Split-Treatment Analysis to Rank Heterogeneous Causal Effects for Prospective Interventions

Yanbo Xu
Georgia Institute of Technology
yxu465@gatech.edu

Divyat Mahajan
Microsoft Research India
t-dimaha@microsoft.com

Liz Manrao
Microsoft
elmanrao@microsoft.com

Amit Sharma
Microsoft Research India
amshar@microsoft.com

Emre Kıcıman
Microsoft Research AI
emrek@microsoft.com

ABSTRACT

For many kinds of interventions, such as a new advertisement, marketing intervention, or feature recommendation, it is important to target a specific subset of people for maximizing its benefits at minimum cost or potential harm. However, a key challenge is that no data is available about the effect of such a *prospective intervention* since it has not been deployed yet. In this work, we propose a *split-treatment* analysis that ranks the individuals most likely to be positively affected by a prospective intervention using past observational data. Unlike standard causal inference methods, the split-treatment method does not need any observations of the target treatments themselves. Instead it relies on observations of a *proxy treatment* that is caused by the target treatment. Under reasonable assumptions, we show that the ranking of heterogeneous causal effect based on the proxy treatment is the same as the ranking based on the target treatment's effect. In the absence of any interventional data for cross-validation, `Split-Treatment` uses sensitivity analyses for unobserved confounding to eliminate unreliable models. We apply `Split-Treatment` to simulated data and a large-scale, real-world targeting task and validate our discovered rankings via a randomized experiment for the latter.

CCS CONCEPTS

• **Computing methodologies → Causal reasoning and diagnostics**.

KEYWORDS

Causal inference; observational studies; split treatment; longitudinal data; heterogeneous treatment effect; sensitivity analysis

1 INTRODUCTION

Identifying individuals who will benefit from an intervention or treatment is an important challenge in many domains. We focus on this problem in the context of computing applications. For instance, many modern applications and devices display educational messages or other recommendations to users, telling them about undiscovered features, advertising related products, or providing other tips that may aid them. In complex software and scenarios, not all such messaging interventions are useful for every individual, and some may even be detrimental. For example, a message recommending an advanced feature may provide significant help to some advanced users, but also confuse and thus harm others.

There are several practical approaches for identifying individuals who will benefit from a specific messaging intervention based on reinforcement learning, such as contextual bandits [15] and off-policy learning [8, 33]. However, these approaches face limitations when applied to targeting *prospective* interventions where the potential harms or other costs of the intervention limit our ability to run experiments. Contextual bandits require active experimentation for every new message or intervention to select the individuals to assign treatment. That is, they may require exposing individual users to detrimental messages and causing negative effects at scale, including negative effects of irrelevant ads like ad blindness [21] and annoyance; and negative effects of users taking action on mismatched recommendations like user confusion and frustration. In addition, they depend on feedback (reward) signals in relatively shorter time-frames, such as click signals for online ads. However in many setting, the outcome of interest is longer-term such as sustained usage of a product or a feature. While off-policy learning does not require experimentation, it assumes that the target treatment has been deployed and therefore observed data for the intervention is available. Therefore, existing approaches either require active experimentation or assume that a new intervention's effects can be estimated based on the effects of past interventions.

We present and evaluate an alternative approach that uses observational (non-experimental) data for identifying individuals who are likely to benefit from a prospective, novel intervention (Fig. 1). To do so, we realize that it is often possible to split the effect of a messaging treatment into two parts: (1) how persuasive a treatment is in encouraging an individual to take a short-term action (e.g., click-through rate); and (2) the long-term effect of the action (e.g.,

sustained usage of the product). While we cannot know the first part of our split-effect without performing the prospective intervention, we can identify the second part of our split-effect when the action we are encouraging people to take has been performed by others in our non-experimental data. We show that, in many situations, this is sufficient for ranking individuals most likely to benefit from the otherwise novel intervention. Beyond this analytical approach, we make the additional contribution of demonstrating how challenges that commonly threaten the validity of observational studies can be addressed in this setting through the careful application of refutation and sensitivity analyses [29].

We apply `Split-Treatment` to estimate the heterogeneous causal effect of a novel messaging intervention in a large application suite. Each messaging intervention, Z, is a recommendation for a feature, A. The outcome of interest is the feature's long-term adoption, Y. Our method enables ranking of individuals based on the estimated effect of the intervention. We implement methods for causal effect estimation based on inverse propensity weighting and machine learning models like random forests and CNN, and use sensitivity analysis to arrive at robust models. Finally, we validate these models based on a randomized experiment and demonstrate that our methods and sensitivity analyses correctly identify, *a priori* the most accurate of the ranking models. We make 3 contributions:

(1) **Split-Treatment:** We propose an identification technique for *ranking* the effect of a prospective treatment without any access to the data with the treatment. We contrast our approach with instrumental variable and front-door analyses that require observed data for the treated population.

(2) **Sensitivity Analysis:** How can we practically check the assumptions and ensure reliable conclusions from our observational data? We show how to adapt sensitivity methods for this problem and demonstrate that their violations correlate with empirical errors.

(3) **Validation through Active Experiments :** We validate our conclusions using an A/B experiment of recommendation efficacy for a feature on over 1M users. Our findings show that `Split-Treatment` rankings of the individuals that are most likely to benefit from the recommended message match the results from A/B experiments.

2 BACKGROUND AND RELATED WORK

2.1 Causal analysis

Conventionally, the problem of showing recommended items to people is considered an outcome prediction problem: what would be the predicted outcome (e.g., rating) for an item by an individual? Based on this prediction, a ranked list of recommended items is shown to the individual [31]. However, recent work [26, 28] frames recommendation as a causal inference problem: what is the effect on the outcome (e.g., usage metrics) of showing a recommended item to an individual? Rather than simply verifying if a person achieves the predicted outcome metric, it is important to understand if the desired outcome was achieved *because* of the recommendation. If the outcome would have been the same without the recommendation, then it does not have a causal effect on the outcome.

One of the fundamental challenges in causal analysis is we only observe one of the outcomes: either an intervention was shown,
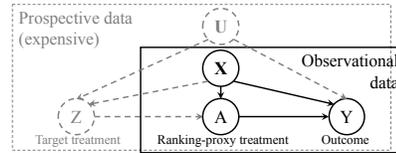

Figure 1: The goal is to estimate the effect of a prospective treatment that is never observed. `Split-Treatment` uses a proxy treatment A that lies on the causal path between Z and Y. Observed variables (X) confound the effect of A on Y (and possibly also Z's effect). Under proper assumptions, individuals for which the proxy treatment has a positive effect will also have a positive effect to the target treatment.

or not. Following the Structural Causal Model (SCM) described in [17, 18], we can define the interventions and counterfactuals through a mathematical operator called $do(z)$: $Y|do(z = 1)$ refers to the outcome measure had intervention (or treatment) Z been enforced, and $Y|do(z = 0)$ refers to the outcome had there been no treatment. The causal effect of treatment Z is:

$$\texttt{Causal Effect} = \mathbb{E}[Y|do(z = 1)] - \mathbb{E}[Y|do(z = 0)].$$

To address the causal question, one needs to account for *confounding* in the observed data. Formally, confounding is due to the intervention (recommendation) and outcome having common causes. In addition, we expect a treatment to have varying effects across individuals. There is recent research [6, 11, 16, 35] on how to estimate heterogeneous causal effect for different subgroups, known as the Conditional Average Treatment Effect (CATE). In the limit, we may consider the causal effect for each individual, known as the Individual Treatment Effect (ITE) [27]. We extend this line of work by proposing a method to rank causal effect for prospective treatments, for which we have no data as yet.

That said, since counterfactual outcomes are never observed simultaneously, evaluating a causal estimate is non-trivial. Unlike prediction, it is not possible to validate a causal effect estimate based only on observed outcomes. Typically, any causal estimation method makes certain assumptions on the counterfactuals and their robustness depends on how plausible those assumptions are. There is past work on validating assumptions of causal models to the extent possible, often by exploring the sensitivity of the obtained estimate to unobserved variables [7, 24, 25]. We utilize sensitivity analyses to rank different candidate methods for causal estimation.

2.2 Effects of Ads or Recommendations

Previous work for identifying individuals who will benefit from messages, ads, or recommendations mainly fall into two groups. The first group of work [13, 22, 32, 37] rely on campaign studies to analyze the effect of the known treatment ads through A/B testing or reinforcement learning-based methods like contextual bandits [12, 15, 34] that target treatment to specific individuals during an experiment. Brodersen et al. [4] applies Bayesian structural time-series models to infer causal impact of market interventions on an outcome metric over time. They illustrate the statistical properties of posterior inference on simulated data, and demonstrate their approach practically in an online ad campaign. Wang et al.

[36] enables analysis for more complex ad treatments using campaign data, such as measuring the impact of ads with different frequencies rather than of the randomized binary ads exposures. They employ tree structure in modeling the complex treatment propensities for achieving robust and unbiased estimations. Li and Pearl [14] integrates both experimental data and observational data, using Pearl's structural causal model [19] to formulate a unit selection problem that aims to identify a set of individuals who would most likely respond positively to the ads in short term. [38] targets for long-term outcomes by identifying a surrogate index in the short-term campaign data to impute the potential long-term values. Our work differs from these previous analysis of the causal impact of ads campaigns in that we are trying to predict the impact of an advertisement *before* starting a campaign.

The second group of work do not require treatment ads to be randomized, but need them exposed in the data, and then adjust selection bias in such observational data. For example, Chan et al. [5] utilizes doubly robust methods to evaluate the effectiveness of online ads from large observational data, and validates results using simulations based on realistic scenarios. Gordon et al. [9] assesses empirically how the variation in data availability can affect the ability of observational methods to recover the causal effects of online advertising learnt from randomized trials. Our work differs from these observational studies in that we estimate the impact of a prospective message that does not exist in the current data.

3 IDENTIFICATION BY SPLIT-TREATMENT

In this section we describe how the Split-Treatment method can identify the ranking of individuals most likely to benefit from a novel treatment or recommendation. Suppose Z is the prospective treatment of interest and Y is the outcome we aim to measure its effect on. If we run an A/B experiment where half of the population is treated with Z, then we can assess the conditional average treatment effect (CATE) of Z within a subpopulation $\mathcal{G}$ by taking the difference between the two outcome means respectively measured in the treatment and control groups:

$$\text{CATE}_{\mathcal{G}}^{(z)} = \mathbb{E}_{\mathcal{G}|Z=1}[Y \mid Z = 1] - \mathbb{E}_{\mathcal{G}|Z=0}[Y \mid Z = 0], \quad (1)$$

However, our observational data does not contain the prospective (target) treatment Z, as it has not been deployed yet. To achieve the goal, we propose splitting the target treatment's effect into two parts. The first part is the effect of the target treatment on some short-term user behavior A, such as the first immediate action encouraged by the treatment Z. The second part is the effect of this immediate action A on the outcome measure Y that we are optimizing. The requirements for our choice of A are: (1) The effect of Z on Y should be mediated through A. That is, Z does not directly affect the outcome; (2) Z does (or would) effect the use of A; (3) A exists, with some natural variation, in our observational logs.

Note that the first two requirements are exactly the requirements for a valid instrumental variable (exclusion and relevance [2]), except in this case Z is not observed. In addition, Z needs not to be a valid instrumental variable. It may be caused by the same variables that confound the effect of A on Y (however there cannot be any unobserved common cause for Z and A). For example, the prospective

intervention may be assigned based on the features of each individual. The causal graph from Fig. 1 is also similar to the front-door identification criterion (Definition 3.3.3, Pearl [18]), however the criterion cannot be directly applied since Z is unobserved. Instead we use the following assumptions to use A as a proxy for ranking the most promising individuals for assigning the target treatment.

Assumption 1 (Ignorability): There exists no unobserved confounding between the proxy treatment A and the outcome measure Y: $P(Y|do(a), \boldsymbol{x}) = P(Y|a, \boldsymbol{x})$, for some observed variables $\boldsymbol{X}$.

Theorem 3.1. Under the causal graph G from Figure 1 and given Assumption 1, the post-intervention distribution of Y resulting from the intervention $do(z)$ conditional on $\boldsymbol{x}$ can be identified as,

$$P(Y|do(z), \boldsymbol{x}) = \sum_a P(a|z, \boldsymbol{x})P(Y|a, \boldsymbol{x}).$$

Proof. For any set of nodes Y, R, S and T, the rules of the do-calculus [17] can be written as (where $G_{\overline{B}/\underline{B}}$ denotes the graph by deleting from G all arrows pointing to/emerging from the set of nodes B respecively).

Rule 2: $P(Y|do(r), do(s), t) = P(Y|do(r), s, t)$, if $(Y \perp\!\!\!\perp S|R, T)_{G_{\overline{R}, \underline{S}}}$;

Rule 3: $P(Y|do(r), do(s), t) = P(Y|do(r), t)$, if $(Y \perp\!\!\!\perp S|R, T)_{G_{\overline{R}, \overline{S(T)}}}$, where $S(T)$ are the S-nodes that are not ancestors of T in $G_{\overline{R}}$.

$$
\begin{aligned}
P(Y|do(z), \boldsymbol{x}) &= \sum_a P(a|do(z), \boldsymbol{x})P(Y|do(z), a, \boldsymbol{x}) \\
&= \sum_a P(a|z, \boldsymbol{x})P(Y|do(z), do(a), \boldsymbol{x}) \quad \text{... Using } Rule\ 2 \\
&= \sum_a P(a|z, \boldsymbol{x})P(Y|do(a), \boldsymbol{x}) \quad\quad \text{... Using } Rule\ 3 \\
&= \sum_a P(a|z, \boldsymbol{x})P(Y|a, \boldsymbol{x}) \quad\quad\ \text{... Using } Assumption\ 1
\end{aligned}
$$

Using the above theorem, for individuals characterized by $\boldsymbol{x}$, we can write their individualized treatment effect (ITE) of Z as,

$$\text{ITE}^{(z)}(\boldsymbol{x}) = \mathbb{E}[Y|do(z = 1), \boldsymbol{x}] - \mathbb{E}[Y|do(z = 0), \boldsymbol{x}] \quad (2)$$
$$= \underbrace{\left(P(a = 1|z = 1, \boldsymbol{x}) - P(a = 1|z = 0, \boldsymbol{x}) \right)}_{\text{Compliance}(\boldsymbol{x})}$$
$$\cdot \underbrace{\left(\mathbb{E}[Y|a = 1, \boldsymbol{x}] - \mathbb{E}[Y|a = 0, \boldsymbol{x}] \right)}_{\text{ITE}^{(a)}(\boldsymbol{x})}.$$

While we cannot identify the first term, Compliance$(\boldsymbol{x})$, as Z does not exist in our data, we can still estimate ITE$^{(a)}$ from the observational data. If we take the group $\mathcal{G}$ containing the most effective individuals to the proxy treatment A to be the same group of the most promising individuals to the target treatment Z, we can derive another format of Eq. (1) based on Eq. (2) into the following:

$$\text{CATE}_{\mathcal{G}}^{(z)} = \mathbb{E}_{\boldsymbol{x} \in \mathcal{G}}\left[\text{ITE}^{(z)}(\boldsymbol{x})\right] \quad (3)$$
$$\propto \mathbb{E}_{\boldsymbol{x} \in \mathcal{G}}\left[\text{ITE}^{(a)}(\boldsymbol{x})\right], \text{ for } \mathbb{E}_{\boldsymbol{x} \in \mathcal{G}}\left[\text{Compliance}(\boldsymbol{x})\right] > 0$$

Assumption 2 (Compliance): We assume the target treatment Z has a positive average effect to encourage the proxy treatment A within the selected group containing the most effective individuals to A. That is, $\mathbb{E}_{\boldsymbol{x} \in \mathcal{G}}[P(a = 1|z = 1, \boldsymbol{x}) - P(a = 1|z = 0, \boldsymbol{x})] > 0$.
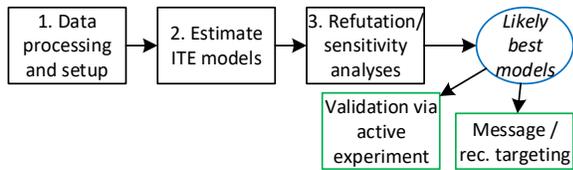
Figure 2: An end-to-end analysis pipeline of using `Split-Treatment` in feature recommendation. Validation is added only if experimental data having Z is available.

Figure 3: Data timeline in observational and experimental regimes

We cannot test this assumption in observational data due to the absence of Z, so we must use external domain knowledge to ensure that the assumption is satisfied.

4 ESTIMATION USING SPLIT-TREATMENT

The key challenge in feature recommendation, such as marketing software features to existing users, is to identify who would benefit from using a feature. Here, we show how `Split-Treatment` can be used to rank individuals using only observational data.

Following the formulation of `Split-Treatment` in Eq. (3), we say the target treatment Z is the given message recommending a software feature. We pick the proxy treatment A to be an individual's first trial of the feature from a target group that is assumed to have positive comparable compliance to Z; and we define the outcome measure Y to be an observation that the individual is receiving benefit from A. In our analysis, Y is sustained usage of the recommended feature measured in a post-treatment window, indicating the individual found the feature useful enough to keep engaging with it. Other examples (not included in this paper) include proactive churn management, where the target treatment Z is a price discount given to the existing subscribers, the outcome measure Y is long-term (e.g., three or five years) revenue; we can pick the proxy treatment A to be a subscriber's continuous subscription.

We split the treatment effect, use only observational data to learn the ITE function $\text{ITE}^{(a)}$ for treatment A, and rank the individuals characterized by $\boldsymbol{x}$'s as the most likely to keep using the feature if they have tried it now. We select the top K individuals into the recommendation group $\mathcal{G}$. By making the assumption that a well-designed message will encourage these individuals to try out the product (i.e., positive compliance of A to Z in the selected group), we claim that the individuals in group $\mathcal{G}$ are most likely to be positively affected by the recommendation such that they will use the feature and sustain using it once they got exposed.

Our end-to-end observational analysis and validation pipeline is shown in Figure 2. First, we collect and analyze observational data containing the proxy treatment A; we use this to estimate the function $\text{ITE}^{(a)}$. Using this ITE function, we rank individuals by the effect that A (and thus in turn, Z) will have on them. In our analysis of this data, we might use different algorithms, and even different feature sets. Second, we test our causal assumptions through sensitivity analyses. These help us eliminate the analysis design that appears to be unreliable. Finally, in our validation stage, we use experimental data to check that our `Split-Treatment` observational analysis has correctly identified the recommendation group $\mathcal{G}$. Note that in practice, experimental data for validation will not be available for all prospective treatments. As we will show
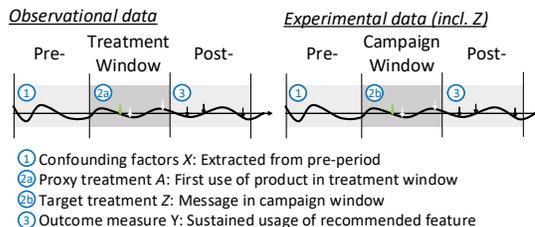
in the empirical analysis, sensitivity analyses can be useful to eliminate unreliable models in the absence of experimental validation (or when experiments are available only for a limited number of interventions). Below we provide further details of the three major components in the pipeline.

Step 1. Data preprocessing and set up

Given a cohort of users, we divide the observational data temporally into three windows. Observations made during a *pre-treatment window* provide a behavioral baseline of users (X). We use this X to address confounding. We look for observations of A during the *treatment window*, labeling a user as treated if we observe A, and untreated otherwise. In the *post-treatment window*, we measure the outcome Y that indicates a user has benefited from the action A.

For target data that our proposed method will be deployed on, we will only need to extract the user features X in the Pre-treatment window prior to the time point when the feature recommendation decision should be made. But for experimental data that we use for validation, we need to align the data timeline between the observational and experimental regimes. Instead of specifying a Treatment window, the experimental data has a *Campaign window* during which the ad treatment Z was randomly assigned and actively tested; the proxy treatment A of the first usage is also measured in this window. We exclude those users who still got exposed to the ad in the *Post-campaign window*, and count the sustained usage Y in that window. The user features are again extracted from the *Pre-Campaign window*. The three windows are aligned with the same length between the two regimes, as shown in Figure 3.

Step 2. Estimating Individual Treatment Effects

The key challenge of estimating causal effects from observational data is to adjust the factors X that can potentially confound the causal relationship between treatment A and outcome measure Y. Given unconfoundedness in Assumption 1, existing causal methods fall into three main categories: inverse probability of treatment weighting (IPTW) based methods that predicts propensity scores ($P(a = 1|\boldsymbol{x})$) to reweight individuals and obtain unbiased ATE estimates; g-formula based methods that predict the two arm potential outcomes $\mathbb{E}[Y^{(a)}|\boldsymbol{x}]$ for individuals and obtain unbiased ITE estimates; doubly robust methods that combine the two methods so that only one of the model need to be correctly specified. We develop a generic doubly robust framework in `Split-Treatment` by fitting a logistic regression model for generating propensity scores and using different machine learning techniques for estimating the potential outcomes. For predicting the amount of usage in the

post-treatment window, our regression choices include Poisson Regression, linear regression with Stochastic Gradient Descent, Fast Tree Regression, Fast Forest Regression, and a 2-layer Convolutional Neural Network (CNN). These methods can be switched to classification models if the outcome measure is binary.

We define an outcome regression function $f : \mathbb{R}^K \times \{0, 1\} \to \mathbb{R}$, where K is the cardinality of the input feature X and treatment A is considered binary, and compute ITE for an individual $\boldsymbol{x} \in \mathbb{R}^K$ as

$$\hat{\text{ITE}}^{(a)}(\boldsymbol{x}) \equiv f(\boldsymbol{x}, a = 1) - f(\boldsymbol{x}, a = 0). \tag{4}$$

Given observational data $\mathcal{D}^{(n)} = \{(y_i, \boldsymbol{x}_i, a_i) \,|_{i=1}^n\}$, we can learn f by minimizing the following loss function

$$\sum_{i=1}^{n} w_i \mathcal{L}(y_i, f(x_i, a_i)), \tag{5}$$

where $\mathcal{L}$ is either Poisson loss or mean squared error in the selected regression models; w_i is the IPT weight computed from the predicted propensity score. Note that loss in Eq. (5) goes back to ordinary regression loss if $w_i \equiv 1$, and becomes no longer a causal model with confounding adjustments. In particular, we use the Stabilized IPTW [23] to compute the weight w_i:

$$w_i = a_i \cdot \frac{\hat{P}(a_i = 1)}{\hat{P}(a_i = 1 | \boldsymbol{x}_i)} + (1 - a_i) \cdot \frac{1 - \hat{P}(a_i = 1)}{1 - \hat{P}(a_i = 1 | \boldsymbol{x}_i)}. \tag{6}$$

Step 3. Sensitivity Analyses

In general, multiple treatment effect estimation methods will provide varying results about the causal effect. These can be due to faults in different parts of the analysis pipeline: mis-specification of the underlying causal model, due to errors in identification, or due to limits in estimation. Here we provide methods to evaluate the estimates returned by different methods. Note that similar to statistical hypothesis tests, the tests below can refute an estimator that may be unsuitable, but cannot find the "correct" estimator. In other words, we can weed out some unreliable estimators, but cannot prove that a certain estimator is the closest to the true estimate.

We consider two tests: *placebo* test and *unobserved-confounding* test. In the first, we introduce a random variable and rerun the analysis assuming it is the treatment variable [3, 29]. Our expectation is that an estimator should return a zero causal effect. To the extent that estimate varies significantly from zero, we can assess the bias of the estimator and prune out estimators that show substantial bias. In the unobserved-confounding test, we wish to estimate how sensitive the models are to the presence of unobserved confounders. In practice, it is plausible that some confounders were missed, so we would like to prefer methods whose estimates are relatively stable in the presence of new confounds, especially not vary substantially with small changes in confounding. To verify this, we add a new confounder to the feature set with varying degrees of its effect on A and Y, and re-compute the causal effect. While we expect the causal estimate to change as the degree of effect of the confounder is increased, a better estimator is expected to be less sensitive to such changes. Thus, we can rank estimators based on the variation in their estimate for the same amount of confounder's effect added.

To implement the second test, we model the confounder as sampled from a Gaussian distribution and use the Bayes rule to arrive at the posterior distribution that shows correlation with both the outcome and the treatment. Let us denote the new confounder as U, treatment variable as A and the outcome as Y. Intuitively, we choose the degree of effect of U on Y and A to a desired value, and then use Bayes Rule to obtain the distribution of such a U.

$$P(U|Y, A) = \frac{P(Y|U, A)p(U|A)}{P(Y|A)}$$

Since we are interested in modeling the direct effect of U on Y, we can ignore the causal association between Y and A and obtain,

$$P(U|Y, A) = \frac{P(Y|U)p(U|A)}{P(Y)}$$

For our experiments, we use a parametric form for the distributions. To implement the relationship between U and A, $P(U|A)$, we assume the following the gaussian prior on U: $U \sim \mathcal{N}(u|u_0, \epsilon)$; where $u_0 = \alpha + a$, depending on a given realisation of treatment variable $A = a$ and tunable parameters α and ϵ. For the relationship between Y and U, we then define the likelihood of observing all the outcome samples using gaussian distribution as follows: $\prod_{i=1}^{N_a} \mathcal{N}(y_i|u, \epsilon)$ where N_a are the total of data points with treatment $A = a$, and ϵ is a tunable parameter denoting the strength of effect (higher ϵ corresponds to a weaker effect of U on Y) . Substituting the gaussian prior and the likelihood, we obtain:

$$\mathcal{N}(u|u^*, \epsilon^*) = \frac{(\prod_{i=1}^{N_a} \mathcal{N}(y_i|u, \epsilon)) * \mathcal{N}(u|u_0, \epsilon)}{\int \prod_{i=1}^{N_a} \mathcal{N}(y_i|u, \epsilon) du} \tag{7}$$

Based on the above solution, we obtain,

$$u^* = \frac{u_0 + N_a * \sum_{i=1}^{N_a} y_i}{N_a + 1} \qquad \epsilon^* = \frac{\epsilon}{N_a + 1}$$

Note that the posterior distribution $\mathcal{N}(u|u^*, \epsilon^*)$ is dependent on the treatment value via $u_0 = \alpha + a$. Hence, for each data point; we sample value for the new confounder U from the posterior distribution corresponding to the particular treatment class. The mean of the posterior distribution is a weighted combination of the prior mean (dependent on treatment) and the average of outcome variables, which ensures U is correlated with both A and Y.

Once the confounder is generated, we re-compute the estimate with this additional confounder and compare the estimate to its original value for each method. Since our goal is to select individuals for treatment (rather than estimating individual treatment effect), we evaluate each method on what fraction of the individuals selected for treatment stay the same between the original estimate and the new estimate. Specifically, we consider individuals whose estimate was higher than the median in both analyses, and compute the number of common individuals between the two analyses.

5 SIMULATION

We conduct a simulation study to demonstrate how `Split-Treatment` can be used to rank the effect of a target treatment that does not exist in the data. We simulate $10,000$ individuals, characterized by a 50-dimensional variable X, from a multivariate normal distribution with mean zero. We randomly assign them into 4 groups, and assume that the ground truth CATE of target treatment Z on the outcome Y is $10, 20, 30$ and 40 respectively in each group. Given each individual $\boldsymbol{x}$, the outcome measure $Y^{z=0}$ with no treatment is generated from a linear model over $\boldsymbol{x}$ (coefficients are integers
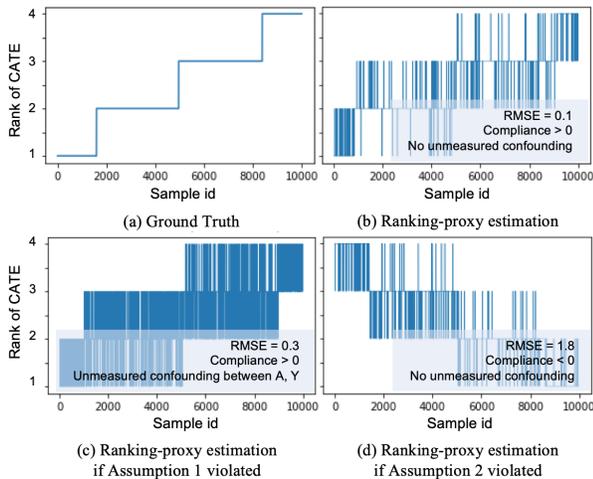
Figure 4: Comparison between the ground-truth rank and the proxy-estimated rank in simulations with or without violations of the two assumptions made in `Split-Treatment`.

randomly picked from 0 to 4 with certain probabilities), and the outcome measure $Y^{z=1}$ with treatment is generated by adding the assigned CATE value to the base measure $Y^{z=0}$. Lastly, we generate a binary proxy treatment A conditional on Z from a contingency table producing positive compliance for satisfying Assumption 2[1].

The step function in Figure 4(a) shows the ground-truth ranking of CATE of the target treatment Z. After masking Z from the simulated data, we fit a doubly robust model with outcome model being a linear regression (namely IPTW-LR), and obtain estimates of ITE for the proxy treatment A. We rank the ITE estimates and group them into 4 equally sized buckets assigned to level 1 to 4. We compute the root mean squared error (RMSE) between the estimated rank and the ground truth rank. Figure 4(b) show the estimated ranks; the obtained RMSE for our IPTW-LR method is low (0.1).

Next we study how the estimation would be biased if we violated the two untestable assumptions made in Section 3. To create a dataset that violates Assumption 1, we introduce a confounding variable $U \sim \mathcal{N}(0, 1)$, generate outcome measure $Y^{z=0}$ based on X concatenating U, and then mask U from the data. We fit the same model above, plot the estimated proxy rank in Figure 4(c) showing that it obtains a larger RMSE of 0.3. Similarly, for violating Assumption 2, we manipulate the contingency table of A given Z to produce negative compliance and fit the same model to the new simulated data. We plot the estimated proxy rank in Figure 4(d) showing that rank was reversed and RMSE increases to 1.8. This analysis shows the importance of the identifying assumptions (especially Assumption 2) for obtaining the correct estimate.

Finally, we demonstrate the value of sensitivity analysis for model elimination. Consider another doubly robust method that uses Support Vector Machine (SVM) as the outcome regression (namely IPTW-SVM). We use the placebo and the unobserved confounder methods described in Section 4 to test sensitivities of the

---

[1]Code for the simulation is provided at https://github.com/yanboxu/split-treatment-simulation
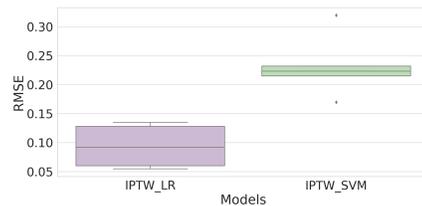

Figure 5: *Unobserved-confounding* analysis. Comparing RMSE between estimated causal effect with and without unobserved confounding, for two causal models. IPTW-LR is less sensitive to unobserved confounding. Box plots are for 5 runs with different degrees of confounding.

two methods. We find that placebo test does not help us differentiate between the two models, as both of them give high RMSEs (1.6 vs. 1.4 for IPTW-LR and IPTW-SVM respectively) which is expected due to the placebo treatment. We plot results from the unobserved confounder method in Figure 5, in which we vary the hyper parameters defined in Eq. (7) for changing the amount of effect from U to A and Y. Since ground-truth CATE is not available in practice, we compare the two models based on RMSEs between the *estimated* proxy rank under the no-confounder case and those from 5 different simulations under the confounder case. IPTW-LR obtains lower empirical RMSE than IPTW-SVM and thus it is less sensitive (more robust) to the varying degrees of confounding. This result indicates the IPTW-LR should be chosen. Sensitivity analysis , therefore allows us to select between competing models for estimating CATE.

6 APPLICATION TO REAL-WORLD DATA

6.1 Dataset

We apply our method to feature and product recommendations in a large software ecosystem. Feature and product recommendations are occasionally shown within the ecosystem as short, clickable messages that lead to web pages that educate individuals about features they already have access to in the ecosystem, and new software products that are available for download. We report on results studying a specific product recommendation encouraging individuals to try a new software product. Prior to running a recommendation campaign, we analyze existing logged user data through the `Split-Treatment` method to identify individuals most likely to benefit from the software product. In this scenario, the treatment Z is the message that encourages individuals to try the software. The `Split-Treatment` or proxy treatment A is the individual's first usage of the software during the treatment window. The outcome Y is sustained usage of the software during the post-treatment window.

We run our method on an observational data collected from $2.2M$ users within the ecosystem. For the purpose of validation, we ran a separated randomized experiment on another $1.1M$ users who are randomly exposed to the treatment Z, but had not used the recommended software during the pre-campaign window. The campaign was run from March 29 to April 27 in 2019, during which 66.1% were randomly exposed to Z. Exposure to the treatment Z encouraged the proxy treatment A as it caused a 7.5% increase in

software adoption. We align the timeline between the two datasets, take both pre-treatment and treatment window to be 30 days, and obtain the measure of sustained usage of the recommended software in varying from 0 to 30 days during the post-treatment window.

6.2 Feature extraction

We first extract two sets of snapshots on each day within the pre-treatment window. One snapshot contains 25 features summarizing the users' total usage in the past 4 weeks of different software within the ecosystem and collaboration patterns. The other contains 106 of similar features but finer grained usage information from software ecosystem. As a result, we reach at two static feature sets, namely *Features−25* and *Features−106*, by using only the last day's snapshots in the pre-treatment window; and two dynamic feature sets, namely *Features−25−Seq* and *Features−106−Seq*, by aggregating (i.e., taking average, min, max, number of increases and decreases) the 30 snapshots within the pre-treatment window.

6.3 Baselines and Algorithmic Details

Given each feature set, we explore the following models for outcome regression. Each of these models can be used in a causal mode, addressing confounding issues through use of an IPTW adjustment, or in a predictive mode that does not do such adjustment. Model hyperparameters were tuned via grid search by splitting a 10% validation set from the training data.

- Fast-Tree Regression (FTR): An efficient tree regression with gradient boosting [20].
- Fast-Forest Regression (FFR): An efficient random forest regression using the Fast-Tree learners.
- Poisson Regression (PR): A linear regression with respect to minimizing Poisson loss instead of mean squared errors.
- Convolutional Neural Network (CNN): A 2-layer 1-D convolutional network with Poisson loss.

6.4 Results

6.4.1 Evaluation on propensity scores. We build a logistic regression propensity model that predicts the likelihood that a user gets treated by A over each of the four feature sets respectively. For each propensity model, we discard the extreme propensity scores they generate that are below the .01 quantile or above the 0.99 quantile. Comparing the proxy treatment and control group distributions, we find the two groups are well overlapped across all the four feature sets. The purpose of propensity score model is for ensuring balance of potential confounds. We compute the standardized mean differences (SMD) [30] across the features in the proxy treatment and control groups to evaluate the balance. SMD calculates the divergence of the mean feature values between two groups as a fraction of the summed standard deviation of the two groups. To validate that our propensity weighting is reducing the confounding, we compare the computed SMDs before IPTW weighting and after. We find that the SMDs are all reduced after weighting, indicating that IPTW is reducing confounding. Moreover, we find that, with the exception of one feature, the SMDs of IPTW-weighted Features-106 and Features-106-Seq are all reduced to lower than 0.2, considered a threshold for well-balanced covariates [10, 30].
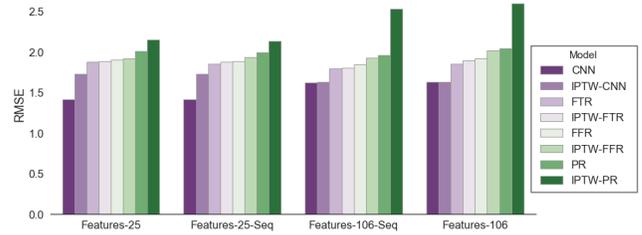

Figure 6: RMSE of outcome predictions on target data from the baseline models over different feature sets; the range of the outcome measure Y varies from 0 to 30.

6.4.2 Evaluation on outcome prediction. Now we evaluate how accurate the regression models listed in Section 6.3 can generalize their outcome predictions to an independent target dataset. In Figure 6, we see CNN models give the lowest RMSE across all the four feature sets while Poisson regression models give the highest error; all the other models perform similarly in between.

6.4.3 Model Elimination via Sensitivity Analysis. We first try the placebo-treatment test by replacing the proxy treatment in each dataset with a Bernoulli (p=0.5) random variable. We find that all methods pass the test: no method reports an estimate significantly away from zero for a placebo treatment. For the second test on sensitivity of a method to unobserved confounding, we generate multiple runs over different configurations of the hyperparameters α and ϵ in Eq. (7), as below. In choosing these hyper parameters, the goal is to generate a confounder variable U such that it is correlated with both A and Y. Different configurations capture different degrees of this correlation, and thus confounding due to U.

Features-106:
- α: 10^5, ϵ: 40*α, Corr. T: 0.38, Corr. Y: 0.18
- α: 10^5, ϵ: 100*α, Corr. T: 0.25, Corr. Y: 0.12
- α: 10^3, ϵ: 1700*α, Corr. T: 0.53, Corr. Y: 0.25

Features-25:
- α: 10^5, ϵ: 10*α, Corr. T: 0.49, Corr. Y: 0.18
- α: 10^5, ϵ: 50*α, Corr. T: 0.22, Corr. Y: 0.08
- α: 10^3, ϵ: 600*α, Corr. T: 0.58, Corr. Y: 0.22

We ensure that the correlations across the two feature sets, `Features-106` and `Features-25` are similar. Further, we only consider causal models for this analysis since the other methods do not account for confounding adjustment. Figure 7 shows the results across all three configurations, by comparing the fraction of individuals who stay in the top-50 percentile ranking of effect, even after adding the confounder. As reported in Section 4, methods that have a higher fraction of top-50 individuals that are consistent in the original data and the simulated data (with an additional confounder) are desirable. Note that we do expect that the estimates from each model will change: our goal is to find the methods that change the least with the same amount of additional confounding introduced, and thus are likely to capture more stable relationships. Our first observation is that models based on Features-106 are less sensitive to additional confounding than Features-25, except the IPTW-PR model. The two best models w.r.t. low sensitivity are IPTW-FFR
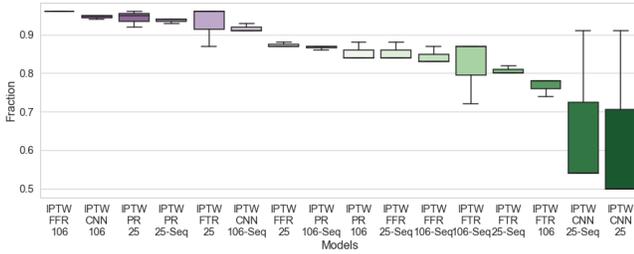
Figure 7: Sensitivity Analysis: Fraction of the top 50-percentile individuals that remain in the top 50-percentile after adding an unobserved confounder. Rankings from IPTW-FFR106 and IPTW-CNN106 stay most consistent when unobserved confounding is introduced. Box plots are for 3 runs with different degrees of confounding.

and IPTW-CNN based on Features-106. Among models, we observe models based on IPTW-FFR and IPTW-PR tend to achieve the highest fraction of consistent top-50 individuals, across both feature sets. While the IPTW-CNN model performs well for Features-106, its sensitivity is the worst among all models for Features-25. These results indicate that models based on Features-25 (and models using FTR method) may not be as robust to unobserved confounding as other methods. On balance, we therefore conclude that IPTW-FFR and IPTW-CNN methods on Features-106 may be most suitable to estimate the causal effect.

6.4.4 Validation through Active Experiments. While we may not frequently have the ability to run active experiments in practice, we do run an A/B experiment in this case, to provide an end-to-end evaluation of the proposed `Split-Treatment` method. Through our sensitivity analyses, we have identified that the methods most likely to be reliable are the IPTW-FFR on Features-106 and IPTW-CNN on Features-106; and the worst models are likely to be IPTW-CNN on Features-25-Seq and IPTW-CNN on Features-25. While we focus the presentation of our results on these best and worst models, our evaluation of the rankings produced by other models is consistent with their quality as identified by our sensitivity experiments.

The treatment assignment randomized in the A/B test is the message treatment Z, while our observational study provides an ITE estimate of the effect of the split-treatment, $P(Y|A)$. To validate the quality of the rankings implied by our observational models, we treat our randomized experiment as an instrumental variable (IV) analysis [2]: the message Z acts as a (strong) instrument in our problem because 1) it encourages A the choice to first tryout of the software (with treated individuals being 7.5% more likely to do A than untreated users); and 2) its effect on Y the sustained usage of the software is also fully mediated via its effect on the choice of A. For further details of IV analysis see [1, 2].

As an IV experiment does not allow us to calculate ITEs, we use the following strategy to validate the rankings implied by our observational studies. First, we apply the ITE prediction model learned from our observational study to predict the ITEs for individuals observed in the course of our randomized experiment. We then split individuals into two sets, $\mathcal{G}_{\text{high},k}$ and a $\mathcal{G}_{\text{low},k}$. $\mathcal{G}_{\text{high},k}$ consists of all individuals having an ITE in the top k^{th} percentile of predicted

ITEs, and $\mathcal{G}_{\text{low},k}$ includes all remaining individuals. Note that the average treatment effect in $\mathcal{G}_{\text{high},k}$ is necessarily higher than the average treatment effect in $\mathcal{G}_{\text{low},k}$. Secondly, we use our IV experiment to calculate the conditional average treatment effect (CATE) for the both $\mathcal{G}_{\text{high},k}$ and $\mathcal{G}_{\text{low},k}$.[2] By conditioning on instrument Z being randomized, we can fit a two-stage least squares (2SLS) [1] to obtain the CATE estimates $\text{CATE}^{(a)}_{\mathcal{G}_i}$ within a specified group $\mathcal{G}_i$.

If the ranked causal effects given by our observational study model are accurate, we will expect, for any splitting threshold k, that $\text{CATE}^{(a)}_{\mathcal{G}_{\text{high},k}}$ will be higher than $\text{CATE}^{(a)}_{\mathcal{G}_{\text{low},k}}$. Thus, despite the challenges of gaining ground-truth information about treatment effects, this procedure effectively allows a validation of our model's estimated rankings of the causal effect of the split-treatment A.
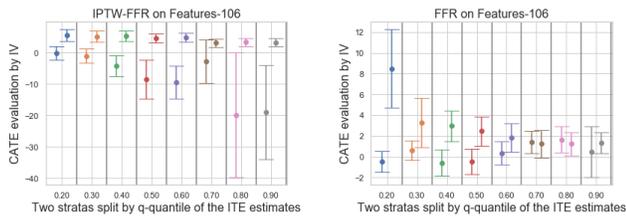
Figure 8 (a) shows the results of this validation procedure over our best causal model, as well as its non-causal counterparts. We see that the best model consistently ranks the *low* and *high* groups correctly. That is, for each threshold k, $\text{CATE}^{(a)}_{\mathcal{G}_{\text{high},k}} > \text{CATE}^{(a)}_{\mathcal{G}_{\text{low},k}}$. We see that this ranking holds particularly strongly at high thresholds. The quality of our ranking at high thresholds is particularly important in our recommendation domain as, given the multitude of possible feature recommendations, we are likely to display any particular message only to a small portion of users who are most likely to benefit. Figure 8 (b) shows the results of our validation over our worst model, identified as such by our sensitivity analyses. We see that the worst model clearly fail our experimental validation procedure. Note that the non-causal counterparts for each of these models perform worse in each case. Validation results for our other observational models (not shown) is consistent with the above results for our best and worst models—i.e., not as good as our best models, but not as bad as our worst models. Overall, we find that our validation procedure confirms that our refutation and sensitivity analyses are useful in helping to identify the most likely best models based on observational data.
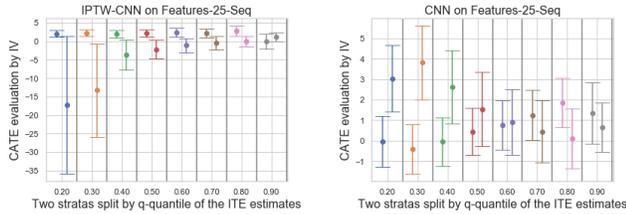
7 CONCLUSION

We presented a practical, observational analysis method for identifying individuals likely to benefit from a novel message or recommendation Z that encourages people to take action A. Through causal analysis of existing logs that contain observations of A, though not Z, we identify people who benefit from A, as measured by a target outcome Y. Under a simple assumption that Z is a positive encouragement that increases the likelihood of A, this allows targeting of the message Z to individuals most likely to benefit. A key contribution of our analysis is that our use of refutation tests and sensitivity analyses enables a principled *a priori* identification of the feature selection and elimination of unreliable algorithmic design. We validate our analysis procedure with an A/B experiment in a large real-world setting.

Promising future work includes development of additional refutation and sensitivity analyses to provide further protection against validity threats; using characteristics of the identified individuals to aid writers and marketers in the crafting of messages that better

---

[2]The CATE, when conditioned on group membership, may also be known as a Local-ATE over the group.

(a) Best model: Our validation of IPTW-FFR on Features-106 (left) shows that CATE for *low* and *high* groups are correctly separated across all values of k. The non-causal counterpart, FFR on Features-106, is not well-separated.

(b) Worst model: Our validation of IPTW-CNN on Features-25-seq (left) shows that CATE for *low* and *high* groups are not correctly separated most values of k. The non-causal counterpart, CNN on Features-25-seq, is worse.

Figure 8: Our CATE evaluation of the best model vs. the worst model chosen by sensitivity analysis (left), as well as their non-causal counterparts (right). The x-axis represents the threshold k, and the y-axis the CATE estimate. For each k, we show the CATE estimate, with standard error, for the paired *low* and *high* groups in the same color.

express the benefits individuals may receive; and expansion of our procedures to jointly analyze multiple treatments and outcomes.

REFERENCES

[1] Joshua D Angrist and Guido W Imbens. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association* 90, 430 (1995), 431–442.
[2] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association* 91, 434 (1996), 444–455.
[3] Susan Athey and Guido W Imbens. 2017. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives* 31, 2 (2017), 3–32.
[4] Kay H Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, Steven L Scott, et al. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* 9, 1 (2015), 247–274.
[5] David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. 2010. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 7–16.
[6] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. Double/debiased/neyman machine learning of treatment effects. *American Economic Review* 107, 5 (2017), 261–65.
[7] Iván Díaz and Mark J van der Laan. 2013. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics* 9, 2 (2013), 149–160.
[8] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 198–206.
[9] Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science* 38, 2 (2019), 193–225.
[10] Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Twelfth*

[11] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.
[12] Finnian Lattimore, Tor Lattimore, and Mark D Reid. 2016. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*. 1181–1189.
[13] Randall A Lewis and David H Reiley. 2014. Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on Yahoo! *Quantitative Marketing and Economics* 12, 3 (2014), 235–266.
[14] Ang Li and Judea Pearl. 2019. Unit Selection Based on Counterfactual Logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 1793–1799. https://doi.org/10.24963/ijcai.2019/248
[15] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.
[16] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. 2019. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*. 4932–4941.
[17] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.
[18] Judea Pearl. 2000. Causality: models, reasoning, and inference. *Cambridge University Press. ISBN 0* 521, 77362 (2000), 8.
[19] Judea Pearl. 2009. *Causality*. Cambridge university press.
[20] Korlakai Vinayak Rashmi and Ran Gilad-Bachrach. 2015. DART: Dropouts meet Multiple Additive Regression Trees.. In *AISTATS*. 489–497.
[21] Marc Resnick and William Albert. 2014. The impact of advertising location and user task on the emergence of banner ad blindness: An eye-tracking study. *International Journal of Human-Computer Interaction* 30, 3 (2014), 206–219.
[22] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental economics* 9, 2 (2006), 79–101.
[23] James M Robins. 2000. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 95–133.
[24] James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. 2000. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 1–94.
[25] Paul R Rosenbaum et al. 2010. *Design of observational studies*. Vol. 10. Springer.
[26] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).
[27] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3076–3085.
[28] Amit Sharma, Jake M Hofman, Duncan J Watts, et al. 2018. Split-door criterion: Identification of causal effects through auxiliary outcomes. *The Annals of Applied Statistics* 12, 4 (2018), 2699–2733.
[29] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* (2020).
[30] Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.
[31] Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009 (2009).
[32] S Shyam Sundar, Sunetra Narayan, Rafael Obregon, and Charu Uppal. 1998. Does web advertising work? Memory for print vs. online media. *Journalism & Mass Communication Quarterly* 75, 4 (1998), 822–835.
[33] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. 2017. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*. 3632–3642.
[34] Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. 2013. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1587–1594.
[35] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.
[36] Pengyuan Wang, Wei Sun, Dawei Yin, Jian Yang, and Yi Chang. 2015. Robust tree-based causal inference for complex ad effectiveness analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 67–76.
[37] Russell S Winer. 2001. A framework for customer relationship management. *California management review* 43, 4 (2001), 89–105.
[38] Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. 2020. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835* (2020).