_{MT})$$

For simplicity, assume $w_t = w_f = 1$. Then:

$$\begin{aligned} 0 &< TPR_{ML} - TPR_{MT} - FPR_{ML} + FPR_{MT} \\ &< \delta_T - \delta_T(1-\xi_U) - \frac{\alpha\delta_T + (1-\alpha-\beta)\xi_T}{1-\beta} \\ &\quad + \frac{\alpha\delta_T(1-\delta_U) + (1-\alpha-\beta)\xi_T(1-\xi_U)}{1-\beta} \end{aligned}$$

After doing some term rearrangement, we find that:

$$0 < (1-\alpha-\beta)\xi_U(\delta_T - \xi_T) - \alpha\delta_T(\delta_U - \xi_U)$$

Let N and A represent the proportions of Never-Takers and Always-Takers respectively:

$$0 < N * FPR_U * Quality_T - A * TPR_T * Quality_U$$

The formulation above yields several interesting results:

(1) The trade-off is contingent on the proportion of Never-Takers (N) and Always-Takers (A). If the former proportion increases, it becomes more attractive to use ML, and if the latter increases, it becomes more attractive to use MT.
(2) The trade-off is contingent on the quality of the outcome classifiers, where the quality is a number between -100% and 100%. Quality 100% means that the classifier is perfect, quality 0% means that the classifier is as good as random, and quality -100% means that the classifier perfectly discriminates positives as negatives and negatives as positives.
(3) An untreated classifier with **lower** quality makes ML more attractive, particularly when there is a large FPR_U because MT mistakes more Compliers for Always-Takers.
(4) A treated classifier with **higher** quality makes ML more attractive, **contingent** on a comparatively lower quality of the untreated classifier and a small proportion of Always-Takers. This occurs because a high TPR_T also increases the bias towards Always-Takers.
(5) Intuitively, we can think of the left-hand term as the reduction in variance error and of the right-hand term as the increase in bias error.

In practice, however, it is difficult to work with TPRs and FPRs because they will depend on a threshold τ. One way to circumvent this is to use the area under the ROC curve (AUC) to represent classifier quality. The AUC measures a classifier's ability to discriminate between classes; for a particular decision setting the TPR and FPR increase and decrease (respectively) with the AUC. Moreover, suppose we have a binary classifier such that $\delta_i = (1-\xi_i) \geq 0.5$. The ROC curve will be triangular and the AUC corresponds to the sum of two right triangles and a square:

$$AUC_i = 2 * \frac{\delta_i\xi_i}{2} + \delta_i^2 = \delta_i(\xi_i + \delta_i) = \delta_i(\xi_i + 1 - \xi_i) = \delta_i$$

Therefore:

$$AUC_i = \delta_i = (1-\xi_i)$$

Thus, for such classifiers the AUC actually represents the quantities exactly; however, in practice, for a particular decision threshold the actual TPR and FPR may be asymmetric, thus in general we are using AUC as a proxy. Then, we would choose ML over MT whenever this inequality is met:

$$0 < N(1-AUC_U)(2AUC_T - 1) - A(1-AUC_T)(2AUC_U - 1)$$

If we assume that both classifiers have the same AUC and we rearrange some terms, then:

$$0 < -2(N+A)AUC^2 + (3N+A)AUC - N$$

The above is a concave quadratic equation. Its two solutions are 0.5 and $N/(N+A)$. Therefore, we would prefer to use the ML approach instead of MT whenever the AUC is below $N/(N+A)$, but above 50% (i.e., better than random). Interestingly, we also found that we would prefer LL over MT when the AUC is below $A/(N+A)$, but we spare the details given the limited space. Considering that $N + A = 1 - C$, where C is the proportion of Compliers, our results can be summarized in four main conclusions:

(1) A lower AUC for both classifiers disfavors MT.
(2) MT is preferable as the number of Compliers increases.
(3) ML is preferable as the number of Never-Takers increases.
(4) LL is preferable as the number of Always-Takers increases.

These guidelines can be used to decide whether to address causal classification using outcome prediction (i.e. ML or LL) or treatment effect estimation (i.e., MT). Assuming there are no Defiers, we would only need an estimate of the following variables: the AUC when predicting the outcome of unobserved instances (AUC), the average treatment effect (β), and the percentage of positive outcomes when there is no treatment (α). Then, these estimates could be used with the inequality we developed in this section to decide which approach to use. Assuming the AUC is the same for treated and untreated observations, we would:

(1) Choose ML when $AUC < \frac{1-\alpha-\beta}{1-\beta}$.
(2) Choose LL when $AUC < \frac{\alpha}{1-\beta}$.
(3) Choose MT otherwise.

5 SIMULATOR & EXPERIMENTAL RESULTS

One way to test the analytical results we developed in the previous section is through simulations, as then we can know the ground truth and also test across a wide variety of settings. The simulator should meet several requirements. We should be able to: (1) adjust the proportion of different observation types (e.g., Compliers) in the population; (2) generate samples from a wide range of distributions, and (3) simulate various degrees of "irreducible error" in the outcome predictions to test across different levels of predictability. This section presents a simulator that meets all of these requirements and that could be quite useful for future research.

The main design assumption is that each observation is based on a set of factors, some possibly unobserved, that can be combined into a latent variable that determines whether the outcome is positive or not. This is consistent with discrete choice theory [29], where the latent variable often represents the utility that an agent would get from performing an action. Thus, for illustrative purposes, we discuss the simulation in terms of an agent deciding on a positive or a negative outcome (e.g., buy or not buy) depending

on a utility function. Each observation is based on the following information, some of which is unobserved to modeling:

(1) **Influencing variables (X):** A set of variables that influence the outcome. Models may observe all or only some of them.
(2) **Treatment (γ):** Binary variable that indicates if the agent was treated.
(3) **Utility (J):** The utility that the agent would get if he chooses a positive outcome without the treatment. This is a function of X and is not observed by the models.
(4) **Treatment effect (K):** The additional utility that the agent would get if he chooses a positive outcome when treated. This is a function of X and is not observed by the models.
(5) **Outcome (Y):** Binary variable that indicates if the outcome is positive. The agent chooses a positive outcome only if the utility he would get from that decision is greater than 0. Thus, the outcome is a function of X and the treatment (γ):

$$Y(X,\gamma) = \mathbb{I}(J(X) + K(X) * \gamma > 0)$$

To meet the first requirement (the ability to adjust the proportion of different observation types) the simulator receives two parameters:

(1) **Positive outcome base rate without treatment (α).**
(2) **Average additive treatment effect (β)**—the average increase in the probability of a positive outcome when there is a treatment.

The simulator meets this requirement by imposing two restrictions:

$$E[Y(X,\gamma=0)] = E[\mathbb{I}(J(X) > 0)] = \alpha$$
$$E[Y(X,\gamma=1)] = E[\mathbb{I}(J(X) + K(X) > 0)] = \alpha + \beta$$

$J(X)$ is defined as:

$$J(X) = G - \nu = g(X) - \nu$$

where ν is a constant and G is a random variable determined by an arbitrary function $g(\cdot)$ of X. Then, the first restriction is met when we set ν to be the $(1-\alpha)$th percentile of G:

$$\begin{aligned}\alpha &= E[\mathbb{I}(J(X) > 0)] \\ &= E[\mathbb{I}(G > \nu)] \\ &= \int_{\nu}^{\infty} f_G(x)dx \\ &= 1 - F_G(\nu) \\ \nu &= F_G^{-1}(1-\alpha)\end{aligned}$$

Similarly, in the second restriction $K(X)$ is defined as:

$$K(X) = H + \omega = h(X) + \omega$$

where ω is a constant and H a random variable determined by an arbitrary function $h(\cdot)$ of X. Thus, the second restriction is met when we set ω to be the difference between ν and the $(1-\alpha-\beta)$th percentile of $G+H$:

$$\begin{aligned}\alpha + \beta &= E[\mathbb{I}(J(X) + K(X) > 0)] \\ &= E[\mathbb{I}(G + H > \nu - \omega)] \\ &= 1 - F_{G+H}(\nu - \omega) \\ \omega &= \nu - F_{G+H}^{-1}(1-\alpha-\beta)\end{aligned}$$

The simulator satisfies both restrictions by generating samples for G and H, and then computing percentiles to set ν and ω.

Notice that the simulator does not impose any restrictions on the distribution of X or the functional forms of $g(X)$ and $h(X)$, which means that the data generation process is extremely flexible. Thus, the second fundamental requirement for the simulator is also met. A computational advantage of this approach is that samples for G and H only need to be generated once, and then the same samples can be used to create data for multiple experiments using different values for α and β.

For this paper, the simulator generated samples for G, H, and X using Gibbs sampling [10] on a randomly generated Bayesian Network (BN) [20] over the variables described above (with 10 variables for X). The arcs in the BN are created using Algorithm 1 for random BN generation presented by Ide and Cozman [14]. The only modification is the restriction that G and H cannot have descendants. Each node is assigned a second-degree polynomial in terms of its parents, and each term in the polynomial is assigned a randomly and uniformly generated coefficient, using smaller coefficients for the second-degree terms. Finally, except for G and H, we define the probability distribution of every node as a normal distribution with a variance of one and a mean equal to the node's polynomial. On the other hand, G and H are completely determined given their parents; their values correspond to their respective polynomials.

To meet the third simulation requirement (i.e., various degrees of irreducible error), we use the noisy channel proposed by Domingos [6]. This is accomplished by introducing two new elements (N and η) in the definition of the outcome (Y):

$$L_\gamma = \mathbb{I}(J(X) + K(X) * \gamma > 0)$$
$$Y(X,\gamma,\eta) = N_{\eta,\gamma}(L_\gamma)$$

N is a function that introduces noise to the outcome, and η is a number between zero and one that represents the magnitude of the noise. N receives a label L (i.e., the outcome when there is no noise) and is defined as follows:

$$N_{\eta,\gamma}(L) = L(1-C) + C * B$$

Here, C is a Bernoulli random variable that has probability of success η, and it represents the probability that the label is noisy. Moreover, B is also a Bernoulli random variable that has probability of success $\alpha + \gamma * \beta$, and it represents the outcome if the label is noisy. It is easy to show that the simulation restrictions hold:

$$\begin{aligned}E[Y(X,\gamma=0,\eta)] &= E[N_{\eta,\gamma=0}(L_{\gamma=0})] \\ &= E[L_{\gamma=0}(1-C) + C * B_{\gamma=0}] \\ &= \alpha(1-\eta) + \eta\alpha \\ &= \alpha \\ E[Y(X,\gamma=1,\eta)] &= E[N_{\eta,\gamma=1}(L_{\gamma=1})] \\ &= E[L_{\gamma=1}(1-C) + C * B_{\gamma=1}] \\ &= (\alpha+\beta)(1-\eta) + \eta(\alpha+\beta) \\ &= \alpha + \beta\end{aligned}$$

As a result, a small value of η implies that the functional form of the outcome remains as it is in most of the cases, and a large η implies that the outcome is a "coin-toss" in most of the cases. Notice that this does not prevent the simulator user from introducing other sorts of error terms, e.g., by including additional unobserved covariates as part of X. Therefore, our simulator meets all three requirements.

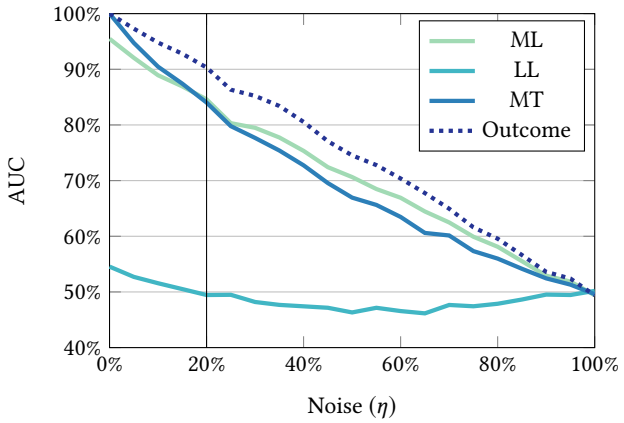

Figure 1: Predictive performance with $\alpha = 8\%$ and $\beta = 12\%$.

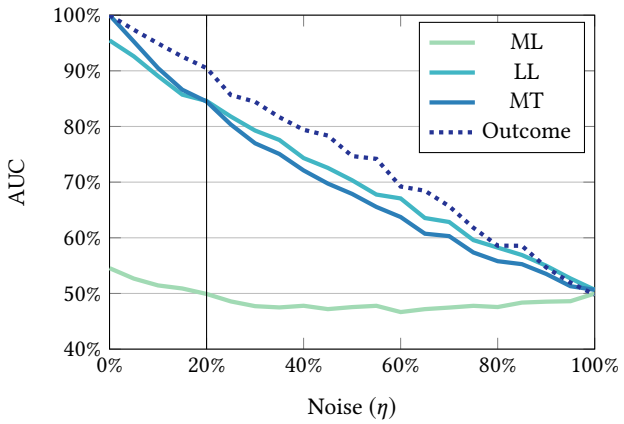

Figure 2: Predictive performance with $\alpha = 80\%$ and $\beta = 12\%$.

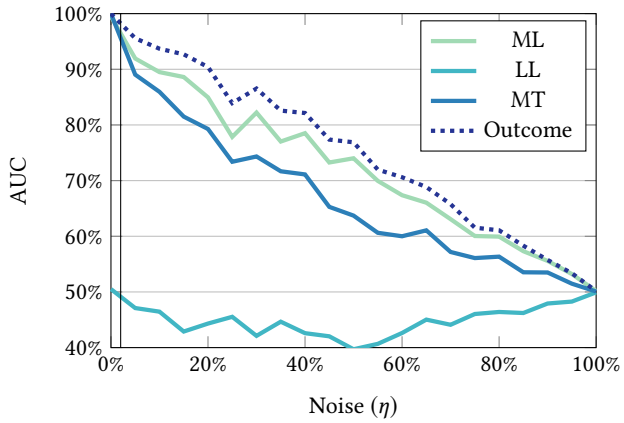

Figure 3: Predictive performance with $\alpha = 1\%$ and $\beta = 1\%$.

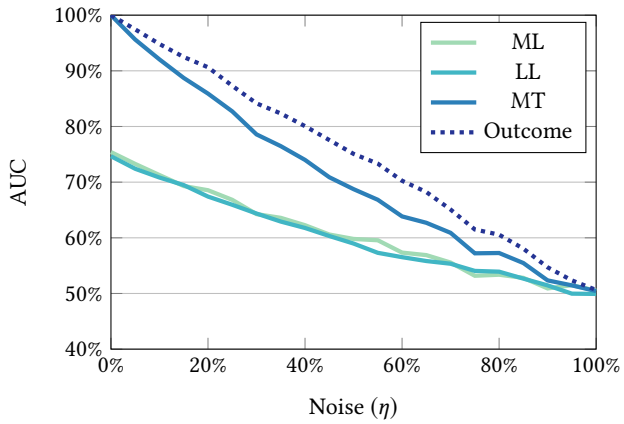

Figure 4: Predictive performance with $\alpha = 33\%$ and $\beta = 33\%$.

One can (1) specify different proportions of observation types, (2) simulate a wide range of distributions, (3) and introduce noise.[7]

To test our analytical results, we ran a large number of simulations to create about 250 test sets representing various settings, each set containing 10,000 observations. Figures 1-4 display the performance of all causal classification approaches under different settings, defined by α, β and various degrees of noise (η). The outcome classifiers use the true model that is generating the data, and so they predict the utility J and the treatment effect K perfectly; they are not subject to estimation errors or bias. This implies a setting without confounding where the classifiers are the best possible models to predict the outcome, and there is no need for sophisticated causal estimation methods (such as propensity score matching or instrumental variables). This also implies that it's actually unnecessary to use different random BNs in these experiments; regardless of the complexity of $g(\cdot)$ and $h(\cdot)$, these functions will be known to the classifiers. Thus, the source of all misclassification errors is irreducible error coming from noise, and our results are not specific to the particular statistical dependencies between the Xs and Ys. However, the simulator's flexibility allows extensions of this work; for example, not assuming that the true model is known (we give a concrete example below).

The "Outcome" (dotted) curves in the figures show the AUCs of predicting the outcome, while all the other curves represent the AUCs of predicting if an observation is a Complier, for the different approaches. The vertical line in each figure represents the point at which the analytical results (from above) say we should prefer either ML or LL over MT according to the classifier's AUC. For example, in Figures 1 and 2 we would prefer ML and LL respectively when the outcome AUC is below 90%. The figures illustrate four different scenarios: one where ML is preferable for higher levels of noise (Figure 1), one where LL is similarly preferable (Figure 2), an extreme case where ML is strongly preferable in almost all cases (Figure 3), and one where MT is always preferable (Figure 4).

The settings were chosen to illustrate our analytical results: ML should be preferable over MT whenever $AUC < N/(1-C) = \frac{1-\alpha-\beta}{1-\beta}$, and LL should be preferable when $AUC < A/(1-C) = \frac{\alpha}{1-\beta}$. Thus Figure 1 shows results for $\alpha = 8\%$ and $\beta = 12\%$ and Figure 2 shows

[7] The simulated data and the code for the simulator is available in the following link: https://github.com/ferlocar/causal_predictions/

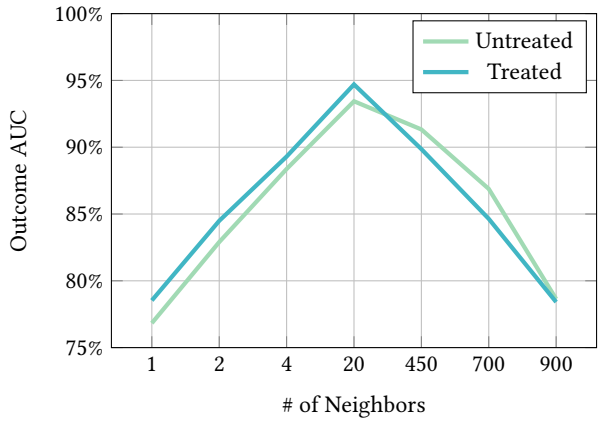

Figure 5: kNN Performance by # of Neighbors.

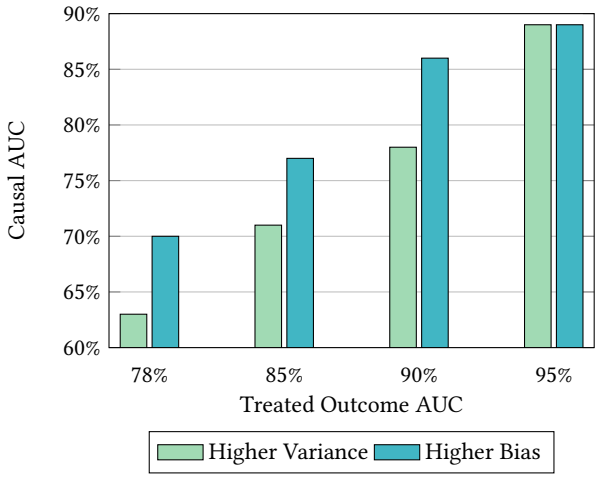

Figure 6: MT Performance with respect to error types.

results for $\alpha = 80\%$ and $\beta = 12\%$. In these settings, the analytical results predict that ML and LL will be preferable in these settings whenever $AUC < \frac{8\%}{88\%} = 90\%$, which is what we observe. We also observe that the gap in performance is much larger in the more extreme case where the number of Always-Takers and Compliers is very small (Figure 3). Here, the difference between ML and MT can be up to 10.3% in AUC. This is the same loss in outcome predictive performance that is experienced when noise increases from 0% to 20%. Figure 4 supports the conclusion that MT will always be preferable when the proportion of Always-Takers and Never-Takers is the same (and thus there is no vertical line).

6 DISCUSSION AND FOLLOW-UP

These findings imply that fundamentally different approaches to causal targeting are called for under different settings. Specifically, depending on α, β, and the predictive power of the models, it may be better to target with outcome prediction models even when causal targeting is the goal. This holds even when the causal modeling is not challenged by selection bias or other confounding.

At the outset we motivated the paper by observing that practitioners "stubbornly" continue to use outcome prediction. Our results allow the examination of problem settings to assess whether this practice might actually be appropriate. As an example, consider the display ad targeting work presented by Perlich et al. [21], which focuses on ad targeting—where causal targeting is very likely the true goal. The study contains a large collection of different ad campaigns, and the key aspects of the settings (α, β, AUC) all either are reported or can be bounded. These are very-low-base-rate problems; $\alpha + \beta$ varies between 0.001 and 0.0000001, which means that assuming no Defiers, both $\alpha \leq 0.001$ and $\beta \leq 0.001$. According to the analytical results, we would need models with $AUC > 0.99$ in this setting in order for MT to be preferable! The paper also reports AUCs for different models, none of which is close to 0.99, giving some support that using outcome prediction indeed is justified.

This is to our knowledge the first work to examine this question systematically, and so the paper does not answer every question and also makes various assumptions. For example, the paper shows that even if the actual data-generating processes ("true models") are known, it still may be better to choose outcome targeting over causal targeting. An important topic for future research is to ask what will happen when we actually are learning the models from data—specifically, to look more deeply into the trade-offs between different types of *supervised learning errors*, namely bias error, variance error, and irreducible error [17].[8] For example, if a particular bias is present in both the treated classifier *and* the untreated classifier, the scores from the two classifiers would be correlated. In this case, the variance of the combined scores might tend to be smaller if the scores are closely correlated:

$$Var(S_T - S_U) = Var(S_T) + Var(S_U) - 2Cov(S_T, S_U)$$

As an example, suppose that the classification models tend to systematically overestimate the effect of a covariate X on the outcome, regardless of the treatment status. Then, even though the performance of each individual classifier will be lower than it could potentially be without the bias, this will not affect much the distance between scores (i.e., the degree to which we are sure that the observation is a Complier) because the outcome predictions given X will be higher for both classifiers.

To illustrate, consider the following additional experiment where we learn a classifier for our data, rather than using the true models. Specifically, we apply k-NN, for which a smaller k implies more variance error and less bias error, and a larger k implies less variance error and more bias error. Thus, increasing k has a nonlinear effect on predictive performance. Importantly, it is possible to have two k-NN models with the same AUC but a different number of neighbors (as shown in Figure 5 for data from the simulator).

If our hypothesis about the effects of bias error vs variance error is correct, we would expect that given two k-NN models with the same AUC, the one with larger k would work substantially better for MT because it would have more bias and less variance. This is shown in Figure 6, using the following conditions for the simulation: no noise (so no irreducible error), a positive outcome rate of 5%, an average treatment effect of 5%, and a set of 1,000 observations under each treatment condition (which implies a training set of 2,000 observations). Given a treated outcome AUC, we observe

[8] Bias and variance errors are also known as approximation and estimation errors.

that the higher variance alternative (the one with smaller k) indeed performs worse at causal classification. The results are qualitatively the same if we repeat the experiment given an untreated outcome AUC. Thus, considering that we are comparing the same modeling procedure but under different sources of error, this provides some confirmation that bias errors indeed do have less negative impact than variance errors for causal classification.

7 CONCLUSIONS AND LIMITATIONS

This paper's main contribution is to reveal, analyze and illustrate the bias/variance tradeoff when predicting the counterfactual to target treatments based on predictive models. Specifically, because treatment effect estimation depends on two outcome estimates instead of one, the larger variance may lead to higher misclassification error than the (biased) outcome prediction approach. The results of a theoretical analysis show that outcome prediction—ignoring the causal treatment effect—is preferable when positive outcomes are (1) very rare, (2) difficult to predict, and when (3) treatment effects are small. We also introduce a flexible simulation environment for experimenting with causal classification; results from large-scale simulations support the analytical results.

These analytical results may offer a justification for the seemingly naive, common practice of targeting "treatments" such as online advertisements based simply on outcome models, rather than treatment effect models—and we show as an example that the problem of targeting display ads falls into the regime where outcome targeting would be preferable to treatment-effect targeting. This has important implications for practitioners because acquiring the data to estimate the counterfactual is complicated and expensive, and causal confounding is insidious.

The study of course has limitations. For example, what if the score distributions are not conditionally independent of other factors (given the outcomes)? How do such non-independencies manifest in real causal targeting? Do the analytical results hold, at least qualitatitvely, under typical independence violations? As related example, we showed that even if you were to have the best possible outcome models, it still is sometimes preferable to ignore the estimations of the treatment effects. However, you never (know that you) have the best possible outcome models, and you have particular problems estimating them fairly under ubiquitous selection bias and confounding. Can the effects of biased modeling be modeled analytically to extend the results we have presented? Moreover, in non-simulated data it is rare to be able to know what would have been the counterfactual outcome, so empirical work will need either to aggregate, ideally with data from experiments [16] or to be clever [9]. The work is also limited because we assumed away the Defiers—those for whom the treatment causes a negative flip of the outcome. When Defiers are present, α and β are not enough to describe the proportion of all four observation types (Table 1). This is just a first study and these limitations present directions for future research. Both the analytical framework and the simulator are quite general, and can be easily adapted to support future studies.

Finally, what about the design of new methods focusing on causal classification? An upshot of the findings is that regularization may be even more important in causal classification tasks, since these tasks are even more sensitive to variance errors. In fact, it might be worthwhile to design causal classification models that specifically focus on limiting the variance of the combined scores. In targeting learning [30], for example, the MT approach exposed here is thought of as a *non-targeted* semiparametric model for treatment effect estimation. The underlying premise behind targeted learning is that the estimation can be improved by applying targeted maximum likelihood estimation (TMLE) to create a *targeted* semiparametric model that optimizes the bias-variance trade-off for the causal effect of interest (not for the outcome). These ideas have been discussed mostly in the context of estimating statistical parameters (e.g., a causal effect), but we believe it would also be valuable to apply them in the context of causal classification.

ACKNOWLEDGEMENTS

We thank Claudia Perlich, Brian Dalessandro and Ori Stitelman for many discussions motivating a deeper examination of causal targeting. Claudia pointed out the potential for increased variance in treatment effect estimation for targeting. Foster Provost thanks Andre Meyer for a Faculty Fellowship.

REFERENCES

[1] E. Ascarza et al. 2016. In Pursuit of Enhanced Customer Retention Management. *Customer Needs and Solutions* (2016), 1–17.
[2] P. Auer et al. 2002. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32, 1 (2002), 48–77.
[3] A. Beygelzimer and J. Langford. 2009. The offset tree for learning with partial labels. *KDD'09* (2009), 129.
[4] D. Bhattacharya and P. Dupas. 2012. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics* 167, 1 (2012).
[5] R. Dehejia. 2005. Program evaluation as a decision problem. *Journal of Econometrics* 125 (2005), 141–173.
[6] P. Domingos. 2000. A unified bias-variance decomposition. *International Conference on Machine Learning* (2000), 231–238.
[7] M. Dudik, J. Langford, and L. Li. 2011. Doubly Robust Policy Evaluation and Learning. *International Conference on Machine Learning* (2011).
[8] J.H. Friedman. 1997. On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery* 1, 1 (1997), 55–77.
[9] D. Garant and D. Jensen. 2016. Evaluating Causal Models by Comparing Interventional Distributions. *KDD'16 Workshop on Causal Discovery* (2016).
[10] A. Gelfand and A. Smith. 1990. Sampling-Based Approaches to Calculating Marginal Densities. *J. Amer. Statist. Assoc.* 85, 410 (1990), 398–409.
[11] L. Han et al. 2012. Overlapping Decomposition for Gaussian Graphical Modeling. *KDD'12* (2012), 114–122.
[12] D. Hill et al. 2015. Measuring Causal Impact of Online Actions via Natural Experiments. *KDD'15* (2015), 1839–1847.
[13] K. Hirano and J. R. Porter. 2009. Asymptotics for Statistical Treatment Rules. *Econometrica* 77, 5 (2009), 1683–1701.
[14] J. Ide and F. Cozman. 2002. Random generation of Bayesian networks. *Advances in Artificial Intelligence* (2002), 366–376.
[15] D. Jensen et al. 2008. Automatic identification of quasi-experimental designs for discovering causal knowledge. *KDD'08* (2008).
[16] K. Kane, V. Lo, and J. 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.
[17] R. Kohavi and D. Wolpert. 1996. Bias plus variance decomposition for zero-one loss functions. *International Conference on Machine Learning* (1996), 275–283.
[18] V. Lo. 2002. The true lift model: a novel data mining approach to response modeling in database marketing. *KDD Explorations Newsletter* 4, 2 (2002).
[19] C. F. Manski. 2004. Statistical Treatment Rules for Heterogeneous Populations. *Econometrica* 72, 4 (2004), 1221–1246.
[20] J. Pearl. 2014. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann.
[21] C. Perlich et al. 2014. Machine learning for targeted display advertising: Transfer learning in action. *Machine Learning* 95, 1 (2014), 103–127.
[22] F. Provost and T. Fawcett. 2001. Robust classification for imprecise environments. *Machine Learning* 42, 3 (2001), 203–231.
[23] F. Provost and T. Fawcett. 2013. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. (2013).
[24] N. Radcliffe and P. Surry. 2011. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions* (2011).

[25] D. B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688–701.
[26] G. Shmueli. 2011. To Explain or to Predict? *Statist. Sci.* 25, 3 (2011), 289–310.
[27] O. Stitelman et al. 2011. Estimating The Effect Of Online Advertising On Browser Conversion In Observational Data. *ADKDD-2011: Data Mining and Audience Intelligence for Advertising* (2011), 1–7.
[28] M. Taddy et al. 2016. A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation. *Journal of Business & Economic Statistics* 34, 4 (2016), 661–672.
[29] K. E. Train. 2009. *Discrete choice methods with simulation.* Cambridge U. press.
[30] M. Van der Laan and S. Rose. 2011. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media.
[31] S. Wager and S. Athey. 2017. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J. Amer. Statist. Assoc.* (2017).
[32] H. Weisberg and V. Pontes. 2015. Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials* 12, 4 (2015), 357–364.
[33] I. Yahav, G. Shmueli, and D. Mani. 2016. A Tree-Based Approach for Addressing Self-Selection in Impact Studies With Big Data. *MIS Quarterly* 40, 4 (2016).