Journal of Machine Learning Research 23 (2022) 1-35 Submitted 6/19; Revised 1/22; Published 2/22

Causal Classification: Treatment Effect Estimation vs. Outcome Prediction

Carlos Fernández-Loría imcarlos@ust.hk
HKUST Business School
Hong Kong University of Science and Technology
Hong Kong

Foster Provost fprovost@stern.nyu.edu
Stern School of Business
New York University
New York, NY, USA

Editor: David Blei

Abstract

The goal of causal classification is to identify individuals whose outcome would be positively changed by a treatment. Examples include targeting advertisements and targeting retention incentives to reduce churn. Causal classification is challenging because we observe individuals under only one condition (treated or untreated), so we do not know who was influenced by the treatment, but we may estimate the potential outcomes under each condition to decide whom to treat by estimating treatment effects. Curiously, we often see practitioners using simple outcome prediction instead, for example, predicting if someone will purchase if shown the ad. Rather than disregarding this as naive behavior, we present a theoretical analysis comparing treatment effect estimation and outcome prediction when addressing causal classification. We focus on the key question: "When (if ever) is simple outcome prediction preferable to treatment effect estimation for causal classification?" The analysis reveals a causal bias–variance tradeoff. First, when the treatment effect estimation depends on two outcome predictions, larger sampling variance may lead to more errors than the (biased) outcome prediction approach. Second, a stronger signal-to-noise ratio in outcome prediction implies that the bias can help with intervention decisions when outcomes are informative of effects. The theoretical results, as well as simulations, illustrate settings where outcome prediction should actually be better, including cases where (1) the bias may be partially corrected by choosing a different threshold, (2) outcomes and treatment effects are correlated, and (3) data to estimate counterfactuals are limited. A major practical implication is that, for some applications, it might be feasible to make good intervention decisions without any data on how individuals actually behave when intervened. Finally, we show that for a real online advertising application, outcome prediction models indeed excel at causal classification.

Keywords: Bias–variance tradeoff, causal inference, treatment assignment

1. Introduction

Predictive modeling increasingly is applied to improve tasks such as predicting whether customers will leave after their contracts expire or purchase after seeing an ad, whether fraud is present on an account, and 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 instead the customers for whom our incentive will cause them to stay when they would leave otherwise. For many advertising settings, the advertiser's goal is not simply to target ads to people who will purchase after seeing the ad, but to target people who will purchase because they saw the ad. Ideally, instead of simply predicting the likelihood of a positive outcome, for tasks such as these, we would like to estimate the likelihood of *changing* a particular individual's outcome with a treatment (i.e., the ad or the retention incentive in our examples). Let's call these **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 (Ascarza et al., 2018). For example, although targeted online ad campaigns are sometimes evaluated in terms of their causal effect (e.g., through A/B testing), causal-effect models seldom are used for targeting the ads! Instead, potential customers are usually targeted based on predictions of how likely they are to convert, which are upwardly biased estimates of individual causal effects because they do not account for the counterfactual (e.g., a customer may buy even without the ad). Researchers and savvy practitioners see this as due either to naivety, lack of modeling sophistication, or pragmatics (Ascarza et al., 2018; Provost and Fawcett, 2013).[1] However, we also see continued use of outcome prediction models even in cases where practitioners have looked carefully at the estimation of treatment effects (cf., Stitelman et al. (2011) and Perlich et al. (2014) for a clear example).

This paper examines the question: might targeting based on outcome prediction be more effective (in terms of optimizing the outcome of interest) 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 unconfounded data—meaning that our results are not due to the difficulty of learning causal models in the presence of selection biases or other confounding.

Instead, the result is the consequence of a causal bias–variance tradeoff. First, when the treatment effect estimation depends on two outcome predictions, then the larger sampling variance may lead to more errors than the (biased) outcome prediction approach that ignores counterfactuals. Second, if outcomes and effects are correlated, a strong signal-to-noise ratio in outcomes may imply that the bias is actually helpful for treatment assignments. Which strategy is theoretically better depends on a combination of aspects of the application setting, including the variance of the outcome predictions, whether it is possible to set a threshold that partially corrects the bias, and the correlation between outcomes and

[1] The pragmatics of using outcome prediction instead of treatment effect estimation vary based on the setting. For targeted advertising, brands often are reluctant to incur the opportunity cost associated with experimentation, such as not advertising to good prospects. Thus, large-scale training data on one "side" of the counterfactual may not be available. Even if they are persuaded to run an experiment, usually it will be designed to estimate whether the campaign is having a significant effect rather than to obtain sufficient labeled data to learn a high-quality model. For churn control, new retention offers suffer from a cold-start problem—there is no data on individuals who have been given the offer. Once again, large-scale training data on one side of the counterfactual is not available.

treatment effects. Therefore, a major contribution of this paper is to show the bias–variance tradeoff between ignoring the counterfactual and trying to estimate it (i.e., treatment effect estimation) when the goal is to perform better causal classification.

We also provide additional support for the theoretical results using simulations, allowing us to know all aspects of the application setting with certainty. We cover the following settings of interest to illustrate the conditions under which outcome prediction should outperform treatment effect estimation: (1) situations where the data to estimate one of the counterfactuals are limited and (2) application settings where outcomes are informative of treatment effects (e.g., likely buyers are more susceptible to ads). Thus, another important contribution of the paper is to showcase commonplace settings where outcome prediction may work better than treatment effect estimation for causal classification.

Finally, we discuss multiple practical implications of the theoretical results. Notably, in applications like advertising, it might be feasible for practitioners to make good intervention decisions even if they do not have data on how people actually behave when intervened upon. This can be accomplished by intervening based on a good proxy of the intervention effect, such as the probability of purchase when the intervention is an advertisement. In fact, decision making may not improve at all even when the necessary data to estimate intervention effects can be acquired. Appendix D provides an empirical example of this based on a real online advertising application. Therefore, the theoretical results in this study have practical implications for other stages in the data mining process besides modeling, such as data acquisition and model evaluation.

2. Problem Definition

In this paper, we will consider problems with binary outcomes (e.g., purchase or not) and binary treatments (e.g., show an ad or not).[2] **Causal classification** consists of identifying **Persuadables**, those individuals whose outcome would change from negative to positive if they were treated. We use the potential outcomes framework[3] (Rubin, 1974) to define the causal classification 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.[4] So, in addition to Persuadables, we have individuals who would have a positive outcome with or without treatment (Sure Things), individuals who would have a negative outcome even with treatment (Lost Causes), and to complete the picture, individuals for whom treatment would have a negative effect (Do-Not-Disturbs).[5]

The approach we will consider in this paper is to use **outcome classification** models to identify Persuadables by predicting the probability of a positive outcome for treated and

[2] Note that some of the business problems that motivate this paper may not be strictly binary in practice. For instance, if the outcome of interest is whether a customer will leave, then customer lifetime value may also be important. In the context of churn management, Lemmens and Gupta (2020) found that incorporating such information in the loss function of predictive models may have a significant impact on the profitability of incentive targeting campaigns.

[3] For readers familiar with causal calculus, the potential outcomes when treated and untreated, $Y_{T=1}$ and $Y_{T=0}$, are equivalent to $Y|do(T=1)$ and $Y|do(T=0)$, respectively (Pearl, 2009).

[4] As a result, this paper considers the actual ground truth for each individual to be categorical, but the conditional expectation of the classes given the individual's feature vector is probabilistic; this aligns causal classification with traditional machine-learning-based class probability estimation.

[5] The names of these observation types come from the uplift modeling literature (Kane et al., 2014).

Table 1: Causal classification labels

Obs. Type	Outcome (no treatment)	Outcome (treatment)	Causal Label
Lost Cause	Negative	Negative	Negative
Sure Thing	Positive	Positive	Negative
Do-Not-Disturb	Positive	Negative	Negative
Persuadable	**Negative**	**Positive**	**Positive**

untreated instances. It is worth clarifying that we have two different classification tasks here: outcome prediction and Persuadable prediction. Ideally, we would like to use the predictive models to treat Persuadables and not waste resources treating the other three categories. More specifically, causal classification is a classification task in which the goal is to identify Persuadables, usually by assigning scores to each instance and giving higher scores to Persuadables to separate them from other instances. Subsequently, we could treat the instances (e.g., individuals) with the largest Persuadable scores.

The main challenge for causal classification, as for other causal inference tasks, is that we only observe one of the potential outcomes, so the causal label is not visible. For example, if we observe that a consumer made a purchase after seeing an ad, we do not know if that consumer would have purchased *even without* the ad. This has obvious implications for both training and evaluation. First, we cannot label historical cases as Persuadable or not. Second, if a model's prediction leads to a different treatment decision from the one observed in the data, we cannot tell if the outcome would have changed.

We can deal with this challenge in two ways. 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 (Lo, 2002; Provost and Fawcett, 2013). However, as the assumption will be violated in the cases we care about here, these predictions are likely to include Sure Things and therefore seem wasteful. A second, seemingly more sensible alternative is to make targeting decisions based on estimated treatment effects at the individual level. This can be achieved by training classification models to predict the outcomes for the treated and the untreated and then scoring each unlabeled observation based on the difference between the predictions of the two models.[6] This approach, known as "true lift" or "uplift modeling" (Lo, 2002), is often used for **causal targeting**: targeting observations based on estimated treatment effects.

Causal targeting generally will take into account costs and benefits that may affect treatment decisions (Provost and Fawcett, 2013). For example, a simple and very common targeting policy is based on a budget constraint: given a targeting budget, maximize the "return" from the targeting. If benefits and costs are constant across instances (or are assumed to be), then this translates into targeting the k individuals estimated to be most

[6] 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 when modeling for treatment effect estimation. For the purposes of this paper, the treated and untreated models in this case would be this single model with the appropriate setting for the treatment covariate.

likely to be Persuadables (presuming their expected net benefit is positive), where k is determined by the costs of targeting. A similar common policy is to target some top quantile of individuals (top-1%, top-decile)—for example, due to perceived-but-unquantified costs of false positives (e.g., we do not want to bombard our customers with cross-sell offers; only make offers to those most likely to be affected by the offer).

As mentioned in the introduction, despite the first approach (simple outcome prediction) seeming to be naive and the second approach (treatment effect prediction) seeming to be well considered, practitioners nonetheless broadly follow the first approach, even when they have the resources to follow the second. Therefore, we should think carefully about why the first might actually be preferable. Targeting based simply on outcome prediction will clearly be biased when used for Persuadable prediction because ignoring the counterfactual leads to overestimation in the predictions. However, there is a causal bias–variance tradeoff. First, when treatment effect prediction is based on two predictive models rather than one, then the sampling variance in the combined predictions will tend to be larger. Second, outcomes are often easier to predict than causal effects, so if outcomes are informative of treatment effects, the bias may help to distinguish good treatment targets. 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. As our theoretical analysis shows, the systematic bias may even help with causal classification.

We model these ideas theoretically and assess them experimentally. We find that each approach is preferable in different situations depending on three factors: (1) errors due to variance in the probability estimates, (2) whether the bias can be partially corrected by setting a different decision boundary, and (3) the correlation between outcomes and treatment effects. Considering that the treatment effect approach is often more expensive in data acquisition and more technically challenging to implement, our findings are particularly relevant because they describe the conditions under which a simple outcome prediction model can outperform a causal model on an inherently causal task: making optimal interventions.

3. Related Work

Much of the potential outcomes literature focuses on estimating the so-called average treatment effect (ATE) in settings where the treatment is binary and the outcome of interest is continuous (rather than binary, as in our case). The ATE, as the name suggests, corresponds to the average causal effect in a well-defined population, and a large part of the causal inference literature is devoted to the unbiased estimation of the ATE[7] by adjusting for measured or latent confounders. However, the ATE does not discriminate at all between the individuals in the population because it does not account for the fact that effects may vary across individuals, so it does not support making fine-grained targeting decisions.

3.1 From Individual Causal-Effect Estimation to Causal Targeting

Recently there has been a surge in the development of methods for the estimation of individual treatment effects, also referred to as heterogeneous treatment effect (HTE) estimation (Hill, 2011; Wager and Athey, 2018; Weisberg and Pontes, 2015; Taddy et al., 2016;

[7] Or other aggregated estimates, such as the average treatment effect on treated (ATT).

Yahav et al., 2016). HTE studies differ from the present paper because their main object of study is the understanding of treatment effects in heterogeneous subpopulations, not effective treatment assignment.[8] Therefore, these studies are concerned with the accurate estimation of treatment effects, whereas in this paper we care about effect estimates only to the extent that they help to discriminate Persuadables.

Aspects of causal classification have also been considered from the perspective of optimal treatment policies (OTP) (Dudík et al., 2011; Beygelzimer and Langford, 2009; Bhattacharya and Dupas, 2012) and uplift modeling (UM) (Lo, 2002; Kane et al., 2014; Radcliffe and Surry, 2011). UM and OTP studies are concerned with treatment assignment, which is an analogous problem to causal classification. 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 (Hirano and Porter, 2009). As a result, several authors have proposed parametric or semi-parametric models for optimal policies (Bhattacharya and Dupas, 2012; Dehejia, 2005; Manski, 2004). In the setting of contextual bandits (Auer et al., 2002), Beygelzimer and Langford (2009) 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. (2014) describe uplift modeling 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 differs from these studies in at least three ways. First, we analyze and evaluate causal classification with the goal of identifying Persuadables, rather than maximizing 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 too complex to model easily or not readily available (e.g., when deciding which employees to train, online advertising for automobiles), and because it allows us to focus cleanly on a crucial dimension of the real goal: to target individuals for whom the treatment will change the outcome into a positive one.

Second, the focus of our study—and the main contribution of this paper—is to present and analyze the bias–variance tradeoff that exists between ignoring the counterfactual (i.e., outcome prediction) and trying to estimate it (i.e., treatment effect estimation). While the tradeoff between bias and variance errors in classification tasks has received significant research attention since Friedman (1997), to our knowledge this is the first paper that has transferred these ideas to causal classification. More specifically, thinking about treatment assignment as causal classification provides a useful analog to conventional classification tasks with zero-one loss compared to tasks with squared loss. Thus, by focusing on a limited task (classification of Persuadables) rather than on a more general task (estimating treatment effects), we show theoretically that good treatment assignments may be possible even in the presence of biased estimates of causal effects. So, by comparing to straightforward outcome prediction, we quantify mathematically the intuitive notion that "treatment effect estimation may suffer from high variance because it is a complicated estimation."

[8] Compare with discussions about differences between explaining and predicting (Shmueli, 2011).

Third, we run a set of simulation-based experiments to showcase the conditions under which it is preferable to address causal classification using outcome prediction rather than causal estimation, and vice versa, depending on the specific characteristics of the setting and the quality of the models. These simulations show common application settings in which outcome prediction should work better than treatment effect estimation to make intervention decisions, and Appendix D shows an example of such an application with non-simulated data. Radcliffe and Surry (2011) provide a list of general suggestions on when to use uplift modeling that closely aligns with the results we show in this paper, but their suggestions are motivated by their professional experience deploying uplift models rather than a theoretical analysis.[9] We return to this in Section 7.

Overall, our study has important implications for the way we should think about modeling causal classification tasks, as well as for other stages in the data mining process, such as business understanding, data acquisition, and model evaluation.

3.2 Challenges of Causal Targeting

Thoughtful researchers and practitioners have pointed out the potential for increased variance in treatment effect estimation when making targeting decisions. For example, Radcliffe and Surry (2011) note that when predicting treatment effects based on two outcome models (also called the "two-model approach"), the errors may combine in unfortunate ways and degrade the quality of the estimations when the signal-to-noise ratio is low.[10] They also discuss other concerns with the two-model approach: (1) the fitting objective is defined in terms of outcomes rather than effects; (2) because outcome signals tend to be much stronger than effect signals, the fitting procedure may prioritize the former; (3) the fitting procedure may prioritize features that are predictive of outcomes but not of treatment effects.

Many researchers have also proposed methods that directly model effects (rather than outcomes) to address these issues (see Dorie et al., 2019, for a recent survey). Nie and Wager (2021), for example, develop a general class of two-step algorithms for HTE estimation and note that effect predictions are "unstable" when made using outcome models trained separately for the treated and the untreated. Similarly, Künzel et al. (2019) compare several meta-learning approaches that use off-the-shelf supervised learning algorithms for HTE estimation. They show through simulations that the two-model approach (which they call "T-learner") is prone to overfitting (particularly when the treated and untreated samples are of different sizes). Therefore, several methods have been specifically designed to deal with the large variance often encountered in causal-effect estimation. These methods, discussed in more detail in Section 4.8, may not necessarily suffer from larger variance than outcome prediction. However, because these studies are concerned with HTE estimation, their com-

[9] In fact, their paper points out a theory/practice paradox: "*While we would argue that, in principle, uplift modeling is almost always, from a theoretical perspective, the correct way to formulate marketing response modeling, it is not the case that in practice, it will always produce superior results, nor that even when it will, the improvement always justifies the extra complexity.*" Our analysis provides one answer to this paradox, giving theoretical support to their claim that uplift modeling may not always be the correct approach for treatment assignment.

[10] Radcliffe and Surry (2011) also acknowledge that "*We are not, however, able to make this idea mathematically precise.*" Our work in this paper adds this precision.

parisons are with respect to squared (rather than zero-one) loss and do not consider the performance of the proposed methods for treatment assignment.

Interestingly, regarding treatment assignment, Jaskowski and Jaroszewicz (2012) note that if the treatment effect is strongly correlated with the outcome, the two-model approach may perform very well, and in two out of their three examples the two-model approach beats other methods that model effects directly. Olaya et al. (2020) report similar results when assigning non-binary treatments. In their study, the separate-model approach (SMA) corresponds to predicting treatment effects based on multiple outcome prediction models. They find that none of the evaluated techniques, which include state-of-the-art methods for treatment effect estimation, consistently outperforms the SMA in terms of treatment assignment performance. Hence, they conclude that performance largely depends on the context and problem characteristics. However, these studies do not consider the possibility of a simple outcome model beating an uplift (or treatment effect) model.

4. Causal Targeting, Bias & Variance

In this section, we introduce the different methods for causal classification that we will analyze and the simplifying assumptions we make to carry out the analysis. We then go on to define *causal* bias and variance and derive their values for the different causal classification methods. This sets us up to analyze the paper's key question: When (if ever) is simple outcome prediction preferable to treatment effect estimation for causal classification? After presenting the mathematical condition that answers this question, we discuss the implications for choosing one causal classification method over another.

4.1 Causal Classification Approaches

Consider these three approaches to address causal classification:

1. **Outcome Most (OM)**: Target the observations with the highest estimated probability of a positive outcome when *treated*.

2. **Outcome Least (OL)**: Target the observations with the least estimated probability of a positive outcome when *untreated*.

3. **Treatment Difference (TD)**: Target the observations with the largest difference between the estimated probabilities of a positive outcome when treated and untreated.

The first two ignore the counterfactual, while the latter explicitly estimates it.[11] In this section, we present a theoretical comparison of these approaches. We focus on OM and TD, but the results for OL are directly analogous to the ones we present. We also discuss the alternative of targeting the observations with the highest probability of a positive outcome when untreated—the opposite of OL and a common advertising scenario.

We assume that there are no Do-Not-Disturbs, which allows us to interpret average treatment effects as probabilities of being Persuadable.[12] However, this assumption is not

[11] In the uplift modeling literature, OM and OL would usually be referred to as response modeling, while TD would be referred to as uplift modeling.

[12] If $P(Y_T = 0, Y_U = 1) = 0$, then $P(Y_T = 1) - P(Y_U = 1) = P(Y_T = 1, Y_U = 0)$.

Table 2: Intended targets by approach

	Scoring Function	Intended Targets
Outcome Most (OM)	$\hat{P}(Y_T = 1\|X)$	Persuadables Sure Things
Outcome Least (OL)	$\hat{P}(Y_U = 0\|X)$	Persuadables Lost Causes
Treatment Difference (TD)	$\hat{P}(Y_T = 1\|X)$ $-\hat{P}(Y_U = 1\|X)$	Persuadables

Y_T and Y_U are the (potential) outcomes for treated and untreated.

necessary to prove any of our theoretical derivations, and we return to the practical implications of this assumption in Section 7. This assumption, also known as monotone treatment response (Manski, 1997), is common in many causal inference methods, such as when using instrumental variables (Angrist et al., 1996). In our previous example, this would mean that we are assuming that advertising does not reduce the probability of someone buying the product. Also, as detailed in the description of TD, we assume that treatment effects are estimated as the difference between the outputs of two outcome prediction models (i.e., using a two-model version of treatment effect estimation). This assumption is not necessary for our theoretical derivations either, and we discuss its implications in Section 4.8. In the analysis that follows, making these assumptions allows us to derive elegant conditions that differentiate the scenarios where we would prefer TD over OM.

Classification problems are usually 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 chosen threshold. Typically, estimated probabilities of class membership are used as scores, and a default threshold of 0.5 is used to discriminate observations. In practice, we would choose a threshold appropriate for the problem at hand, based on costs, benefits, budget constraints, etc. (Provost and Fawcett, 2013). Table 2 shows how the different causal classification approaches would score and discriminate observations. To simplify the notation, we will assume that scores are conditioned on a specific X for the following analysis (i.e., assume $X = x$ so that this analysis is specific to individuals with the feature vector x).[13]

4.2 Causal Bias

In causal classification, the optimal (albeit infeasible) solution would be to rank observations according to their probability of being a Persuadable: $P(Y_T = 1, Y_U = 0)$, where Y_T and Y_U are the (potential) outcomes when treated and untreated, respectively. Both OM and TD favor Persuadables, but OM gives a high score to Sure Things, too. Thus, when being used

[13] Returning to the second approach for causal classification (OL), the scoring function for TD could also be defined as $\hat{P}(Y_U = 0|X) - \hat{P}(Y_T = 0|X)$. From this formulation, one can see that analyzing the tradeoff between OL and TD is analogous to analyzing the tradeoff between OM and TD. The only difference is that OM has a bias towards Sure Things, while OL has a bias towards Lost Causes.

to rank Persuadables, OM will be biased. As this sort of bias is different from the bias in the underlying estimation of the scoring function used by OM, we will call it **causal bias**. Thus, due to causal bias, OM may have a higher causal-classification false-positive rate than the optimal solution. We can view TD as correcting for the causal bias in OM because, in principle, it reduces the scores of the Sure Things. Nevertheless, as discussed in more detail in the following sections, correcting for the causal bias may actually be undesirable for the purposes of decision making.

Additionally, both OM and TD may suffer from **estimation bias**, meaning that their scoring functions may be biased with respect to what they intend to estimate (e.g., $E[\hat{P}(Y_T)] \neq P(Y_T)$). This type of bias would include the "model bias" typically discussed in machine learning (e.g., using a linear model in the presence of non-linear relationships), but it is not limited to that. For example, consider confounding: if the treated are systematically different from the untreated in unobserved ways, then $E[\hat{P}(Y_T|X, T=0)] \neq P(Y_T|X, T=0)$. Thus, as defined here, estimation bias would also include bias due to confounding and other selection biases in the data generating process. So, let ξ_T and ξ_U be the estimation biases in the scoring function for the treated and the untreated, respectively:

$$\xi_T = E[\hat{P}(Y_T)] - P(Y_T) \tag{1}$$

$$\xi_U = E[\hat{P}(Y_U)] - P(Y_U). \tag{2}$$

For now, however, let us assume that the scoring functions are unbiased with respect to what they are intending to estimate (i.e., $\xi_T = \xi_U = 0$). Then, OM, as an estimate of the probability of being a Persuadable, has an upward causal bias equal to the probability of being a Sure Thing:

$$\begin{aligned}\text{Causal Bias (OM)} = &\; E[\hat{P}(Y_T = 1)] - P(Y_T = 1, Y_U = 0) \\ = &\; P(Y_T = 1, Y_U = 1).\end{aligned}$$

At the same time, however, the causal bias of TD is:

$$\begin{aligned}\text{Causal Bias (TD)} &= E[\hat{P}(Y_T = 1) - \hat{P}(Y_U = 1)] - P(Y_T = 1, Y_U = 0) \\ &= -P(Y_T = 0, Y_U = 1).\end{aligned}$$

Under the assumption that there are no Do-Not-Disturbs:

$$\begin{aligned}\text{Causal Bias (OM)} &= P(Y_U = 1) \\ \text{Causal Bias (TD)} &= 0.\end{aligned}$$

4.3 Causal Variance

However, comparing these approaches requires more than just accounting for their biases because their scoring functions use *estimates* of the true outcome probabilities induced from a data sample. So, even if the estimators are unbiased, the variance in the estimation procedure may lead to inaccurate probability estimates. Assuming the variances of $\hat{P}(Y_T)$ and $\hat{P}(Y_U)$ come from sampling error in the training data, the errors in the estimates would not be correlated. If we define the **causal variance** as the variance in the estimation of the probability of being a Persuadable (which is different from the estimation of the outcome probabilities), then we can see that TD will suffer from larger causal variance than OM because it uses two probability estimates rather than one:[14]

$$\begin{aligned} Var(TD) &= Var[\hat{P}(Y_T = 1) - \hat{P}(Y_U = 1)] \\ &= Var[\hat{P}(Y_T = 1)] + Var[\hat{P}(Y_U = 1)] \\ &\geq Var[\hat{P}(Y_T = 1)] = Var(OM). \end{aligned}$$

Given that most of our analysis is specific to causal (rather than outcome) classification, for the remainder of the paper we will refer to causal variance simply as variance unless its meaning is ambiguous from the context.

4.4 Causal Bias–Variance Tradeoff

Because one approach has higher causal bias and the other higher causal variance, it is unclear which approach is preferable. Let's consider this tradeoff in terms of how it affects optimal decision-making. Returning to the work of Friedman (1997) on the bias–variance tradeoff for normal classification, bias and variance affect how often a classifier disagrees with the Bayes-optimal model. For causal classification, suppose $A \in \{0, 1\}$ represents whether someone is classified as a Persuadable. Let τ be the optimal decision threshold for a scoring function f that returns the probability of being a Persuadable, meaning that τ considers the costs of incorrectly classifying Persuadables and non-Persuadables. Then, A^* is a Bayes-optimal causal classifier:

$$A^* = \mathbf{1}(f \geq \tau),$$

where $f = P(Y_T = 1, Y_U = 0)$. Suppose we have a classifier $\hat{A}$ that uses a scoring function $\hat{f}$ estimated from training data:

$$\hat{A} = \mathbf{1}(\hat{f} \geq \tau).$$

Let $p(\hat{f})$ be the probability density function of $\hat{f}$, ω_1 be the cost of a false negative, and ω_2 be the cost of a false positive. Then, we can express causal-classification model error as:

$$\begin{aligned} \text{Error } (\hat{A}) = {} & \text{Cost}(\hat{A} \neq A^*) \\ = {} & A^* \omega_1 P(\hat{A} = 0) + (1 - A^*) \omega_2 P(\hat{A} = 1) \\ = {} & A^* \omega_1 \int_{-\infty}^{\tau} p(\hat{f}) d\hat{f} + (1 - A^*) \omega_2 \int_{\tau}^{\infty} p(\hat{f}) d\hat{f}. \end{aligned} \tag{3}$$

Furthermore, suppose we approximate $p(\hat{f})$ using a normal distribution:

$$p(\hat{f}) = \frac{1}{\sqrt{2\pi Var[\hat{f}]}} \exp\left(-\frac{(\hat{f} - E[\hat{f}])^2}{2\, Var[\hat{f}]} \right).$$

[14] TD does not necessarily suffer from larger variance than OM when the estimation is not based on the difference between two model predictions (e.g., when using some of the methods for treatment effect estimation introduced in Section 3.2). This does not change any of our subsequent theoretical derivations or the main conclusions of the study. We discuss this in more detail in Section 4.8.

Friedman argued that such an approximation is reasonable because the computation of $\hat{f}$ often involves a "sometimes complex" averaging procedure and because the qualitative conclusions are still generally valid even if this is not the case. Indeed, we will show in later sections that our qualitative conclusions hold even in the absence of normality. Under this assumption, we have the following lemma (proof based on Friedman, 1997, in Appendix A).

Lemma 1 *Let f be the average treatment effect given a specific set of feature values, and let $\hat{f}$ be the effect estimate provided by a scoring model. If $\hat{f}$ is normally distributed (conditional on the set of feature values), then the causal-classification model error for an individual with those feature values may be expressed as:*

$$Error\ (\hat{A}) = \tilde{\Phi}\left[sign(f-\tau)\frac{E[\hat{f}]-\tau}{\sqrt{Var[\hat{f}]}}\right](A^*\omega_1 + (1-A^*)\omega_2), \tag{4}$$

where A^* is the optimal classification for that individual, τ is the optimal threshold when making classifications with f, ω_1 is the cost of a false negative, ω_2 is the cost of a false positive, and $\tilde{\Phi}$ is the upper tail area of the standard normal distribution:

$$\tilde{\Phi}(z) = \frac{1}{\sqrt{2\pi}}\int_z^\infty \exp\left(-\frac{u^2}{2}\right) du.$$

Lemma 1 implies that bias and variance affect model error in very different ways. Statistically speaking, bias is defined as the difference between $E[\hat{f}]$ and f, and in traditional regression problems, a larger bias means worse predictions. However, (as shown in Equation 4) for classification problems, bias only hurts when it goes in the "opposite direction" of the correct classification: $(f-\tau)*(E[\hat{f}]-f) < 0$. Bias may help when it goes in the same direction as the correct classification because it can lessen errors due to variance. Therefore, correcting for the bias can be counterproductive for decision making! In contrast, larger variance hurts when the expected score is on the "correct side" of the threshold (i.e., $(f-\tau)*(E[\hat{f}]-\tau) > 0$), but it helps otherwise because the variance may push the estimates back to the "correct side" by chance.

4.5 When is OM Preferable to TD?

Returning to our original motivation of comparing different approaches for causal classification, when comparing two models, the better model is the one with the lower model error (as defined in Equation 3). So, under the assumption of normality, the following theorem determines which is the better model (proof in Appendix B).

Theorem 2 *Let f be the average treatment effect given a specific set of feature values, and let $\hat{f}_1$ and $\hat{f}_2$ be the effect estimates provided by two scoring models. If $\hat{f}_1$ and $\hat{f}_2$ are normally distributed (conditional on the set of feature values), then $\hat{f}_1$ leads to lower causal-classification model error than $\hat{f}_2$ for individuals with those feature values if:*

$$\frac{b_1}{m} < 1 + \frac{\frac{b_2}{m}-1}{\sqrt{\gamma}}, \tag{5}$$

where $b_1 = f - E[\hat{f}_1]$ is the (negation of the) bias in $\hat{f}_1$, $b_2 = f - E[\hat{f}_2]$ is the (negation of the) bias in $\hat{f}_2$, $\gamma = \frac{Var[\hat{f}_2]}{Var[\hat{f}_1]}$ is the ratio of the variances of $\hat{f}_2$ and $\hat{f}_1$, and $m = f - \tau$ is the causal margin—the distance between the conditional average treatment effect and the optimal decision boundary (or threshold) when f is used to make classifications.

Returning to our comparison of TD and OM, we can examine specific values for b_1, b_2, and γ to characterize the scenarios where OM ($\hat{f}_1$) may perform better than TD ($\hat{f}_2$) due to the causal bias–variance tradeoff between them. For starters, define $\gamma \geq 1$; that is, TD suffers from higher variance than OM (see Section 4.3). This addresses the variance.

The bias, in contrast, requires a more detailed look. Given a feature vector $X = x$, let α be the base rate for positive outcomes when untreated and $\beta = f$ be the expected treatment effect (the expected change in the probability of a positive outcome for the $X = x$ on which we are conditioning):

$$\alpha = P(Y_U = 1)$$
$$\beta = P(Y_T = 1) - P(Y_U = 1) = f.$$

Note that α is equal to the causal bias in OM (i.e., the ignored counterfactual). Under the assumption of no Do-Not-Disturbs, α and β can be interpreted as the probabilities of being a Sure Thing and a Persuadable. Thus, we may use them to define the expected values of the scoring functions of OM ($\hat{f}_1$) and TD ($\hat{f}_2$) as follows:

$$E[\hat{f}_1] = \beta + \alpha + \xi_T$$
$$E[\hat{f}_2] = \beta + \xi_T - \xi_U,$$

where ξ_T and ξ_U are the estimation biases in the scoring function for the treated and the untreated, respectively (see Equations 1 and 2).

While this formulation is useful for comparing OM and TD, such a comparison is not quite fair because it does not reflect how one would actually use the models. Recall that τ is the threshold on the scoring function (in our formulation, the probability decision threshold). Because the causal bias in OM is systematically upwards, OM could be partially corrected by choosing a higher threshold. To make the two approaches comparable using threshold τ, we will assume that a constant δ_1 is subtracted from $\hat{f}_1$, and a constant δ_2 is subtracted from $\hat{f}_2$ (which is equivalent to using a different scoring threshold for each approach), without designating any specific values for δ_1 and δ_2. This adjustment addresses situations in which OM and TD classify Persuadables using different probability thresholds and situations in which τ is only used implicitly. For example, when targeting a fixed number of individuals with the largest scores based on model $\hat{f}_j$, we may be using some $\tau + \delta_j$ without knowing τ or δ_j. We can then define biases b_1 and b_2 as:

$$b_1 = f - E[\hat{f}_1] = \delta_1 - (\alpha + \xi_T) \tag{6}$$
$$b_2 = f - E[\hat{f}_2] = \delta_2 - (\xi_T - \xi_U). \tag{7}$$

Therefore, b_1 and b_2 may be interpreted, respectively, as OM's and TD's **uncorrectable bias**. Further, recall that $m = \beta - \tau$ is the **causal margin** (see Equation 5), the distance between the expected treatment effect and the decision boundary (or threshold). For mathematical clarity, Theorem 2 casts uncorrectable bias in relative terms, (b/m)—the **causal boundary bias**. This quantity can be interpreted as the relative distance that the estimates are pushed in the direction of the decision boundary due to uncorrectable bias (so a larger quantity is worse). Whether OM is better than TD will thus depend on the causal boundary bias of each approach and the relative increase in variance when using TD instead of OM. For the rest of the paper, we refer to causal boundary bias simply as boundary bias.

4.6 Causal Bias–Variance Tradeoff: Three Scenarios for Causal Classification

We discuss first the case in which scoring functions are unbiased ($\xi_T = \xi_U = 0$), and no threshold correction is applied to TD ($\delta_2 = 0$).[15] Doing so allows us to focus on whether OM can perform better than TD even in the absence of confounding (e.g., due to selection bias). Unless controlled for perfectly, confounding will reduce the quality of TD's causal-effect estimates, so (at least on the surface) the unconfounded setting gives the most advantage to TD (however, see the more nuanced perspective in Section 4.8 below). Note that the unconfounded setting is not of purely theoretical interest, as firms can gather data from A/B tests (at a cost) to build unconfounded models for TD. Under these assumptions, $b_2 = 0$, and the following corollary follows from Theorem 2.

Corollary 3 *If $b_2 = 0$ then $\hat{f}_1$ leads to lower classification model error than $\hat{f}_2$ if:*

$$\frac{b_1}{m} < 1 - \frac{1}{\sqrt{\gamma}}, \tag{8}$$

Equation 8 allows us to clearly delineate the approaches and provides a quantification of when we should heed the motivating intuition that "TD may be worse because of higher variance." Recall $\gamma \geq 1$ when comparing OM and TD. Thus, according to Corollary 3:

1. If the boundary bias in OM is larger than 1, TD is preferable.
2. Otherwise, if the boundary bias in OM is larger than 0, TD is preferable *only if* the increase in variance is small compared to the boundary bias.
3. If the boundary bias in OM is less than or equal to 0, OM is preferable.

As expected, TD is preferable if the boundary bias is very large. Recall that boundary bias represents the distance that estimates are pushed in the direction of the causal-classification decision boundary. Therefore, considering that boundary bias is the ratio between uncorrectable bias and the causal margin, a boundary bias larger than 1 implies that the bias is so strong that OM's scores are pushed to the "wrong" side of the decision boundary on average. In this scenario, we would always choose TD no matter the increase in variance. However, if the boundary bias is positive but smaller than 1, we choose TD only if the increase in variance (due to choosing TD) is small compared to the boundary bias. This increase in variance is captured by γ, the ratio between the TD and OM variances. Figure 1 shows the maximum increase in variance that would justify choosing TD. For a given boundary bias, TD is preferable as long as its increase in variance is not above the curve. Otherwise, OM is preferable.

[15] The threshold correction is not necessary for TD when the scoring functions are unbiased.

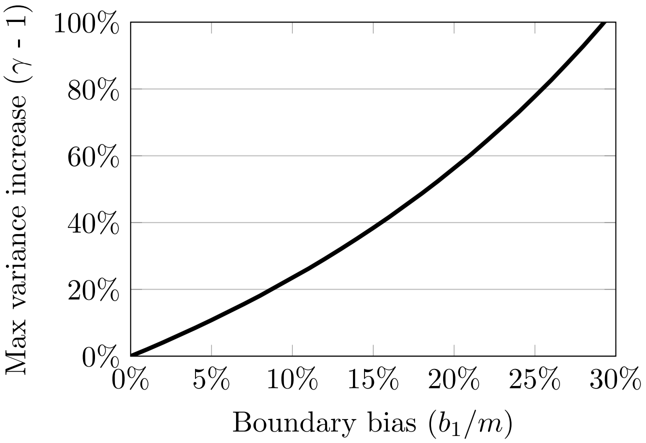

Figure 1: **Maximum variance increase to prefer TD over OM**. For a given boundary bias, a variance increase above the curve implies it is better to target based on outcome prediction (OM) rather than treatment effect estimation (TD).

Equation 8 reveals other situations in which OM may be preferable. For instance, boundary bias will be close to zero (favoring OM) if most of the causal bias in OM scores can be corrected by using a higher threshold. Because δ_1 is a correction that is chosen, suppose as an example that we set δ_1 as the base rate across the entire untreated population (i.e., the average bias):

$$\delta_1 = E[P(Y_U = 1|X)] = E[\alpha].$$

Using this δ_1, OM's uncorrectable bias (see Equation 6) would depend on the dispersion among the probabilities of a positive outcome when untreated (α). If the probabilities are similar across all observations (i.e., the bias is more-or-less uniform), then most of the bias can be corrected with the higher threshold. The intuition behind this is that while OM systematically overestimates the probability of being a Persuadable, the ranking of the observations will tend to be the same as that of the optimal model if α does not vary much, leading to low boundary bias. Alternatively, if the causal margin is large, boundary bias will also be smaller, ceteris paribus. Note the key difference between causal classification and precisely estimating treatment effects: for causal classification, ranking by likelihood of being a Persuadable is most important. Thus we can tolerate certain sorts of causal bias without negatively affecting our choices of whom to target.

If boundary bias is negative, counterintuitively, we choose OM. Boundary bias is negative when the uncorrectable bias and the causal margin have different signs. This happens when the causal bias pushes the scores further in the direction of the correct decision, for example, when large causal effects are overestimated. This might occur if the individuals who are more likely to have a positive outcome are also more susceptible to treatments, which makes sense in certain applications. For example, people more likely to buy a product organically may also be more likely to be influenced by advertisements for the product. Mathematically speaking, this implies a correlation between α and β (when not conditioning on a specific

$X = x$), so boundary bias will tend to be negative. In this case, OM would be preferable. We discuss the further implications of this and other types of bias in the next sections.

4.7 Other Implications of Causal Bias

OM targets based on a model of who is most likely to have a positive outcome (e.g., to purchase) when treated. A popular alternative in modern marketing settings is to target individuals based on models that predict the probability of having a positive outcome when *untreated*—especially when considering a new treatment. This type of modeling is also known as **purchase modeling (PM)** (Radcliffe and Surry, 2011). This approach seems strange on the surface because these models are estimating who is most likely to be a Sure Thing! However, a key factor to keep in mind is that because the models estimate the probability of being a Sure Thing, even the highest probabilities might be very low. As a result, most individuals with the highest probabilities of being a Sure Thing would not actually be Sure Things. However, they may well be the most likely to be Persuadables.

In the previous setting, we showed that OM is preferable when there is a positive correlation between α and β because the boundary bias would be negative. If this is the case, then PM may also be a suitable alternative. It may be even more so if there is substantially more data for the untreated than for the treated, which may be the case even beyond the cold-start period for small targeting budgets. The intuition behind this is that it is usually much easier to predict outcomes than to predict causal effects because the signal for the former is (very often) stronger. Therefore, to the extent that outcomes (the probability of a Sure Thing) and treatment effects (the probability of a Persuadable) are correlated, targeting based on outcomes rather than effects may lead to better treatment assignments. In other words, the ranking of α and β will be similar if they are correlated. However, directly analogous to the analysis above, because α is easier to predict, it may be better to perform the causal classification task by ranking by $\hat{\alpha}$ instead of $\hat{\beta}$, even though what we really care about is the ranking by β.

Indeed, the estimands of OM, PM, and TD have the same ranking under the assumption that the treatment amplifies the outcome. That is, assuming that the treatment has an effect that is increasing in the probability of a positive outcome without treatment, then a larger probability of being a Sure Thing also implies a larger probability of being a Persuadable. For example, if you are one of the most likely to buy a pizza from us this Friday evening without being targeted, you might also be one of the most likely to be affected by our Friday evening special offer coupon—especially if everyone's likelihood without targeting is rather low. Similarly, if you have one of the lowest likelihoods of buying a pizza organically, you may indeed be a lost cause for the special offer (maybe you're low carb or gluten free). Thus, ranking according to $\alpha + \beta$, α, or β would be equivalent (which are the estimands of OM, PM, and TD, respectively). However, as mentioned above, recall that the models provide *estimates* of these quantities, not the true values. Therefore, estimating the ranking using OM might be easier because of its stronger signal ($\alpha + \beta \gg \beta$), and ranking using PM might be even better when there are limited data for the treated (because $\hat{\alpha}$ would be less noisy than $\widehat{\alpha + \beta}$).

4.8 Implications of Estimation Bias and Treatment Effect Modeling

We discussed in detail the implications of OM's causal bias under the assumption of no other types of bias—such as confounding bias—to show that, even without confounding, OM may be preferable. It would be easy to say that things would be worse with confounding, given that causal-effect modeling (TD) would be even more difficult (and thus more error-prone). However, as revealed by Equation 5, things might not actually be worse for causal classification: confounding helps if it leads to a negative boundary bias! Confounding may lead to better treatment assignments if the confounding bias is positively correlated with the treatment effect (Fernández-Loría and Provost, 2019). This may happen when the selection bias that produced the confounding is driven by the magnitude of the treatment effect. Specifically, this would translate to a positive correlation between β and $\xi_T - \xi_U$.

Similarly, the presence of model bias might affect the predictions of OM more than those of TD. As an example, suppose the classification models tend to systematically overestimate the effect of a feature X on the outcome, regardless of the treatment assignment. Then, the individual outcome probability estimates will be worse. However, as TD is computed as the difference between these scores, it may not be affected much—because the outcome predictions given X will be shifted by the same amount (on average) for both classifiers. Note though that in this particular example, if this overestimation is systematic across all individuals, then the common use of budget- or quantile-based thresholding would automatically deal with this shift for OM as well (i.e., by implicitly setting a larger δ_1).

Another related point is that TD could perform better if implemented with a learning procedure specifically designed to estimate effects, instead of using paired outcome prediction models. This would correspond to the methods we discussed in Section 3.2. In our analysis, TD has a larger variance than OM because we assume that its estimation is done using two outcome prediction models (one for the treated and another for the untreated), but TD could potentially achieve a lower variance if the model estimation is optimized to predict effects. For example, suppose that features X_1 and X_2 are predictive of outcomes, but only feature X_1 is predictive of effects. This implies that, for the purposes of identifying Persuadables, segmenting individuals using exclusively X_1 is more statistically efficient than segmenting them according to X_1 and X_2. Hence, by focusing statistical power on features that are predictive of effects, models optimized for treatment effect estimation can achieve lower bias and lower variance than models optimized for outcome prediction (Nie and Wager, 2021; Foster and Syrgkanis, 2019).

The implication for our analysis is that when using such methods, in Equation 5 and Equation 8, γ could possibly have a value smaller than 1. So, our theoretical derivations do not change even if TD is implemented with a method specifically designed to estimate effects. However, if γ is indeed smaller than 1, then OM can only outperform TD when the boundary bias is negative, which occurs when effects and outcomes are correlated. This can also benefit modeling because identifying features that predict effects can be significantly harder than identifying features that predict outcomes when effect sizes are small. So, if effects and outcomes are correlated, segmenting according to outcomes may provide better results than segmenting according to effects. As discussed in more detail in Section 7, this can be cast as a specific form of transfer learning or proxy modeling.

5. Simulator

The theoretical analysis in the previous section reveals the regimes in which a simple outcome prediction model may be preferable to a causal-effect model. Extending the same analysis to a collection of decisions is not straightforward because the results will depend on the distributions of the causal effects, the bias, and the variance across the population. Empirically testing these analytical results is also challenging, as (i) we cannot know the counterfactual truth, and (ii) any single empirical setting would likely only give a narrow view of the possible scenarios (perhaps not even including an interesting one).

Therefore, we use simulations to address both of these challenges, allowing us to know the ground truth and extend the analysis to multiple intervention decisions under a variety of settings. With this purpose in mind, we now present a simulator designed to meet two major requirements. First, the simulator should allow us to adjust the distributions of the probabilities of different observation types (e.g., Persuadables) to assess the implications of causal bias. Second, it should also allow us to adjust how good the scoring functions are at estimating these probabilities to assess the implications of causal variance.

The simulator's main design assumption is that the probabilities of different observation types are distributed according to a logistic-normal distribution, which is the logistic transformation of a multivariate normal distribution. One of the main advantages of using this distribution is that we can model the median, variance and correlation of probabilities with great flexibility, allowing us to satisfy the first requirement of the simulator. Thus, the simulator receives the following five parameters to meet the first requirement:

1. **Median of the probability of a Sure Thing ($\tilde{\alpha}$)**
2. **Median of the probability of a Persuadable ($\tilde{\beta}$)**

Thus, assuming that probabilities follow a logistic-normal distribution (defined by the multivariate normal $[A, B]$), these parameters can be used to solve the following equations and recover the location parameters of the logistic-normal distribution (μ_A and μ_B):

$$e^{\mu_A}(1 + e^{\mu_A} + e^{\mu_B})^{-1} = \tilde{\alpha}$$
$$e^{\mu_B}(1 + e^{\mu_A} + e^{\mu_B})^{-1} = \tilde{\beta}.$$

3. **Variance of A (σ_A^2)**
4. **Variance of B (σ_B^2)**
5. **Pearson correlation coefficient of A and B (ρ)**

We simulate the probabilities of each observation type by sampling from the following multivariate normal distribution:

$$\begin{pmatrix} A \\ B \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix}\right],$$

and then applying a logistic transformation as shown in Table 3.

The second requirement of the simulator relates to the estimation of the probabilities in Table 3. One alternative to meet this requirement would be to simulate features along

Table 3: Simulator probabilities

Obs. Type	Probability	Simulator Output
Sure Thing	$P_1 = P(Y_T = 1, Y_U = 1)$	$e^A(1 + e^A + e^B)^{-1}$
Persuadable	$P_2 = P(Y_T = 1, Y_U = 0)$	$e^B(1 + e^A + e^B)^{-1}$
Lost Cause	$P_3 = P(Y_T = 0, Y_U = 0)$	$(1 + e^A + e^B)^{-1}$
Do-Not-Disturb	$P_4 = P(Y_T = 0, Y_U = 1)$	0

The vector $[A, B]$ follows a multivariate normal distribution.

these probabilities, then apply machine learning algorithms to learn models that predict causal effects and outcomes based on those features, and finally apply the models on new data to assess their performance. However, a key disadvantage of this approach is that traditional supervised learning errors would come into play, preventing us from giving a clean assessment of the implications of causal bias and variance. Furthermore, the analysis would be specific to the learning procedure used to learn the models.

Thus, rather than simulating features to subsequently learn models, we simulate model predictions directly. Our simulator generates predictions free from estimation bias to facilitate a clean analysis of causal bias and variance. However, its formulations could be easily adjusted to generate biased predictions (and thus simulate confounding or model bias). More formally, let P_1 be the probability of a Sure Thing and P_2 be the probability of a Persuadable (both defined in Table 3). Let $\hat{P}(Y_T)$ be the estimated probability of a positive outcome when treated, and $\hat{P}(Y_U)$ be the estimated probability of a positive outcome when untreated. Then:

$$E[\hat{P}(Y_T)] = P(Y_T) = P_1 + P_2$$
$$E[\hat{P}(Y_U)] = P(Y_U) = P_1.$$

We simulate the probability estimates using beta distributions. The benefits of doing this are threefold: (1) it is easy to manipulate the variance (and thus the quality) of the estimates, (2) it bounds the scoring function between 0 and 1 so that scores can be interpreted as probabilities, and (3) it allows us to compare the models under alternatives to the normality assumption made in the previous section, because beta distributions can take on a wide variety of "shapes." The simulator receives the following two parameters to determine the quality of the probability estimates:

6. **Quality of probability estimates when treated (η_T)**

7. **Quality of probability estimates when untreated (η_U)**

The quality parameters must be larger than zero. Higher values tighten the beta distributions around the true probabilities, so probability estimates are closer to the true probabilities when quality increases. Probability estimates are sampled from:

$$\hat{P}(Y_T) \sim Beta\ (\eta_T P(Y_T), \eta_T(1 - P(Y_T)))$$
$$\hat{P}(Y_U) \sim Beta\ (\eta_U P(Y_U), \eta_U(1 - P(Y_U))).$$

Table 4: Simulator default parameters

$\tilde{\alpha}$	$\tilde{\beta}$	σ_A^2	σ_B^2	ρ	η_U	η_T	λ	N
0.001	0.001	1	1	0	10^3	10^3	5%	10^7

We use the results from the simulated observations to compare OM and TD as follows. First, we create a set of simulated observations based on a choice of simulation parameters, rank the observations according to P_2 (the optimal model), and select the observations in the top λ percent of the ranking. Then, we rank observations according to OM and TD and also select the observations in the top λ percent. We compare OM and TD by computing the match between the observations in their top λ percent and the observations in the top λ percent of the optimal model. Importantly, λ implicitly defines the probability threshold for each approach, which corresponds to the score on the top λ^{th} percentile. Thus we can compare both approaches without penalizing any of them for not choosing an optimal threshold. The simulator, therefore, also receives these parameters:

8. **Number of simulated observations (N)**

9. **Percentage of targeted observations (λ)**

6. Experimental Results

We ran simulations to illustrate conditions that may arise in practice and in which outcome prediction could outperform treatment effect estimation to test our analytical results. The first setting represents cases where the data to estimate one of the counterfactuals are limited. The second setting relates to situations in which outcomes and treatment effects are correlated. Table 4 shows the default values used for the simulation parameters. For each setting, we specify which parameters have values different from the defaults. The low values for $\tilde{\alpha}$ and $\tilde{\beta}$ were chosen to be representative of targeting for online advertising (Perlich et al., 2014), and the other values were chosen arbitrarily. We also generated results with other default values; the results concurred qualitatively with what we present here. The simulator code is available in Appendix C.

Our experiments focus primarily on comparing OM with TD, to remain consistent with the theoretical derivations above. However, comparing OL and TD produces results that are directly analogous to the ones discussed here. Therefore, the results are relevant to settings in which it is difficult to obtain (training) data for either the treated or the untreated.

6.1 Limited Training Data

This subsection considers first a setting in which the causal bias (the probability of a Sure Thing) is uncorrelated with the treatment effect and varies substantially from one individual to another. So, the bias is generally unhelpful, and a threshold correction is not enough to remove it. Under these circumstances, we would expect TD to easily beat OM. However, building machine learning models for TD requires training data on "both sides" of the counterfactual (treated individuals and untreated individuals), and acquiring such

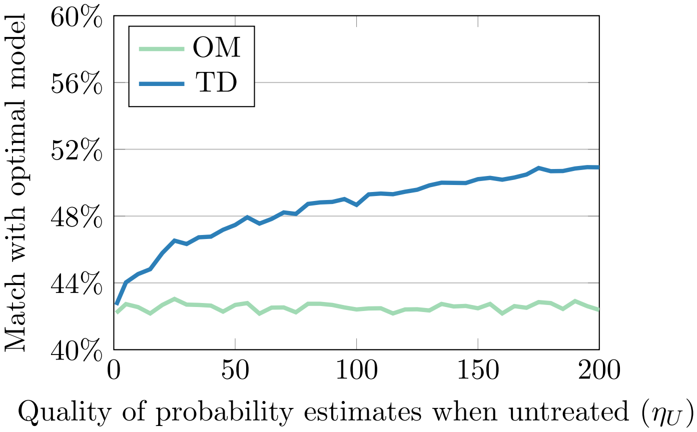

Figure 2: **Typical performance**. Treatment effect estimation (TD) should dominate simple outcome prediction (OM) in most settings, but the degree of dominance will depend on the quality of the counterfactual estimates. The horizontal axis also represents increasing investment in data, in settings where one must invest in the data necessary to build models for both sides of the counterfactual.

data is often substantially more expensive than gathering data just on “one side” of the counterfactual. This has particular relevance when learning models because if the training data set used to estimate TD is small, the sampling error of TD may be large. As a result, the gap in performance between OM and TD should decrease if the causal bias of OM is small relative to the larger sampling error (causal variance) of TD.

Importantly, having few training data for “one side“ of the counterfactual is endemic in many causal modeling settings due to the costs of experimentation. For example, advertisers may resist withholding promotions to good candidates due to potential lost revenue, leading to limited data about the behavior of consumers when untreated—exactly the data required to accurately estimate the counterfactual for the most likely purchasers. This is also common in settings where there is plenty of data for the untreated but not for the treated, such as when firms decide to target a new retention offer to the customers least likely to stay if not given the offer (i.e., the OL approach); given that the offer is new, the firm would have little or no data for the treated. Thus, this is a common situation in a wide variety of settings.

To simulate the conditions described above, we use the default parameters and compare the performance of OM and TD using different values for η_U (the quality of the probability estimates when untreated) to simulate different sampling errors from having fewer or more training data. A smaller η_U implies a larger sampling error for the probability estimate of a positive outcome when untreated. Therefore, TD performs worse when η_U is small (e.g., when only limited training data on the untreated are available). Figure 2 shows the results.

We can see that TD is always preferable to OM, but the gap in performance depends on how good the probability estimates for the untreated are. So, practitioners should carefully consider whether the potential benefit of correcting for the bias is sufficient to compensate for the investment in the training data. Moreover, the marginal value of improving quality

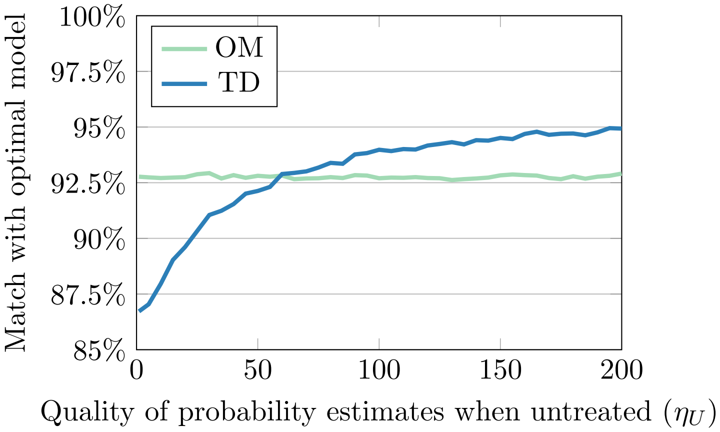

Figure 3: **Causal bias–variance tradeoff in action**. Outcome prediction (OM) outperforms treatment effect estimation (TD) if its boundary bias is small relative to the variance in TD. The horizontal axis also represents increasing investment in data in settings where one must invest in the data necessary to build models for both sides of the counterfactual.

to correct for the causal bias (e.g., by using a larger training sample) is decreasing. Indeed, the curve for TD in Figure 2 is directly analogous to a standard learning curve, which generally increases steeply at first but has decreasing marginal improvements with more training data. Therefore, more data is not necessarily better from a business perspective.

By changing some of the default parameters, we can also illustrate cases in which the variance in TD leads to worse treatment assignments than the bias in OM. More specifically, we set $\sigma_A^2 = 0.2$, $\sigma_B^2 = 2$, $\tilde{\alpha} = 0.1$ and $\tilde{\beta} = 0.1$. Setting a lower value for σ_A^2 (relative to $\sigma_B^2 = 2$) implies that the causal bias in OM does not vary as much from one individual to another, so the choice of the threshold for OM may correct (implicitly or explicitly) a substantial part of the boundary bias. Moreover, we set $\tilde{\alpha} = 0.1$ and $\tilde{\beta} = 0.1$ to make more noticeable how the performance of TD increases as the sampling error decreases.[16]

Notice that with these parameters, the bias in OM's Persuadable predictions remains quite large. For many individuals, more than half of the "effect" in OM's predictions may be attributed to the ignored counterfactual. Nevertheless, recall that here we are targeting only the top 5% of individuals ($\lambda = 5\%$ in Table 4), so the causal bias is partially corrected by the effectively higher probability threshold for OM. Therefore, this example also illustrates how large bias errors in causal-effect estimation do not necessarily imply large bias errors in causal classification. As Figure 3 shows, OM may outperform TD under this new regime if the sampling error in TD is large relative to the boundary bias in OM.

[16] If $\tilde{\alpha}$ and $\tilde{\beta}$ are extremely small, then the sampling error is also smaller because most outcomes are zero.

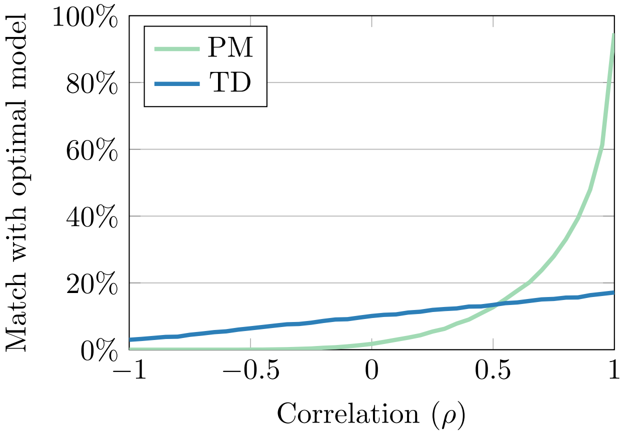

Figure 4: **Performance with correlated biases and effects**. When causal effects and outcomes are correlated (e.g., the ad is more likely to affect the people more likely to buy), the bias in outcome prediction (PM) improves targeting.

6.2 Correlated Outcomes and Effects

A setting with great practical relevance for causal classification is when there is a positive correlation between the outcome of interest and the treatment effect. In this case, the bias in OM may actually help. There are countless examples where this might happen: education may be more profitable for highly self-motivated individuals who would earn better wages even without going to college; an ad about a particular car brand is likely to be more effective on someone who is looking to buy a car; a fitness product may show better results for people who are into exercise, etc.

Prior research studies also show evidence that targeting based on outcome predictions can be effective. For example, in an advertising campaign of a major food chain, Stitelman et al. (2011) found that displaying ads based on the estimated likelihood of conversion increased the volume of prospective new customers by 40% more than displaying ads to the general population. Huang et al. (2015) deployed a churn prediction model in a telecommunications firm that increased the recharge rate of potential churners by more than 50% according to an A/B test. MacKenzie et al. (2013) estimated that 35% of what consumers purchase on Amazon and 75% of what they watch on Netflix come from product recommendations based on non-causal predictive models (as we understand it). Importantly, none of these targeting approaches modeled more than one potential outcome, so it would be wrong to interpret their estimates as causal effects. However, their estimates were useful to identify individuals positively affected by interventions.

To illustrate this type of setting, we set $\tilde{\alpha} = 0.1$ to show how substantial is the role of correlation when the outcome base rate is much larger than the treatment effect. We model different levels of correlation by changing the parameter ρ and show the results of comparing TD and PM in Figure 4. Recall that PM is the opposite of OL and consists of targeting based on the probability of a positive outcome when untreated. Given the very small causal effects, PM and OM are virtually equivalent in this setting, so we exclude OM from the comparison to avoid cluttering the plots.

Notice that different levels of correlation between the bias and the effect can dramatically change the decision of which approach to choose. In this case, the average treatment effect is quite small ($\sim 0.1\%$) compared to the base rate ($\sim 10\%$). As a result, given a finite number of observations to train models, it's much easier to discriminate Sure Things than to discriminate Persuadables. Therefore, if there is a high correlation between the probability of being a Sure Thing and the probability of being a Persuadable, we could predict who is a Persuadable much better by fitting a model for the Sure Things and then applying it to identify Persuadables. This is what PM is doing. TD, in contrast, destroys most of the signal when "correcting" for the bias. Importantly, this suggests that (in some settings) we may be able to make good targeting decisions using machine learning even if we have no data about how people actually behave when targeted!

7. Discussion and Limitations

These findings imply that we should choose fundamentally different approaches for causal classification depending on the setting. Specifically, even when causal targeting is the goal, it may be better to target with outcome prediction models when:

1. It is straightforward to choose a different threshold to partially correct the bias;
2. Probability estimates for one counterfactual suffer from large variance, and
3. Outcomes and treatment effects are correlated.

Thus, targeting based on outcome prediction models may be particularly suitable: when only a fixed number or percentage of the top individuals are targeted (e.g., due to a budget), so a threshold that corrects the bias is chosen implicitly; when there are limited training data for one of the counterfactuals (so causal-effect estimates have larger variance), and when the treatment often has an amplifying effect on the outcome, so there is a correlation between effects and outcomes. If this is the case, then outcome prediction can be preferable *even when* the causal modeling is not challenged by selection bias or other confounding. Appendix D shows a real-world online advertising application where outcome prediction indeed outperforms treatment effect estimation for causal classification. Therefore, in line with recent research that shows that accurate estimates of causal effects are not necessary to make good causal decisions (Fernández-Loría and Provost, 2022), we show that even non-causal predictions can potentially be useful for causal decision making.

This work is, to our knowledge, the first to compare outcome prediction and causal modeling systematically when targeting, and so the paper, of course, has limitations. For example, as does Friedman (1997), we assume normality in our theoretical analysis to gain intuition about the bias–variance tradeoff. However, because we assume a beta distribution for the scoring functions in our experiments, our simulations show that the qualitative results can hold even in the absence of normality. Our simulation analysis also assumed no confounding—to show that OM may be preferable even without confounding. However, our analytical results imply that confounding may not negatively affect causal classification to the same extent that it affects causal-effect estimation if the bias it confers is positively correlated with treatment effects. Fernández-Loría and Provost (2019) extend our theoretical derivations to analyze such settings.

Many other ideas presented in this paper could also be extended further. For example, the fact that predicting outcomes may be useful to predict optimal assignments could potentially be cast as a specific form of transfer learning, which focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. The outcome model can be seen as knowledge that can be transferred to aid in estimating a treatment assignment policy. Transfer learning already is common in the applications we have used as illustrative examples. For example, when building targeting models for online ads, often insufficient data are available on the true outcome of interest, so auxiliary data are brought in to transfer knowledge from a related learning task to the task at hand (Perlich et al., 2014; Dalessandro et al., 2014).

Alternatively, using OM rather than TD could be viewed as proxy modeling, which is based on identifying a suitable alternative (proxy) target variable when data on the true objective are in short supply (or even completely nonexistent). For example, in settings such as online advertising, often few data are available on the ultimate desired conversion (e.g., purchases). This can pose a challenge for accurate campaign optimization (due to low conversion rates, cold start periods, etc). Dalessandro et al. (2015) demonstrate the effectiveness of proxy modeling and find that predictive models built based on brand site visits do a remarkably good job of predicting which browsers will make a purchase. Our study provides conditions under which outcomes could be a good proxy for treatment effects.

Our study shows several conditions under which we argue it is preferable not to base targeting decisions on estimates of the treatment effect. However, most of these conditions depend on parameters that we typically do not know in non-simulated data, such as the distribution of the bias. Future empirical work testing these results should ideally use data from randomized experiments, as we do in Appendix D. Additionally, it would be interesting to characterize applications in which one approach should dominate the other, particularly because domain knowledge is crucial to decide whether the conditions discussed in this paper are likely to be met (e.g., correlation between outcomes and effects).

Regarding this last point, Radcliffe and Surry (2011) provide a list of general advice (based on their professional experience) about when to use uplift modeling; this advice is largely aligned with the theoretical and experimental results we have presented. Specifically, they advise that uplift modeling is worthwhile when there is a large volume of data for the control group and when outcomes are "anticorrelated" with the incremental impact of the marketing activity.[17] They also report that the benefits of uplift modeling are often smaller in retail environments because direct marketing activities usually have the most positive effect on high-spending customers. Thus, our study provides a strong theoretical foundation for previously published practical observations and guidelines.

Alternatively, one empirically driven approach to choose in practice between an uplift model and an outcome prediction model is to use data from an A/B test to compare both alternatives; we show a practical example of this in Appendix D. This approach may be particularly useful in settings where the modeler is still undecided on whether to invest in

[17] Note that our results imply that a strong negative correlation can also favor an outcome model over a causal-effect model for causal classification: if outcomes are negatively correlated with effects, then target the individuals with the smallest predicted outcomes.

additional data to improve the uplift model. The reason is that comparing an outcome prediction model with a preliminary uplift model typically requires a substantially lower data investment than what would be necessary to build a full, robust uplift model. For example, Radcliffe and Surry (2011) recommend that sample sizes should be at least ten times larger for modeling than for simple measurement of incremental response. Moreover, when modeling binary outcomes, they recommend that the product of the overall uplift and the size of each population should be at least 500. So, if the overall uplift is 0.1% (as in our simulation settings), this means that both the treated and the untreated group in the training data would need to be at least 500,000.

Our analysis also relies on the assumption that we can set a threshold to partially correct for the bias that results from ignoring the counterfactual. This correction is made automatically when targeting only the top-k or top-quantile individuals because a systematic upward bias implies using a higher probability threshold (as in our simulations). Thus, our simulation results show that outcome prediction models can be useful to identify the individuals with the largest treatment effects. It is important to note that this is different from identifying the individuals with a positive treatment effect. Critically, the latter would require estimating the magnitude of the bias, which could be a non-trivial challenge. In cases where this is necessary, one alternative is to conduct experiments to gather data to estimate δ_1 because it may be possible to estimate δ_1 with significantly less data than would be needed to build full-blown causal-effect models.

Another related limitation is that our study focuses on settings with binary outcomes and no Do-Not-Disturbs—those for whom the treatment causes a negative outcome. We make this assumption because it enables us to understand the phenomena better. However, the theoretical results in Section 4 can also be derived with continuous outcomes and Do-Not-Disturbs; what changes is our interpretation of the quantities in the equations. For example, we can no longer use the categorizations in Table 1 when using continuous outcomes, so α and β may only be interpreted as the expected outcome when untreated and the expected treatment effect, respectively. Similarly, if Do-Not-Disturbs are not assumed away, α and β can no longer be interpreted as the probability of being a Sure Thing and the probability of being a Persuadable, respectively, but the equations remain the same.

Nevertheless, from a practical standpoint, we should be careful not to presume that these limitations are negligible. Similar to analyses of regular classification under different practical scenarios, we would have to consider the relative (causal) misclassification costs because misclassifying Do-Not-Disturbs is generally more costly. Similarly, the cost of misclassifying a Sure Thing might be different from the cost of misclassifying a Lost Cause (e.g., when a retention incentive is offered). Radcliffe and Surry (2011) also mention that the presence of negative effects is a strong reason to consider uplift modeling, and that in the area of retention it is not uncommon for uplift modeling to deliver more value by identifying populations where negative effects are prevalent. The literature provides evidence both in favor (Huang et al., 2015) and against (Ascarza, 2018) the use of outcome prediction models for churn incentive targeting, so practitioners should proceed with caution when applying our results to settings where they suspect a significant fraction of individuals may be Do-Not-Disturbs. This paper presents a first study, and these limitations present directions for future research. Both the analytical framework and the simulation approach are general and can be adapted to support future studies.

Our study also assumes that treatment effects are estimated using two outcome prediction models, but more efficient estimation approaches exist (Dorie et al., 2019). Importantly, if the treatment effect estimation procedure is directly optimized for causal effect prediction, then it is not necessarily the case that an outcome prediction model will suffer from less variance than a causal-effect model (Nie and Wager, 2021; Foster and Syrgkanis, 2019). Nevertheless, our theoretical derivations do not require us to assume that causal-effect models will necessarily suffer from larger sampling variance. Furthermore, our analytical framework implies that a simple outcome prediction model can still beat these more efficient methods in cases where outcomes and effects are correlated (because then the causal bias in the outcome prediction model would generally be helpful for causal classification).

As mentioned in Section 4.8, another important topic for future research is to look deeper into what will happen when we are learning models from data—specifically, to examine the tradeoffs between different types of supervised learning errors, namely bias error, variance error, and irreducible error (Kohavi et al., 1996).[18] As mentioned above, biases that are consistent in both the treated classifier and the untreated classifier may be canceled out, so causal classification tasks are likely to be more sensitive to variance errors. Thus, it might be worthwhile to design causal classification algorithms that specifically focus on limiting the variance of the combined scores.

In targeted learning (Van der Laan and Rose, 2011), for example, the TD approach we analyze is considered a *non-targeted* semiparametric model for treatment effect estimation. The premise behind targeted learning is that causal estimation can be improved by applying targeted maximum likelihood estimation (TMLE) to create a *targeted* semiparametric model that optimizes the bias–variance tradeoff for the causal effect of interest (not for the outcome). These ideas have been discussed mostly in the context of estimating coarse statistical parameters (viz., average treatment effects). However, it would also be valuable to study their application in the context of causal classification. We hope that this paper will highlight the importance of designing new methods for causal classification.

Viewing causal classification in this way also clarifies issues with certain ways of modeling causal effects. Recall that in Section 4.7, we discussed the possibility of the treatment having an amplifying effect on the outcome, which may lead us to prefer OM. This assumption is implicit in certain models typically used for targeting! For instance, in a logistic regression model where the treatment is a covariate, people with a higher probability of a positive outcome when untreated will also have a larger treatment effect if the treatment has a positive coefficient, as long as the outcome base rate is low (as with applications like advertising and churn modeling). If one were to use such a logistic regression for causal classification, the model *itself* would assume that the treatment has an amplifying effect on the outcome when untreated, and thus that effects and outcomes are positively correlated.

8. Conclusions

This paper's main contribution is to reveal, analyze, and illustrate the *causal* bias–variance tradeoff when predicting the counterfactual to target treatments based on predictive models. Specifically, when the treatment effect estimation depends on two outcome predictions, then the larger sampling variance may lead to more misclassifications than the (causally

[18] Bias and variance errors are also known as approximation and estimation errors, respectively.

biased) outcome prediction approach. Additionally, because predicting outcomes is often easier than predicting effects, the bias may actually help if outcomes are correlated with treatment effects. The results of our theoretical analysis show that outcome prediction—ignoring the counterfactual—may be preferable when (1) probability threshold selection partially corrects for the bias that results from ignoring the counterfactual, (2) probability estimates for the ignored counterfactual are imprecise, and (3) outcomes and treatment effects are positively correlated. Simulations support the analytical results and illustrate common settings in which outcome prediction may be a better alternative than treatment effect estimation when making intervention decisions. Condition (1) is satisfied automatically for decision procedures that select the top-k (or top-quantile) highest-probability individuals, as is common when targeting under budget constraints or imprecision in costs and benefits (Provost and Fawcett, 2001).

These analytical results also explain the seemingly naive but common practice of targeting "treatments," such as online advertisements, based simply on outcome models rather than causal-effect models. Our simulations show that the problem of targeting display ads may fall into the regime where outcome targeting would be preferable to causal-effect targeting, and Appendix D shows an example with data from a real online advertising application where this is actually the case. This has important implications for practitioners because acquiring data to estimate counterfactuals is complicated and expensive.

Acknowledgments

We thank Brian Dalessandro, Rob Moakler, Claudia Perlich, Ori Stitelman, Huanci Wang, and the participants in seminars where we presented this work for many discussions motivating a deeper examination of causal targeting. Claudia pointed out the potential for increased variance in treatment effect estimation for targeting. We also thank David Blei, Charles Elkan and the anonymous reviewers for their valuable suggestions and feedback. Charles suggested using the amplifying treatment effect example to motivate the approach of targeting based on the probabilities of a positive outcome when untreated. Foster Provost thanks Andre Meyer and Ira Rennert for their support. We thank the NYU/Stern Fubon Center for supporting research on AI and data analytics for business.

Appendix A. Proof of Lemma 1

Lemma 1. *Let f be the average treatment effect given a specific set of feature values, and let $\hat{f}$ be the effect estimate provided by a scoring model. If $\hat{f}$ is normally distributed (conditional on the set of feature values), then the causal-classification model error for an individual with those feature values may be expressed as:*

$$Error\ (\hat{A}) = \tilde{\Phi}\left[sign(f-\tau)\frac{E[\hat{f}]-\tau}{\sqrt{Var[\hat{f}]}}\right](A^*\omega_1 + (1-A^*)\omega_2)$$

where A^* is the optimal classification for that individual, τ is the optimal threshold when making classifications with f, ω_1 is the cost of a false negative, ω_2 is the cost of a false *positive, and $\tilde{\Phi}$ is the upper tail area of the standard normal distribution:*

$$\tilde{\Phi}(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty \exp\left(-\frac{u^2}{2}\right) du$$

Proof. Equation 3 defines causal-classification model error as:

$$\begin{aligned} \text{Error } (\hat{A}) = & \ A^*\omega_1 P(\hat{A}=0) + (1-A^*)\omega_2 P(\hat{A}=1) \\ = & \ A^*\omega_1 \int_{-\infty}^{\tau} p(\hat{f}) d\hat{f} + (1-A^*)\omega_2 \int_{\tau}^{\infty} p(\hat{f}) d\hat{f} \end{aligned}$$

Let F be the cumulative distribution function (CDF) of $\hat{f}$. Then:

$$= A^*\omega_1 F(\tau) + (1-A^*)\omega_2(1-F(\tau))$$

Let $\theta = \frac{\tau - E[\hat{f}]}{\sqrt{Var[\hat{f}]}}$, and let $\tilde{\Phi}$ be the upper tail area of the standard normal distribution. Then, under the assumption that $\hat{f}$ follows a normal distribution:

$$\begin{aligned} &= A^*\omega_1(1-\tilde{\Phi}(\theta)) + (1-A^*)\omega_2\tilde{\Phi}(\theta) \\ &= A^*\omega_1\tilde{\Phi}(-\theta) + (1-A^*)\omega_2\tilde{\Phi}(\theta) \end{aligned}$$

Recall that $\text{sign}(f-\tau) = 1$ if $A^* = 1$ and $\text{sign}(f-\tau) = -1$ if $A^* = 0$. Then:

$$\begin{aligned} &= A^*\omega_1\tilde{\Phi}(-\text{sign}(f-\tau)\theta) + (1-A^*)\omega_2\tilde{\Phi}(-\text{sign}(f-\tau)\theta) \\ &= \tilde{\Phi}(-\text{sign}(f-\tau)\theta)(A^*\omega_1 + (1-A^*)\omega_2) \\ &= \tilde{\Phi}\left[\text{sign}(f-\tau)\frac{E[\hat{f}]-\tau}{\sqrt{Var[\hat{f}]}}\right](A^*\omega_1 + (1-A^*)\omega_2) \end{aligned}$$

■

Appendix B. Proof of Theorem 2

Theorem 2. *Let f be the average treatment effect given a specific set of feature values, and let $\hat{f}_1$ and $\hat{f}_2$ be the effect estimates provided by two scoring models. If $\hat{f}_1$ and $\hat{f}_2$ are normally distributed (conditional on the set of feature values), then $\hat{f}_1$ leads to lower causal-classification model error than $\hat{f}_2$ for individuals with those feature values if:*

$$\frac{b_1}{m} < 1 + \frac{\frac{b_2}{m} - 1}{\sqrt{\gamma}},$$

where $b_1 = f - E[\hat{f}_1]$ is the (negation of the) bias in $\hat{f}_1$, $b_2 = f - E[\hat{f}_2]$ is the (negation of the) bias in $\hat{f}_2$, $\gamma = \frac{Var[\hat{f}_2]}{Var[\hat{f}_1]}$ is the ratio of the variances of $\hat{f}_2$ and $\hat{f}_1$, and $m = f - \tau$ is the causal margin—the distance between the conditional average treatment effect and the optimal decision boundary (or threshold) when f is used to make classifications.

Proof. Let $\hat{A}_1$ and $\hat{A}_2$ be the causal classifications made by $\hat{f}_1$ and $\hat{f}_2$ when τ is the decision boundary or threshold for both scoring functions:

$$\hat{A}_1 = \mathbf{1}(\hat{f}_1 \geq \tau)$$
$$\hat{A}_2 = \mathbf{1}(\hat{f}_2 \geq \tau)$$

Then, $\hat{f}_1$ leads to lower causal-classification model error than $\hat{f}_2$ if:

$$\text{Error }(\hat{A}_1) < \text{Error }(\hat{A}_2)$$

Assume that $\hat{f}_1$ and $\hat{f}_2$ each follow a normal distribution, and let $\omega = A^*\omega_1 + (1 - A^*)\omega_2$. Then, according to Lemma 1, the inequality above is equivalent to:

$$\tilde{\Phi}\left[\text{sign}(f-\tau)\frac{E[\hat{f}_1]-\tau}{\sqrt{Var[\hat{f}_1]}}\right]\omega < \tilde{\Phi}\left[\text{sign}(f-\tau)\frac{E[\hat{f}_2]-\tau}{\sqrt{Var[\hat{f}_2]}}\right]\omega$$
$$\text{sign}(f-\tau)\frac{E[\hat{f}_1]-\tau}{\sqrt{Var[\hat{f}_1]}} > \text{sign}(f-\tau)\frac{E[\hat{f}_2]-\tau}{\sqrt{Var[\hat{f}_2]}}$$

Let $\gamma = \frac{Var[\hat{f}_2]}{Var[\hat{f}_1]}$. Then:

$$\text{sign}(f-\tau)\frac{E[\hat{f}_1]-\tau}{\sqrt{Var[\hat{f}_1]}} > \text{sign}(f-\tau)\frac{E[\hat{f}_2]-\tau}{\sqrt{\gamma Var[\hat{f}_1]}}$$
$$\text{sign}(f-\tau)\sqrt{\gamma}(E[\hat{f}_1]-\tau) > \text{sign}(f-\tau)(E[\hat{f}_2]-\tau)$$

Let $b_1 = f - E[\hat{f}_1]$ and $b_2 = f - E[\hat{f}_2]$. Then:

$$\text{sign}(f-\tau)\sqrt{\gamma}(f-\tau-b_1) > \text{sign}(f-\tau)(f-\tau-b_2)$$

Let $m = f - \tau$. Then:

$$\text{sign}(m)\sqrt{\gamma}(m-b_1) > \text{sign}(m)(m-b_2)$$
$$\sqrt{\gamma}\frac{m-b_1}{m} > \frac{m-b_2}{m}$$
$$1-\frac{b_1}{m} > \frac{1-\frac{b_2}{m}}{\sqrt{\gamma}}$$

Therefore, $\hat{f}_1$ leads to lower causal-classification model error than $\hat{f}_2$ if:

$$\frac{b_1}{m} < 1 + \frac{\frac{b_2}{m}-1}{\sqrt{\gamma}}$$

■

Appendix C. Simulator Code

We present here the Python code we used to generate the data in Section 6.

```
import numpy as np

def eval_models(alpha=0.001, beta=0.001, var_a=1, var_b=1, corr=0,
                qual_u=1000, qual_t=1000, target=0.05, N=1000000):
    # Set means and covariance for the multivariate normal
    e_b = beta * (1+alpha/(1-alpha))/(1-beta*(1+alpha/(1-alpha)))
    mean_b = np.log(e_b)
    mean_a = np.log(alpha*(1+e_b)/(1-alpha))
    cov = corr * np.sqrt(var_a) * np.sqrt(var_b)
    # Generate probabilities for Sure Things and Persuadables
    means = [mean_a, mean_b]
    cov_mat = [[var_a, cov], [cov, var_b]]
    exps = np.exp(np.random.multivariate_normal(means, cov_mat, N))
    probs_a = exps[:, 0] / (exps.sum(axis=1) + 1.0)
    probs_b = exps[:, 1] / (exps.sum(axis=1) + 1.0)
    # Probabilities for the treated and untreated
    y_t = probs_a + probs_b
    y_u = probs_a
    # Generate probability estimates
    scores_u = np.random.beta(y_u * qual_u, (1-y_u) * qual_u)
    scores_t = np.random.beta(y_t * qual_t, (1-y_t) * qual_t)
    # Generate rankings
    opt_ranking = probs_b.argsort().argsort()
    om_ranking = (scores_t).argsort().argsort()
    td_ranking = (scores_t - scores_u).argsort().argsort()
    # Evaluate what percent matches the top in the optimal
    rank_threshold = N*(1-target)
    om_chosen = opt_ranking[om_ranking >= rank_threshold]
    td_chosen = opt_ranking[td_ranking >= rank_threshold]
    om_match = (om_chosen >= rank_threshold).mean()
    td_match = (td_chosen >= rank_threshold).mean()
    # Return results
    return [om_match, td_match]
```

Appendix D. Practical Example of Choosing Between Treatment Effect Estimation and Outcome Prediction

We present here an example to illustrate how to choose between the approaches discussed in Section 4 (e.g., OM, TD) in practice. We use data made available by Criteo (an advertising platform) based on randomly targeting advertising to a large sample of users (Diemert

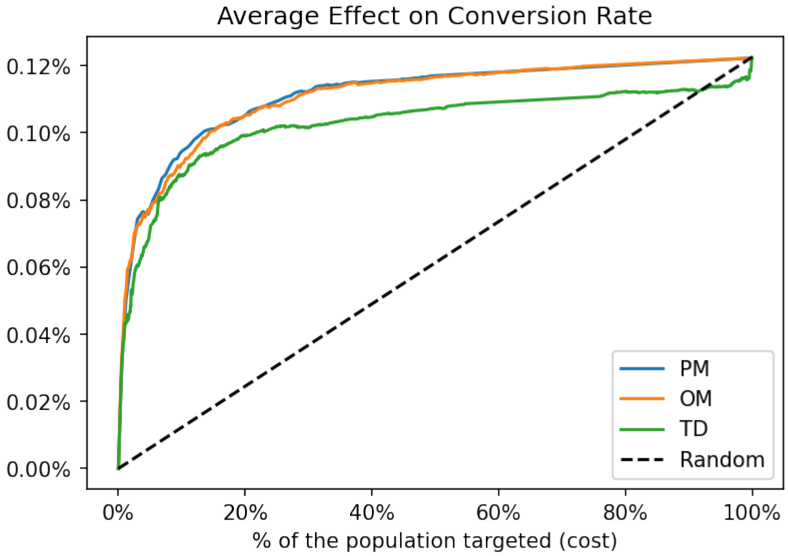

Figure 5: **Increase in conversions produced by non-causal approaches (OM, PM) and a causal approach (TD)**. The non-causal approaches in this example can produce at least as much benefit as the causal approach for any given cost.

Eustache, Betlei Artem et al., 2018).[19] Conversions are the outcome of interest. The data include 13,979,592 instances, each representing a user with 11 features, a treatment indicator for the ad, and the label (i.e., whether the user converted or not). The treatment rate is 85%, the average conversion rate when untreated is 0.19%, and the average treatment effect (ATE) is 0.12%.

We consider three different targeting approaches: PM (discussed in Section 4.7), OM, and TD. All the approaches were implemented using decision tree models. The models were trained and tuned with cross-validation using 80% of the sample (the training set). The targeting approaches were evaluated using the remaining 20% of the sample (the test set).

We evaluated the targeting approaches by using them to score individuals and plotting the average effect on the conversion rate as a function of the percentage of the individuals with the largest scores who are targeted, as shown in Figure 5; this is known as an uplift curve. Given some targeting approach, the average effect on the conversion rate is:

$$(E[Y|T=1,D=1]-E[Y|T=0,D=1])\times P[D=1], \tag{9}$$

where D is a binary variable that indicates whether the user is targeted, T is a binary variable that indicates whether the user was treated in the data, and Y is a binary variable that indicates whether the user converted. The test set is used to estimate Equation 9 for each targeting approach.

Figure 5 shows that the non-causal approaches, PM and OM, are indeed better than TD at identifying the users for whom the ad is most effective. In theory, TD should eventually

[19] See `https://ailab.criteo.com/criteo-uplift-prediction-dataset/` for details and access to the data. We use the version of the data set without leakage.

outperform the other approaches with more data. However, collecting additional training data is costly because it requires randomly targeting and withholding treatments, which is substantially worse than making targeting decisions based on OM or PM. Therefore, a decision-maker analyzing these results may conclude that investing in additional data to build a better uplift (TD) model is not worthwhile.

References

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Eva Ascarza. Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98, 2018.

Eva Ascarza, Scott A. Neslin, Oded Netzer, Zachery Anderson, Peter S. Fader, Sunil Gupta, Bruce Hardie, Aurélie Lemmens, Barak Libai, David Neal, Foster Provost, and Rom Schrift. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*, 5(1):65–81, 2018.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 129–138, 2009.

Debopam Bhattacharya and Pascaline Dupas. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196, 2012.

Brian Dalessandro, Daizhuo Chen, Troy Raeder, Claudia Perlich, Melinda Han Williams, and Foster Provost. Scalable hands-free transfer learning for online advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1573–1582, 2014.

Brian Dalessandro, Rod Hook, Claudia Perlich, and Foster Provost. Evaluating and optimizing online advertising: Forget the click, but there are good proxies. *Big data*, 3(2): 90–102, 2015.

Rajeev H. Dehejia. Program evaluation as a decision problem. *Journal of Econometrics*, 125(1-2):141–173, 2005.

Diemert Eustache, Betlei Artem, Christophe Renaudin, and Amini Massih-Reza. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London,United Kingdom, August, 20, 2018*. ACM, 2018.

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

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, pages 1097–1104, 2011.

Carlos Fernández-Loría and Foster Provost. Observational vs experimental data when making automated decisions using machine learning. *Available at SSRN: https://ssrn.com/abstract=3444678*, 2019.

Carlos Fernández-Loría and Foster Provost. Causal decision making and causal effect estimation are not the same... and why it matters. *INFORMS Journal on Data Science*, 2022.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Jerome H. Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

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

Keisuke Hirano and Jack R. Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701, 2009.

Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, and Jia Zeng. Telco churn prediction with big data. In *International Conference on Management of Data*, pages 607–618. ACM, 2015.

Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, 2012.

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

Ron Kohavi, David H Wolpert, et al. Bias plus variance decomposition for zero-one loss functions. In *Internacional Conference on Machine Learning*, volume 96, pages 275–83, 1996.

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

Aurélie Lemmens and Sunil Gupta. Managing churn to maximize profits. *Marketing Science*, 39(5):956–973, 2020.

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

Ian MacKenzie, Chris Meyer, and Steve Noble. How retailers can keep up with consumers. *McKinsey & Company*, 2013.

Charles F. Manski. Monotone treatment response. *Econometrica*, 65(6):1311–1334, 1997.

Charles F. Manski. Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246, 2004.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

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

Judea Pearl. *Causality*. Cambridge university press, 2009.

Claudia Perlich, Brian Dalessandro, Troy Raeder, Ori Stitelman, and Foster Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95(1):103–127, 2014.

Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine learning*, 42(3):203–231, 2001.

Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.", 2013.

Nicholas J. Radcliffe and Patrick D. Surry. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, 2011.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 00220663.

Galit Shmueli. To Explain or to Predict? *Statistical Science*, 25(3):289–310, 2011. ISSN 0883-4237.

Ori Stitelman, Brian Dalessandro, Claudia Perlich, and Foster Provost. Estimating the effect of online display advertising on browser conversion. *Data Mining and Audience Intelligence for Advertising (ADKDD 2011)*, 8, 2011.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

Mark J. Van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*, volume 4. Springer, 2011.

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

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

Inbal Yahav, Galit Shmueli, and Deepa Mani. A tree-based approach for addressing self-selection in impact studies with big data. *MIS Quarterly*, 40(4):819–848, 2016.