Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation*

Günter J. Hitsch

University of Chicago Booth School of Business

Sanjog Misra

University of Chicago Booth School of Business

January 2018

Abstract

We discuss how to construct optimal targeting policies and document the difference in profits from alternative targeting policies by using estimation approaches that are based on recent advances in causal inference and machine learning. We introduce an approach to evaluate the profit of *any* targeting policy using only one single randomized sample. This approach is qualitatively equivalent to conducting a field test, but reduces the cost of multiple field tests because all comparisons can be conducted in only one sample. The approach allows us to compare many alternative optimal targeting policies that are constructed based on different estimates of the conditional average treatment effect, i.e. the incremental effect of targeting. We draw a conceptual distinction between methods that predict the conditional average treatment effect *indirectly* via the conditional expectation function trained on the outcome level, and methods that *directly* predict the conditional average treatment effect. We propose a new direct estimation method, called *treatment effect projection*. The empirical application is to a catalog mailing with a high-dimensional set of customer features. We find that the optimal targeting policies based on the direct estimation methods typically outperform the indirect estimation methods, both in the validation sets from the same population from which the training set is drawn and in the data obtained one year after the training set was collected. In particular, the treatment effect projection performs similar to the recently introduced causal forest of Wager and Athey (2017). We also compare targeting policies based on conditional average treatment effects with a sophisticated application of the traditional CRM approach that is based on a prediction of the outcome level. Even though based on a conceptually incorrect metric—outcome levels—the sophisticated application of the traditional CRM approach often yields larger profits than the targeting policies based on the indirect estimation methods.

Keywords. Targeting, customer relationship management (CRM), causal inference, machine learning, field experiments

*We thank Thomas Otter and Bernd Skiera for helpful comments and suggestions. We also benefitted from the comments of seminar participants at the Goethe University Frankfurt and the 2017 Marketing Science Conference. We are particularly grateful to Walter Zhang for his superb research assistance, and to the data science team at the company that is the source of our data for their help and numerous insights. All correspondence may be addressed to the authors by e-mail at guenter.hitsch@chicagobooth.edu or sanjog.misra@chicagobooth.edu.

1 Introduction

Despite the fundamental importance of targeting in marketing and the recent industry focus on targeting policies based on big data and machine learning methods, the empirical evidence that documents the profitability differences of alternative targeting policies is sparse. We conjecture that the high cost of evaluating multiple policies is a main reason for this paucity. The conventional approach in the extant literature is to conduct multiple field experiments to compare proposed targeting policies (Simester et al. 2016). Conducting such tests entails the direct cost of the field implementation and the opportunity cost of not treating all customers with the optimal targeting effort. The overall cost of testing increases in the number of policies to be compared, and the number of policies that can be evaluated is limited by the size of the customer population available to a firm relative to the desired sample size in each test.

This paper introduces an approach to evaluate the profit of *any* targeting policy using only one single randomized sample. This approach allows us to compare arbitrarily many different targeting policies without incurring the cost of an equal number of large-scale field experiments. The approach predicts the total profit of a policy based on the *usable* observations where the proposed and realized targeting assignments agree, and it employs inverse probability weighting to account for the rate at which observations are not usable because the proposed and realized targeting assignments differ. We use this approach to conduct the following evaluations: (i) We provide an exhaustive comparison of optimal targeting policies that we construct using the estimated incremental effect of targeting based on estimation methods that combine ideas from the causal inference and machine learning literatures. A key finding is that estimation methods that are trained to *directly* predict the incremental effect of targeting yield larger profits than conventional methods that *indirectly* predict the incremental effect based on the conditional expectation function that is trained on the outcome *level*. (ii) We evaluate the predicted profits based on a new estimation method, called *treatment effect projection*, that we propose in this paper. The treatment effect projection is trained to directly predict the incremental effect, and in our empirical application it performs similarly to the recently introduced causal forest of Wager and Athey (2017). (iii) We compare the targeting policies based on the estimated incremental effect of targeting to a sophisticated application of traditional CRM techniques that are widely used in industry. This more traditional approach *scores* customers based on the predicted outcome level (e.g. spending), which, for the purpose of developing optimal targeting policies, is a conceptually incorrect metric. The application of the traditional approach, as used by the firm in our empirical application, ranks customers according to the predicted score and then finds the optimal percentage of customers to be targeted using a randomized training sample. We find that this conceptually incorrect approach yields larger profits than most of the policies based on incremental effects that are predicted by the indirect estimation methods. However, policies based on the direct estimation methods using recent developments in the causal inference and machine learning literatures outperform the sophisticated application of the traditional CRM approach.

The overarching goal of this paper is to construct and compare *optimal targeting policies*.

An optimal targeting policy satisfies two principles. First, focusing on the binary case where a customer is targeted or not, a customer is targeted if and only if the incremental value of targeting—the expected difference in the profit contribution between being targeted and not being targeted—exceeds the targeting cost.[1] Second, we recognize that in general customers differ in their purchasing behavior and in their response to a targeting effort. Therefore, the prediction of the incremental value of targeting should allow for customer heterogeneity and thus be based on the observed customer features.

The principle of targeting based on the incremental effect of a marketing action, not on the response level, is widely understood in some areas of modern marketing research, such as digital marketing (e.g. Goldfarb and Tucker 2011, Lambrecht and Tucker 2013, Lewis and Reiley 2014, Blake et al. 2015, Narayanan and Kalyanam 2015, Sahni 2015). More generally, much of the literature on demand estimation in marketing and economics during the last two decades has been concerned with the problem of estimating the causal effect of a marketing action, such as pricing or advertising, in the presence of endogeneity (see, for example, Ackerberg et al. 2007 and Rossi 2014 and the references therein). In contrast, the principle of targeting based on incremental effects has had comparatively little impact on the extant CRM literature. In the *traditional CRM approach* (our terminology), a customer is targeted if the predicted response *level* (e.g. expected spending) exceeds a threshold. For example, Blattberg et al. (2008) provide an in-depth discussion of the response level-based approach (in particular, see Chapter 10, "The Predictive Modeling Process"), whereas a discussion of incrementality is absent. Similarly, the recent survey by Ascarza et al. (2017) presents research that is largely based on the traditional CRM approach, with the exceptions of Gönül et al. (2000) and Ascarza et al. (2016b). The recent papers by Ascarza (2017), Simester et al. (2016), and Zantedeschi et al. (2017) that study incrementality-based targeting are closely related to our work. We will discuss the relationship between these papers and our research below. Also, targeting based on incremental effects is rarely employed in the current industry practice. This is our subjective assessment based on anecdotal evidence and interactions with companies that engage in CRM. We believe that the traditional CRM approach has remained dominant due to the conceptual challenges of causal inference, and in particular due to the perception that the prediction of incremental effects is too difficult or may not improve profits compared to a traditional CRM approach. Correspondingly, a key goal of this work is to show that targeting based on incremental effects is feasible in practice, and that such policies can substantially improve profits compared to targeting using response levels, especially if the prediction of the incremental effects employs recent methodological innovations from the causal inference and machine learning literatures.

Following the literature on causal inference (Imbens and Rubin 2015), we formally define the incremental effect of targeting as the conditional average treatment effect (CATE), i.e. the average causal effect of targeting for a sub-population of customers with identical features. The methods in this paper are most applicable if a firm conducts a field experiment and randomizes

[1] In a continuous treatment setting, the targeting policy is based on equating the marginal effect of the treatment to its marginal cost. This paper focuses on the binary setting, but generalizations to other settings are conceptually straightforward.

the targeting assignment. In a randomized sample the conditional average treatment effect is identified unless there are social effects resulting from the targeting effort, which are unlikely in the applications that we consider.[2] Identification means that in principle the conditional average treatment effect can be recovered from the data with an infinite population. However, identification tells us nothing about the performance of different estimation methods in a finite sample. One contribution of this work is a comparison of the performance of different estimation approaches in an actual targeting application. We employ two large, fully randomized samples that we obtained through a collaboration with a company that targets customers using catalogs. The first sample contains 293 thousand observations from a targeting campaign in 2015, and the second sample contains 148 thousand observations from a similar targeting campaign in 2016. In both samples we observe a high-dimensional set of customer features, the randomized treatment assignment, and customer spending.

We find that some key ideas emphasized in the machine learning literature, in particular regularization and non-parametric estimation, affect the performance of the different estimators. However, the main distinction that we draw between the methods concerns the approach to predict the conditional average treatment effect. We distinguish between two classes of estimators. The first class includes estimators for the regression function that are trained using a squared-error loss based on the difference between the observed and predicted outcome *levels*. These estimators provide an *indirect* prediction of the conditional average treatment effect from the predicted difference in outcome levels between treated and untreated units with identical features. The second class includes estimators that are *directly* trained to predict the conditional average treatment effect. Such methods appear to be infeasible due to the fundamental problem of causal inference—we never observe the individual treatment effect, but only the outcome corresponding to the assigned treatment. However, the recently developed causal forest and treatment effect projection (TEP) estimators can solve this problem. The causal forest by Wager and Athey (2017) is based on an ensemble of causal trees that are constructed using a splitting technique that mimics the approach that would be used if the individual treatment effects were observed. The treatment effect projection (TEP) is a new method introduced in this paper that is based on a different idea than the causal forest. In a first stage, the treatment effect projection predicts the conditional average treatment effect based on a causal KNN regression using the K nearest treated and untreated neighbors for a given unit. In this stage, the novel idea is to select the tuning parameter K to minimize the out-of-sample *transformed outcome loss*, a loss function that can be computed from the observed data. Choosing K to minimize the expected transformed outcome loss is equivalent to minimizing the expected mean squared error loss based on the difference between the KNN prediction of the treatment effect and the actual yet unobserved individual treatment effect. Hence, this method of selecting K directly links the KNN prediction to the conditional average treatment effect. In a second stage, we project the KNN predictions of the conditional average treatment effect onto the feature variables to allow for some regularization. In this stage, the choice of the regression method is entirely flexible

[2]Formally, the conditional average treatment effect is identified if three conditions, unconfoundedness, overlap, and the stable unit treatment value assumption (SUTVA) are satisfied.

and can be adapted to the data. For example, the Lasso is suitable if the conditional average treatment effect function is linear in the features or a basis expansion of the features, and a random forest is suitable if we have no prior information on the functional form of the regression function.

We compare all estimation methods out-of-sample either using cross-validation or based on 1000 bootstrap test sets. All methods predict a substantial degree of heterogeneity in the estimated conditional average treatment effects. Because the individual treatment effects are unobserved, we cannot directly calculate a measure of model fit. However, we can assess model fit using the transformed outcome loss, which in expectation is equivalent to ranking the different estimators using the mean squared difference between the predicted and actual, unobserved treatment effects. We also compare the estimation methods by examining the percentage of negative predicted treatment effects, with the assumption being that in our application a targeting effort does not cause a customer to purchase less, and using lift factors, which are widely used for visual model validation in CRM. We find that in general, the direct estimation methods have better model fit than the indirect estimation methods of the conditional average treatment effect.

Ultimately, profits, not model fit, are the criterion that determines the value of a targeting method. Hence, we construct the optimal targeting policy for each of the estimation methods of the conditional average treatment effect and we compare the corresponding distribution of expected profits using our proposed inverse probability-weighted profit estimator. The intuition behind the inverse probability-weighted estimator is as follows. A given targeting policy cannot be directly evaluated in a randomized sample, because for many units (customers) the proposed treatment assignment by the targeting policy will differ from the realized treatment assignment. However, for units where the proposed and realized treatment assignment agree, we can scale the realized profit contribution by the inverse of the propensity score to account for the percentage of similar units that are "missing" in the sense that the proposed and realized treatment assignment are different. Thus, the sum of the weighted profit contributions for units where the proposed and realized treatment assignment agree is an *unbiased estimate* of the total profit from the targeting policy. Inverse probability-weighted estimators are known in statistics (Cao et al. 2009) and have been proposed for policy evaluation, but, to the best of our knowledge, have not been used for the analysis of targeting policies or any other form of policy evaluation in marketing before.

We find that, on average, the expected profits are larger if an optimal targeting policy is predicted using a direct estimation method for the conditional average treatment effect compared to an indirect estimation method. The difference in profits is statistically significant and represents an important, economically significant increase in the return of the targeting policies. We then compare the incrementality-based targeting policies to a sophisticated application of the traditional CRM approach using a prediction of outcome (spending) levels. This "sophisticated" approach is inspired by the targeting process employed by the firm that is the source of our data. Using this approach, we first predict the expected customer-level spending level. Then, we rank customers according to expected spending, and we predict the profit from targeting a given percentage of all customers according to expected spending. The profit prediction is based

on the inverse probability-weighted estimator of targeting profits, and thus provides a correct, unbiased estimate of the targeting profit despite using a conceptually incorrect and potentially sub-optimal targeting policy. Since the method can be employed to predict the profit for any percentage of customers to be targeted, we can also predict the optimal, profit-maximizing targeting percentage. This sophisticated application of traditional CRM improves profits over a policy that either targets all or no customers, and it also yields larger profits than most of the indirect estimation methods. However, the targeting policies predicted based on the direct estimation methods yield profits that are substantially larger than those from the traditional CRM approach. This comparison illustrates that the sophisticated application of traditional targeting techniques can work under the right conditions, in particular if the outcome level and treatment effects are correlated as in our application. However, we also show that the traditional approach, even if applied in a sophisticated manner, is inferior compared to a targeting policy based on some of the best estimation methods that directly predict the incremental effect of targeting.

All our comparisons of model fit and targeting profits demonstrate the external validity of the proposed methods. Initially, we compare the methods in test sets drawn from the same 2015 population that is also used to obtain the training sets. Thus, we establish the external validity of the different methods while holding the data-generating process fixed. We then provide a stricter test of the external validity or transportability of the results when we estimate our models in the 2015 data and predict in the 2016 data, which was collected one year after the initial targeting campaign. The latter approach corresponds to how the models are built and then implemented in practice. Although there are quantitative differences across the 2015 and 2016 comparisons, all our main results, especially the better predictive power of the direct estimation methods, are robust.

Finally, in addition to the individual contributions of this work that we highlighted, we intend the structure of our paper to provide a blueprint for an easily implementable approach that provides guidelines on how to systematically choose and compare different estimation methods which can be used to predict an optimal targeting policy. Above and beyond targeting, this blueprint should be generally of value for applications of policy evaluation.

2 Relationship to the literature

As already mentioned in the introduction, recent work by Ascarza (2017) and Simester et al. (2016) is closely related to this paper. Ascarza (2017) studies churn management and emphasizes the difference between the churn probability (the outcome level) and the incremental effect on churn of a targeted intervention. Indeed, in one of her applications the churn probability appears to be negatively correlated with the treatment effect, i.e. the reduction in the churn probability due to a targeting effort. In this situation, targeting customers in order of the predicted churn probability, as is typically done in the industry, will be counterproductive.

Simester et al. (2016) evaluate different estimation methods to predict the treatment effect of two targeted customer acquisition efforts. The methods are estimated using a training set and

then tested using large-scale field experiments that separately implement the targeting policies corresponding to each of the estimation methods.

Unlike this paper, Simester et al. (2016) do not draw the distinction between indirect and direct estimation methods to estimate conditional average treatment effects. Their best-performing method is a linear model estimated using a Lasso, which, in our application, yields substantially lower profits compared to the causal forest and treatment effect projections that directly predict the treatment effects.[3] Ascarza (2017) estimates the treatment effects using the uplift random forest of Guelman et al. (2015), which is based on the uplift trees of Rzepakowski and Jaroszewicz (2012). The uplift random forest is a precursor of the causal forest of Wager and Athey (2017) and an example of a direct estimation method. Ascarza does not compare different estimation methods for the conditional average treatment effects. Furthermore, the uplift tree and random forest is only developed for the case of binary outcomes, and thus is not applicable to data with continuous outcomes as in our work. Most similar to the comparison between indirect and direct estimation methods in this paper is Guelman et al. (2015), who compare the "subtraction of two models" to "modeling uplift directly." In their empirical application to churn management, they compare the incremental gains curve and the "Qini coefficient" for the estimated models. However, as standard errors or confidence intervals are not provided, it is unclear if the two measures of fit are statistically different. More importantly, the paper neither predicts nor compares the profits implied by the estimated models.

Our work is also related to Zantedeschi et al. (2017), who estimate the treatment and carryover effects of catalog and e-mail targeting from a sequence of randomized targeting campaigns. Their model allows for *unobserved heterogeneity* using a Bayesian hierarchical model, whereas all estimation methods in this paper are based on observed heterogeneity.

The inverse probability-weighted estimator that we use to predict expected targeting profits has, to the best of our knowledge, not been used in marketing before. Inverse probability-weighted estimators have been known in statistics (especially biostatistics) for a long time. See Horvitz and Thompson (1952) for the origins and, for example, Cao et al. (2009) and the references therein. See Kitagawa and Tetenov (2017) for a recent application to policy evaluation.

3 Targeting policy evaluation

We cast our discussion in the framework of the Neyman-Rubin potential outcomes model (see Imbens and Rubin 2015 for a comprehensive introduction). A company interacts with a population of customers that are indexed by i. Each customer i receives a treatment, $W_i \in \{0, 1\}$. In the applications that we consider, $W_i = 1$ if the customer is targeted and $W_i = 0$ otherwise. The treatment may represent a targeted price, a display ad, or—as in our empirical application—a direct mail marketing effort such as a catalog.

[3] Note that the Lasso in Simester et al. (2016) is fitted separately for the treated and untreated units, whereas our Lasso includes interactions with the treatment indicator. The predicted treatment effects from these two models will not generally be identical.

$Y_i(0)$ and $Y_i(1)$ are the potential outcomes corresponding to either treatment. In the application in this paper, $Y_i(0)$ is the dollar amount spent by customer i during an observation period if the customer is not targeted, whereas $Y_i(1)$ is the amount spent if the customer is targeted. Alternatively, $Y_i(0)$ and $Y_1(1)$ may indicate a purchase, click on a search link, account cancellation, etc. Throughout this paper we assume that the potential outcomes are affected only by the treatment assignment for customer i, but not by the treatment assignment for any other customer in the population. This assumption rules out social effects from a targeting campaign, and is called the stable unit treatment value assumption (SUTVA) in the literature on causal inference.

We also observe a potentially high-dimensional vector of features for each customer, $X_i \in \mathbb{X}$. X_i captures observed customer attributes, such as the number and type of past purchases, visits to a company's webpage, and demographics.

Based on these variables we define the (potential) profit contribution depending on the targeting status:

$$\pi_i(W_i) = \begin{cases} mY_i(0) & \text{if } W_i = 0, \\ mY_i(1) - c & \text{if } W_i = 1. \end{cases}$$

Here, m is the profit margin and c is the cost of the targeting effort.[4] $\pi_i(W_i)$ is the profit that accrues from customer i given the targeting status $W_i \in \{0, 1\}$.

A targeting policy is a function $d : \mathbb{X} \to \{0, 1\}$ that indicates if a customer with attributes X_i should be targeted, $d(X_i) = 1$, or not, $d(X_i) = 0$. The observed profit of a targeting policy $d : \mathbb{X} \to \{0, 1\}$ implemented in the field on a sample of N customers with features $X_1, \ldots, X_N$ is given by

$$\Pi(d) = \sum_{i=1}^{N} \left((1 - d(X_i)) \cdot \pi_i(0) + d(X_i) \cdot \pi_i(1) \right). \tag{1}$$

The *expected* profit from a targeting policy d that is implemented in a sample of customers with observed features $X_1, \ldots X_N$ is

$$\mathbb{E}[\Pi(d)|X_1, \ldots, X_N] = \sum_{i=1}^{N} \mathbb{E}\left[(1 - d(X_i)) \cdot \pi_i(0) + d(X_i) \cdot \pi_i(1) | X_i \right]. \tag{2}$$

The expectation in (2) is conditional on the customer features, which are observed by a firm before implementing a targeting policy. In contrast, $\mathbb{E}[\Pi(d)] = \mathbb{E}_X[\mathbb{E}[\Pi(d)|X_1, \ldots, X_N]]$ is the unconditional expectation of the profit from targeting policy d, taken over the population distribution of X_i. $\mathbb{E}[\Pi(d)]$ is the expected profit before the identity of the N customers in the field implementation of the policy d is known.

Our goal it to evaluate different targeting policies based on the level of expected profits, $\mathbb{E}[\Pi(d)]$. Testing for the difference in the effectiveness of two targeting policies, d_A and d_B, is conceptually straightforward: We randomly select two sets of N customers each from the

[4] A generalization to the case of heterogeneous profit margins and targeting costs is straightforward.

population of customers that can be targeted, and we implement policy d_A in one and policy d_B in the other set. Although easy to implement in principle, in practice this approach will be costly if the goal is to compare many different targeting policies. This cost includes not only the cost of implementing each field test but also the opportunity cost of imposing a sub-optimal targeting policy. Furthermore, the number of policies that can be evaluated using such field tests is limited by the total number of customers that a firm can target relative to the size of each field test.

We propose an alternative approach that allows us to compare arbitrarily many targeting policies using only one fixed, fully randomized sample of customers. To obtain this sample, we first randomly select N customers, and then independently target each customer i with probability $\Pr\{W_i = 1|X_i = x\} = e(x)$, $0 < e(x) < 1$. It may appear that it is impossible to evaluate the profit from an arbitrary targeting policy in such a randomized sample. In particular, the randomly assigned treatment will not generally coincide with the proposed targeting policy, $W_i \neq d(X_i)$, and hence we cannot calculate the realized profit in equation (1) from the randomized sample. However, we can calculate the realized profits using only the *usable observations* i, where the actually assigned treatment and the proposed policy agree, $W_i = d(X_i)$. To account for the rate at which any observation i is not usable in the calculation of profits, we use weights based on the inverse of the probability of being usable, $\Pr\{W_i = d(X_i)|X_i\}$. The resulting inverse probability weighted profit estimator is defined as follows:

$$\hat{\Pi}(d) = \sum_{i=1}^{N} \left(\frac{1 - W_i}{1 - e(X_i)} (1 - d(X_i)) \cdot \pi_i(0) + \frac{W_i}{e(X_i)} d(X_i) \cdot \pi_i(1) \right). \tag{3}$$

The quantities required to construct this estimator include the propensity score $e(X_i)$ and the targeting policy d, which are chosen by a firm or an analyst, as well as data on the realized treatment W_i and the individual profit outcomes $\pi_i(w)$ that correspond to the realized treatment, $W_i = w$. As such, all elements in (3) estimator are known or observed. Below we discuss some of the properties of our proposed estimator.

Remark 1. $\hat{\Pi}(d)$ is an *unbiased estimator* for the expected profit $\mathbb{E}[\Pi(d)|X_1, \ldots, X_N]$:

$$\begin{aligned}
\mathbb{E}[\hat{\Pi}(d)|X_1, \ldots, X_N] &= \sum_{i=1}^{N} \mathbb{E}\left[\frac{1-W_i}{1-e(X_i)}(1-d(X_i)) \cdot \pi_i(0) + \frac{W_i}{e(X_i)} d(X_i) \cdot \pi_i(1)|X_i\right] \\
&= \sum_{i=1}^{N} \left(\frac{1-e(X_i)}{1-e(X_i)}(1-d(X_i)) \cdot \mathbb{E}[\pi_i(0)|X_i] + \frac{e(X_i)}{e(X_i)} d(X_i) \cdot \mathbb{E}[\pi_i(1)|X_i]\right) \\
&= \sum_{i=1}^{N} \left((1-d(X_i)) \cdot \mathbb{E}[\pi_i(0)|X_i] + d(X_i) \cdot \mathbb{E}[\pi_i(1)|X_i]\right) \\
&= \sum_{i=1}^{N} \mathbb{E}\left[(1-d(X_i)) \cdot \pi_i(0) + d(X_i) \cdot \pi_i(1)|X_i\right] \\
&= \mathbb{E}[\Pi(d)|X_1, \ldots, X_N].
\end{aligned}$$

Note that the second line in this derivation follows because $\pi_i(0)$ and $\pi_i(1)$ are independent of W_i conditional on X_i, which holds because each W_i is a Bernoulli draw with success probability $e(x)$, $x = X_i$. Furthermore, because the N customers in the calculation of (3) are randomly drawn from the customer population, $\hat{\Pi}(d)$ is also an unbiased estimator of the expected profit of d in the whole customer population,

$$\mathbb{E}[\hat{\Pi}(d)] = \mathbb{E}_X[\mathbb{E}[\hat{\Pi}(d)|X_1, \ldots, X_N]] = \mathbb{E}_X[\mathbb{E}[\Pi(d)|X_1, \ldots, X_N]] = \mathbb{E}[\Pi(d)].$$

We emphasize that in addition to the randomized sampling and treatment assignment mechanism, the stable unit treatment value assumption is necessary to derive these results.

The intuition for why $\hat{\Pi}(d)$ is an unbiased estimator of the expected profit from the targeting policy d is straightforward in the case when the targeting probability in the randomized sample is constant, $e(x) \equiv e$ for all $x \in \mathbb{X}$. The realized profit if the policy d is implemented in the sample consists of the sum of the profits from customers who should be targeted, $d(X_i) = 1$, and the profit from customers who should not be targeted, $d(X_i) = 0$. With a constant targeting probability e the estimated profit in the randomized sample, $\hat{\Pi}(d)$, is based on a $100 \cdot e$ percent random sample of the profits from customers who should be targeted and a $100 \cdot (1-e)$ random sample of the profits from customers who should not be targeted. Scaling the observed profits of customer who should and should not be targeted by $\frac{1}{e}$ and $\frac{1}{1-e}$, respectively, restores the correct profit level of the intended targeting policy d. For example, if $e(x) \equiv e = \frac{1}{2}$, then we observe the profits of approximately one half of customers who should be targeted and one half of customers who should not be targeted according to the policy d. The other half of all observations is missing at random. To predict the correct overall profit level when targeting N customers, we multiply the measured profit from the usable observations by $\frac{1}{e} = 2$.

Remark 2. The proposed estimator is *qualitatively equivalent* to a field implementation. We emphasize that there is no qualitative difference between the evaluation of the expected profit, $\mathbb{E}[\Pi(d)]$, using either $\Pi(d)$, the observed profit in a field experiment where the targeting policy

d is directly implemented, or using $\hat{\Pi}(d)$ in a randomized sample. In particular, whether we intended to test a policy d and hence implemented a corresponding field test, or if we decided ex-post, after obtaining the randomized sample, that we wanted to evaluate the profits from a policy that we had not even considered before collecting the data does not change the fact that either approach yields an unbiased estimate of the expected profit. Quantitatively, the two approaches differ in the precision of the profit estimates because the effective number of observations that can be used to calculate $\hat{\Pi}(d)$ is generally less than N. However, we can ensure an effective target sample size of N_T by choosing the actual sample size N according to the equation

$$(1-e)(1-\omega_d)N + e\omega_d N = N_T, \tag{4}$$

where ω_d is the percentage of customers targeted based on the strategy d and the targeting probability used to create the randomized sample is constant, $e(x) \equiv e$.

For example, if $e(x) \equiv \frac{1}{2}$ and we want to obtain an effective sample size of N_T, then the size of the randomized sample needs to be $N = 2N_T$. We compare the variance of the profit estimator $\Pi(d)/N_T$ from a field test with N_T observations to the variance of $\hat{\Pi}(d)/N$ from a randomized sample with $2N_T$ observations. Note that we make the profit levels comparable by scaling the predicted profit levels of the two estimators to the same per-customer level. To predict $\Pi(d)/N_T$ from a field test, we first randomly select N_T customers from the population, and then we implement the policy d in the selected sample. To predict $\hat{\Pi}(d)/N$ from a randomized sample, we randomly select $2N_T$ customers from the population, randomly target approximately one half of the customers, and then calculate

$$\begin{aligned}\frac{\hat{\Pi}(d)}{N} &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\frac{1}{2}}(1-W_i)(1-d(X_i))\cdot\pi_i(0) + \frac{1}{\frac{1}{2}}W_i d(X_i)\cdot\pi_i(1)\right) \\ &= \frac{1}{N_T}\sum_{i=1}^{2N_T}\left((1-W_i)(1-d(X_i))\cdot\pi_i(0) + W_i d(X_i)\cdot\pi_i(1)\right).\end{aligned}$$

The number of usable observations, $W_i = d(X_i)$, in this sum will be approximately N_T, and these usable observations represent a random sub-sample of the randomly selected N customers. Hence, the two estimators are almost identical, with the only difference being that the number of usable observations in the calculation of $\hat{\Pi}(d)/N$ is randomly distributed according to a binomial distribution with mean $\frac{1}{2}N = N_T$.

Remark 3. The proposed estimator is *cost-effective*. In particular, as we already discussed, beyond the initial cost of obtaining a randomized sample with N observations, the marginal cost of evaluating an additional estimator is zero. Furthermore, unlike in the case of field tests, there is no limit on the number of targeting policies that can be evaluated.

In summary, we have proposed a cost-effective, unbiased estimator of profits from any targeting policy d that is qualitatively equivalent to a field implementation. This estimator enables us to

evaluate the profits of a large number of alternative targeting policies in this paper.

4 Optimal targeting policies and conditional average treatment effects

A key goal of this paper is to compare many different estimates of an optimal targeting policy using the inverse probability weighted profit estimator $\hat{\Pi}(d)$ that we introduced in the previous section. We now show how optimal targeting policies can be constructed based on estimates of heterogeneous treatment effects. We also review the conditions under which the conditional average treatment effect is identified.

4.1 Optimal targeting policies

An optimal policy, $d^* : \mathbb{X} \to \{0, 1\}$, maximizes the expected profit $\mathbb{E}[\Pi(d)]$. As discussed in Section 3, we consider applications where the treatment received by customer i, W_i, has no impact on the behavior of any other customer (stable unit treatment value assumption). Because W_i only affects the behavior of customer i, a policy is optimal if and only if it maximizes the expected profit from each individual customer with features X_i,

$$d^*(X_i) = \underset{d(X_i)\in\{0,1\}}{\arg\max}\ \mathbb{E}\left[(1 - d(X_i)) \cdot \pi_i(0) + d(X_i) \cdot \pi_i(1)|X_i\right].$$

Hence, an optimal policy targets a customer, $d^*(X_i) = 1$, if and only if $\mathbb{E}\left[\pi_i(1)|X_i\right] > \mathbb{E}\left[\pi_i(0)|X_i\right]$. Equivalently,

$$\begin{aligned}\mathbb{E}\left[\pi_i(1) - \pi_i(0)|X_i\right] > 0 &\Leftrightarrow \mathbb{E}\left[(mY_i(1) - c) - (mY_i(0))|X_i\right] > 0 \\ &\Leftrightarrow m\mathbb{E}\left[Y_i(1) - Y_i(0)|X_i\right] - c > 0 \\ &\Leftrightarrow m\tau(x) > c. \end{aligned} \tag{5}$$

Here, $\tau(x)$ is the conditional average treatment effect (CATE),

$$\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]. \tag{6}$$

The conditional average treatment effect represents the expected *causal effect* of being targeted versus not being targeted for a customer with observed features $X_i = x$. In the application in this paper, this causal effect represents the *incremental* effect on spending that can be attributed to a targeting effort.

The analysis above shows that knowledge of the true conditional average treatment effect is sufficient to construct an optimal targeting policy. The rule described in equation (5) simply states that a customer should be targeted if and only if the incremental profit contribution from targeting exceeds the targeting cost.

4.2 Identification of conditional average treatment effects

The observed outcome, Y_i, is such that $Y_i = Y_i(0)$ if $W_i = 0$ and $Y_i = Y_i(1)$ if $W_i = 1$, and the data set consists of the observations $\mathcal{D} = (Y_i, X_i, W_i)_{i=1}^N$, where N is the number customers. The fundamental problem of causal inference is that only one of the potential outcomes is observed for the each unit (customer) i, whereas the individual treatment effect $Y_i(1) - Y_i(0)$ is not observed. Hence, the conditional average treatment effect (6) cannot be generally inferred from the data. However, the conditional average treatment effect is identified if the data satisfy the following three conditions. First, under *unconfoundedness,* the treatment assignment is random within each sub-group of customers with identical features, $X_i = x$. More formally,

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \mid X_i.$$

Second, the *overlap* assumption requires that the targeting probability or the *propensity score*, $e(x) \equiv \Pr\{W_i = 1 | X_i = x\}$, is strictly between 0 and 1 for all x,

$$0 < e(x) < 1.$$

Third, we assume that the stable unit treatment value assumption (SUTVA) holds, which rules out social interactions or equilibrium effects. Under unconfoundedness, overlap, and SUTVA, the conditional average treatment effect is identified based on the observed conditional expectation function $\mathbb{E}[Y_i | X_i, W_i]$:

$$\tau(x) = \mathbb{E}[Y_i | X_i, W_i = 1] - \mathbb{E}[Y_i | X_i, W_i = 0]. \tag{7}$$

In practice, a company can always ensure that unconfoundedness and overlap are satisfied by conducting an A/B test or a randomized controlled trial. Otherwise, if only observational data are available, knowledge of the targeting policy that was implemented by a company may indicate that unconfoundedness is satisfied (see, for example, Nair et al. 2017). Even if unconfoundedness holds, however, overlap will be violated if a company implements a targeting policy that partitions the feature space into regions where either no customer is targeted or all customers are targeted. SUTVA, the stable unit treatment value assumption, may be violated if targeting a customer increases word-of-mouth and thus influences the purchase decisions of other customers (for example see Ascarza et al. 2016a).

In our application, customers were randomly targeted with a constant, known targeting probability, $e(x) \equiv \frac{2}{3}$. Also, SUTVA is likely satisfied because economically significant social effects due to the receipt of a catalog appear implausible.

5 Estimation of heterogeneous treatment effects

We consider applications of optimal targeting where the conditional average treatment effect, $\tau(x)$, is identified given the conditions stated in the previous section. However, there are many

different available methods to estimate the conditional average treatment effect. A central question asked in this paper is how the predictive accuracy and corresponding expected targeting profits of the policies based on the different estimation methods compare.

Estimation based on the sample analog of (7), the mean difference in outcomes between the treated and untreated units with identical features, is generally not feasible. This is because in practice only few units will have the exact same features, especially when using "big data" with a high-dimensional feature space. We hence need to consider other estimation approaches.

5.1 Direct versus indirect estimation methods

We classify the estimation methods into two groups based on an important conceptual distinction in how the predictions of the conditional average treatment effect is obtained.

The first group includes *indirect estimation methods* that are trained to minimize a squared-error loss based on the observed and predicted outcome *levels*, $\mathbb{E}[(Y_i - \hat{\mu}(X_i, W_i))^2]$. The corresponding best predictor is the regression function,

$$\mu(x, w) = \mathbb{E}[Y|X = x, W = w].$$

Due to unconfoundedness,

$$\mu(x, w) = \mathbb{E}[Y|X = x, W = w] = \mathbb{E}[Y(w)|X = x, W = w].$$

Hence, having obtained an estimate of regression function, we can indirectly predict the conditional average treatment effect based on the predicted expected outcome difference

$$\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0).$$

Indirect estimation appears inefficient as the main object of interest for optimal targeting is a prediction of the conditional average treatment effect, not the outcome level. Correspondingly, we should use *direct estimation methods* based on the loss $\mathbb{E}[(\tau(X_i) - \hat{\tau}(X_i))^2]$. However, due to the fundamental problem of causal inference this loss is infeasible, and hence direct estimation of the treatment effect appears impossible. Despite this challenge, some recent methods developed in the literature on causal inference are effectively direct estimators of the conditional average treatment effect. One such method is the causal forest of Wager and Athey (2017) that is built on the causal trees of Athey and Imbens (2016). Two alternative direct estimation methods that we introduce in this paper are the causal KNN regression and the treatment effect projection (TEP) that builds on the causal KNN predictions.

5.2 Indirect estimation of $\tau(x)$ via $\mu(x, w)$

Linear models

Assuming that the conditional expectation of Y_i can be approximated using a linear function, both for the treated and the untreated units, we can estimate $\mu(x, w)$ using the regression model for $x \in \mathbb{R}^p$

$$\mathbb{E}[Y_i|X_i = x, W_i = w] = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \delta_0 w + \sum_{k=1}^{p} \delta_k x_{ik} w_i. \tag{8}$$

It follows that $\tau(x)$ is a linear function of the features,

$$\begin{aligned} \tau(x) &= \mathbb{E}[Y_i|X_i = x, 1] - \mathbb{E}[Y_i|X_i = x, 0] \\ &= \delta_0 + \sum_{k=1}^{p} \delta_k x_{ik}. \end{aligned} \tag{9}$$

The most basic estimator for (8) is the method of ordinary least squares (OLS). Equivalently, we could estimate two separate linear regressions for the treated and for the untreated units. When estimated using OLS, the difference in the prediction of these two regressions will be identical to (9).

We also estimate (8) using the Lasso (Tibshirani 1996, Hastie et al. 2015). The Lasso is a regularized regression method that shrinks the regression coefficients and performs variable selection. The Lasso adds a penalty term to the squared-error loss that is increasing in the absolute value of the regression coefficients. The weight of this penalty (the tuning parameter) is chosen based on the cross-validated prediction error. Because the penalty is based on the cross-validated error, an out-of-sample measure of fit, the Lasso mitigates overfitting and typically has higher predictive accuracy than OLS. Unlike in the case of OLS, estimating (8) using the Lasso will not generally be equivalent to estimating two separate regressions for the treated and untreated units.

Logit/OLS model

We can decompose the conditional expectation of Y_i into the purchase incidence and the conditional expectation of Y_i if a purchase occurs. Spending, Y_i, is non-negative, and a purchase occurs if $Y_i > 0$. Expected spending is then given by

$$\mathbb{E}[Y_i|X_i = x, W_i = w] = \mathbb{E}[Y_i|Y_i > 0, X_i = x, W_i = w] \cdot \Pr\{Y_i > 0|X_i = x, W_i = w\}.$$

This approach is a variant of a model that is widely employed by practitioners in the marketing industry to score customer based on the expected spending *level* (see Section 10). In practice, the purchase probability is typically estimated using logistic regression. The expectation of spending conditional on a purchase, or—more commonly—the expectation of the log of spending condi-

tional on a purchase, $\mathbb{E}[\log(Y_i)|Y_i > 0, X_i = x, W_i = w]$, is estimated using a linear regression.[5] We refer to this estimation method as the logit/OLS model.

Non-parametric models

The machine learning literature has provided us with some flexible estimation methods that are feasible to implement in practice. Such methods may provide better predictions when the models that we discussed so far are mis-specified due to their parametric assumptions. Random forests (Breiman 2001) are one leading method that allow for model flexibility and—due to the availability of high-quality software—are straightforward to implement. Random forests are ensembles of trees. Trees are flexible estimation methods but tend to overfit the training data and hence predict poorly out-of-sample. Random forests are based on two key ideas to reduce the variance of the prediction from one single tree. First, *bagging* (bootstrap aggregation) is an algorithm that creates a large number of training sets using the bootstrap, fits a tree to each of the training sets, and then forms a prediction based on the average of all predictions from the individual trees. Second, when growing a tree, each candidate split is based on a randomly chosen subset of all features. Bagging and random feature selection reduce overfitting and improve the out-of-sample predictive accuracy.

When estimating a random forest, we provide the algorithm with the set of features and the treatment indicator, W_i. Specifying the interactions between X_i and W_i is not be necessary because random forests, a non-parametric estimation method, in principle should be able to automatically detect any relevant interactions.

We also consider an alternative approach that uses two separate random forests to estimate the conditional expectation function for the treated and untreated units. This approach enforces that the treatment indicator is taken into account when predicting $\tau(x)$ based on the conditional expectation function, and hence *may* provide a more accurate prediction of the CATE than the random forest that jointly uses x and w as features.

5.3 Direct estimation of $\tau(x)$

All estimation methods in Section 5.2 predict the conditional average treatment effect indirectly, based on the conditional expectation function that is trained using a squared-error loss with respect to the outcome level. In this section, we consider methods that directly predict the conditional average treatment effect, which is the main object of interest for the purpose of optimal targeting.

We first introduce two new methods, causal K-nearest neighbor (KNN) regression and treatment effect projection (TEP), which builds on the causal KNN predictions. These methods select the tuning parameter K using a feasible loss that, in expectation, attains its minimum at the same K value as the infeasible loss based the treatment effect, $\mathbb{E}[(\tau(X_i) - \hat{\tau}(X_i))^2]$. We then review the recently developed causal forest of Wager and Athey (2017). The causal forest's

[5] Assuming normally distributed error terms, we can use the formula for the mean of a log-normally distributed random variable to predict expected spending based on the estimated expectation of the log of spending.

splitting rule builds trees and mimics the process that would be used if the treatment effects were observed.

Causal KNN regression

For any feature vector x, we find the K nearest *treated neighbors* and separately the K nearest *untreated neighbors*.[6] We then estimate the conditional average treatment effect using the mean difference between the nearest treated and untreated units,

$$\hat{\tau}_K(x) = \frac{1}{K} \sum_{i \in \mathcal{N}_K(x,1)} Y_i - \frac{1}{K} \sum_{i \in \mathcal{N}_K(x,0)} Y_i.$$

Here, $\mathcal{N}_K(x, w)$ is the set of the K nearest neighbors with treatment status $w \in \{0, 1\}$.

To predict $\hat{\tau}_K(x)$ we need to choose a value of K. If we tuned K based on a squared-error loss with respect to the outcome levels, then we would obtain an indirect estimator of the treatment effect. To directly predict the conditional average treatment effect we would ideally choose K to minimize a loss with respect to the treatment effect, but, as already discussed, this approach is infeasible because the treatment effect is unobserved. Instead, we tune K based on the *transformed outcome loss*, the mean-squared difference between $\hat{\tau}_K(x)$ and the *transformed outcome*, Y_i^*, which is a proxy for the true conditional average treatment effect. Following Athey and Imbens (2016), we define the transformed outcome as follows:

$$Y_i^* = W_i \cdot \frac{Y_i(1)}{e(X_i)} - (1 - W_i) \cdot \frac{Y_i(0)}{1 - e(X_i)}.$$

The transformed outcome only depends on the potential outcome that corresponds to the realized treatment level and can hence be calculated from the observed outcome, Y_i:

$$Y_i^* = \frac{W_i - e(X_i)}{e(X_i)(1 - e(X_i))} Y_i.$$

Therefore, if the propensity score $e(x)$ is known as in the applications that we consider, the transformed outcome Y_i^* is observed for all units i. Under the maintained assumption of unconfoundedness,

$$\begin{aligned}
\mathbb{E}[Y_i^* | X_i = x] &= \mathbb{E}[W_i | X_i = x] \cdot \frac{\mathbb{E}[Y_i(1) | X_i = x]}{e(X_i)} - \mathbb{E}[1 - W_i | X_i = x] \cdot \frac{\mathbb{E}[Y_i(0) | X_i = x]}{1 - e(X_i)} \\
&= e(X_i) \cdot \frac{\mathbb{E}[Y_i(1) | X_i = x]}{e(X_i)} - (1 - e(X_i)) \cdot \frac{\mathbb{E}[Y_i(0) | X_i = x]}{1 - e(X_i)} \\
&= \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x] \\
&= \tau(x).
\end{aligned}$$

[6] We measure the distance between two feature vectors using the Euclidean metric after standardizing the features.

Hence, the transformed outcome is an unbiased estimate of the conditional average treatment effect, $Y_i^* = \tau(X_i) + \nu_i$, where $\mathbb{E}[\nu_i|X_i] = 0$ and ν_i is orthogonal to any function of X_i.

To choose K we minimize the transformed outcome loss, $\mathbb{E}\left[(Y_i^* - \hat{\tau}_K(X_i))^2\right]$. In particular, we obtain the prediction $\hat{\tau}_K(X_i)$ using a training set and then evaluate the transformed outcome loss in a separate test set. Hence, for the purpose of evaluating the expectation, $\hat{\tau}_K(x)$ is a given, non-random function. Therefore, because ν_i is orthogonal to any function of X_i:

$$\begin{aligned}\mathbb{E}\left[(Y_i^* - \hat{\tau}_K(X_i))^2|X_i\right] &= \mathbb{E}\left[(\tau(X_i) + \nu_i - \hat{\tau}_K(X_i))^2|X_i\right] \\ &= \mathbb{E}\left[(\tau(X_i) - \hat{\tau}_K(X_i))^2 + 2(\tau(X_i) - \hat{\tau}_K(X_i)) \cdot \nu_i + \nu_i^2|X_i\right] \\ &= \mathbb{E}\left[(\tau(X_i) - \hat{\tau}_K(X_i))^2|X_i\right] + \mathbb{E}[\nu_i^2|X_i].\end{aligned}$$

Hence, the transformed outcome loss can be decomposed into the infeasible loss based on the true treatment effect and the variance of the residual ν_i,

$$\mathbb{E}\left[(Y_i^* - \hat{\tau}_K(X_i))^2\right] = \mathbb{E}\left[(\tau(X_i) - \hat{\tau}_K(X_i))^2\right] + \mathbb{E}[\nu_i^2].$$

$\mathbb{E}[\nu_i^2]$ does not depend on K. Hence, the value of K that minimizes the transformed outcome loss also minimizes the infeasible loss, $\mathbb{E}\left[(\tau(X_i) - \hat{\tau}_K(X_i))^2\right]$. In practice, we choose K to minimize the sample analog of $\mathbb{E}\left[(Y_i^* - \hat{\tau}_K(X_i))^2\right]$.

Treatment effect projection (TEP)

The treatment effect projection builds on the causal KNN regression. It adds an additional step where we project the causal KNN treatment effect estimates, $\hat{\tau}_K(X_i)$, onto the features, X_i. The purpose of this additional step is to regularize the causal KNN treatment effect estimates and thus to reduce the variance of the prediction. One advantage of the regularization step is that it can be performed using any regression method and thus be adapted to the specific data-generating process in an application. In our empirical application, we compare a linear projection estimated using the Lasso and we use a random forest as a non-parametric alternative.

Transformed outcome regression

Recall the relationship between the transformed outcome and the conditional average treatment effect: $Y_i^* = \tau(X_i) + \nu_i$, where $\mathbb{E}[\nu_i|X_i] = 0$ and ν_i is orthogonal to any function of X_i. Hence, we can estimate the conditional average treatment effect function $\tau(x)$ based on a regression of Y_i^* on X_i. While feasible in principle, this estimation approach will yield noisy estimates if the variance of the residual, ν_i, is large. Note that

$$\mathbb{E}\left[\nu_i^2|X_i\right] = \frac{1}{1 - e(X_i)} \cdot \mathbb{E}\left[{Y_i}^2|X_i, W_i = 0\right] + \frac{1}{e(X_i)} \cdot \mathbb{E}\left[{Y_i}^2|X_i, W_i = 1\right] - \tau(X_i)^2.$$

Hence, if the scale of the conditional average treatment effects is small compared to the scale of the outcome levels, then the signal-to-noise ratio will be small, and the variance of the estimated

conditional average treatment effect function, $\hat{\tau}(x)$, will be large.

Causal forest

Causal forests (Wager and Athey 2017) are ensembles of causal trees. To understand how causal trees are grown using the approach proposed by Athey and Imbens (2016), we first review the approach to grow standard trees. Standard trees are trained to predict outcome levels. The algorithm to grow a tree employs recursive binary splits to obtain a partition of the feature space into regions (leaves) $R_1, R_2, \ldots$ Let $\mathcal{R}_k = \{i : X_i \in R_k\}$ be the set of observations in leaf R_k, and let N_k be the corresponding number of observations. For any feature $x \in R_k$, the predicted outcome is the average over all observations in R_k:

$$\hat{\mu}(x) = \hat{\mu}_{R_k} = \frac{1}{N_k} \sum_{i \in \mathcal{R}_k} Y_i.$$

The algorithm adds a split to a current terminal node, R_k, based on a feature, X_l, and a cutoff, κ, such that R_k is divided into the regions $R_{k1} = \{x \in R_k : x_l < \kappa\}$ and $R_{k2} = \{x \in R_k : x_l \geq \kappa\}$. The residual sum of squares that results from such a split is given by

$$\text{RSS} = \sum_{i \in \mathcal{R}_{k1}} (Y_i - \hat{\mu}_{R_{k1}})^2 + \sum_{i \in \mathcal{R}_{k2}} (Y_i - \hat{\mu}_{R_{k2}})^2. \tag{10}$$

The splitting rule chooses a feature and a cutoff such that the resulting split yields the smallest RSS among all possible binary splits. The residual sum of squares (10) can be written as

$$RSS = \sum_{i \in \mathcal{R}_k} Y_i^2 - \sum_{i \in \mathcal{R}_k} \hat{\mu}(X_i)^2 = \sum_{i \in \mathcal{R}_k} Y_i^2 - N_{k1} \hat{\mu}_{R_{k1}}^2 - N_{k2} \hat{\mu}_{R_{k2}}^2. \tag{11}$$

Here, $\hat{\mu}(X_i)$ is the outcome prediction after splitting R_k into R_{k1} and R_{k2}. Hence, finding a split that minimizes the residual sum of squares (10) is equivalent to finding a split that maximizes the sum of the squared predictions, $\sum_{i \in \mathcal{R}_k} \hat{\mu}(X_i)^2$. Furthermore, maximizing the sum of the squared predictions is equivalent to finding a split that maximizes the variance of the predictions $\hat{\mu}(X_i)$ across the observations in the two new leaves, $i \in \mathcal{R}_{k1} \cup \mathcal{R}_{k2}$.

Like standard trees, causal trees are also grown using binary splitting rules. Causal trees predict the conditional average treatment effect for a feature $x \in R_k$ based on the mean difference between the outcome levels of the treated and untreated units,

$$\hat{\tau}(x) = \hat{\tau}_{R_k} = \frac{1}{N_k(1)} \sum_{i \in \mathcal{R}_k(1)} Y_i - \frac{1}{N_k(0)} \sum_{i \in \mathcal{R}_k(0)} Y_i.$$

Here, $\mathcal{R}_k(w)$ is the set of observations i such that $X_i \in R_k$ and $W_i = w$, and $N_k(w)$ is the corresponding number of observations.

Adding a split to a causal tree based on an analog of the residual sum of squares (10) appears infeasible, once again because the treatment effects are unobserved. However, for standard trees

we can see from the equivalence between (10) and (11) that any split affects the RSS only through its impact on the sum of squared predictions, $\sum_{i \in \mathcal{R}_k} \hat{\mu}(X_i)^2$. Hence, analogous to the approach to grow a standard tree, Athey and Imbens (2016) propose a splitting rule that maximizes $\sum_{i \in \mathcal{R}_k} \hat{\tau}(X_i)^2$, which maximizes the variance of the predicted treatment effects $\hat{\tau}(X_i)$ across the observations in the two new leaves. This splitting rule is feasible given the data and mimics the approach that would be used if the treatment effect was observed.

6 Empirical application and data description

The data stem from a collaboration with a U.S. company that sells a large assortment of consumer products. The company has a sophisticated data science team that plans and evaluates customer targeting efforts, including catalogs, e-mails, and Facebook ads. A large fraction of the product volume is sold through direct channels that includes phone orders, mail orders, and purchases made on the company's website. The company also has established retail stores in some U.S regions.

6.1 Data set

The company maintains an extensive data base that captures customer-level transactions, detailed records of customer behavior, and the marketing efforts targeted at the customers. The company frequently employs randomized customer samples to evaluate the performance of its campaigns. In our application we use two randomized samples that were created to measure the effectiveness of a catalog mailing. The catalogs were mailed in spring of 2015 and 2016 at the approximately same date within each year. The format of the catalogs and the *types* of products (not the exact product assortments) included in the catalogs are identical. From the company's perspective, these catalogs are two instances of the same campaign that is repeated annually at the same time within the calendar year. The randomized samples include approximately 293 thousand customers in 2015 and 148 thousand customers in 2016. To draw these samples the company determines the sample size, typically as a percentage of the customer base that qualifies for receipt of the catalog, and then randomly selects a corresponding number of customers. The selected customers then receive a catalog with probability $\frac{2}{3}$.

All customers who are not selected into randomized sample are mailed according to the company's targeting model. This model is an application of the traditional CRM approach to targeting. The company first predicts a *score* for each customer, the expected spending *level* conditional on some customer features and conditional on being targeted, $W_i = 1$. The targeting decision is then based on a cutoff rule, such that customers are mailed if their predicted score exceeds a threshold level. We will discuss this approach in detail in Section 10.

Our data set is at the customer level and includes a treatment (mailing) indicator, an outcome variable that measures dollar spending, and a set of close to 500 features. Customer spending is measured based on all phone orders, mail orders, and website transactions that were made using the product codes included in the catalog during the three-month period after receiving

the catalog. In addition, spending also includes all web purchases that were not made using the product codes in the thirty-day period after the catalog was received. This spending measure is the variable used by the company to build its scoring model. According to the company most mail, phone, and website transactions made using product codes occur in the month after receipt of the catalog, and the spending measure only allows us to estimate the *short-run* causal effect of a targeting effort. This short-run effect may be larger than the long-run effect if the catalog leads to forward-buying or smaller than the long-run effect if the catalog affects purchases that are not made using the product codes beyond the thirty-day period after the catalog is received.

The feature set includes demographic variables, information on past purchasing behavior, customer behavior on the company's website, and customer actions after the receipt of a target e-mail. The features that capture past purchasing behavior are RFM (recency, frequency, and monetary value) variables that are widely employed in customer analytics. The data capture the time elapsed since the most recent transactions and the number of transactions and dollar volume over various time periods (the last six months, the last 7-12 months, the last 13-18 months, etc.). The data are provided both in the aggregate and at a disaggregated level that distinguishes among transactions for certain product types and transactions made through different sales channels, such as phone and web orders. At the most granular level, we can observe transactions at the time period/product type/channel level, such as the number of purchases during the last twelve months for products in a specific category that were made online. The features that capture customer-level website behavior include the number of page views and the number of clicks on product pages, captured at the time period/product type level. The usual caveats apply: page views and clicks that cannot be linked to a customer ID are not recorded. The target (promotional) e-mail data record the number of e-mails received, the number of e-mails that a customer opened, and the number of click-throughs resulting from these e-mails. Again, the data are recorded separately for different time periods. The 2015 and 2016 data come with separate feature sets that provide up-to-date customer-level information as of a few weeks before the respective catalog mailing.

6.2 Data description

We provide summary statistics of the outcome variable (spending) and the treatment indicator, separately for 2015 and 2016, in Table 1. We first focus on the 2015 data. 6.2 percent of customers made a purchase, and mean spending is 7.3 dollars. Figure 1 displays the distribution of spending conditional on a purchase, i.e. spending > 0. The distribution is skewed, with mean spending of 117.7 dollars and median spending of 79.9 dollars. Mean spending in 2016 is smaller than in 2015, 6.5 dollars versus 7.3 dollars. The difference reflects both a smaller purchase incidence, 5.6 versus 6.2 percent, and less spending on average conditional on a purchase, 115.1 versus 117.7 dollars. Overall, however, the differences in the distribution of spending across the 2015 and 2016 data appear small, which is also visible when comparing the conditional spending distributions in Figure 1.

6.3 Average treatment effect

We summarize the average treatment effect (ATE) of a catalog mailing on spending in Table 2. The ATE is 2.56 dollars in 2015 and 2.38 dollars in 2016. Both estimates are statistically different from zero at any conventional significance level. The ATE on the purchase incidence represents an increase in the average purchase probability by 2.2 percentage points. This effect is identical in both years and statistically different from zero. We also compare the difference in spending among the treated and untreated units conditional on a purchase, $Y_i > 0$, although we emphasize that this difference cannot be interpreted as a causal effect. The difference in average conditional spending across the two groups is -0.80 dollars in the 2015 data and -2.17 dollars in the 2016 data. This suggests that the marginal customers (those who would not have made a purchase if they had not been targeted) spend less on average than the customers who make a purchase irrespective of being targeted, although the difference in conditional spending is not statistically different from zero.

The company's margin and targeting cost data imply that a customer should be targeted if the conditional average treatment effect exceeds 2.003. Hence, if there was no heterogeneity in the treatment effects across customers, or if the company assumed that all treatment effects were identical, the company should use a blanket targeting strategy and mail all customers in its data base.

6.4 Randomization and covariate balance

We test if the mean of each feature is equal across the treated and untreated units using Welch's t-test. Figure 2 shows the distribution of the corresponding p-values, separately for 2015 and 2016. In 2015 15.9 percent of the tests have a p-value less than 0.05, whereas 5.7 percent of the p-values in 2016 are less than 0.05. These results are consistent with a completely randomized treatment assignment mechanism in 2016, but in 2015 the percentage of features for which we reject the equality of means at a 5 percent level is larger than the percentage (5 percent) that we expect under complete randomization.

To assess if the difference in means across the treated and untreated units is quantitatively important we assess the covariate balance (the degree of overlap) in the data. Generally, for each unit i in a balanced sample there is a similar unit k ("similar" means that $\|X_i - X_k\|$ is small) with the opposite treatment status. This ensures that the estimates of the conditional average treatment effects rely on the data, not on an extrapolation from units with the opposite treatment status that have substantially different features. To quantify the difference in the distributions of the features between the treated and untreated units we first calculate the normalized difference for each component l of the feature vector,

$$\frac{\bar{X}^{(l)}(1) - \bar{X}^{(l)}(0)}{\sqrt{\frac{S_l^2(0)+S_l^2(1)}{2}}}.$$

Here, $\bar{X}^{(l)}(w)$ is the estimated mean and $S_l(w)$ is the estimated standard deviation for all ob-

servations of $X^{(l)}$ with treatment status $W_i = w$. We also assess the difference in the dispersion of the distributions for the treated and untreated units, based on the log-ratio of the standard deviations, $\log(S_l(1)) - \log(S_l(0))$. We summarize the distribution of the absolute value of the two measures in Table 3. Overall, the difference in the means and standard deviations across the treated and untreated units is small. In 2015, the median difference in means is 0.5 percent, and the median log ratio of the standard deviations is 2.2 percent. Focusing on the subset of features for which we reject the equality of means, the median normalized difference in means is 0.9 percent and the median log ratio of the standard deviations is 2.2 percent. Hence, even though there is a statistically significant difference in means for this subset of features, the quantitative difference between the distributions of $X^{(l)}$ across the treated and untreated units is small. We also check if the treatment probability systematically differs across units with different feature vectors. We estimate a logistic regression model of the propensity score using the Lasso, and we plot the distribution of the predicted propensity score in Figure 3, separately for the treated and untreated units. We find that the variance of the predicted treatment probabilities is very small, and there is virtually no difference in the propensity score across the treated and untreated units. Hence, the data are highly balanced.

We return once more to the issue that in 2015 the percentage of features with statistically significant differences in means is larger than expected under full randomization. The company that we are collaborating with has a long history of evaluating its targeting tactics using completely randomized samples, and the data science team assured us that the treatment assignment in 2015 was completely randomized. Suppose, however, that part of the 2015 data was "contaminated," for example with customers who were targeted according to the firm's targeting policy. Note that the targeting policy of the firm is based on a strict subset of the *observed* features that we use in our analysis, and thus unconfoundedness is satisfied. Hence, the only remaining concern would be that overlap was violated. However, our analysis reveals a high degree of covariate balance. Based on this evidence, even if there was a tiny degree of "contamination" it would be negligible for our analysis.

7 Estimation and prediction: Preliminaries

Before we discuss the estimation results, an assessment of model fit, and an evaluation of the targeting profits, we provide some specifics of the empirical implementation.

7.1 Data samples

We already indicated that we employ two data sets that capture customer spending, pre-treatment customer features, and the randomized treatment indicator. One data set was collected in 2015 and the other was collected in 2016. All the results that we present—model fit and the targeting policy evaluation—indicate the external validity of the proposed methods. The results are always obtained for a test set that is distinct from the training set used to obtain the model estimates. However, the 2015 and 2016 data serve distinct purposes. In the 2015 data we assess

the external validity and value of the proposed methods in test sets of customers that are drawn from the same data-generating process as the training sets. The results in Sections 8 and 9 are obtained for the 2015 data only. In Section 11 we assess the transportability of the estimation results obtained from the 2015 data to the 2016 customer population. The analysis based on the 2015 data takes precedence because it allows us to compare the value of the different estimation methods without any potential confounds that may arise if customer behavior systematically changes over time. A separate question is if the model predictions from the 2015 data still apply in 2016. The latter situation corresponds to an actual targeting problem where the training set is obtained in advance of the implementation of the targeting policy.

7.2 Feature sets

The data set includes 472 features. Many features are highly correlated, and the full set of features is linearly dependent. Hence, OLS and logit estimates cannot be obtain from the full data set. However, OLS and logit are widely used in the industry and hence serve as an important baseline for the model comparison. Hence, we create a reduced feature set. To obtain this smaller set we use the Lasso to estimate a linear regression of customer-level spending on the full set of features using the 2015 data. We perform the estimation *separately* for the treated and untreated customers, and then choose all variables that are selected by the Lasso in at least one of the two regressions. Four pairs of the selected variables have a correlation coefficient that exceeds 0.975, and we drop one of these highly correlated features in each pair.[7] Using this procedure we end up with a reduced feature set that includes 64 linearly independent variables.

7.3 Selection of K in the causal KNN regression

Following the discussion in Section 5.3, we choose a value of K to implement the causal KNN estimator $\hat{\tau}_K(X_i)$ based on the K that minimizes the sample analog of the transformed outcome loss, $\mathbb{E}\left[(Y_i^* - \hat{\tau}_K(X_i))^2\right]$. Recall that the K that minimizes the feasible transformed outcome loss also minimizes the squared-error loss with respect to the conditional average treatment effect, $\mathbb{E}\left[(\tau(X_i) - \hat{\tau}_K(X_i))^2\right]$.

Figure 4 shows a plot of K versus the mean-squared difference between the transformed outcome and $\hat{\tau}_K(X_i)$. We predict the criterion function based on 1000 bootstrap samples from the 2015 data. Because the optimal K depends on the sample size, we choose samples with an equal number of observations as the training sets that we will employ. In each bootstrap sample we predict $\hat{\tau}_K(X_i)$ using the K nearest treated and untreated neighbors of observation i, not including i itself, and we then calculate the mean-squared prediction error with respect to the transformed outcome Y_i^*. The criterion function in Figure 4 is the average over all 1000 criterion functions obtain for each bootstrap draw.

The criterion function is minimized at $K = 1700$ over the range of K values considered, $K = 25, 50, 75, \ldots, 5000$. To assess the importance of the choice of K for the accuracy of the

[7] We keep the variable that is more highly correlated with the outcome.

model predictions we also consider an alternative value of $K = 425$. This is the K that one would choose heuristically as the value at which the initial steep decline of the criterion function becomes substantially less pronounced when using bootstrap samples equal in size to the *whole* 2015 data.[8] This sub-optimally chosen K prioritizes small bias in the prediction.

7.4 Bootstrap algorithm

To account for the inherent variability in the model predictions across data sets we summarize the distribution of the predictions based on $B = 1000$ bootstrap replications. For each replication b, we first randomly split the original data into two groups of equal size, $\mathcal{A}_b$ and $\mathcal{B}_b$. We then create a training set, $\mathcal{E}_b$, by sampling $|\mathcal{A}_b|$ observations from $\mathcal{A}_b$ with replacement, and we similarly create a test set $\mathcal{T}_b$ from $\mathcal{B}_b$.[9] We use this particular sampling approach to ensure that no single observation is contained in *both* the training set and the test set. A more standard approach is to first draw a bootstrap sample with replacement from the original data, and then to randomly split the bootstrap sample into a training set and a test set. However, using this standard approach some observations will be contained in both the training set and the test set. We find that the random forest in particular is often able to perfectly predict the outcome for some of these common observations, which provides a deceptive appearance of good model fit. To avoid this problem we hence use the bootstrap sampling approach that rules out any common observations across the training and test sets.

Note that we choose the reduced feature set and the value of K in an initial step using the entire 2015 data. We then use the same reduced feature set and K value across all bootstrap replications.

8 Estimation results and model fit

We predict the conditional average treatment effects, $\hat{\tau}(X_i)$, for the 2015 data using ten-fold cross validation.[10] Figure 5 shows the distribution of the predicted conditional average treatment effects separately for nine key estimation methods. The corresponding summary statistics are presented in Table 4. All estimation methods predict a substantial degree of heterogeneity in the conditional average treatment effects across customers. For example, the causal forest (at the bottom of Figure 5 and Table 4) predicts that 90 percent of the predicted $\hat{\tau}(X_i)$ values are between 0.68 and 8.18 dollars. Furthermore, for most of the methods the median of the CATE predictions is less than 1.5. As we pointed out in Section 6.3, the company should target customers with conditional average treatment effects above 2.003. If the company ignored the heterogeneity in the treatment effects and based its targeting strategy on the average treatment

[8] We chose K in this manner at an early stage in the project before we understood that we could select K using the transformed outcome loss.

[9] Because the words "training" and "test" have the same initial letter, we use the letter $\mathcal{E}$, as in "estimation," to refer to the training set.

[10] In particular, we randomly split the data into ten folds, $k = 1, \ldots, 10$, and we obtain the treatment effect predictions for all observations in fold k based on the estimates using all other folds as a training set.

effect (2.56 dollars in 2015), it would use a blanket mailing strategy. In contrast, based on the estimated distribution of the customer-level treatment effects, less than half of all customers should receive a catalog.

We devote the remainder of this section to a discussion of model fit. The fit of a standard regression model can be assessed based on the mean distance between the observed and predicted outcomes, which is typically calculated using the mean-squared error in a test set, $|\mathcal{T}|^{-1} \sum_{i \in \mathcal{T}} (Y_i - \hat{\mu}(X_i))^2$. An analogous measure of fit cannot be calculated when predicting treatment effects because $\tau(X_i)$ is unobserved. Instead, we use three alternative approaches to assess the fit of the different estimation methods.

8.1 Percentage negative $\hat{\tau}(X_i)$ values

Even though we cannot directly compare the treatment effect predictions to the true treatment effects, we can ask if the predicted distributions are plausible. In particular, in our application it is unlikely that $\tau(x)$ is negative—the receipt of a catalog mailing will not *cause* a customer to spend less. If this assumption is correct, then a smaller percentage of negative predicted $\hat{\tau}(X_i)$ values indicates better model fit, and especially large negative $\hat{\tau}(X_i)$ values suggest poor model fit.

Table 4 summarizes the percentage of predicted negative $\hat{\tau}(X_i)$ values. The indirect estimation methods all predict a fairly large percentage of negative $\hat{\tau}(X_i)$ values. Among the linear models the percentage of negative values is 20.3 percent for OLS and 18.4 percent for the Lasso. The logit/OLS predictions yield a smaller percentage of negative values (7.7 percent), whereas the flexible random forest predictions provide no advantage over the linear models when evaluated based on the negative predictions.

Most of the direct estimation methods yield better predictions. The causal KNN regression ($K = 2225$) predicts only 4.3 percent negative values. Note that $K = 2225$ is the value that minimizes the transformed outcome criterion for the ten-fold cross validation case. Instead, if we choose $K = 525$ based on the heuristic that prioritizes low bias (Section 7.3), the percentage of negative values is 15.6. This finding demonstrates the importance of tuning the K value based on the transformed outcome criterion. The results improve further when we regularize the causal KNN predictions using the treatment effect projections (TEP). The TEP using the Lasso yields 1.3 percent and the TEP using the random forest yields 1.2 percent of negative values. The performance of the causal forest is similar, with 1.8 percent of negative $\hat{\tau}(X_i)$ values. The predictions from the transformed outcome regressions, however, provide no advantage over the indirect estimation methods. Based on the discussion in Section 5.3 this result is not unexpected, due to the large variance in the transformed outcome values relative to the variance of the conditional average treatment effects.

Comparing the distributions of the predicted treatment effects in Figure 5 we also find more dispersion and a higher incidence of large, negative treatment effects for the indirect estimation methods than for the direct estimation methods, except for the causal KNN regression with the

improperly tuned value of $K = 525$.

Overall, the results suggest that the estimation methods that are directly trained to predict the conditional average treatment effect (except the transformed outcome regressions) provide a better model fit than the methods that indirectly predict the conditional average treatment effect via the regression function, $\mu(x, w)$.

8.2 Mean-squared prediction error based on transformed outcome

Analogous to the approach that we use to tune the value of K in the causal KNN regression, we can evaluate the accuracy of the conditional average treatment effect predictions based on the mean-squared prediction error with respect to the transformed outcome. We use the bootstrap sampling approach described in Section 7.4 to create 1000 training and test sets. For each bootstrap sample b, we estimate all models using the training set and then evaluate the prediction error in the test set based on

$$\text{MSE}_b = \frac{1}{|\mathcal{T}_b|} \sum_{i \in \mathcal{T}_b} (Y_i^* - \hat{\tau}_b(X_i))^2. \tag{12}$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\text{MSE}_b | (X_i)\right] &= \frac{1}{|\mathcal{T}_b|} \sum_{i \in \mathcal{T}_b} \mathbb{E}\left[(Y_i^* - \hat{\tau}_b(X_i))^2 | X_i\right] \\
&= \frac{1}{|\mathcal{T}_b|} \sum_{i \in \mathcal{T}_b} \mathbb{E}\left[(\tau(X_i) - \hat{\tau}_b(X_i))^2 | X_i\right] + \frac{1}{|\mathcal{T}_b|} \sum_{i \in \mathcal{T}_b} \mathbb{E}\left[\nu_i^2 | X_i\right].
\end{aligned}$$

The variance $\mathbb{E}\left[\nu_i^2 | X_i\right]$ does not depend on the treatment effect predictions. Hence, in expectation, if the mean-squared error (12) is smaller for estimation method A than for estimation method B, then the mean-squared prediction error with respect to the true conditional average treatment effect will also be smaller for method A than for method B.

Table 5 provides summary statistics of the mean-squared prediction error across the 1000 bootstrap samples for all estimation methods. The relatively large standard error of the mean reflects the high variability of MSE across bootstrap samples, not that the mean MSE is indistinguishable across estimation methods. Indeed, Table 6 documents the mean differences in MSE's across methods and shows that most differences are statistically different from zero. Similar to the findings in Section 8.1, the results indicate that the direct estimation methods (except the transformed outcome regressions) typically have smaller mean-squared prediction errors than the indirect estimation methods.

8.3 Lift factors

Lift factors are widely used in the marketing analytics industry to visualize the correspondence between predicted and observed outcome levels. Lift factors can easily be adapted for incremental effects. For each estimation method and bootstrap sample b, we predict the conditional average

treatment effect for customers in the test set $\mathcal{T}_b$. We then split the observations into twenty segments of equal size, $s = 1, \ldots, 20$, such that segments s includes all observations where $\hat{\tau}_b(X_i)$ falls between the percentiles $5(s-1)$ and $5s$ of the predicted treatment effect distribution. $\mathcal{T}_{sb}(1)$ is the set of treated and $\mathcal{T}_{sb}(0)$ is the set of untreated units in segment s. The average treatment effect for segment s in the test set $\mathcal{T}_b$ is

$$\bar{\tau}_{sb} = \frac{1}{|\mathcal{T}_{sb}(1)|} \sum_{i \in \mathcal{T}_{sb}(1)} Y_i - \frac{1}{|\mathcal{T}_{sb}(0)|} \sum_{i \in \mathcal{T}_{sb}(0)} Y_i.$$

Figure 6 displays lift charts for nine estimation methods. Each lift chart shows the mean, $\bar{\tau}_s = B^{-1} \sum_{b=1}^{B} \bar{\tau}_{sb}$, and 95 percent range of $\bar{\tau}_{sb}$ over all 1000 bootstrap samples. Corresponding summary statistics are documented in Table 7. All lift charts reveal a positive association between s and $\bar{\tau}_s$, indicating that segments with larger predicted conditional average treatment effects have larger average treatment effects in the test set. Lift charts cannot generally be used to rank estimation methods according to model fit. However, the lift charts reveal some specific differences in fit across the different methods. The indirect estimation methods correctly classify the observations with the largest treatment effects, but are less able to predict observations with small or medium treatment effects. In particular, for the indirect methods the relationship between s and $\bar{\tau}_s$ exhibits a slight U-shape. On the other hand, the direct estimation methods (except the causal KNN regression with the improperly tuned $K = 425$) predict a steeper, monotonic relationship between s and $\bar{\tau}_s$.

In summary, all approaches to assess model fit in this section yield a consistent message. First, there are clear differences in the distribution of predicted conditional average treatment effects across the different estimation methods. Second, the direct estimation methods (except the transformed outcome regressions) have better model fit than the indirect methods. The treatment effect projections and the causal forest in particular provide the best model fit.

9 Targeting policy evaluation

The results in Section 8 indicate differences in the predictive fit across the estimation methods, with an overall advantage of the methods that are trained to directly predict the conditional average treatment effect. However, the ultimate criterion to evaluate the different methods is not model fit, but the distribution of the targeting profits that can be obtained using these methods. In this section we compare different targeting policies that are based on estimates of the conditional average treatment effect, $\tau(x)$. Given margin (m) and targeting cost (c) data, we predict an optimal targeting policy, d^*, such that $d^*(X_i) = 1$ if and only if $m\hat{\tau}(x) > c$. In our application, the targeting cost is identical across customers and we assume a constant margin since customer-specific margin data are not available. Because $c/m = 2.003$, a customer should

be targeted if and only if $\hat{\tau}(x) > 2.003$.[11]

The different estimation methods yield different predictions of the optimal targeting policies. We evaluate the profits of these policies using the inverse probability-weighted estimator of targeting profits, $\hat{\Pi}(d)$, that we introduced in Section 3:

$$\hat{\Pi}(d) = \sum_{i=1}^{N} \left(\frac{1 - W_i}{1 - e(X_i)} (1 - d(X_i)) \cdot \pi_i(0) + \frac{W_i}{e(X_i)} d(X_i) \cdot \pi_i(1) \right).$$

To account for the statistical uncertainty in the predictions, we employ the bootstrap algorithm outlined in Section 7.4.

9.1 Profit levels

Table 8 provides summary statistics of the distribution of profit levels implied by the different targeting policies. The profits are evaluated in the bootstrap test sets. Table 8 also provides information on the percentage of customers who are targeted under the alternative policies. To make the profit numbers more easily interpretable, we scale the predicted profits to a base of one million customers. Table 9 shows the mean difference in profit levels across the different targeting policies. Almost all pairwise profit differences are statistically different from zero.

The results indicate that targeting based on estimates of the conditional average treatment effect provides economic value to the company: for *all* such targeting policies, the mean profit level exceeds the mean profit level of no targeting (2.022 million dollars) and the mean profit level of a blanket mailing strategy that targets all customers (2.231 million).

We first compare the profit levels using targeting strategies based on an indirect estimate of the conditional average treatment effect. The mean profit level implied by the OLS estimates (2.365 million dollars) is smaller than the mean profit level based on the Lasso (2.408 million), logit/OLS (2.414 million), and the two forest method (2.425 million). On the other hand, the random forest estimates yield somewhat lower profits (2.356 million) compared to OLS. The difference in the profit predictions across the random forest and the two forest methods suggests that *forcing* each tree to split on the treatment indicator at the root node, which is what the two-forest method implicitly does, impacts on the ability to predict treatment effects and thus targeting profits.

All targeting policies based on the direct estimation methods (except the improperly tuned causal KNN regression and the transformed outcome regressions) outperform the policies based on the indirect estimation methods. The causal KNN regression yields mean profits of 2.454 million dollar when K is tuned using the sample analog of the transformed outcome loss ($K = 1700$). Regularizing the causal KNN predictions using the treatment effect projections further improves profits to 2.490 million when using the Lasso and 2.473 million when using a random forest. Finally, the targeting policy based on the causal forest estimates yields an additional improvement with mean profits of 2.496 million dollars.

[11] Our non-disclosure agreement with the company does not permit us to reveal the levels of the margin and targeting cost data.

The results in Table 8 reaffirm the importance of tuning the K value in the causal KNN regression based on the transformed outcome loss. Using the heuristically chosen value of $K = 425$, mean profits (2.401 million) are substantially smaller than the profits using the properly tuned K value. Also, as already discussed in Section 8.1, the poor performance of the transformed outcome regressions among the direct estimation methods is anticipated due to the large variance of the outcome levels relative to the variance of the conditional average treatment effects.

9.2 Profit curve when targeting the top ϕ percent of customers

To provide further insights on why some of the targeting policies yield larger profits than others, we compare the ability of the estimation methods to *sort* customers according to the predicted conditional average treatment effect. Consider an estimation method s. For any number $\phi \in [0, 100]$, the set $\mathcal{I}_s(\phi)$ includes the ϕ percent of units with the largest predicted conditional average treatment effects according to method s. The expected profit from targeting the top ϕ percent of customers according to method s is

$$\begin{aligned}
\mathbb{E}[\Pi_s(\phi)|X_1, \ldots, X_N] &= m \sum_{i \in \mathcal{I}_s(\phi)} \mathbb{E}[Y_i(1)|X_i] + m \sum_{i \in \mathcal{I}_s^c(\phi)} \mathbb{E}[Y_i(0)|X_i] - \phi N c \\
&= m \sum_{i \in \mathcal{I}_s(\phi)} \mathbb{E}[Y_i(1) - Y_i(0)|X_i] + m \sum_{i=1}^{N} \mathbb{E}[Y_i(0)|X_i] - \phi N c \\
&= m \sum_{i \in \mathcal{I}_s(\phi)} \tau(X_i) - \phi N c + m \sum_{i=1}^{N} \mathbb{E}[Y_i(0)|X_i].
\end{aligned}$$

Here, $\mathcal{I}_s^c(\phi)$ denotes the complement of the set $\mathcal{I}_s(\phi)$. Method s will yield larger expected profits than method s' when targeting the top ϕ percent of customers if and only if method s is better at sorting customers according to the conditional average treatment effect than method s',

$$\sum_{i \in \mathcal{I}_s(\phi)} \tau(X_i) > \sum_{i \in \mathcal{I}_{s'}(\phi)} \tau(X_i).$$

In particular, if we could predict $\tau(X_i)$ without error then we could sort customers perfectly. For any ϕ, we could determine the set $\mathcal{I}^*(\phi)$ or the ϕ percent of units with the largest actual conditional average treatment effects. Because

$$\sum_{i \in \mathcal{I}^*(\phi)} \tau(X_i) \geq \sum_{i \in \mathcal{I}_s(\phi)} \tau(X_i),$$

the expected profit from this perfect sort is at least as large as the expected profit from the sort of any other method, s. Furthermore, if method s is consistently better at sorting customers then method s' for any $\phi \in [0, 100]$, then $\mathbb{E}[\Pi_s(\phi)|X_1, \ldots, X_N]$ will attain a larger maximum than $\mathbb{E}[\Pi_{s'}(\phi)|X_1, \ldots, X_N]$ for a optimal, profit-maximizing targeting percentage for s is strictly between 0 and 100.

Using the inverse probability-weighted profit estimator, we obtain unbiased estimates of $\mathbb{E}[\Pi_s(\phi)|X_1, \ldots, X_N]$. Figure 7 shows the mean and the 95 percent range of $\hat{\Pi}_s(\phi)$ over a range of ϕ from 0 to 100 percent, separately for different estimation methods. As before, the mean and the 95 percent range are taken over all 1000 bootstrap test sets. All plots show the baseline profit when no customer is targeted and also indicate the maximum of the profit curves. The graphs reveal that the direct estimation methods (except the transformed outcome regression) have a superior ability to sort customers according to the predicted conditional average treatment effect compared to the indirect estimation methods. As a consequence, the mean of $\hat{\Pi}_s(\phi)$ attains a higher maximum for the main direct estimation methods.

In Figure 8 we plot the difference in the profit when targeting the top ϕ percent of customers between the causal forest and other estimation methods. The panels display the mean difference across the profit curves and its standard error over all bootstrap test sets. The causal forest dominates all other estimation methods, in the sense that it attains larger profits for any ϕ, although the treatment effect projections come close to the ability of the causal forest to sort customers. In particular, the difference in profits with respect to the treatment effect projections indicates that the causal forest is better at pinpointing the very top customers with the largest conditional average treatment effects, which accounts for the initial spikes in the profit differences with respect to the treatment effect projections.

9.3 Percentage of customers targeted

Table 8 shows that the policies based on the indirect estimation methods typically target at least 42 percent of customers.[12] In contrast, the policies based on the direct estimation methods (except the transformed outcome regressions and the improperly tuned causal KNN regression) target at most 37.4 percent of customers. This reflects the superior ability of the direct estimation methods to sort customers according to the predicted treatment effect, as discussed before in Section 9.2.

To provide some intuition behind this result we inspect the distribution of the conditional average treatment effects predicted by the best-performing methods, the treatment effect projections and the causal forest, in Figure 5. The histograms indicate a significant mass of customers with treatment effects just below the targeting threshold of 2.003. The mass of customers with treatment effects just above the threshold is substantial smaller, reflecting the skewness in the treatment effect distribution. The discussion of model fit in Section 8 suggests a substantial degree of noise in the treatment effect predictions by the indirect estimation methods. If this noise is approximately symmetrically distributed and independent of the true value of $\tau(x)$, then the fraction of "false positives," i.e. customers who are incorrectly classified as being profitable, is larger than the fraction of "false negatives," i.e. customers who are incorrectly classified as not being profitable. Consequently, the misclassification rate of the indirect methods and the percentage of customers targeted is larger compared to the direct estimation methods.

[12]The random forest, which performs poorly compared to all other policies based on customer-level treatment effects, is an outlier and predicts that on average 15.8 percent of customers should be targeted.

10 Comparison to traditional CRM: Targeting based on predicted spending or profit levels

The profit comparisons in Section 9 illustrate the economic value of optimal targeting policies based on conditional average treatment effects. In this section, we compare the value of such optimal targeting policies to a sophisticated application of the traditional CRM approach.

The traditional CRM approach is based on a scoring model that predicts the spending *level* Y_i (or the level of some other outcome of interest) given the observed customer features, X_i. Such a scoring model can be trained using any data with sufficient variation to estimate the conditional expectation $\mathbb{E}[Y_i|X_i]$. More specifically, the company that we collaborate with uses only the treated observations in a randomized sample to predict the customer *score*—the spending level conditional on being targeted, $\mathbb{E}[Y_i|X_i = x, W_i = 1]$. The company decomposes the conditional expectation of spending into the purchase incidence, estimated using logistic regression, and the conditional expectation of log spending conditional on a purchase, estimated using a linear regression model. The company uses a small subset of all available features that were chosen based on experience and experimentation with different model variants.

In a naive application of the traditional CRM approach, a customer will be targeted if the predicted profit level exceeds a threshold such as the cost of targeting,

$$m\mathbb{E}[Y_i|X_i = x, W_i = 1] > c.$$

The data science team at the company that we collaborate with understands that this approach is flawed. However, the customer score, $\mathbb{E}[Y_i|X_i = x, W_i = 1]$, may be correlated with the conditional average treatment effect. If this is the case, then the score is predictive of the conditional average treatment effect and can be used to select a group of customers to be targeted. Consider the following approach. First, estimate a scoring model, such as the model described above. Then sort the customers in the training set according to the estimated score, and predict the expected profit when targeting the top ϕ percent of customers according to the predicted score. We can predict this profit using the approach in Section 9.2, only that we now sort customers according to the predicted score, not the predicted treatment effect. Then, we search for the optimal targeting percentage ϕ^*, i.e. the targeting percentage that maximizes the expected profit in the training set. Finally, we implement a targeting policy that targets the top ϕ^* percent in the whole customer population. This approach is essentially the targeting strategy used by the company that we collaborate with. In practice, the company also considers other factors, such as a sales target or the long-run value of a customer, when choosing the targeting percentage. In the comparison below, however, we choose the optimal targeting percentage as described above.

We train the scoring model in each of the bootstrap training sets and then predict the expected profit in the corresponding test sets over a range of ϕ from 0 to 100 percent. Figure 9 displays the mean and the 95 percent range of the predicted expected profit over all 1000 bootstrap samples. The shape of the expected profit curve is similar to the shape of the expected profit curves when targeting customers based on the conditional average treatment effect (Figure 7). In particular,

for a range of targeting percentages, the company can increase profits by targeting customers according to their expected score. The increase in expected profits when ϕ increases from 0 to approximately 40 percent implies that in our application the spending level must be positively correlated with the conditional average treatment effect. Indeed, the correlation between the score and the causal forest prediction of the treatment effects is 0.269, and the correlation with the TEP-Lasso predictions is 0.699.[13]

Table 8 includes summary statistics of the profit levels obtained using the sophisticated scoring approach, and Table 9 includes the corresponding mean profit differences relative to the targeting strategies based on treatment effects. The sophisticated scoring approach yields larger profits than most targeting strategies based on the indirect estimation methods of the treatment effects, with the exception of the logit/OLS and two forest models. However, the direct estimation methods, excluding the transformed outcome regressions, outperform the scoring model. For example, the mean profit of the causal forest is 2.496 million dollars and the mean profit of the scoring model is 2.413 million dollars.

Hence, perhaps unexpectedly, even though the scoring model is not trained to predict incremental spending, it may yield larger profits than some of the targeting policies based on treatment effect predictions. To provide some intuition for why this occurs, we plot the difference in profits when targeting the top ϕ percent of customers according to the predicted treatment effect versus the top ϕ percent according to the predicted score in Figure 10. When the treatment effects are estimated using one of the indirect estimation methods, the profit difference increase for values of ϕ up to about 20 percent, but then declines and becomes negative over a large range of ϕ. Hence, the indirect estimation methods more accurately pinpoint customers with the largest treatment effects, whereas the scoring model sorts customers more accurately for values of ϕ larger than 20 percent. In contrast, the direct estimation methods mostly dominate the scoring model, i.e. the profit difference is positive for most values of ϕ (the transformed outcome regression is the usual exception). Hence, unlike some of the indirect estimation methods, the direct estimation methods are almost uniformly better at sorting customers according to incremental spending than the scoring model.

We provide some further intuition for the superiority of the direct estimation methods in Figure 11. The graph displays densities of the conditional average treatment effects predicted using the causal forest. We show the densities separately for 20 equal-sized segments based on the rank of the predicted score. For the top segments, i.e. for customers with the largest predicted spending levels, there is a large degree of heterogeneity in the predicted treatment effects. Thus, confirming with the premise of this paper, some customers with a high level of spending do not spend more when being targeted. For the segments with smaller scores, there is still a substantial degree of heterogeneity in the predicted treatment effects, but the distributions are less diffuse than for the top segments. Hence, although expected spending is positively correlated with the conditional average treatment effect, the scoring model is unable to distinguish between customers

[13]The predictions were obtained using ten-fold cross validation in the 2015 data. The 95 percent confidence interval for the correlation with the causal forest predictions is $(0.245, 0.293)$, and the confidence interval for the correlation with the TEP-Lasso predictions is $(0.684, 0.711)$.

with high spending levels who are either unresponsive or highly responsive to a targeting effort. On the other hand, the best-performing methods for the conditional average treatment effect are able to properly distinguish between such groups of customers.

Discussion and caveat

The results in this section show that the traditional CRM approach, when implemented in a sophisticated manner using an evaluation of targeting profits in a randomized sample, can be of value to a firm and improve the profits from a targeting campaign. This is of interest per se and may be one factor that explains why the approach has continued to be widely used in the industry until the present.

An important caveat is that the traditional CRM approach will not work if the predicted score is not correlated with the correct metric—the incremental effect of targeting as predicted by the conditional average treatment effect. A different way to put this is that a company needs to get lucky for the approach to work, because there is no reason why the score and the treatment effects needs to be correlated. More importantly, we have shown in our application that the targeting strategies based on the best-performing treatment effect estimation methods dominate the scoring approach, i.e. yield larger profits. Using available software, estimating the corresponding models is barely costlier in terms of computing time than estimating the scoring model. Based on these considerations, we strongly urge firms to abandon the traditional CRM approach in favor of a modern targeting strategy based on incrementality that is captured by conditional average treatment effects.

11 Transportability of results

So far we have evaluated the predictive fit and the profit differences across different targeting policies using the 2015 data. These results establish the external validity of the predictions because we performed all evaluations in a test set that is distinct from the training set used for estimation. We initially used the 2015 data only to avoid any confounders that reflect idiosyncratic changes in the customer population over time and not systematic differences in the model fit or profits across the methods. However, in an actual implementation of a targeting policy, the training data will be obtained in advance of the implementation data, and the relationship between customer features and outcomes may change over time (reflecting changes in customer preferences, competition, etc.). Hence, to assess a strict form of external validity or *transportability* of the estimation results, we compare the model fit and profit differences across estimation methods in the 2016 data that were collected one year after the 2015 training data. This comparison provides the truest assessment of the value of our proposed methods for customer targeting in practice.

We follow the exact same approach that we used in the previous sections, using a bootstrap algorithm with 1000 replications. The only difference is that now all models are trained using

all of the 2015 data and that all predictions are based on the 2016 data.[14] The distributions of the predicted conditional average treatment effects in 2016 (Figure 12 in the Appendix) and the percentage of negative treatment effects (Table 11 in the Appendix) for the different estimation methods are similar to their respective results in 2015. This similarity is expected, however, because both the 2015 and 2016 treatment effect predictions are obtained using the 2015 data. Hence, any significant difference in the predictions would have to reflect a significant difference in the distributions of the customer features.

Even more informative is a comparison of the mean-squared prediction error based on the transformed outcomes in panels I and II of Table 5 (pairwise mean-squared error differences for the 2016 results are shown in Table 12 in the Appendix). Note that there is a scale difference between the 2015 and 2016 results that appears to reflect a small scale difference in the treatment effects across years. Overall, as in 2015, the mean-squared error is smaller for the direct estimation methods compared to the indirect estimation methods, indicating better model fit. However, there are two notable differences. First, in 2016, the random forest has slightly better model fit than all direct estimation methods except the treatment effect projection using the Lasso, and second, the causal forest has a somewhat worse fit than the causal KNN regression and the treatment effect projections. The 2016 lift charts in Figure 13 in the Appendix also reveal an overall pattern that is similar to the 2015 lift charts in Figure 6. In particular, the direct estimation methods exhibit a steeper, monotonic relationship between the segment index s and the average treatment effect in segment s compared to the indirect estimation methods. Note that the random forest achieves the highest lift in the top segment ($s = 20$) compared to all other estimation methods, which may explain the good model fit based on the mean-squared transformed outcome error. However, apart from this top segment, the random forest appears to have less discriminative power compared to the direct estimation methods. Also, note that in 2016, the causal forest is less able to correctly predict the units with smaller treatment effects compared to its 2015 predictions. This result may explain its slightly worse model fit.

Ultimately, what matters is the comparison of the targeting profits in panels I and II of Table 8 and the corresponding mean difference in profit levels across targeting policies in Table 10. The overall scale of the profit levels is smaller in 2016, but otherwise the ranking of the estimation methods based on the evaluation of targeting profits is similar to the 2015 results. First, the profits from all targeting methods exceed the profits from no targeting or a blanket mailing strategy. Second, among the indirect estimation methods the OLS profits (2.027 million dollars) are smaller compared to the profits achieved using the Lasso (2.046 million) and the logit/OLS model (2.094 million). The ranking of OLS, Lasso, and logit/OLS according to mean profits is hence identical to the ranking in 2015. In contrast, in 2016 the mean profit level of the random forest exceeds the profit level of OLS, and the two forest model yields the smallest profit level among all targeting methods. Thus, the random forest methods trained to predict the outcome level appear less stable than the linear methods and the logit/OLS model. Third, the mean profit levels achieved using the direct estimation methods exceed the profit levels based on the

[14] The properly tuned value of K in the causal KNN regression is now larger ($K = 2475$) because we use training sets equal in size to the whole 2015 data.

indirect estimation methods, except for the improperly tuned causal KNN regression and the transformed outcome regressions. The treatment effect projection using the Lasso yields mean profits of 2.124 million dollars, which is marginally larger than the mean profits of the causal forest (2.121 million) and larger than the profits based on the treatment effect projection using the random forest (2.113 million). Although the difference across treatment effect projection using the Lasso and the causal forest is statistically significant at a 5 percent level, the profits are identical for all practical matters. Fourth, the sophisticated application of the traditional CRM approach predicts mean targeting profits of 2.089 million dollars, which—as in 2015—is only slightly less than the mean profits of the logit/OLS method. Regardless, all direct estimation methods (other than the transformed outcome regressions) predict mean targeting profits that are larger than the profits based on the scoring model.

We provide further comparisons of the targeting policies in the Appendix. Analogous to the analysis in Section 9.2, in Figure 14 we compare the ability of the methods to sort customers according to the predicted conditional average treatment effect. In Figure 15 we show the difference in profits when targeting the top ϕ percent using the causal forest predictions versus the other estimation methods, and in Figure 16 we show the corresponding difference in profits between the methods that predict treatment effects versus the scoring model.

In summary, we find that the differences in the mean profit levels across the different targeting policies in 2016 are almost identical to those that we documented in the 2015 data. In particular, the treatment effect projections and the causal forest predict optimal targeting policies achieve the largest profits. Hence, in our application, the main predictions pass a strict test of external validity.

12 Conclusions

We have presented a systematic, generally applicable approach to designing and evaluating optimal targeting policies. Our results clearly demonstrate that the economic value of an optimal targeting policy depends on the method to estimate the conditional average treatment effect from which the optimal targeting policy is constructed.

While machine learning methods may generally improve the profits from a targeting policy, a pronounced increase in profits only occurs when we use methods that draw on recent ideas from both the causal inference and machine learning literatures. In particular, the estimation methods that directly predict the conditional average treatment effect, including the causal forest of Wager and Athey (2017) and the treatment effect projection that we introduce in this paper, yield significantly larger profits than the conventional indirect estimation methods that have been used to predict treatment effects in the extant literature and in industry practice.

A central idea introduced in this paper to the field of marketing is that any targeting policy can be evaluated in a randomized sample using an inverse probability-weighted estimator of targeting profits. Without this approach the systematic comparison of a large number of targeting policies would have been prohibitively costly.

Note that the differences in model fit and profits across the targeting policies based on the direct and indirect estimation methods is an empirical finding, not a theorem. Indeed, the *no free lunch theorem* in machine learning (see Murphy 2012, Section 1.4.9 and the more formal treatment in Shalev-Shwartz and Ben-David 2014, Section 5.1) states that there is no universally best estimation algorithm. Consequently (and as always), the replication of the main results in our work using different applications and in different targeting contexts is necessary to validate our results and to further improve our understanding of the value and applicability of different methods for target marketing.

References

Ackerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007): "Econometric Tools for Analyzing Market Outcomes," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier B.V., vol. 6A, 4711–4276.

Ascarza, E. (2017): "Retention Futility: Targeting High Risk Customers Might be Ineffective," *Journal of Marketing Research (forthcoming)*.

Ascarza, E., P. Ebbes, O. Netzer, and M. Danielson (2016a): "Beyond the Target Customer: Social Effects of CRM Campaigns," *manuscript*.

Ascarza, E., P. S. Fader, and B. G. S. Hardie (2017): "Marketing Models for the Customer-Centric Firm," in *Handbook of Marketing Decision Models*, ed. by B. Wierenga and R. van der Lans, Springer, 297–329.

Ascarza, E., R. Iyengar, and M. Schleicher (2016b): "The Perils of Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment," *Journal of Marketing Research*, 53, 46–60.

Athey, S. and G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

Blake, T., C. Nosko, and S. Tadelis (2015): "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment," *Econometrica*, 83, 155–174.

Blattberg, R. C., B.-D. Kim, and S. A. Neslin (2008): *Database Marketing*, Springer.

Breiman, L. (2001): "Random Forests," *Machine Learning*, 45, 5–32.

Cao, W., A. A. Tsiatis, and M. Davidian (2009): "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data," *Biometrika*, 96, 723–734.

Goldfarb, A. and C. Tucker (2011): "Online Display Advertising: Targeting and Obtrusiveness," *Marketing Science*, 30, 389–404.

Gönül, F. F., B.-D. Kim, and M. Shi (2000): "Mailing Smarter to Catalog Customers," *Journal of Interactive Marketing*, 14, 2–16.

Guelman, L., M. Guillén, and A. M. Pérez-Marín (2015): "Uplift Random Forests," *Cybernetics and Systems: An International Journal*, 46, 230–248.

Hastie, T., R. Tibshirani, and M. Mainwright (2015): *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press.

Horvitz, D. G. and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Imbens, G. W. and D. R. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Kitagawa, T. and A. Tetenov (2017): "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *manuscript*.

Lambrecht, A. and C. Tucker (2013): "When Does Retargeting Work? Information Specificity in Online Advertising," *Journal of Marketing Research*, 50, 561–576.

Lewis, R. A. and D. H. Reiley (2014): "Online ads and offline sales: Measuring the effects of retail advertising via a controlled experiment on Yahoo!" *Quantitative Marketing and Economics*, 12, 235–266.

Murphy, K. P. (2012): *Machine Learning: A Probabilistic Perspective*, MIT Press.

Nair, H. S., S. Misra, W. J. Hornbuckle, IV, R. Mishra, and A. Acharya (2017): "Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation," *manuscript*.

Narayanan, S. and K. Kalyanam (2015): "Position Effects in Search Advertising and their Moderators: A Regression Discontinuity Approach," *Marketing Science*, 34, 388–407.

Rossi, P. E. (2014): "Even the Rich Can Make Themselves Poor: The Dangers of IV and Related Methods in Marketing Applications," *Marketing Science*, 33, 655–672.

Rzepakowski, P. and S. Jaroszewicz (2012): "Decision trees for uplift modeling with single and multiple treatments," *Knowledge and Information Systems*, 32, 303–327.

Sahni, N. S. (2015): "Effect of Temporal Spacing between Advertising Exposures: Evidence from Online Field Experiments," *Quantitative Marketing and Economics*, 13, 203–247.

Shalev-Shwartz, S. and S. Ben-David (2014): *Understanding Machine Learning*, Cambridge University Press.

Simester, D., A. Timoshenko, and S. I. Zoumpoulis (2016): "Customizing Marketing Decisions Using Field Experiments," *manuscript*.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B (Methodological)*, 58, 267–288.

Wager, S. and S. Athey (2017): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association (forthcoming)*.

Zantedeschi, D., E. McDonnell Feit, and E. T. Bradlow (2017): "Measuring Multichannel Advertising Response," *Management Science*, 63, 2706–2728.

Table 1: Summary statistics

	Mean	SD	N	Percentiles						
2015				1%	5%	25%	50%	75%	95%	99%
Spending	7.311	43.549	292657	0.000	0.000	0.000	0.000	0.000	39.950	182.990
Spending if purchased	117.725	132.445	18174	17.950	27.950	45.900	79.900	139.850	322.753	605.727
Purchase	0.062	0.241	292657	0.000	0.000	0.000	0.000	0.000	1.000	1.000
Treatment	0.669	0.471	292657	0.000	0.000	0.000	1.000	1.000	1.000	1.000
2016										
Spending	6.461	39.565	148200	0.000	0.000	0.000	0.000	0.000	36.981	168.000
Spending if purchased	115.066	124.021	8322	19.950	27.950	48.978	79.900	135.000	309.748	585.282
Purchase	0.056	0.230	148200	0.000	0.000	0.000	0.000	0.000	1.000	1.000
Treatment	0.663	0.473	148200	0.000	0.000	0.000	1.000	1.000	1.000	1.000

Table 2: Average treatment effect and differences among treated and untreated customers

	Mean	**SE**
2015		
ATE spending	2.561	0.166
ATE purchase	0.022	0.001
$\mathbb{E}[Y_i \vert Y_i > 0, W_i = 1] - \mathbb{E}[Y_i \vert Y_i > 0, W_i = 0]$	-0.798	2.456
2016		
ATE spending	2.377	0.210
ATE purchase	0.022	0.001
$\mathbb{E}[Y_i \vert Y_i > 0, W_i = 1] - \mathbb{E}[Y_i \vert Y_i > 0, W_i = 0]$	-2.170	3.383

Table 3: Covariate balance between treated and untreated units

Significance	Mean	SD	1%	5%	25%	50%	75%	95%	99%
Normalized difference									
2015									
All	0.5%	0.3%	0.0%	0.1%	0.2%	0.5%	0.7%	1.0%	1.3%
$p < 0.05$	1.0%	0.2%	0.8%	0.8%	0.8%	0.9%	1.1%	1.3%	1.7%
$p > 0.05$	0.4%	0.2%	0.0%	0.0%	0.2%	0.4%	0.6%	0.7%	0.8%
2016									
All	0.5%	0.4%	0.0%	0.0%	0.1%	0.4%	0.7%	1.3%	1.8%
$p < 0.05$	1.5%	0.4%	1.1%	1.1%	1.2%	1.3%	1.8%	2.3%	2.6%
$p > 0.05$	0.4%	0.3%	0.0%	0.0%	0.1%	0.3%	0.6%	1.1%	1.4%
Log ratio of SD									
2015									
All	4.9%	10.4%	0.0%	0.1%	0.8%	2.2%	5.4%	16.2%	41.8%
$p < 0.05$	4.7%	8.4%	0.0%	0.2%	1.4%	2.2%	4.4%	19.8%	43.2%
$p > 0.05$	5.0%	10.8%	0.0%	0.1%	0.7%	2.2%	5.6%	15.9%	29.9%
2016									
All	5.8%	8.2%	0.0%	0.1%	0.9%	2.7%	7.0%	24.1%	34.7%
$p < 0.05$	10.9%	12.7%	0.3%	0.3%	1.6%	2.7%	22.0%	31.1%	38.0%
$p > 0.05$	5.5%	7.8%	0.0%	0.1%	0.9%	2.7%	6.7%	22.3%	33.1%

Table 4: Predicted conditional average treatment effects

	Mean	SE	SD	$\tau(\hat{X}_i) < 0$	Percentiles						
					1%	5%	25%	50%	75%	95%	99%
OLS	2.42	0.19	5.99	20.3%	-10.49	-3.52	0.36	1.64	3.75	10.92	21.90
Lasso	2.34	0.16	4.94	18.4%	-7.24	-2.05	0.41	1.41	3.30	9.90	19.68
Logit/OLS	2.43	0.30	9.34	7.7%	-7.98	-0.45	0.64	1.55	3.09	8.70	23.61
Random forest	1.30	0.09	2.83	22.3%	-1.77	-0.72	0.04	0.52	1.50	6.05	13.75
Two forests	2.79	0.24	7.71	19.7%	-10.02	-2.32	0.15	1.34	4.02	13.14	27.49
Causal KNN (K = 2225)	2.17	0.09	2.79	4.3%	-0.43	0.04	0.62	1.31	2.58	7.57	14.43
Causal KNN (K = 525)	2.30	0.11	3.57	15.6%	-2.62	-0.82	0.38	1.37	3.00	8.81	17.33
TOR (Lasso)*	2.56	0.14	4.44	9.8%	-6.41	-1.22	1.03	1.87	3.19	8.99	17.75
TOR (random forest)*	2.78	0.42	13.27	21.7%	-18.49	-4.29	0.09	1.38	4.39	15.25	34.20
TEP (Lasso, K = 2225)†	2.17	0.08	2.58	1.3%	-0.03	0.20	0.65	1.39	2.73	6.76	11.98
TEP (random forest, K = 2225)†	2.17	0.08	2.54	1.2%	-0.04	0.21	0.74	1.35	2.55	7.18	13.23
Causal forest	2.46	0.10	3.10	1.8%	-0.67	0.68	1.05	1.45	2.76	8.18	14.92

*TOR - transformed Outcome Regression
†TEP - treatment Effect Projection

Table 5: Mean-squared prediction error (MSE)

	Mean	**SE**	**SD**	**Percentiles**							
				1%	5%	10%	25%	50%	75%	95%	99%
[1] **2015**											
OLS	8256.5	28.7	907.1	6555.8	6925.0	7170.6	7591.0	8183.8	8796.2	9925.5	10818.0
Lasso	8240.2	28.8	909.4	6528.1	6908.6	7154.4	7581.0	8160.6	8784.9	9936.9	10805.1
Logit/OLS	8371.9	28.8	909.2	6645.7	7033.1	7294.7	7698.7	8277.5	8907.9	10045.9	10921.7
Random forest	8229.9	28.8	909.6	6475.4	6902.9	7142.5	7571.6	8148.3	8777.3	9952.2	10821.4
Two forests	8265.6	28.7	907.7	6560.2	6930.2	7178.7	7612.0	8182.3	8801.8	9964.3	10855.8
Causal KNN (K = 1700)	8227.1	28.9	912.7	6463.1	6892.0	7138.0	7563.8	8149.0	8773.5	9955.7	10819.1
Causal KNN (K = 425)	8232.4	28.9	913.3	6467.5	6895.3	7145.3	7565.3	8155.4	8775.7	9958.7	10836.8
TOR (Lasso)*	8258.1	28.6	905.1	6539.1	6925.8	7184.6	7596.1	8174.6	8795.2	9979.7	10791.3
TOR (random forest)*	8420.4	28.2	893.3	6750.2	7091.5	7349.0	7772.6	8336.6	8962.5	10087.2	10894.8
TEP (Lasso, K = 1700)†	8230.6	28.9	913.2	6464.6	6895.8	7141.8	7567.3	8152.7	8782.3	9956.4	10818.8
TEP (random forest, K = 1700)†	8226.7	28.9	912.5	6464.0	6892.0	7137.8	7564.1	8148.4	8775.3	9952.6	10820.2
Causal forest	8221.1	28.8	909.6	6480.4	6889.4	7133.3	7561.6	8141.7	8767.7	9943.1	10805.3
[2] **2016**											
OLS	6723.5	26.0	823.1	5392.9	5600.8	5771.8	6112.3	6586.6	7208.0	8234.6	9122.5
Lasso	6716.4	26.0	823.6	5376.8	5596.3	5766.5	6102.1	6576.5	7202.2	8233.3	9118.8
Logit/OLS	6755.2	25.7	812.4	5427.7	5653.4	5825.7	6139.2	6623.5	7227.2	8243.7	9168.2
Random forest	6701.6	26.1	825.2	5361.6	5580.1	5745.4	6073.0	6559.8	7181.4	8218.9	9102.5
Two forests	6785.5	26.4	833.4	5425.5	5655.8	5824.6	6169.1	6648.1	7279.6	8330.7	9192.3
Causal KNN (K = 2475)	6702.4	26.1	826.0	5363.4	5577.7	5743.1	6069.8	6559.5	7182.7	8223.4	9105.3
Causal KNN (K = 425)	6707.0	26.1	825.4	5368.2	5586.3	5750.5	6079.0	6562.1	7190.8	8227.2	9110.4
TOR (Lasso)*	6738.7	26.0	823.2	5385.4	5628.3	5781.9	6116.8	6607.5	7209.0	8251.0	9136.6
TOR (random forest)*	6978.2	26.9	850.6	5580.6	5847.8	5998.3	6347.3	6845.5	7496.2	8561.6	9469.8
TEP (Lasso, K = 2475)†	6698.8	26.1	826.2	5358.4	5575.4	5741.9	6063.2	6555.5	7179.8	8219.7	9104.3
TEP (random forest, K = 2475)†	6703.0	26.1	826.8	5364.0	5577.8	5743.8	6068.8	6561.1	7184.4	8223.2	9109.4
Causal forest	6707.7	26.1	826.7	5373.1	5586.6	5751.5	6084.3	6566.2	7196.0	8234.9	9107.8

*TOR - transformed outcome regression
†TEP - treatment effect projection

Table 6: Mean-squared prediction error (MSE) differences across estimation methods

	A	B	C	D	E	F	G	H	I	J	K	L
A	-	16.3	-115.4	26.6	-9.1	29.4	24.1	-1.6	-163.9	25.9	29.8	35.4
	-	*(0.4)*	*(2.5)*	*(0.9)*	*(1.2)*	*(0.9)*	*(0.9)*	*(1.2)*	*(3.8)*	*(1.0)*	*(0.9)*	*(0.9)*
B	-16.3	-	-131.7	10.4	-25.3	13.1	7.9	-17.8	-180.1	9.7	13.6	19.1
	(0.4)	-	*(2.7)*	*(0.7)*	*(1.2)*	*(0.8)*	*(0.8)*	*(1.1)*	*(3.8)*	*(0.8)*	*(0.8)*	*(0.7)*
C	115.4	131.7	-	142.0	106.3	144.8	139.5	113.8	-48.5	141.3	145.2	150.8
	(2.5)	*(2.7)*	-	*(3.1)*	*(3.0)*	*(3.1)*	*(3.1)*	*(3.0)*	*(4.5)*	*(3.1)*	*(3.1)*	*(3.0)*
D	-26.6	-10.4	-142.0	-	-35.7	2.8	-2.5	-28.2	-190.5	-0.7	3.2	8.8
	(0.9)	*(0.7)*	*(3.1)*	-	*(1.3)*	*(0.2)*	*(0.3)*	*(1.4)*	*(4.0)*	*(0.3)*	*(0.2)*	*(0.3)*
E	9.1	25.3	-106.3	35.7	-	38.5	33.2	7.5	-154.8	35.0	38.9	44.5
	(1.2)	*(1.2)*	*(3.0)*	*(1.3)*	-	*(1.3)*	*(1.3)*	*(1.6)*	*(3.4)*	*(1.4)*	*(1.3)*	*(1.2)*
F	-29.4	-13.1	-144.8	-2.8	-38.5	-	-5.3	-31.0	-193.3	-3.5	0.4	6.0
	(0.9)	*(0.8)*	*(3.1)*	*(0.2)*	*(1.3)*	-	*(0.1)*	*(1.5)*	*(4.1)*	*(0.2)*	*(0.0)*	*(0.3)*
G	-24.1	-7.9	-139.5	2.5	-33.2	5.3	-	-25.7	-188.0	1.8	5.7	11.3
	(0.9)	*(0.8)*	*(3.1)*	*(0.3)*	*(1.3)*	*(0.1)*	-	*(1.5)*	*(4.1)*	*(0.2)*	*(0.2)*	*(0.4)*
H	1.6	17.8	-113.8	28.2	-7.5	31.0	25.7	-	-162.3	27.5	31.4	37.0
	(1.2)	*(1.1)*	*(3.0)*	*(1.4)*	*(1.6)*	*(1.5)*	*(1.5)*	-	*(3.6)*	*(1.5)*	*(1.5)*	*(1.4)*
I	163.9	180.1	48.5	190.5	154.8	193.3	188.0	162.3	-	189.8	193.7	199.3
	(3.8)	*(3.8)*	*(4.5)*	*(4.0)*	*(3.4)*	*(4.1)*	*(4.1)*	*(3.6)*	-	*(4.1)*	*(4.1)*	*(3.9)*
J	-25.9	-9.7	-141.3	0.7	-35.0	3.5	-1.8	-27.5	-189.8	-	3.9	9.5
	(1.0)	*(0.8)*	*(3.1)*	*(0.3)*	*(1.4)*	*(0.2)*	*(0.2)*	*(1.5)*	*(4.1)*	-	*(0.2)*	*(0.4)*
K	-29.8	-13.6	-145.2	-3.2	-38.9	-0.4	-5.7	-31.4	-193.7	-3.9	-	5.6
	(0.9)	*(0.8)*	*(3.1)*	*(0.2)*	*(1.3)*	*(0.0)*	*(0.2)*	*(1.5)*	*(4.1)*	*(0.2)*	-	*(0.3)*
L	-35.4	-19.1	-150.8	-8.8	-44.5	-6.0	-11.3	-37.0	-199.3	-9.5	-5.6	-
	(0.9)	*(0.7)*	*(3.0)*	*(0.3)*	*(1.2)*	*(0.3)*	*(0.4)*	*(1.4)*	*(3.9)*	*(0.4)*	*(0.3)*	-

Model Key

A: OLS	B: Lasso	C: Logit/OLS
D: Random forest	E: Two forests	F: Causal KNN (K = 1700)
G: Causal KNN (K = 425)	H: TOR (Lasso)*	I: TOR (random forest)*
J: TEP (Lasso, K = 1700)†	K: TEP (random forest, K = 1700)†	L: Causal forest

*TOR - transformed outcome regression

†TEP - treatment effect projection

Note: The values indicate the mean difference between the MSE of the row estimation method and the MSE of the column estimation method. The standard error of the mean difference is in parentheses.

Table 7: Lift factors

	Segment																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
OLS	1.46	1.70	1.49	1.32	1.25	1.16	1.09	1.08	1.17	1.17	1.35	1.41	1.67	1.89	2.21	2.61	3.31	4.28	6.16	12.94
	(0.09)	*(0.05)*	*(0.04)*	*(0.04)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.05)*	*(0.09)*
Lasso	1.00	1.54	1.29	1.16	1.03	0.99	0.99	0.99	1.01	1.13	1.24	1.37	1.56	1.90	2.13	2.67	3.31	4.64	6.79	13.79
	(0.09)	*(0.05)*	*(0.04)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.05)*	*(0.06)*	*(0.09)*
Logit/OLS	2.27	1.20	0.99	0.83	0.84	0.78	0.83	0.89	0.94	1.07	1.24	1.26	1.43	1.67	2.05	2.72	3.60	4.86	7.38	13.20
	(0.10)	*(0.04)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.05)*	*(0.10)*
Random forest	0.86	1.65	1.48	1.26	1.06	1.00	1.13	1.15	1.36	1.40	1.57	1.69	1.77	2.00	2.16	2.62	3.19	4.17	6.24	13.67
	(0.08)	*(0.06)*	*(0.05)*	*(0.05)*	*(0.04)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*
Two forests	0.83	1.41	1.12	0.83	0.68	0.80	0.90	1.03	1.18	1.27	1.44	1.58	1.81	2.10	2.47	3.00	3.66	4.52	6.48	13.52
	(0.09)	*(0.04)*	*(0.04)*	*(0.03)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.05)*	*(0.06)*	*(0.09)*
Causal KNN (K = 1700)	0.79	0.84	0.78	0.80	0.87	0.92	0.97	1.02	1.11	1.15	1.30	1.52	1.68	1.98	2.32	2.88	3.39	4.13	6.48	14.97
	(0.04)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.06)*	*(0.07)*	*(0.11)*
Causal KNN (K = 425)	1.60	1.19	1.12	0.96	0.96	1.04	1.05	1.18	1.23	1.37	1.53	1.61	1.75	1.96	2.34	2.74	3.32	4.30	6.39	12.81
	(0.06)	*(0.04)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.05)*	*(0.06)*	*(0.11)*
TOR (LASSO)	1.79	1.59	1.40	1.17	1.10	1.12	1.13	1.10	1.18	1.26	1.37	1.47	1.67	2.03	2.25	2.68	3.35	4.35	6.21	12.27
	(0.10)	*(0.05)*	*(0.04)*	*(0.04)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.06)*	*(0.09)*
TOR (random forest)	1.31	1.30	1.24	1.05	0.73	0.71	0.84	0.97	1.20	1.30	1.42	1.62	1.81	2.10	2.50	3.06	3.78	4.54	6.30	12.62
	(0.09)	*(0.05)*	*(0.04)*	*(0.03)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.06)*	*(0.10)*
TEP (Lasso, K = 1700)	0.39	0.59	0.62	0.68	0.74	0.73	0.75	0.73	0.92	1.07	1.30	1.47	1.72	2.02	2.33	2.99	4.05	5.29	7.53	13.88
	(0.03)	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.05)*	*(0.05)*	*(0.07)*	*(0.11)*
TEP (random forest, K = 1700)	0.58	0.66	0.71	0.78	0.83	0.84	0.98	0.98	1.10	1.07	1.25	1.43	1.46	1.85	2.19	2.78	3.70	4.49	6.69	15.51
	(0.02)	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.05)*	*(0.05)*	*(0.07)*	*(0.11)*
Causal forest	-0.40	0.87	0.71	0.70	0.75	0.73	0.77	0.83	0.94	1.09	1.19	1.38	1.57	1.90	2.28	2.91	3.79	4.98	7.05	16.25
	(0.09)	*(0.04)*	*(0.03)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.04)*	*(0.05)*	*(0.06)*	*(0.07)*	*(0.09)*

* TOR - transformed outcome regression
† TEP - treatment effect projection
Note: Standard errors are in parentheses.

Table 8: Profit levels

	Profit			% Mailed		Profit Percentiles							
	Mean	SE	SD	Mean	SE	1%	5%	10%	25%	50%	75%	95%	99%
2015													
No mailing	2022.4	2.8	87.5	0.0%	0.0%	1841.0	1879.3	1909.5	1963.3	2020.0	2080.6	2164.6	2234.6
Blanket mailing	2230.5	2.0	63.8	100.0%	0.0%	2082.9	2125.6	2149.3	2187.1	2231.5	2273.6	2337.2	2375.6
OLS	2364.6	2.8	87.6	46.8%	0.1%	2156.4	2223.8	2254.8	2305.6	2360.9	2420.5	2511.9	2582.7
Lasso	2408.1	2.9	90.2	42.1%	0.2%	2204.8	2260.8	2286.4	2348.3	2410.2	2468.2	2558.0	2621.4
Logit/OLS	2414.2	2.5	78.4	42.4%	0.1%	2246.0	2287.9	2317.7	2360.2	2414.1	2467.9	2543.4	2617.2
Random forest	2356.2	2.9	91.6	15.8%	0.1%	2160.7	2216.5	2240.7	2294.1	2350.9	2416.9	2513.3	2585.7
Two forests	2425.1	2.6	83.5	42.6%	0.1%	2236.2	2290.4	2315.8	2367.7	2425.3	2480.5	2560.3	2613.1
Causal KNN (K = 425)	2401.0	2.4	77.0	39.0%	0.1%	2223.8	2271.2	2297.2	2349.2	2403.0	2453.5	2529.8	2569.0
Causal KNN (K = 1700)	2454.0	2.4	74.4	35.0%	0.1%	2285.2	2333.0	2359.7	2402.3	2455.4	2506.3	2573.0	2620.6
TOR (Lasso)*	2355.9	3.1	98.4	49.3%	0.3%	2111.6	2191.2	2230.1	2290.7	2354.2	2424.8	2506.0	2582.0
TOR (random forest)*	2411.0	2.6	81.2	43.3%	0.0%	2225.4	2278.3	2307.8	2356.3	2411.9	2463.3	2552.2	2602.6
TEP (Lasso, K = 1700)†	2489.9	2.2	69.4	36.2%	0.1%	2335.8	2375.2	2400.6	2442.3	2488.8	2540.6	2601.1	2638.8
TEP (random forest, K = 1700)†	2472.5	2.3	71.2	34.6%	0.1%	2307.7	2354.0	2379.1	2424.2	2473.6	2521.0	2588.5	2642.8
Causal forest	2495.5	2.6	82.7	37.4%	0.1%	2285.8	2363.4	2394.1	2440.6	2491.3	2550.6	2632.8	2686.9
Scoring model	2413.2	2.2	68.4	39.3%	0.3%	2260.3	2296.7	2322.9	2367.2	2413.8	2460.2	2524.0	2563.8
2016													
No mailing	1768.4	1.9	60.2	0.0%	0.0%	1640.7	1670.2	1692.7	1727.9	1767.0	1806.9	1867.8	1916.6
Blanket mailing	1904.5	1.5	46.2	100.0%	0.0%	1795.8	1830.0	1846.7	1872.9	1904.2	1934.6	1980.2	2008.4
OLS	2027.1	2.1	64.9	45.5%	0.1%	1876.7	1923.4	1944.3	1983.9	2023.7	2071.1	2133.1	2185.1
Lasso	2046.5	2.1	67.3	43.2%	0.1%	1895.9	1940.0	1961.0	2000.9	2045.3	2091.9	2158.6	2217.9
Logit/OLS	2093.9	2.0	63.7	41.9%	0.1%	1953.5	1991.0	2014.8	2049.5	2091.3	2136.2	2200.4	2246.4
Random forest	2040.8	2.0	64.7	32.0%	0.1%	1890.9	1934.4	1957.3	1996.4	2040.6	2084.2	2148.3	2181.6
Two forests	1965.4	1.9	59.0	62.7%	0.1%	1840.9	1875.1	1891.2	1921.7	1962.4	2005.3	2063.9	2116.0
Causal KNN (K = 425)	2076.9	2.0	62.0	39.7%	0.0%	1949.1	1979.2	1998.9	2034.2	2071.9	2121.9	2182.8	2228.8
Causal KNN (K = 2475)	2099.0	1.7	55.2	34.6%	0.1%	1975.9	2016.0	2034.7	2061.3	2095.0	2131.8	2196.8	2243.4
TOR (Lasso)*	2013.0	2.1	67.6	47.4%	0.2%	1861.3	1905.3	1930.2	1967.2	2010.9	2057.5	2127.9	2194.3
TOR (random forest)*	1977.0	1.8	57.5	60.7%	0.1%	1853.7	1886.1	1906.3	1936.2	1975.7	2016.0	2073.5	2114.8
TEP (Lasso, K = 2475)†	2124.4	1.6	49.8	37.9%	0.1%	2013.6	2044.6	2063.5	2089.5	2122.5	2157.3	2207.9	2247.2
TEP (random forest, K = 2475)†	2113.0	1.5	48.7	36.7%	0.1%	2004.5	2033.6	2051.9	2080.2	2112.2	2144.8	2197.4	2231.8
Causal forest	2121.0	1.8	57.9	37.2%	0.1%	1998.9	2031.2	2048.8	2083.2	2117.6	2157.3	2220.1	2266.2
Scoring model	2089.1	1.5	48.9	39.9%	0.2%	1973.3	2009.4	2026.8	2054.8	2089.8	2122.9	2168.6	2194.5

*TOR - transformed outcome regression
†TEP - treatment effect projection

Table 9: Differences in profit levels across estimation methods

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	-	-208.1	-342.2	-385.7	-391.8	-333.8	-402.6	-378.5	-431.5	-333.5	-388.6	-458.4	-438.9	-467.5	-450.0	-473.0	-390.8
	-	*(3.4)*	*(2.5)*	*(2.7)*	*(2.5)*	*(1.3)*	*(2.7)*	*(2.9)*	*(3.0)*	*(3.1)*	*(2.6)*	*(2.9)*	*(3.0)*	*(2.9)*	*(3.0)*	*(2.8)*	*(3.2)*
B	208.1	-	-134.1	-177.6	-183.7	-125.7	-194.5	-170.4	-223.4	-125.4	-180.5	-250.3	-230.8	-259.4	-241.9	-264.9	-182.7
	(3.4)	-	*(2.7)*	*(2.8)*	*(2.4)*	*(3.3)*	*(2.4)*	*(2.1)*	*(1.9)*	*(3.0)*	*(2.4)*	*(1.8)*	*(1.8)*	*(1.7)*	*(1.7)*	*(2.3)*	*(1.3)*
C	342.2	134.1	-	-43.5	-49.6	8.4	-60.4	-36.3	-89.3	8.7	-46.4	-116.2	-96.7	-125.3	-107.8	-130.8	-48.6
	(2.5)	*(2.7)*	-	*(1.6)*	*(1.9)*	*(2.5)*	*(2.5)*	*(2.7)*	*(2.7)*	*(2.4)*	*(2.6)*	*(2.5)*	*(2.7)*	*(2.6)*	*(2.7)*	*(2.4)*	*(2.8)*
D	385.7	177.6	43.5	-	-6.1	51.9	-16.9	7.2	-45.8	52.2	-2.9	-72.7	-53.2	-81.8	-64.3	-87.3	-5.1
	(2.7)	*(2.8)*	*(1.6)*	-	*(2.2)*	*(2.6)*	*(2.6)*	*(2.8)*	*(2.8)*	*(2.3)*	*(2.7)*	*(2.5)*	*(2.7)*	*(2.6)*	*(2.7)*	*(2.4)*	*(2.8)*
E	391.8	183.7	49.6	6.1	-	58.0	-10.9	13.2	-39.8	58.3	3.2	-66.7	-47.2	-75.7	-58.3	-81.3	1.0
	(2.5)	*(2.4)*	*(1.9)*	*(2.2)*	-	*(2.5)*	*(2.3)*	*(2.4)*	*(2.4)*	*(2.7)*	*(2.3)*	*(2.1)*	*(2.3)*	*(2.2)*	*(2.3)*	*(2.3)*	*(2.4)*
F	333.8	125.7	-8.4	-51.9	-58.0	-	-68.9	-44.8	-97.8	0.3	-54.8	-124.7	-105.2	-133.7	-116.3	-139.3	-57.0
	(1.3)	*(3.3)*	*(2.5)*	*(2.6)*	*(2.5)*	-	*(2.5)*	*(2.7)*	*(2.8)*	*(3.0)*	*(2.5)*	*(2.8)*	*(2.8)*	*(2.8)*	*(2.8)*	*(2.6)*	*(3.1)*
G	402.6	194.5	60.4	16.9	10.9	68.9	-	24.1	-28.9	69.2	14.1	-55.8	-36.3	-64.8	-47.4	-70.4	11.8
	(2.7)	*(2.4)*	*(2.5)*	*(2.6)*	*(2.3)*	*(2.5)*	-	*(2.3)*	*(2.3)*	*(2.9)*	*(1.9)*	*(2.2)*	*(2.3)*	*(2.2)*	*(2.3)*	*(2.1)*	*(2.4)*
H	378.5	170.4	36.3	-7.2	-13.2	44.8	-24.1	-	-53.0	45.1	-10.1	-79.9	-60.4	-88.9	-71.5	-94.5	-12.3
	(2.9)	*(2.1)*	*(2.7)*	*(2.8)*	*(2.4)*	*(2.7)*	*(2.3)*	-	*(1.9)*	*(3.1)*	*(2.2)*	*(1.9)*	*(1.7)*	*(1.9)*	*(1.8)*	*(2.3)*	*(2.1)*
I	431.5	223.4	89.3	45.8	39.8	97.8	28.9	53.0	-	98.1	42.9	-26.9	-7.4	-35.9	-18.5	-41.5	40.7
	(3.0)	*(1.9)*	*(2.7)*	*(2.8)*	*(2.4)*	*(2.8)*	*(2.3)*	*(1.9)*	-	*(3.0)*	*(2.3)*	*(1.6)*	*(1.5)*	*(1.5)*	*(1.3)*	*(2.2)*	*(1.8)*
J	333.5	125.4	-8.7	-52.2	-58.3	-0.3	-69.2	-45.1	-98.1	-	-55.1	-125.0	-105.5	-134.0	-116.6	-139.6	-57.3
	(3.1)	*(3.0)*	*(2.4)*	*(2.3)*	*(2.7)*	*(3.0)*	*(2.9)*	*(3.1)*	*(3.0)*	-	*(3.0)*	*(2.8)*	*(3.0)*	*(2.9)*	*(3.0)*	*(2.8)*	*(3.0)*
K	388.6	180.5	46.4	2.9	-3.2	54.8	-14.1	10.1	-42.9	55.1	-	-69.8	-50.3	-78.9	-61.4	-84.4	-2.2
	(2.6)	*(2.4)*	*(2.6)*	*(2.7)*	*(2.3)*	*(2.5)*	*(1.9)*	*(2.2)*	*(2.3)*	*(3.0)*	-	*(2.2)*	*(2.3)*	*(2.2)*	*(2.3)*	*(2.1)*	*(2.4)*
L	458.4	250.3	116.2	72.7	66.7	124.7	55.8	79.9	26.9	125.0	69.8	-	19.5	-9.0	8.4	-14.6	67.6
	(2.9)	*(1.8)*	*(2.5)*	*(2.5)*	*(2.1)*	*(2.8)*	*(2.2)*	*(1.9)*	*(1.6)*	*(2.8)*	*(2.2)*	-	*(1.5)*	*(1.0)*	*(1.4)*	*(2.1)*	*(1.7)*
M	438.9	230.8	96.7	53.2	47.2	105.2	36.3	60.4	7.4	105.5	50.3	-19.5	-	-28.5	-11.1	-34.1	48.1
	(3.0)	*(1.8)*	*(2.7)*	*(2.7)*	*(2.3)*	*(2.8)*	*(2.3)*	*(1.7)*	*(1.5)*	*(3.0)*	*(2.3)*	*(1.5)*	-	*(1.4)*	*(1.2)*	*(2.1)*	*(1.8)*
N	467.5	259.4	125.3	81.8	75.7	133.7	64.8	88.9	35.9	134.0	78.9	9.0	28.5	-	17.4	-5.6	76.7
	(2.9)	*(1.7)*	*(2.6)*	*(2.6)*	*(2.2)*	*(2.8)*	*(2.2)*	*(1.9)*	*(1.5)*	*(2.9)*	*(2.2)*	*(1.0)*	*(1.4)*	-	*(1.1)*	*(2.1)*	*(1.5)*
O	450.0	241.9	107.8	64.3	58.3	116.3	47.4	71.5	18.5	116.6	61.4	-8.4	11.1	-17.4	-	-23.0	59.2
	(3.0)	*(1.7)*	*(2.7)*	*(2.7)*	*(2.3)*	*(2.8)*	*(2.3)*	*(1.8)*	*(1.3)*	*(3.0)*	*(2.3)*	*(1.4)*	*(1.2)*	*(1.1)*	-	*(2.1)*	*(1.6)*
P	473.0	264.9	130.8	87.3	81.3	139.3	70.4	94.5	41.5	139.6	84.4	14.6	34.1	5.6	23.0	-	82.2
	(2.8)	*(2.3)*	*(2.4)*	*(2.4)*	*(2.3)*	*(2.6)*	*(2.1)*	*(2.3)*	*(2.2)*	*(2.8)*	*(2.1)*	*(2.1)*	*(2.1)*	*(2.1)*	*(2.1)*	-	*(2.2)*
Q	390.8	182.7	48.6	5.1	-1.0	57.0	-11.8	12.3	-40.7	57.3	2.2	-67.6	-48.1	-76.7	-59.2	-82.2	-
	(3.2)	*(1.3)*	*(2.8)*	*(2.8)*	*(2.4)*	*(3.1)*	*(2.4)*	*(2.1)*	*(1.8)*	*(3.0)*	*(2.4)*	*(1.7)*	*(1.8)*	*(1.5)*	*(1.6)*	*(2.2)*	-

Model Key

A: No mailing | B: Blanket mailing | C: OLS | D: Lasso

E: Logit/OLS | F: Random forest | G: Two forests | H: Causal KNN (K = 425)

I: Causal KNN (K = 1700) | J: TOR (Lasso)* | K: TOR (random forest)* | L: TEP (Lasso, K = 425)†

M: TEP (random forest, K = 425)† | N: TEP (Lasso, K = 1700)† | O: TEP (random forest, K = 1700)† | P: Causal forest

Q: Scoring model

*TOR - transformed outcome regression

†TEP - treatment effect projection

Note: The values indicate the mean difference between the profit of the row estimation method and the profit of the column estimation method. The standard error of the mean difference is in parentheses.

Table 10: 2016 differences in profit levels across estimation methods

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	-	-136.1	-258.7	-278.0	-325.5	-272.4	-197.0	-308.5	-330.6	-244.5	-208.6	-345.7	-328.2	-355.9	-344.6	-352.6	-320.7
	-	*(2.4)*	*(1.9)*	*(2.0)*	*(1.8)*	*(1.3)*	*(2.1)*	*(2.1)*	*(2.1)*	*(2.0)*	*(2.1)*	*(2.1)*	*(2.2)*	*(2.1)*	*(2.2)*	*(2.1)*	*(2.2)*
B	136.1	-	-122.6	-141.9	-189.4	-136.3	-60.9	-172.4	-194.5	-108.4	-72.5	-209.6	-192.1	-219.8	-208.5	-216.5	-184.6
	(2.4)	-	*(2.0)*	*(2.1)*	*(1.9)*	*(2.3)*	*(1.7)*	*(1.7)*	*(1.5)*	*(2.1)*	*(1.7)*	*(1.4)*	*(1.3)*	*(1.2)*	*(1.1)*	*(1.7)*	*(1.0)*
C	258.7	122.6	-	-19.3	-66.8	-13.7	61.7	-49.8	-71.9	14.1	50.1	-87.0	-69.5	-97.2	-85.9	-93.9	-62.0
	(1.9)	*(2.0)*	-	*(1.1)*	*(1.5)*	*(2.0)*	*(2.0)*	*(2.1)*	*(2.1)*	*(1.6)*	*(2.1)*	*(1.9)*	*(2.1)*	*(1.9)*	*(2.0)*	*(2.0)*	*(2.1)*
D	278.0	141.9	19.3	-	-47.5	5.7	81.1	-30.5	-52.6	33.5	69.4	-67.7	-50.1	-77.9	-66.6	-74.6	-42.7
	(2.0)	*(2.1)*	*(1.1)*	-	*(1.7)*	*(2.1)*	*(2.1)*	*(2.1)*	*(2.1)*	*(1.6)*	*(2.2)*	*(2.0)*	*(2.1)*	*(2.0)*	*(2.1)*	*(2.0)*	*(2.2)*
E	325.5	189.4	66.8	47.5	-	53.1	128.5	17.0	-5.1	80.9	116.9	-20.2	-2.7	-30.4	-19.1	-27.1	4.8
	(1.8)	*(1.9)*	*(1.5)*	*(1.7)*	-	*(1.8)*	*(1.9)*	*(1.9)*	*(1.9)*	*(1.9)*	*(2.0)*	*(1.7)*	*(1.9)*	*(1.8)*	*(1.9)*	*(1.9)*	*(2.0)*
F	272.4	136.3	13.7	-5.7	-53.1	-	75.4	-36.1	-58.2	27.8	63.8	-73.3	-55.8	-83.6	-72.2	-80.2	-48.3
	(1.3)	*(2.3)*	*(2.0)*	*(2.1)*	*(1.8)*	-	*(2.0)*	*(2.1)*	*(2.1)*	*(2.1)*	*(2.0)*	*(2.1)*	*(2.2)*	*(2.1)*	*(2.1)*	*(2.0)*	*(2.3)*
G	197.0	60.9	-61.7	-81.1	-128.5	-75.4	-	-111.5	-133.6	-47.6	-11.6	-148.7	-131.2	-159.0	-147.6	-155.6	-123.7
	(2.1)	*(1.7)*	*(2.0)*	*(2.1)*	*(1.9)*	*(2.0)*	-	*(2.0)*	*(1.9)*	*(2.1)*	*(1.6)*	*(1.8)*	*(1.8)*	*(1.7)*	*(1.7)*	*(1.7)*	*(1.8)*
H	308.5	172.4	49.8	30.5	-17.0	36.1	111.5	-	-22.1	63.9	99.9	-37.2	-19.7	-47.4	-36.1	-44.1	-12.2
	(2.1)	*(1.7)*	*(2.1)*	*(2.1)*	*(1.9)*	*(2.1)*	*(2.0)*	-	*(1.7)*	*(2.1)*	*(2.0)*	*(1.7)*	*(1.6)*	*(1.6)*	*(1.6)*	*(1.9)*	*(1.7)*
I	330.6	194.5	71.9	52.6	5.1	58.2	133.6	22.1	-	86.0	122.0	-15.1	2.4	-25.3	-14.0	-22.0	9.9
	(2.1)	*(1.5)*	*(2.1)*	*(2.1)*	*(1.9)*	*(2.1)*	*(1.9)*	*(1.7)*	-	*(2.1)*	*(1.9)*	*(1.4)*	*(1.2)*	*(1.2)*	*(1.1)*	*(1.7)*	*(1.4)*
J	244.5	108.4	-14.1	-33.5	-80.9	-27.8	47.6	-63.9	-86.0	-	36.0	-101.1	-83.6	-111.4	-100.0	-108.0	-76.2
	(2.0)	*(2.1)*	*(1.6)*	*(1.6)*	*(1.9)*	*(2.1)*	*(2.1)*	*(2.1)*	*(2.1)*	-	*(2.2)*	*(2.0)*	*(2.1)*	*(2.0)*	*(2.1)*	*(2.1)*	*(2.1)*
K	208.6	72.5	-50.1	-69.4	-116.9	-63.8	11.6	-99.9	-122.0	-36.0	-	-137.1	-119.6	-147.3	-136.0	-144.0	-112.1
	(2.1)	*(1.7)*	*(2.1)*	*(2.2)*	*(2.0)*	*(2.0)*	*(1.6)*	*(2.0)*	*(1.9)*	*(2.2)*	-	*(1.8)*	*(1.8)*	*(1.7)*	*(1.7)*	*(1.7)*	*(1.8)*
L	345.7	209.6	87.0	67.7	20.2	73.3	148.7	37.2	15.1	101.1	137.1	-	17.5	-10.2	1.1	-6.9	25.0
	(2.1)	*(1.4)*	*(1.9)*	*(2.0)*	*(1.7)*	*(2.1)*	*(1.8)*	*(1.7)*	*(1.4)*	*(2.0)*	*(1.8)*	-	*(1.3)*	*(1.0)*	*(1.2)*	*(1.6)*	*(1.4)*
M	328.2	192.1	69.5	50.1	2.7	55.8	131.2	19.7	-2.4	83.6	119.6	-17.5	-	-27.8	-16.4	-24.4	7.5
	(2.2)	*(1.3)*	*(2.1)*	*(2.1)*	*(1.9)*	*(2.2)*	*(1.8)*	*(1.6)*	*(1.2)*	*(2.1)*	*(1.8)*	*(1.3)*	-	*(1.1)*	*(1.0)*	*(1.6)*	*(1.3)*
N	355.9	219.8	97.2	77.9	30.4	83.6	159.0	47.4	25.3	111.4	147.3	10.2	27.8	-	11.3	3.3	35.2
	(2.1)	*(1.2)*	*(1.9)*	*(2.0)*	*(1.8)*	*(2.1)*	*(1.7)*	*(1.6)*	*(1.2)*	*(2.0)*	*(1.7)*	*(1.0)*	*(1.1)*	-	*(0.8)*	*(1.5)*	*(1.0)*
O	344.6	208.5	85.9	66.6	19.1	72.2	147.6	36.1	14.0	100.0	136.0	-1.1	16.4	-11.3	-	-8.0	23.9
	(2.2)	*(1.1)*	*(2.0)*	*(2.1)*	*(1.9)*	*(2.1)*	*(1.7)*	*(1.6)*	*(1.1)*	*(2.1)*	*(1.7)*	*(1.2)*	*(1.0)*	*(0.8)*	-	*(1.5)*	*(1.0)*
P	352.6	216.5	93.9	74.6	27.1	80.2	155.6	44.1	22.0	108.0	144.0	6.9	24.4	-3.3	8.0	-	31.9
	(2.1)	*(1.7)*	*(2.0)*	*(2.0)*	*(1.9)*	*(2.0)*	*(1.7)*	*(1.9)*	*(1.7)*	*(2.1)*	*(1.7)*	*(1.6)*	*(1.6)*	*(1.5)*	*(1.5)*	-	*(1.6)*
Q	320.7	184.6	62.0	42.7	-4.8	48.3	123.7	12.2	-9.9	76.2	112.1	-25.0	-7.5	-35.2	-23.9	-31.9	-
	(2.2)	*(1.0)*	*(2.1)*	*(2.2)*	*(2.0)*	*(2.3)*	*(1.8)*	*(1.7)*	*(1.4)*	*(2.1)*	*(1.8)*	*(1.4)*	*(1.3)*	*(1.0)*	*(1.0)*	*(1.6)*	-

Model Key

A: No mailing | B: Blanket mailing | C: OLS | D: Lasso
E: Logit/OLS | F: Random forest | G: Two forests | H: Causal KNN (K = 425)
I: Causal KNN (K = 2475) | J: TOR (Lasso)* | K: TOR (random forest)* | L: TEP (Lasso, K = 425)†
M: TEP (random forest, K = 425)† | N: TEP (Lasso, K = 2475)† | O: TEP (random forest, K = 2475)† | P: Causal forest
Q: Scoring Model

*TOR - transformed outcome regression
†TEP - treatment effect projection

Note: The values indicate the mean difference between the profit of the row estimation method and the profit of the column estimation method. The standard error of the mean difference is in parentheses.

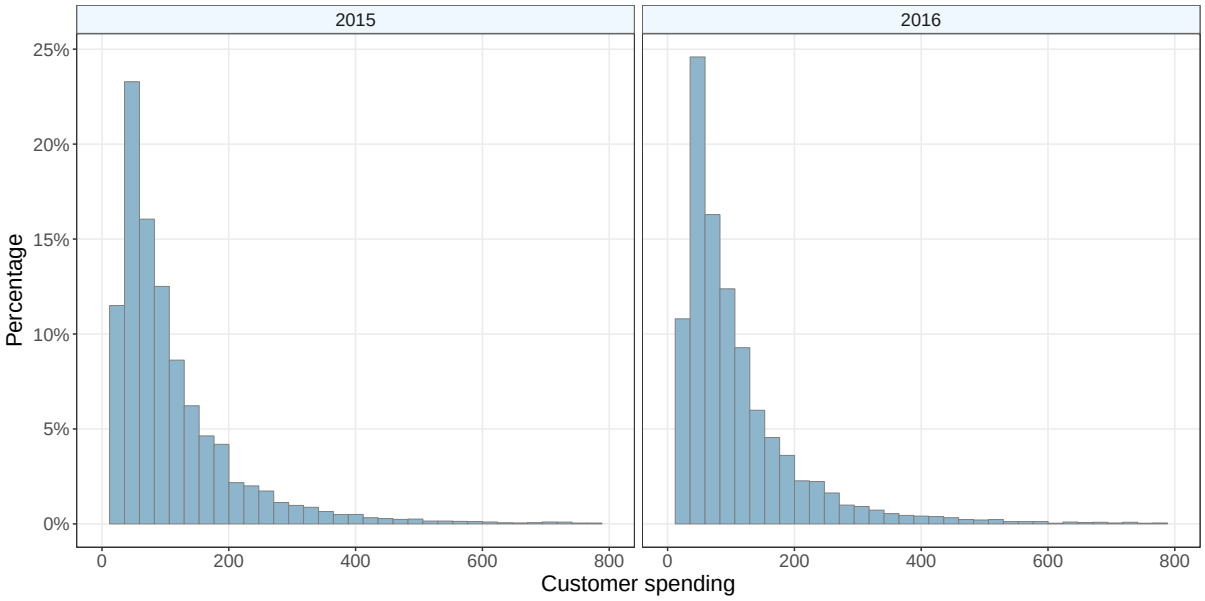


Figure 1: Distribution of dollar spending conditional on spending > 0

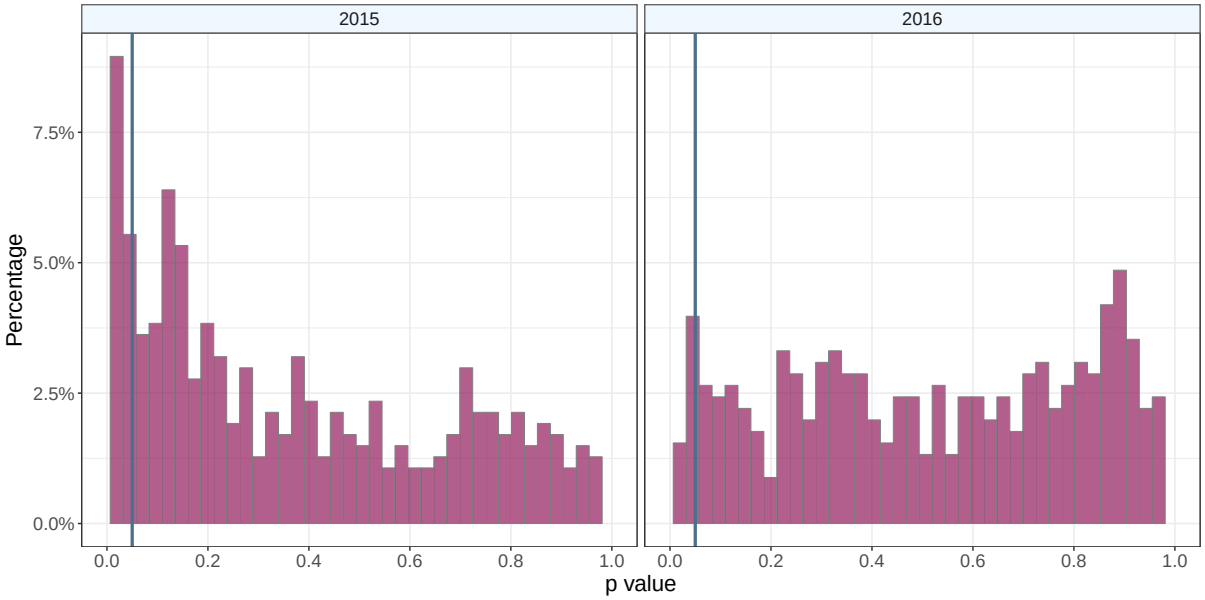


Figure 2: Distribution of p-values for test of equality of means of features

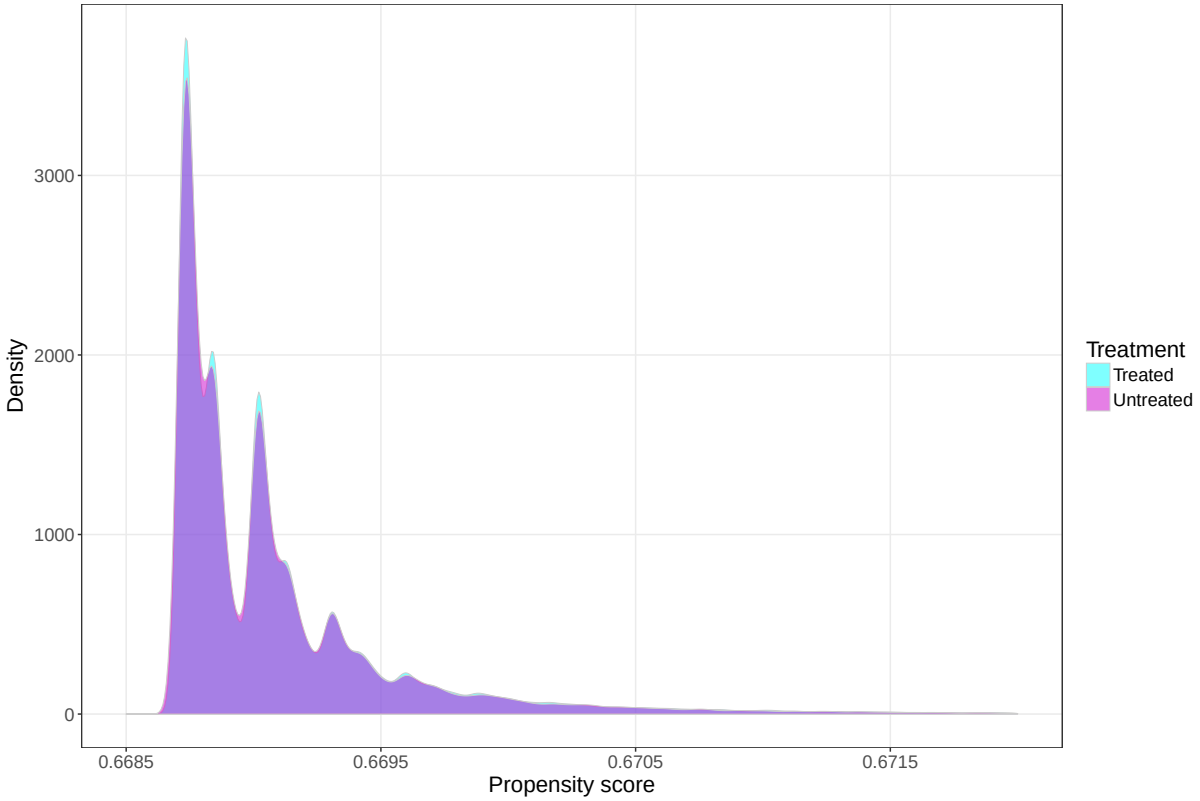


Figure 3: Estimated propensity score

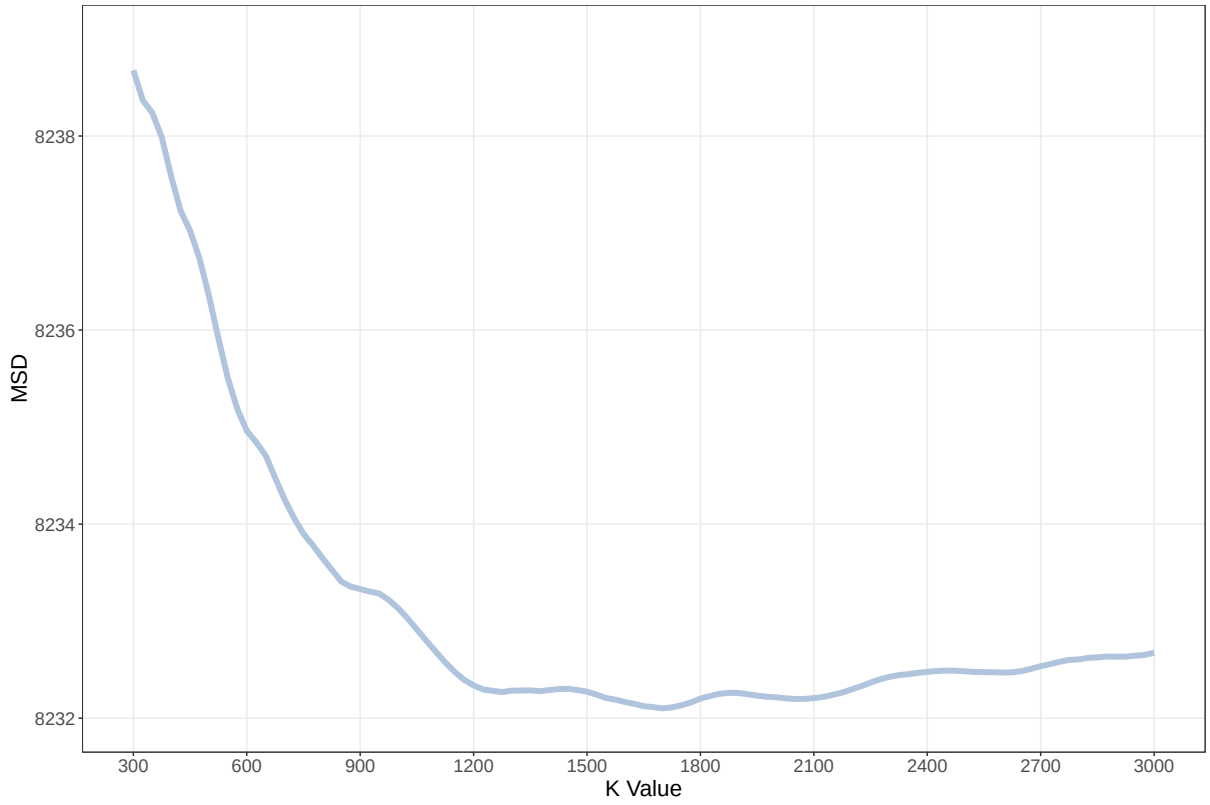
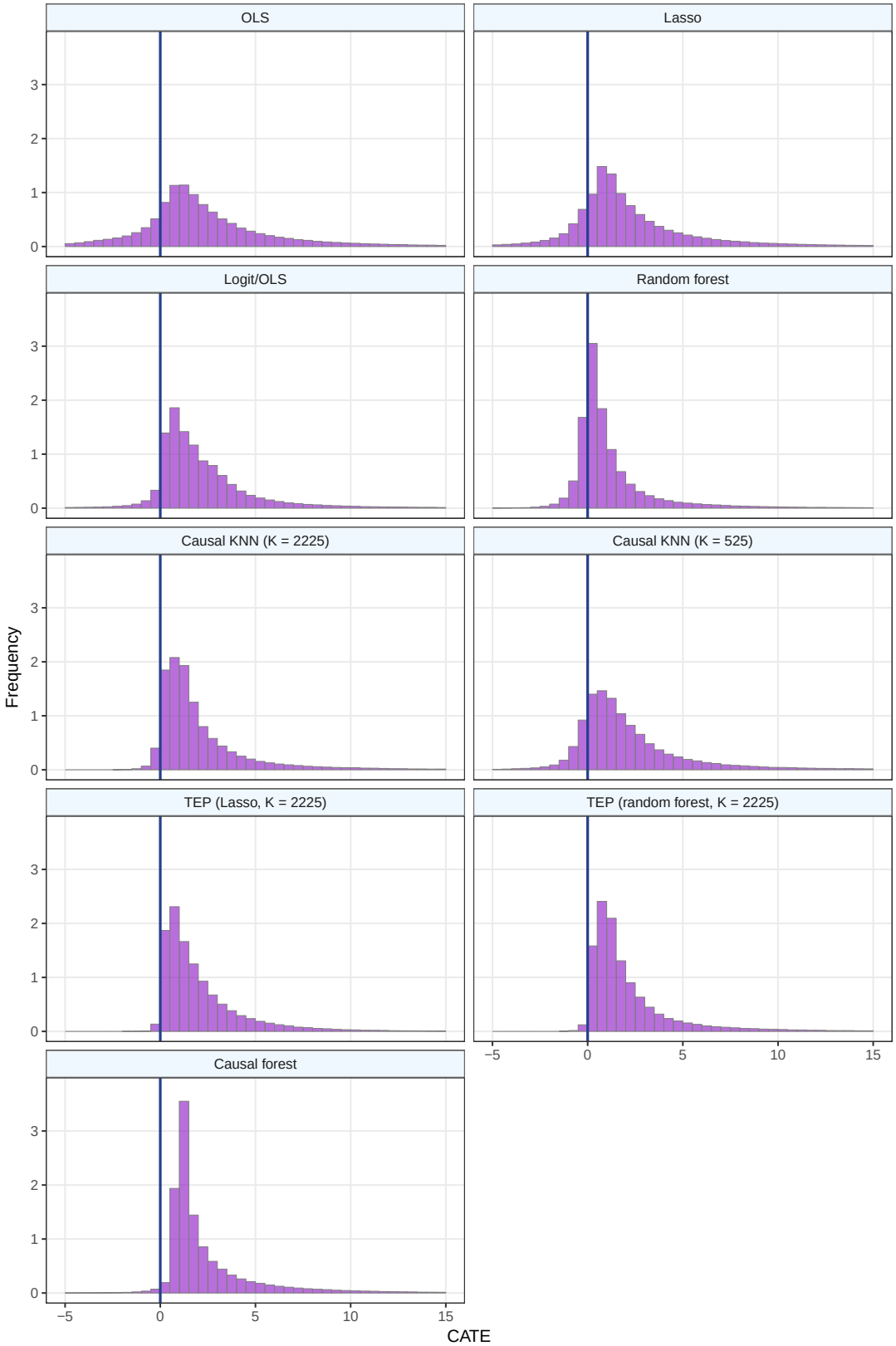
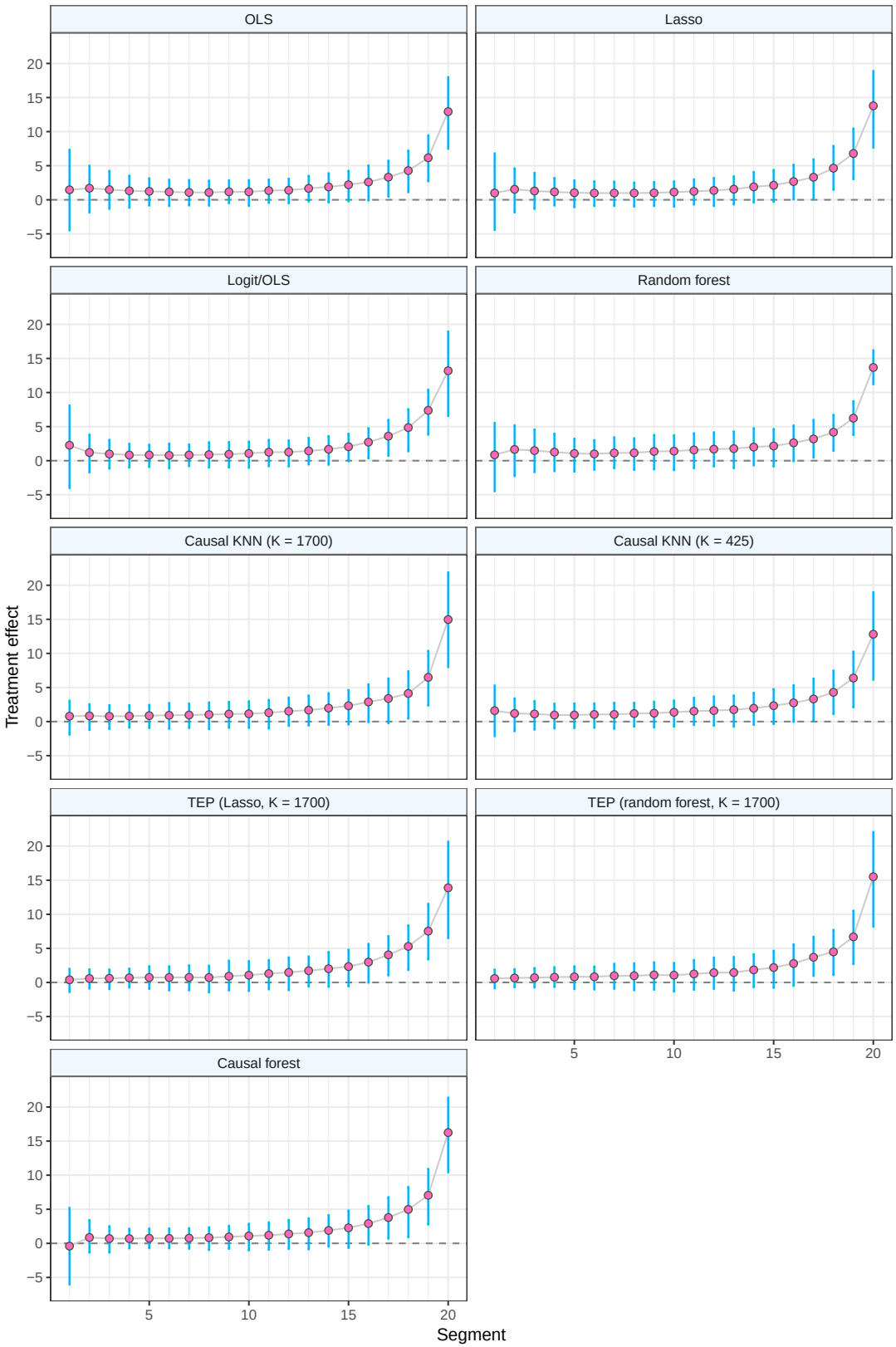


Figure 4: Causal KNN regression: Mean-squared difference between transformed outcome and $\hat{\tau}_K(x)$ for different values of K



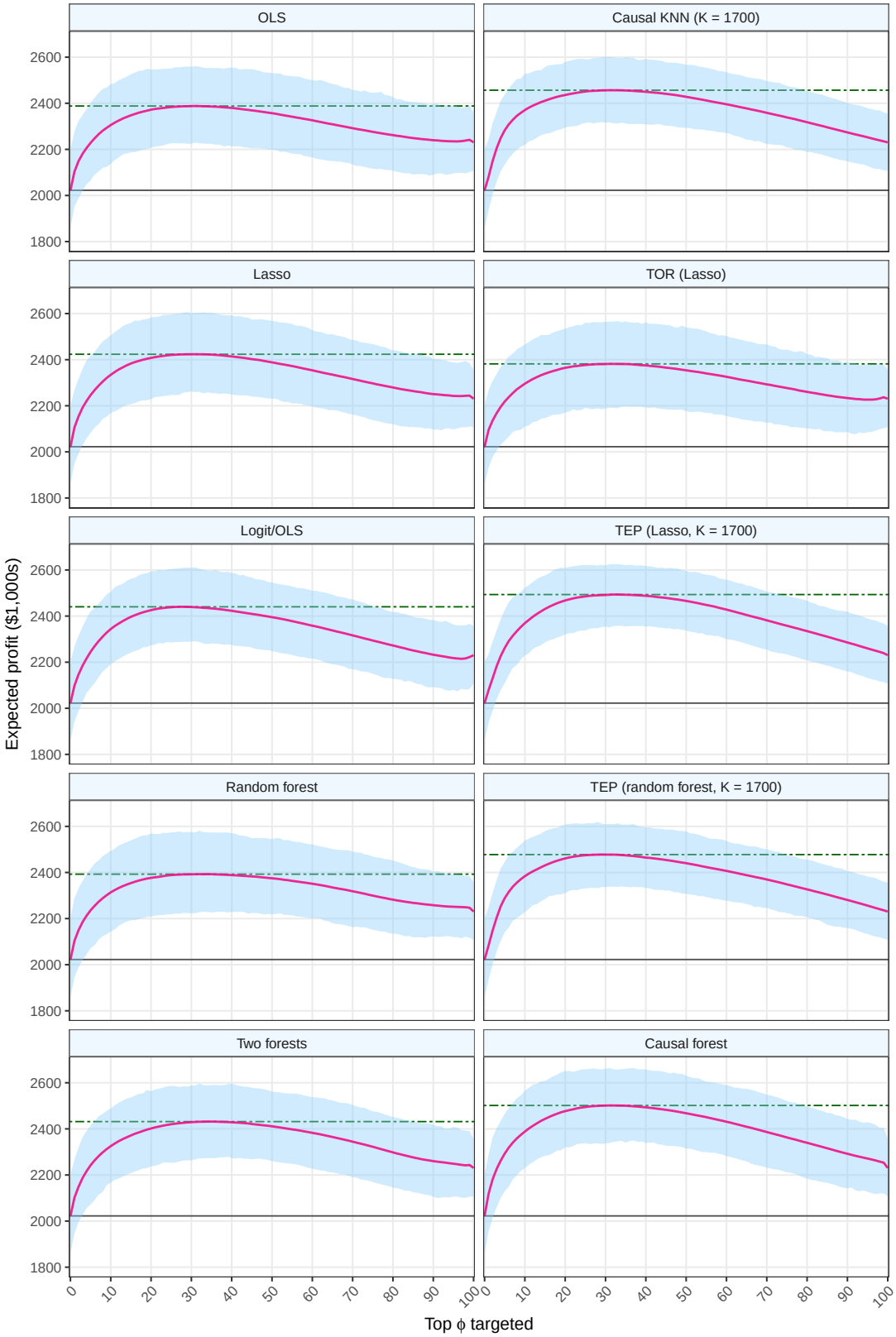
Note: Predictions obtained using ten-fold cross validation.

Figure 5: Predicted conditional average treatment effects



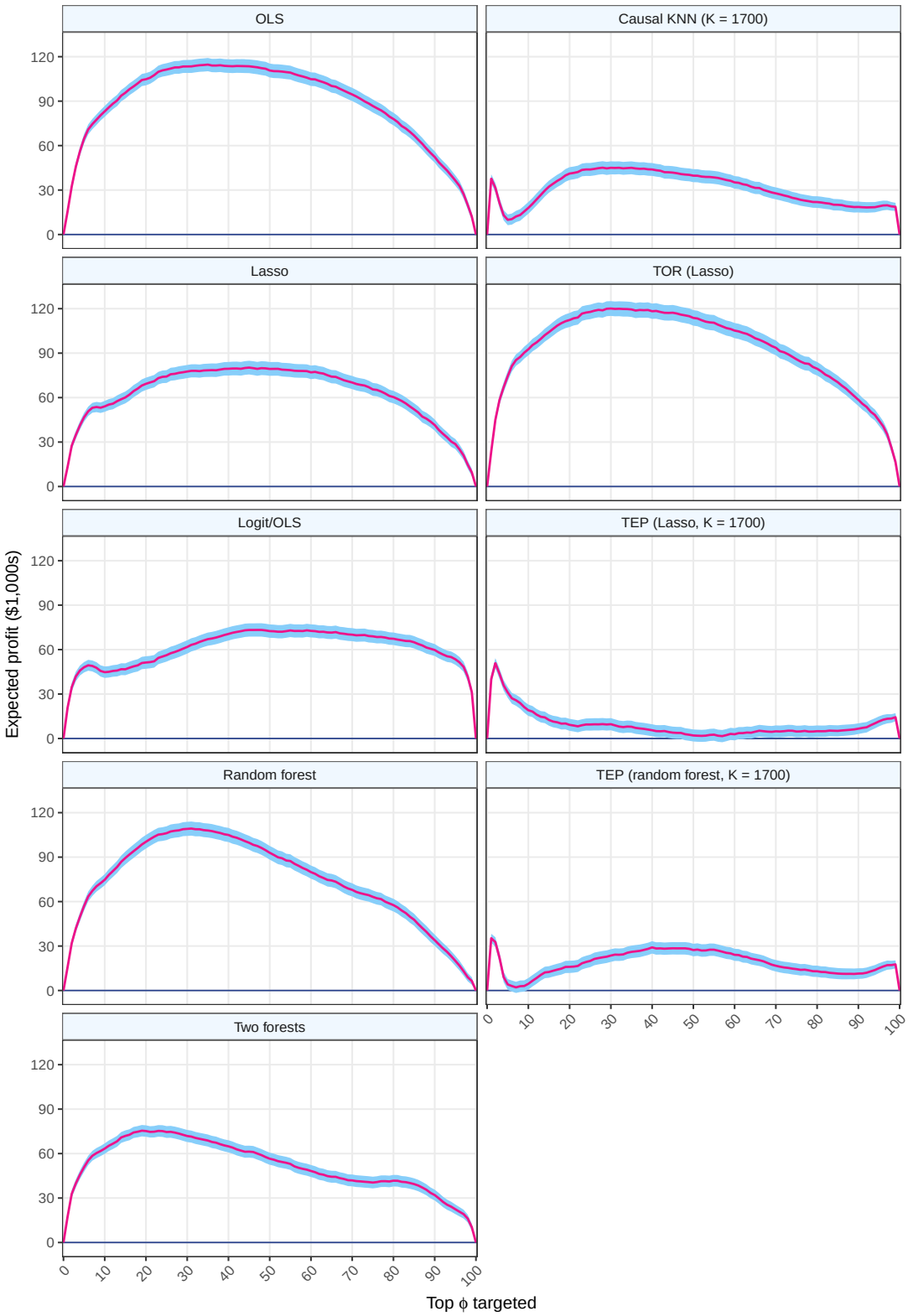
Note: The vertical bars indicate the 95 percent range of values across all bootstrap draws.

Figure 6: Lift factors



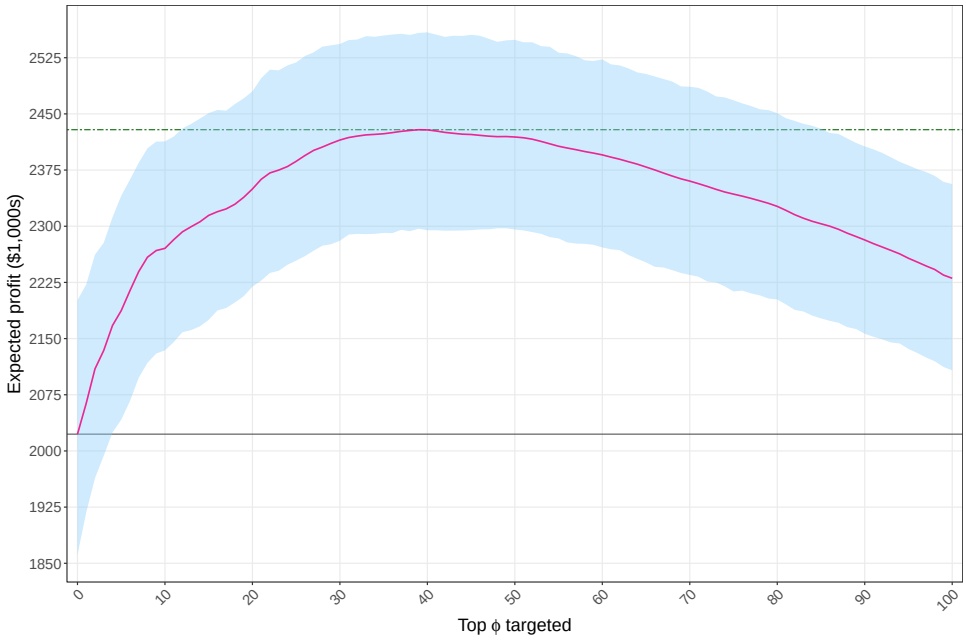
Note: The graphs indicate the mean and 95 percent range of profits levels across all bootstrap draws.

Figure 7: Profit level when targeting top ϕ percent of customers



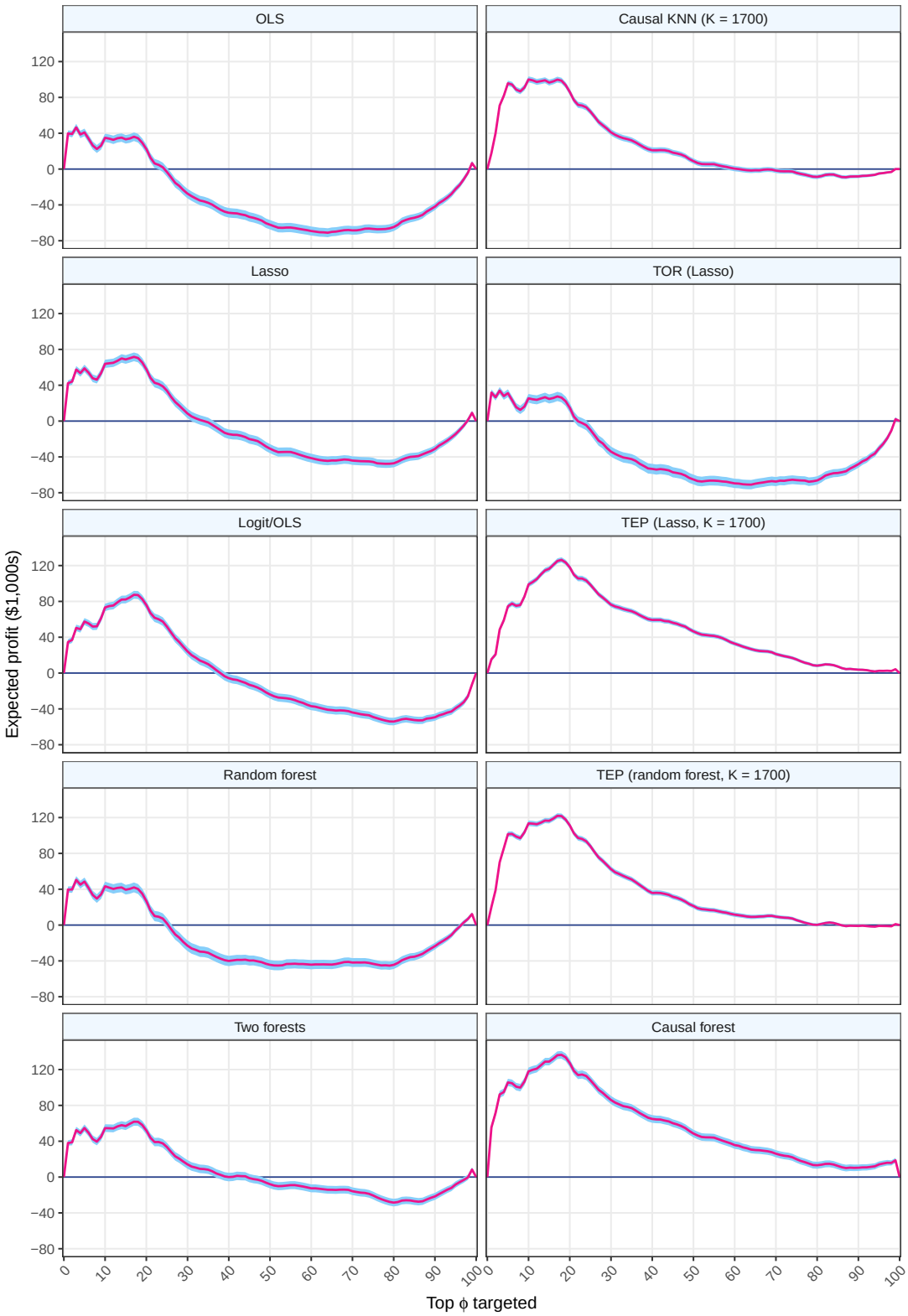
Note: The graph indicates the mean profit difference across bootstrap draws and the corresponding 95 percent confidence interval.

Figure 8: Differences in profit level when targeting top ϕ percent of customers: Causal forest vs. alternative methods



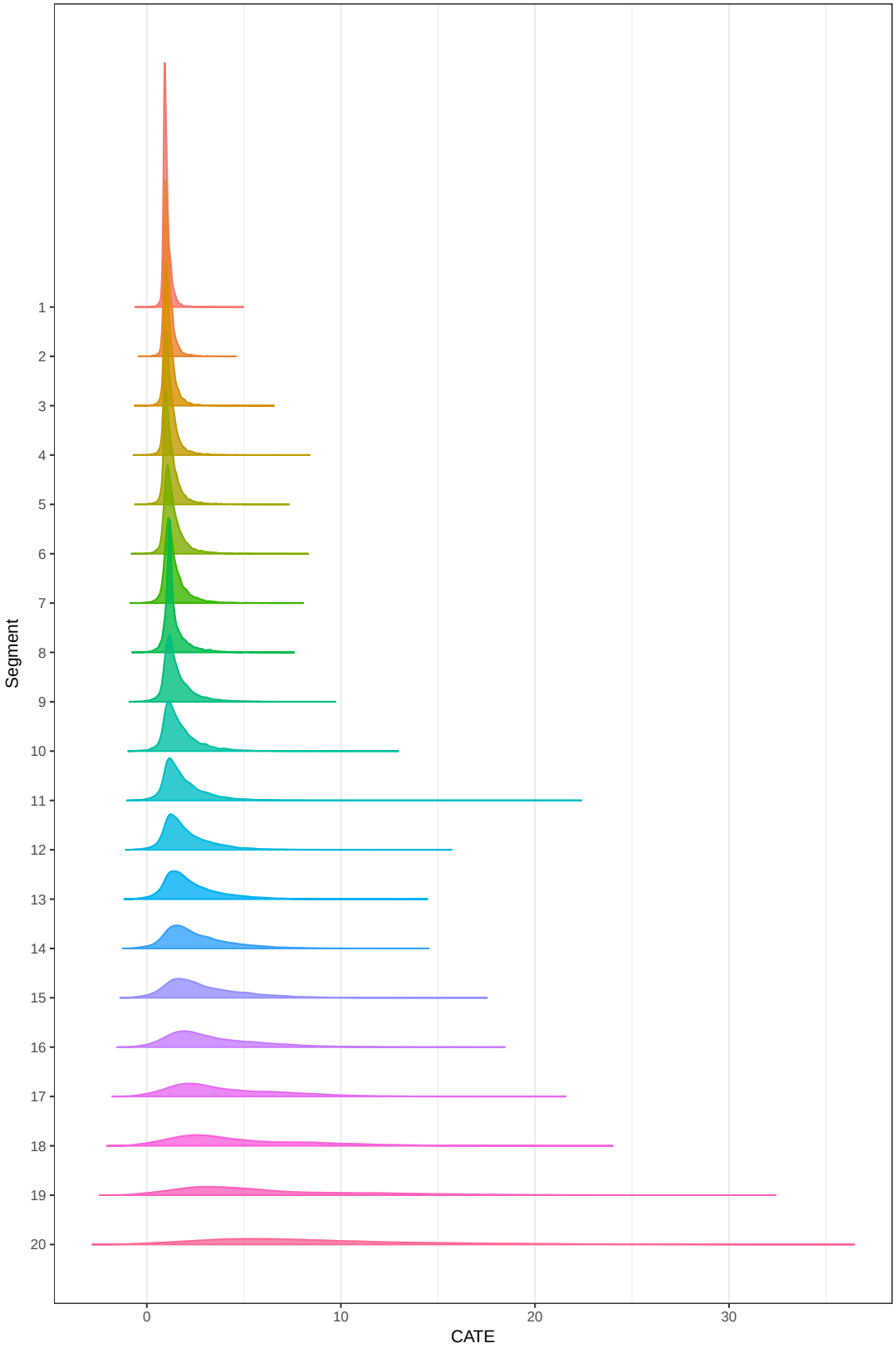
Note: The graphs indicate the mean and 95 percent range of profits levels across all bootstrap draws.

Figure 9: Profit level when targeting top ϕ percent of customers: Scoring model



Note: The graph indicates the mean profit difference across bootstrap draws and the corresponding 95 percent confidence interval.

Figure 10: Differences in profit level when targeting top ϕ percent of customers: Alternative methods vs. scoring model



Note: Predictions obtained using ten-fold cross validation.

Figure 11: Conditional average treatment effect density by score level

Appendix

Table 11: 2016 predicted conditional average treatment effects

	Mean	**SE**	**SD**	$\tau(\hat{X}_i) < 0$	**Percentiles**						
					1%	5%	25%	50%	75%	95%	99%
OLS	2.54	0.18	5.83	19.4	-9.55	-3.19	0.41	1.62	3.80	11.20	22.24
Lasso	2.51	0.16	4.93	16.9	-6.51	-1.70	0.51	1.49	3.41	10.20	20.39
Logit/OLS	2.58	0.28	8.76	7.3	-7.76	-0.39	0.65	1.56	3.18	9.27	25.70
Random forest	1.71	0.09	2.84	17.1	-1.59	-0.62	0.23	0.99	2.31	6.25	13.72
Two forests	3.83	0.22	7.10	13.4	-9.54	-1.87	0.90	2.92	5.63	13.14	27.01
Causal KNN (K = 2475)	2.19	0.09	2.80	3.7	-0.33	0.08	0.63	1.30	2.62	7.64	14.42
Causal KNN (K = 425)	2.35	0.12	3.77	17.1	-3.25	-0.97	0.33	1.39	3.18	9.13	17.90
TOR (Lasso)*	2.70	0.17	5.23	13.7	-8.19	-2.07	0.75	1.70	3.71	10.66	20.92
TOR (random forest)*	3.71	0.39	12.38	17.4	-18.27	-3.95	0.65	2.71	5.59	15.99	36.33
TEP (Lasso, K = 2475)†	2.28	0.08	2.62	0.7	0.05	0.24	0.71	1.46	2.87	6.99	12.47
TEP (random forest, K = 2475)†	2.34	0.08	2.57	0.5	0.10	0.42	0.92	1.45	2.66	7.47	13.50
Causal forest	2.57	0.10	3.18	1.8	-0.53	0.69	1.16	1.51	2.82	8.53	15.52

*TOR - transformed outcome regression
†TEP - treatment effect projection

Table 12: 2016 mean-squared prediction error (MSE) differences across estimation methods

	A	B	C	D	E	F	G	H	I	J	K	L
A	-	7.1	-31.7	22.0	-62.0	21.1	16.5	-15.2	-254.7	24.7	20.5	15.8
	-	*(0.1)*	*(1.1)*	*(0.5)*	*(0.9)*	*(0.5)*	*(0.5)*	*(0.8)*	*(2.8)*	*(0.5)*	*(0.5)*	*(0.4)*
B	-7.1	-	-38.8	14.8	-69.1	14.0	9.4	-22.3	-261.8	17.6	13.4	8.7
	(0.1)	-	*(1.2)*	*(0.4)*	*(0.9)*	*(0.4)*	*(0.4)*	*(0.8)*	*(2.9)*	*(0.5)*	*(0.4)*	*(0.4)*
C	31.7	38.8	-	53.7	-30.3	52.8	48.2	16.5	-223.0	56.4	52.2	47.5
	(1.1)	*(1.2)*	-	*(1.4)*	*(1.5)*	*(1.4)*	*(1.4)*	*(1.5)*	*(3.2)*	*(1.4)*	*(1.4)*	*(1.3)*
D	-22.0	-14.8	-53.7	-	-83.9	-0.9	-5.5	-37.1	-276.6	2.7	-1.4	-6.2
	(0.5)	*(0.4)*	*(1.4)*	-	*(1.0)*	*(0.1)*	*(0.2)*	*(1.0)*	*(3.0)*	*(0.2)*	*(0.1)*	*(0.2)*
E	62.0	69.1	30.3	83.9	-	83.1	78.5	46.8	-192.7	86.7	82.5	77.8
	(0.9)	*(0.9)*	*(1.5)*	*(1.0)*	-	*(1.1)*	*(1.0)*	*(1.2)*	*(2.4)*	*(1.1)*	*(1.0)*	*(0.9)*
F	-21.1	-14.0	-52.8	0.9	-83.1	-	-4.6	-36.3	-275.8	3.6	-0.6	-5.3
	(0.5)	*(0.4)*	*(1.4)*	*(0.1)*	*(1.1)*	-	*(0.1)*	*(1.0)*	*(3.0)*	*(0.1)*	*(0.0)*	*(0.3)*
G	-16.5	-9.4	-48.2	5.5	-78.5	4.6	-	-31.7	-271.1	8.2	4.1	-0.7
	(0.5)	*(0.4)*	*(1.4)*	*(0.2)*	*(1.0)*	*(0.1)*	-	*(1.0)*	*(3.0)*	*(0.2)*	*(0.1)*	*(0.3)*
H	15.2	22.3	-16.5	37.1	-46.8	36.3	31.7	-	-239.5	39.9	35.7	31.0
	(0.8)	*(0.8)*	*(1.5)*	*(1.0)*	*(1.2)*	*(1.0)*	*(1.0)*	-	*(2.7)*	*(1.0)*	*(1.0)*	*(0.9)*
I	254.7	261.8	223.0	276.6	192.7	275.8	271.1	239.5	-	279.4	275.2	270.5
	(2.8)	*(2.9)*	*(3.2)*	*(3.0)*	*(2.4)*	*(3.0)*	*(3.0)*	*(2.7)*	-	*(3.0)*	*(3.0)*	*(2.9)*
J	-24.7	-17.6	-56.4	-2.7	-86.7	-3.6	-8.2	-39.9	-279.4	-	-4.2	-8.9
	(0.5)	*(0.5)*	*(1.4)*	*(0.2)*	*(1.1)*	*(0.1)*	*(0.2)*	*(1.0)*	*(3.0)*	-	*(0.1)*	*(0.3)*
K	-20.5	-13.4	-52.2	1.4	-82.5	0.6	-4.1	-35.7	-275.2	4.2	-	-4.7
	(0.5)	*(0.4)*	*(1.4)*	*(0.1)*	*(1.0)*	*(0.0)*	*(0.1)*	*(1.0)*	*(3.0)*	*(0.1)*	-	*(0.2)*
L	-15.8	-8.7	-47.5	6.2	-77.8	5.3	0.7	-31.0	-270.5	8.9	4.7	-
	(0.4)	*(0.4)*	*(1.3)*	*(0.2)*	*(0.9)*	*(0.3)*	*(0.3)*	*(0.9)*	*(2.9)*	*(0.3)*	*(0.2)*	-

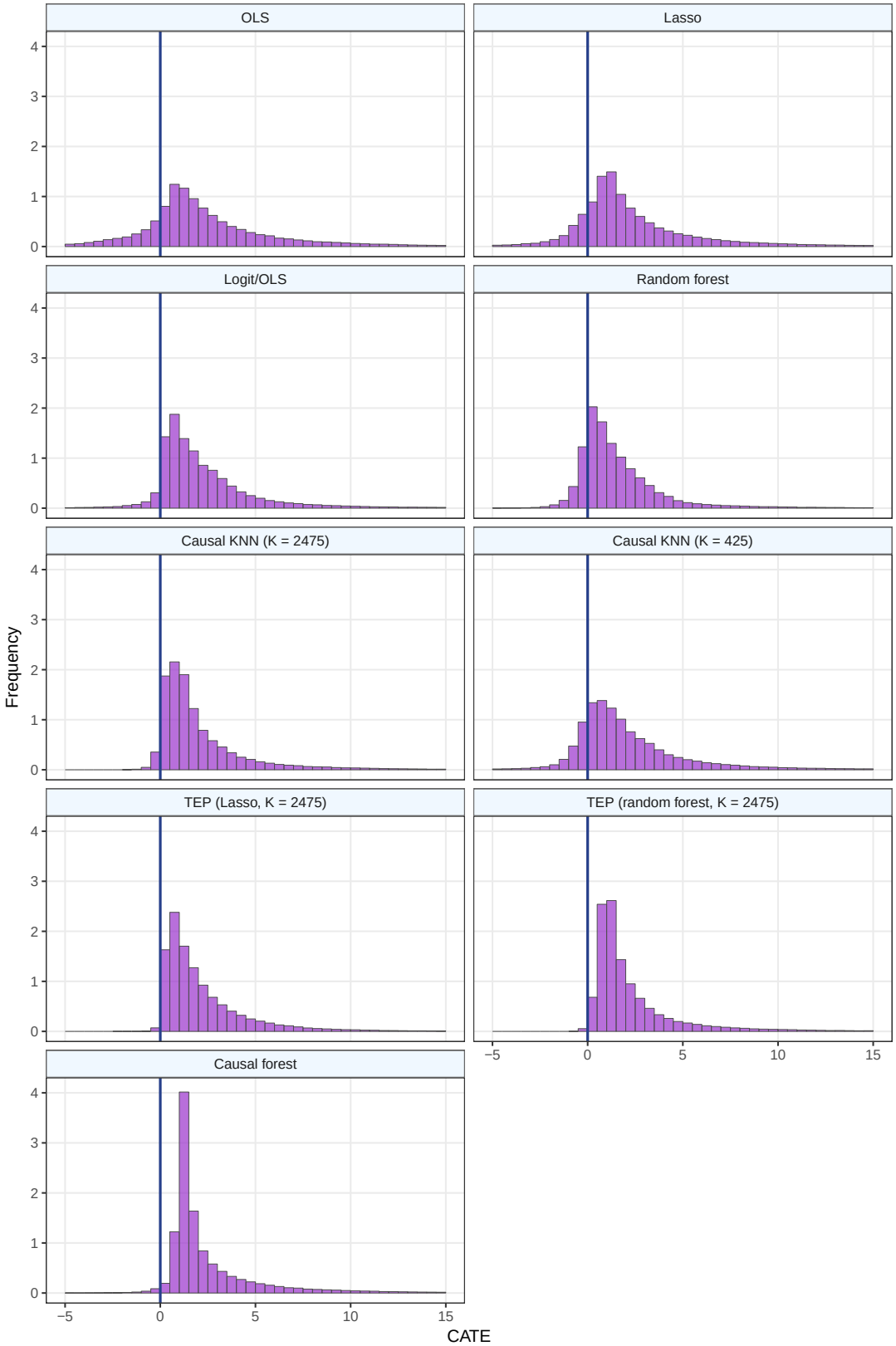
Model Key

A: OLS	B: Lasso	C: Logit/OLS
D: Random forest	E: Two forests	F: Causal KNN (K = 2475)
G: Causal KNN (K = 425)	H: Transformed Outcome (Lasso)*	I: Transformed Outcome (random forest)*
J: TEP (Lasso, K = 2475)†	K: TEP (random forest, K = 2475)†	L: Causal Forest

*TOR - transformed outcome regression

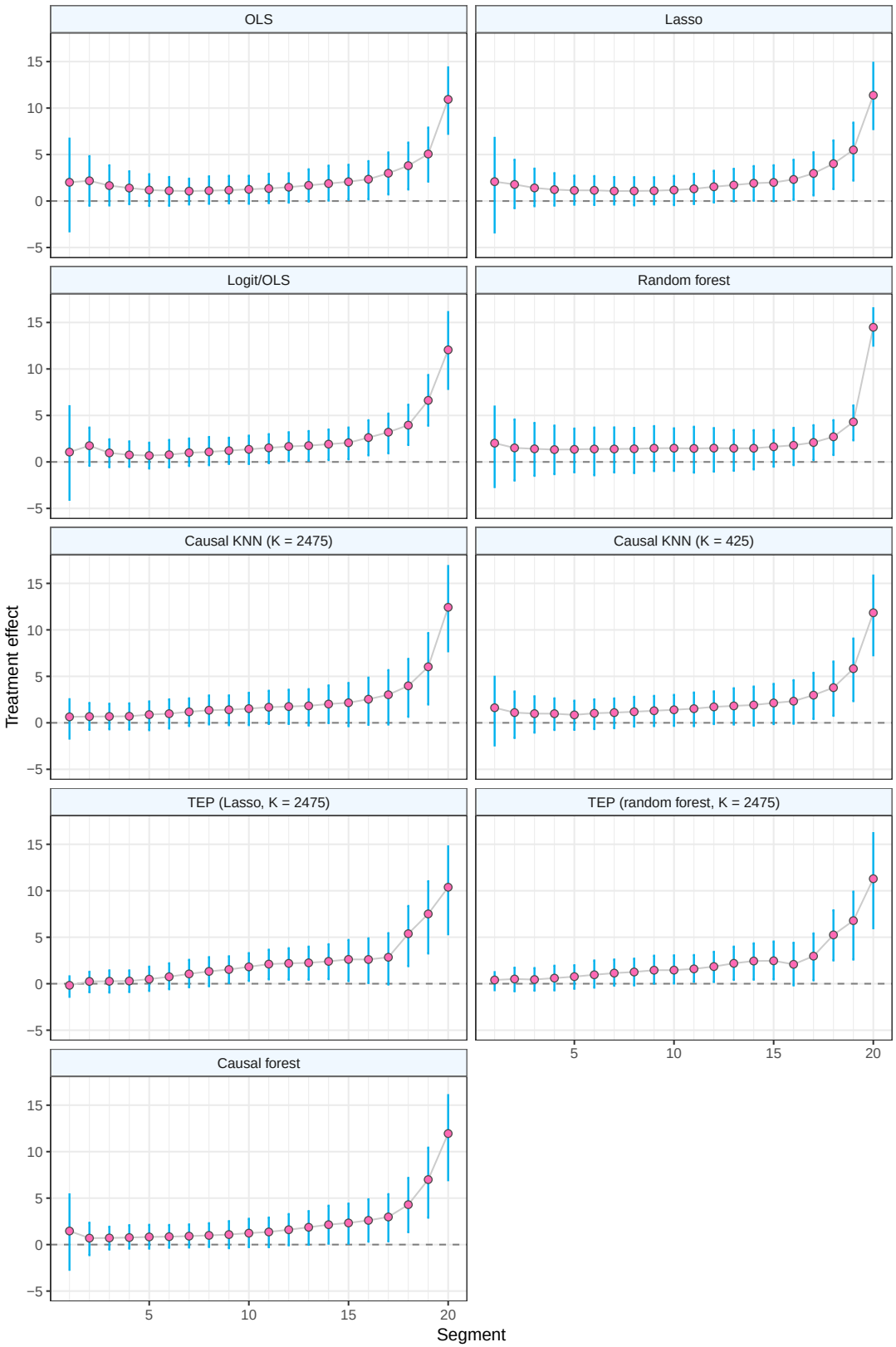
†TEP - treatment effect projection

Note: The values indicate the mean difference between the MSE of the row estimation method and the MSE of the column estimation method. The standard error of the mean difference is in parentheses.



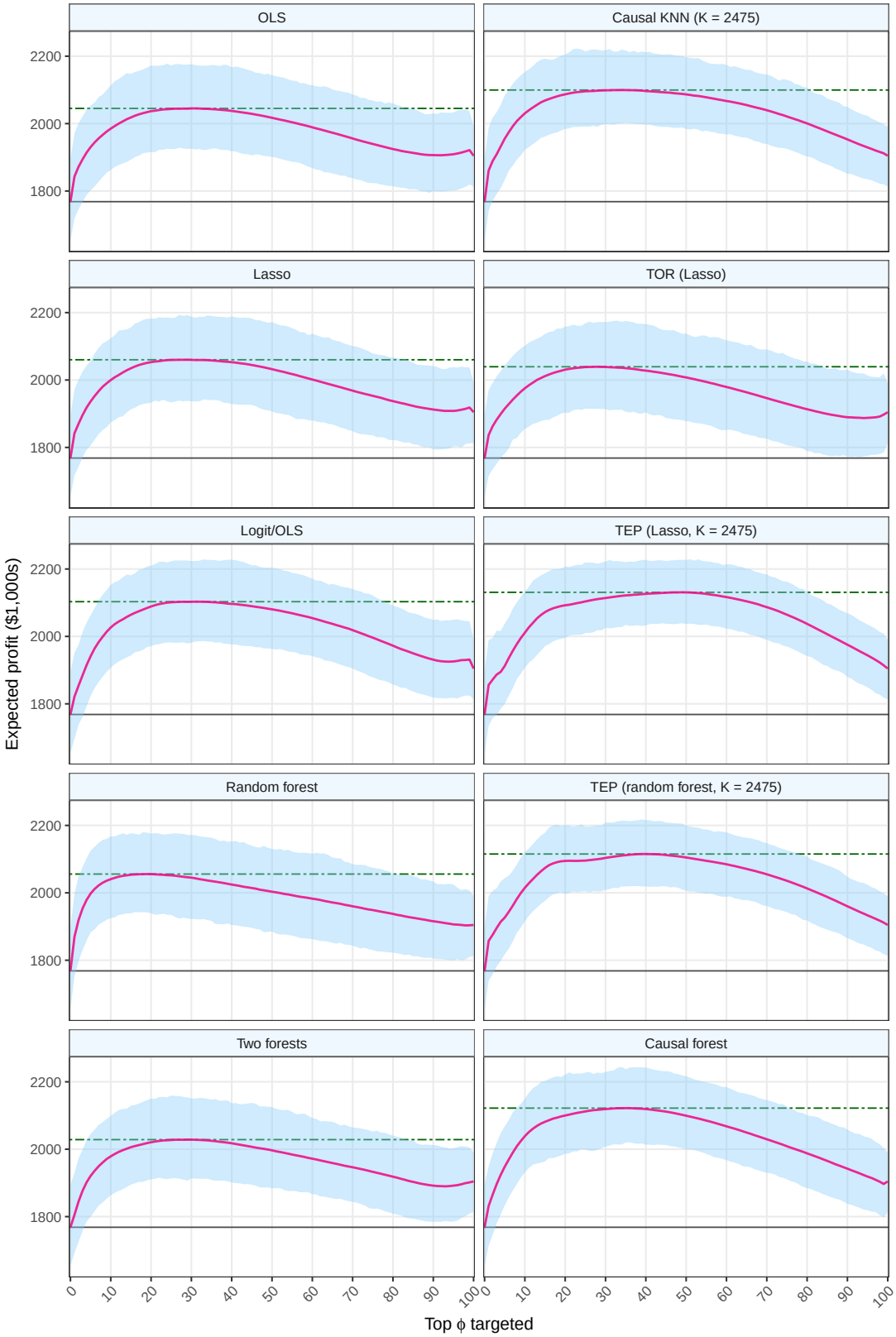
Note: Predictions obtained using ten-fold cross validation.

Figure 12: 2016 predicted conditional average treatment effects



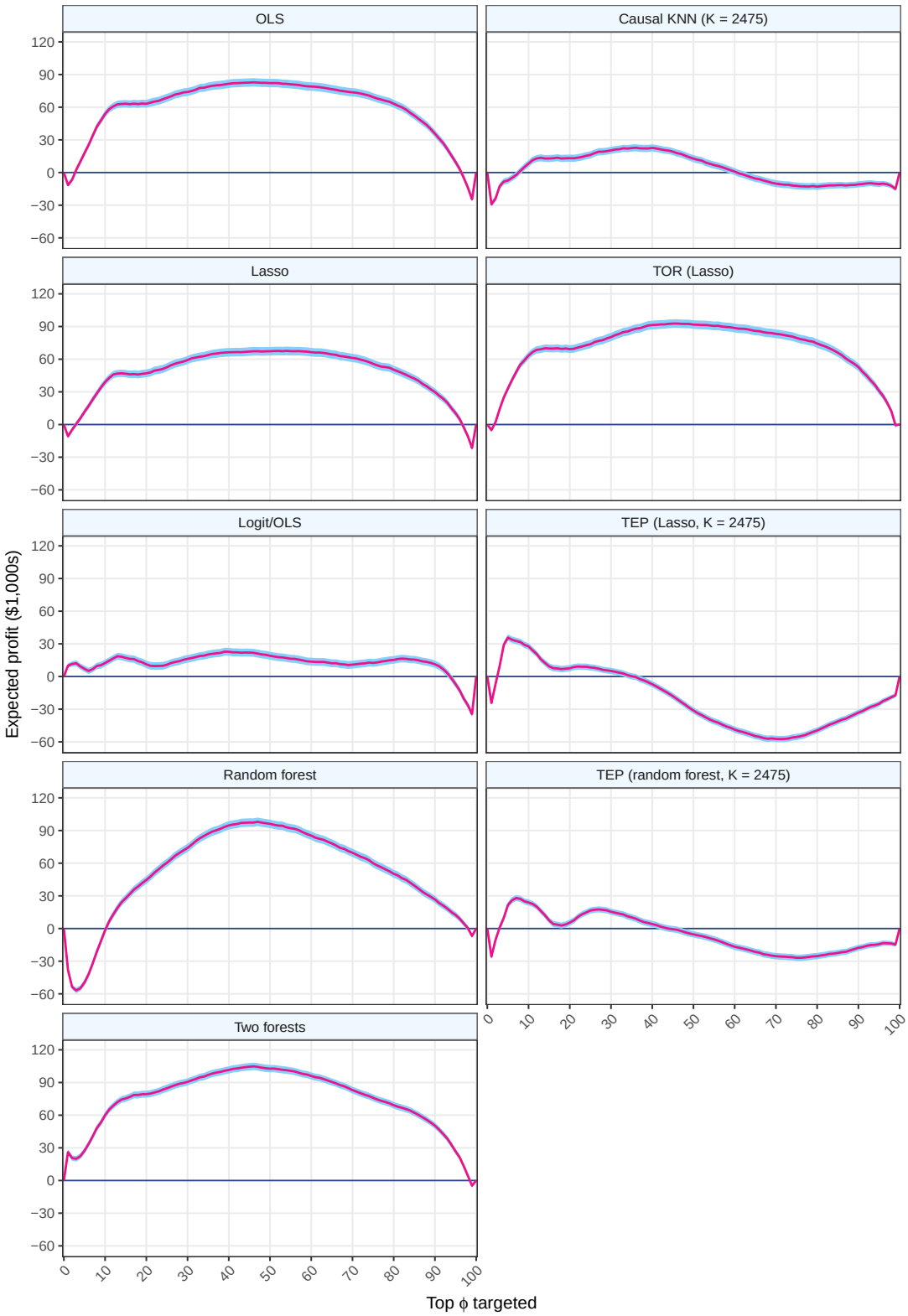
Note: The vertical bars indicate the 95 percent range of values across all bootstrap draws.

Figure 13: 2016 lift factors



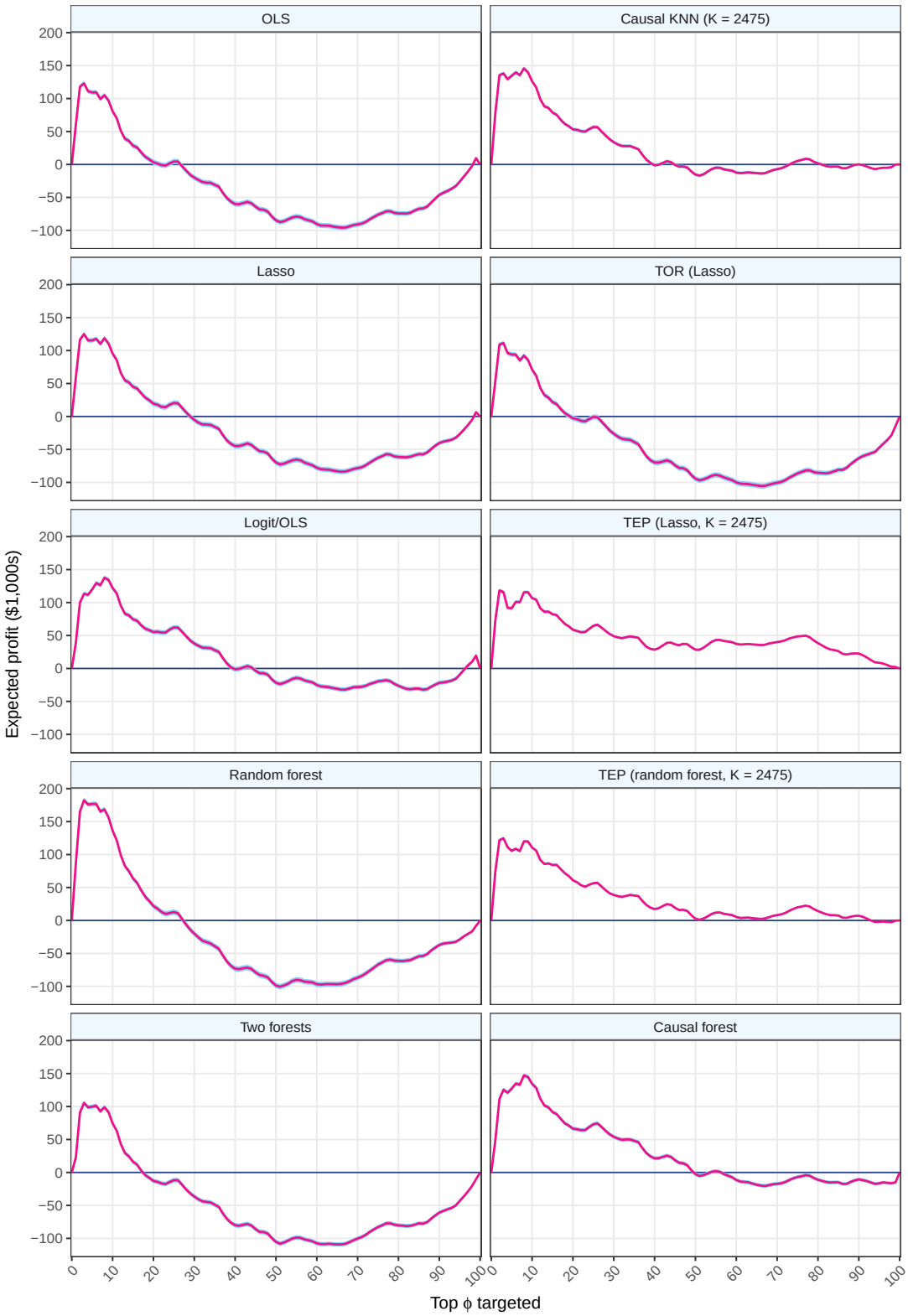
Note: The graph indicates the mean and 95 percent range of profits levels across all bootstrap draws.

Figure 14: 2016 profit level when targeting top ϕ percent of customers



Note: The graph indicates the mean profit difference across bootstrap draws and the corresponding 95 percent confidence interval.

Figure 15: 2016 differences in profit level when targeting top ϕ percent of customers: Causal forest vs. alternative methods



Note: The graph indicates the mean profit difference across bootstrap draws and the corresponding 95 percent confidence interval.

Figure 16: 2016 differences in profit level when targeting top ϕ percent of customers: Conditional average treatment effect methods vs. scoring model