Robust Tree-based Causal Inference for Complex Ad Effectiveness Analysis

Pengyuan Wang† Wei Sun§ Dawei Yin† Jian Yang† Yi Chang†
†Yahoo Labs, Sunnyvale, CA, USA
§Purdue University, West Lafayette, IN,USA
†{pengyuan, daweiy, jianyang,yichang}@yahoo-inc.com §sun244@purdue.edu

ABSTRACT

As the online advertising industry has evolved into an age of diverse ad formats and delivery channels, users are exposed to complex ad treatments involving various ad characteristics. The diversity and generality of ad treatments call for accurate and causal measurement of ad effectiveness, i.e., how the ad treatment *causes* the changes in outcomes without the confounding effect by user characteristics. Various causal inference approaches have been proposed to measure the causal effect of ad treatments. However, most existing causal inference methods focus on univariate and binary treatment and are not well suited for complex ad treatments. Moreover, to be practical in the data-rich online environment, the measurement needs to be highly general and efficient, which is not addressed in conventional causal inference approaches.

In this paper we propose a novel causal inference framework for assessing the impact of general advertising treatments. Our new framework enables analysis on uni- or multi-dimensional ad treatments, where each dimension (ad treatment factor) could be discrete or continuous. We prove that our approach is able to provide an unbiased estimation of the ad effectiveness by controlling the confounding effect of user characteristics. The framework is computationally efficient by employing a tree structure that specifies the relationship between user characteristics and the corresponding ad treatment. This tree-based framework is robust to model misspecification and highly flexible with minimal manual tuning. To demonstrate the efficacy of our approach, we apply it to two advertising campaigns. In the first campaign we evaluate the impact of different ad frequencies, and in the second one we consider the synthetic ad effectiveness across TV and online platforms. Our framework successfully provides the causal impact of ads with different frequencies in both campaigns. Moreover, it shows that the ad frequency usually has a treatment effect cap, which is usually over-estimated by naive estimation.

Categories and Subject Descriptors

G.3 [**PROBABILITY AND STATISTICS**]: Statistical Computing; J.1 [**ADMINISTRATIVE DATA PROCESSING**]: Business, Marketing

WSDM'15, February 2–6, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2684822.2685294.

Keywords

Causal Inference; Advertising; Online Strategy Measurement

1. INTRODUCTION

In the current online advertising ecosystem, user are exposed to ads with diverse formats and channels, and user's behaviors are caused by complex ad treatments combining of various factors. The online ad delivery channels include search, display, mail, mobile and so on. Besides the multi-channel exposure, ad creative characteristics and context may also affect ad effectiveness [33]. Hence the ad treatments are becoming a combination of various factors mentioned above. The complexity of ad treatments calls for accurate and causal measurement of ad effectiveness, i.e., how the ad treatment *causes* the changes in outcomes [1].

Ideally, the gold standard of accurate ad effectiveness measurement is the experiment-based approach, such as A/B test, where different ad treatments are randomly assigned to users. However, the cost of fully randomized experiments is usually very high [10, 25, 35] and in some rich ad treatment circumstances, such fully randomized experiments are even infeasible. The major obstacles to achieve fully randomized experiments are as follows. 1) Implementing a platform for supporting ideal experiments, i.e., perfect randomization, often involves the change of system architecture, which might cause much prohibited engineering effort. 2) When the treatments are a combination of various factors, one might not be able to fully explore all possible combinations of treatments due to the lack of population. 3) The treatment may not be feasible for large-scale experiments, such as the number of ad impressions. In online advertising, it is easy to randomly assign users to see or not see the ad impression, but it is difficult to fully control the number of impressions, except utilizing field experiment [2], which is costly and usually can be conducted only on a relatively small scale. 4) Even if the experiments are perfectly randomized and the ad treatments can fit into an experiment framework, one still should be cautious due to the fact that the randomized experiments may hurt both user experience and ad revenue. Hence it is critical and necessary to provide statistical approaches to estimate the ad effectiveness directly from observational data rather than experimental data.

Previous studies based on observational data try to establish direct relationship between the ad treatment and a success signal, e.g., purchase, brand keyword searching [1], and etc. However, in obser-

[1]An outcome is the user's response in the ad campaigns, such as whether or not the user clicks a link, searches for the brand or visits websites. A success is an action favored by the advertiser, such as click, search or site visitation.

[2]A field experiment is a research study conducted in a natural setting in which the experimenter manipulates one or more independent variables under carefully controlled conditions [17].

vational data, typically the user characteristics may affect both the exposed ad treatment and the success tendency. Such confounding effects of user characteristics are called 'selection biases' [32], and ignoring the confounding effects may lead to biased estimation of the treatment effect [32]. For example, assume in an auto campaign all of exposed users are males and all of the non-exposed users are females. If the males generally have a larger success rate than females, the effectiveness of the campaign could be overestimated because of the confounding effects of the user characteristics—in this case, gender: It might just be that males are more likely to be exposed and perform success actions. Therefore, the relationship between the ad treatments and the success is not causal without eliminating the selection bias. A straightforward approach attempting to eliminate the selection biases is to adjust the outcome with the user characteristics using supervised learning. However the user characteristics may have complex relationships, e.g., nonlinearity, with the treatments and the outcome, and it is not trivial to estimate the causal effect of the treatment by adjusting the outcome with the user characteristics directly.

To address the aforementioned problems, the causal inference has started to attract attention in the online advertising field [10, 13, 35, 38] . Causal inference aims to infer unbiased causality effect of the ad treatment from observational data on chosen outcome metric, e.g., the effect of ads on success rates, by eliminating the impact of the confounding factors such as user characteristics. However, measuring general/complex ad treatment effectiveness is still facing three major challenges. First, the general ad treatment can be much more complex than binary ad treatment. It could be a discrete or continuous, single- or multi-dimensional treatment. To design an analytics framework encompassing so many ad factors is not trivial. Second, the online observational dataset typically has huge volume of records and user characteristics, which demands the methodology to be highly efficient. Traditional statistical causal inference approaches usually cannot reach efficiency required by the advertising industry. Third, when the treatments become more complex, existing methods are usually sensitive to parameter settings (more details in Section 5). To overcome the sensitivity, a robust causal inference approach is also wanted.

In this paper, we propose a computationally efficient tree-based causal inference framework to tackle the general ad effectiveness measurement problem. Our model is well suited for the online advertising datasets which consist of complex treatments, huge volume of users, and high-dimensional features. The novelty and advantages of the proposed method can be summarized as follows:

- Our causal inference is fully general, where the treatment can be single- or multi-dimensional, and it can be binary, categorical, continuous, or a mixture of them. We prove that this framework offers an unbiased estimation of the treatment effect under standard assumptions.
- Compared to previous causal inference work, the proposed approach is more robust and highly flexible with minimal manual tuning. Our tree-based approach automatically determines the important tuning parameters that were chosen arbitrarily in the traditional causal inference methods in a nonparametric way. In addition, it is easy to implement and computationally efficient for large scale online data.
- The tree-based framework is further wrapped in a bagging (boostrapping) procedure to enhance the stability and improve the performance of the final estimator. More importantly, our bagged strategy provides with statistical inference of the obtained point estimators, where the confidence intervals of the estimated treatment effects could be established for hypothesis testing purpose.

We apply the framework to an online advertising campaign and a cross-platform campaign that involves both TV and online platforms, and provide practical guideline to assess advertising strategy on one or more platforms. Our extensive experiments provide the causal impact of ad with different frequencies from one or more platforms, and further show that the ad frequency usually has a treatment effect cap that could have been over-estimated by naive estimations. Hence it is important for the ad providers to make appropriate adjustment for the number of the ads delivered to the users.

Our framework is not limited to online advertising, but is also applicable to other tasks (e.g., social science, and user engagement studies) where causal impact of general treatments (e.g., UI design, content format, ad context, and etc.) needs to be measured with observational data.

2. METHODOLOGY

In this section we briefly review the causal inference theory [21], and then propose our tree-based causal inference framework.

We define the set of potential treatment values to be $\mathcal{T}$, and hence each value $\mathbf{t} \in \mathcal{T}$ indicates a specific treatment, which can be uni- or multi-dimentional. For a specific user, the treatment is a random variable $\mathbf{T}$, which is supported on $\mathcal{T}$. Similarly we define the potential outcome associated with a specific treatment $\mathbf{t}$ as $Y(\mathbf{t})$, which is the random variable mapping the given treatment $\mathbf{t}$ to a potential outcome supported on the set of potential outcomes $\mathcal{Y}$. Since the treatment can be uni- or multi-dimensional, we use the boldface $\mathbf{T}$ and $\mathbf{t}$ to indicate a multivariate treatment variable and T and t to indicate a univariate treatment variable. In this paper, all the methodologies designed for multivariate treatment $\mathbf{T}$ may also be applied to univariate treatment T. In the binary treatment case, $\mathcal{T} = \{0,1\}$ with 1 indicates, for example, ad exposure and 0 indicates no ad exposure. In general, $\mathcal{T}$ could be multivariate and of a mixture of categorical and continuous variables. Typically, one would like to evaluate the effect of treatment $\mathbf{t}$ on the outcome Y, removing the confounding effect of $\mathbf{X}$.

For each user, indexed by $i = 1,2,...,N$, we observe a vector of pretreatment covariates (i.e., user characteristics) $\mathbf{X}_i$ of length p, a treatment $\mathbf{T}_i$, and an univariate outcome Y_i (e.g., purchase indicator) corresponding to the treatment received.

2.1 Primer for Causal Inference

In the causal inference framework for treatment impact measurement, two standard assumptions are usually made in order to unbiasedly evaluate the effect of the treatment [21, 30].

Assumption 1: Stable unit treatment value assumption. "The potential outcome for one unit should be unaffected by the particular assignment of treatments to the other units" [12]. This assumption allows us to model the outcome of one subject independent of another subject's treatment status, given the covariates.

Assumption 2: Strong ignorability of treatment assignment (also called 'Unconfoundedness'). Given the covariates $\mathbf{X}$, the distribution of treatment assignment $\mathbf{T}$ is independent of the potential outcome $Y(\mathbf{t})$ for all $\mathbf{t} \in \mathcal{T}$. This assumption allows us to model the treatment with respect to the covariates, independent of the outcome. It means all the user characteristics that are related to both the treatment assignment and the outcome have been collected.

When making causal inference, the primary interest is the distribution $p(Y(\mathbf{t})|\mathbf{X})$ for each $\mathbf{t} \in \mathcal{T}$ and fixed $\mathbf{X}$, or its average over the population $p(Y(\mathbf{t})) = \int_X p(Y(\mathbf{t})|\mathbf{X})p(\mathbf{X})d\mathbf{X}$. Due to the fact that in observational studies we observe only one of the potential outcome $Y(\mathbf{T} = \mathbf{t}) \in \mathcal{Y}$ for each $\mathbf{t} \in \mathcal{T}$, we must condition on the observed treatment assignment in order to obtain $p(Y(\mathbf{t}))$. As pointed

out in [21], a solution is to condition on the observed covariates. According to Assumption 2, we have $p(Y(\mathbf{t})|\mathbf{T}=\mathbf{t},\mathbf{X}) = p(\mathbf{T}=\mathbf{t}|Y(\mathbf{t}),\mathbf{X})p(Y(\mathbf{t})|\mathbf{X})/p(\mathbf{T}=\mathbf{t}|\mathbf{X}) = p(Y(\mathbf{t})|\mathbf{X})$, and hence

$$p(Y(\mathbf{t})) = \int_{\mathbf{X}} p(Y(\mathbf{t})|\mathbf{T}=\mathbf{t},\mathbf{X})p(\mathbf{X})d\mathbf{X}. \quad (1)$$

In principle, one can model $p(Y(\mathbf{t})|\mathbf{T}=\mathbf{t},\mathbf{X})$ directly, but experience shows that the result can be strongly biased if the relationship between $\mathbf{T}$ and $\mathbf{X}$ is omitted or mis-specified [21]. When the observed covariates $\mathbf{X}$ is low-dimensional, one way to avoid this bias is to classify subjects according to $\mathbf{X}$ and estimate (1) via the weighted average over $\mathbf{X}$. However as the dimension of $\mathbf{X}$ increases, exact sub-classification according to covariates becomes computationally infeasible.

To address the high-dimensionality issue of $\mathbf{X}$, Rosenbaum and Rubin [30] introduced the balancing score to summarize the information required to balance the distribution of covariates and proposed the propensity score method for the binary treatment problem. The balancing score is the random variable such that conditioned on it, the observed covariates and the treatment assignment are independent. By sub-classifying the balancing score, we can obtain a valid causal inference of the treatment effect. The basic algorithm is summarized in Table 1. Later on, Imai and van Dyk [21] introduced the propensity function to generalize the propensity score method to general treatments. Specifically, the propensity function $e(\mathbf{X})$ is defined as the conditional density of the treatment given the observed covariates, i.e., $e(\mathbf{X}) = p(\mathbf{T}|\mathbf{X})$. It was shown that this propensity function is a balancing score, that is, $p(\mathbf{T}|\mathbf{X}) = p(\mathbf{T}|e(\mathbf{X}))$. Hence we can obtain $p(Y(\mathbf{t}))$ in (1) as

$$p(Y(\mathbf{t})) = \int_{e(\mathbf{X})} p(Y(\mathbf{t})|\mathbf{T}=\mathbf{t},e(\mathbf{X}))p(e(\mathbf{X}))de(\mathbf{X}). \quad (2)$$

To compute the integral in (2), Imai and van Dyk [21] assumed that there existed a unique finite-dimensional parameter θ such that the propensity function $e(\mathbf{X})$ depended on $\mathbf{X}$ only through $\theta(\mathbf{X})$. For example, when the conditional distribution of $\mathbf{T}|\mathbf{X} \sim \mathrm{N}(\mathbf{X}^T\beta,\Sigma)$ with some parameter β, the propensity function $e(\mathbf{X})$ is the Gaussian density function which can be fully characterized by $\theta(\mathbf{X}) = \mathbf{X}^T\beta$. In this case, θ is also a balancing score, and hence we can obtain $p(Y(\mathbf{t}))$ in (2) as

$$p(Y(\mathbf{t})) = \int_{\theta} p(Y(\mathbf{t})|\mathbf{T}=\mathbf{t},\theta)p(\theta)d\theta. \quad (3)$$

This integral can be approximated by classifying the subjects into several sub-classes with similar value of θ, estimate the treatment effect within each sub-class, and then average the estimators from each sub-class. Usually θ has a much smaller dimension than X, hence this strategy tackles the high dimensionality issue of the covariates in (1) [21].

Table 1: Algorithm of Propensity Score Method with Sub-classification

Input:	Y_i, $\mathbf{X}_i$, treatment T_i for $i = 1,2,...N$.
Output:	Estimated treatment effect for t.
Step 1:	Find a balancing score B_i for each subject such that $T_i \perp \mathbf{X}_i|B_i$.
Step 2:	Sub-classify the subjects with similar balance score B_i into S sub-classes.
Step 3:	Within each sub-class s, calculate the number of subjects N_s and estimate the treatment effect $R_s(t)$ for each treatment t.
Step 4:	Estimate the population treatment effect as a weighted average of $R_s(t)$, where the weight is proportional to N_s.

2.2 Robust Tree-Based Causal Inference

The approach in (3) is vulnerable to the model misspecification when assuming the parametric form of $\mathbf{T}|\mathbf{X}$, and the final treatment effect estimation is sensitive to the number of sub-classes and the strategy of sub-classification [19]. A larger number of sub-classes leads to a more accurate estimation of the integral in (3) but inevitably implies a less accurate estimation of the inner conditional distribution $p(Y(\mathbf{t})|\mathbf{T}=\mathbf{t},\theta)$ due to limited observations in each sub-class. Furthermore, although equal-frequency strategy is generally used to form the sub-classes [21, 32], experiments showed that this strategy often leads to highly unable estimators for the extreme sub-classes [19]. Therefore, it is in great demand to introduce a model free method which can avoid the choice of the number of sub-classes and the strategy of sub-classification. To overcome the aforementioned issues, we propose a tree-based causal inference method for general treatment problem.

Recall that we can obtain the unbiased estimation of treatment effect by (2). Naturally we can approximate this integral by classifying the subjects into several sub-classes with similar value of $e(\mathbf{X})$, and then average the estimators from each sub-class. We utilize the tree structure to model $e(\mathbf{X})$ nonparametrically and classify the users automatically (Section 2.2.2). The number of sub-classes is also determined by the tree model, and hence we avoid arbitrary selection of the number of sub-classes. Compared to the previous methods, the tree-based model is a nonparametric approach, which requires fewer assumptions. We also propose a bootstrapping aggregated approach to further boost the performance (Section 2.2.2).

2.2.1 Tree-Based Causal Inference

As discussed above, the unbiased treatment estimation $p(Y(\mathbf{t})) = \int_{e(\mathbf{X})} p(Y(\mathbf{t})|\mathbf{T}=\mathbf{t},e(\mathbf{X}))p(e(\mathbf{X}))de(\mathbf{X})$ can be approximated by classifying users with similar $e(X)$, and then one can obtain unbiased estimation of the treatment effect with each sub-class. The tree-based structure naturally partitions the given predictor space into disjoint groups, and hence is ideal to automize the classification and the rest of the causal inference calculation.

Specifically, we hereby adopt a tree-structure on modeling $p(\mathbf{T}|\mathbf{X})$ with the treatment $\mathbf{T}$ as the dependent variable and the covariates $\mathbf{X}$ as the independent variables, which builds the tree in the way that the distribution of treatment becomes homogeneous within each partitioned sub-class. In other words, through tree-based method, we automatically obtain the benefit of the two-step sub-classification method (Table 1), which first models $\mathbf{T}|\mathbf{X}$ to find a balancing score and then partitions the subjects based on the value of the balancing score.

We then estimate the treatment effect within each leaf node, and take the weighted average across all the leaf nodes as the final estimation. The detailed algorithm is described in Table 2. In Figure 8 and Table 7 of Section 4.2, we employ an example to illustrate each step of this algorithm.

Table 2: Algorithm of (Single) Tree-Based Model

Input:	Y_i, $\mathbf{X}_i$, treatment $\mathbf{T}_i$ for $i = 1,2,...N$.
Output:	Estimated treatment effect for $\mathbf{t}$.
Step 1:	Fit a tree-based model with dependent variable $\mathbf{T}_i$ and independent variable $\mathbf{X}_i$.
Step 2:	Within each leaf node s, calculate the number of subjects N_s and estimate the treatment effect $R_s(\mathbf{t})$ for each treatment $\mathbf{t}$.
Step 3:	Calculate the final treatment effect as in (4).

Note that the method to estimate the treatment effect R_s in step 2 may vary with great flexibility. For example, when the treatment $\mathbf{T}$ is discrete, a straightforward nonparametric way to estimate the treatment effect in each node s is to compute the average of outcome Y corresponding to various treatments $\mathbf{T}$, and then subtract the averaged outcome of a baseline treatment. For instance, for a

bivariate and binary treatment $\mathbf{T} = (T_1, T_2)^T$ with $(T_1, T_2) \in \{0,1\}^2$, within each node s, we can estimate the effect of treatment $\mathbf{t}$ as $R_s(\mathbf{t}) = \bar{Y}(\mathbf{t}) - \bar{Y}(\mathbf{t}_0)$ with $\mathbf{t}_0 = (0,0)^T$ as the baseline treatment, where $\bar{Y}(\cdot)$ refers to the averaged outcome. When the treatment $\mathbf{T}$ is continuous, one could choose to fit any proper nonparametric or parametric model for $Y|(\mathbf{T}, \mathbf{X})$ within each sub-class s. The choice of the specific model to fit within leaf node s is not the focus of this paper, but our algorithm is flexible that any proper model can be utilized.

Under the two standard assumptions as in Section 2.1, we prove that the proposed tree-based causal inference estimation is unbiased.

THEOREM 1. *Under Assumptions 1-2 and the condition that the subjects in each leaf node have a homogeneous density of* $\boldsymbol{T}$*, the effect of treatment* $\boldsymbol{t}$ *is equal to the expected outcome corresponding to treatment* $\boldsymbol{t}$ *averaged over the leaf node in the proposed tree-based method.*

Proof of Theorem 1: The condition that in each leaf node the subjects have a homogeneous density of $\mathbf{T}$ implies that conditioning on the leaf node is equivalent to conditioning on $e(\mathbf{X}) = p(\mathbf{T}|\mathbf{X})$. Therefore, the treatment effect induced from the proposed algorithm is $E_{e(\mathbf{X})}\{E[Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, e(\mathbf{X})]\}$ for any given treatment $\mathbf{t}$. Next we will show that this equals $E[Y(\mathbf{t})]$. According to the assumptions 1-2 and the fact that $e(\mathbf{X})$ is a balancing score, we have

$$p(Y(\mathbf{t}), \mathbf{T}, \mathbf{X}|e(\mathbf{X})) = p(Y(\mathbf{t}), \mathbf{X}|e(\mathbf{X}))p(\mathbf{T}|e(\mathbf{X})).$$

Integrating both sides with respect to $\mathbf{X}$ leads to the conclusion that $Y(\mathbf{t}) \perp\!\!\!\perp \mathbf{T}|e(\mathbf{X})$. Therefore, for any $\mathbf{t} \in \mathcal{T}$, we have $E[Y(\mathbf{t})] = E_{e(\mathbf{X})}\{E[Y(\mathbf{t})|e(\mathbf{X})]\} = E_{e(\mathbf{X})}\{E[Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, e(\mathbf{X})]\}$. This ends the proof of Theorem 1. ■

We comment that the assumption of a homogeneous density of $\mathbf{T}$ within each leaf node is standard. Similar condition has also been imposed in Theorem 5 of [36]. An intuitive explanation is that the tree automatically seeks the partition such that the predictor space is the most separable and hence the distribution of $\mathbf{T}$ gets more and more homogeneous within each leaf node as the tree grows [36].

In our algorithm, we follow the CART [7] guideline to construct the single tree in Step 1 of Table 2. The tuning parameters are chosen based on a 10-fold cross validation. After the tree construction, within each leaf node s, we estimate $R_s(\mathbf{t})$ and then estimate the final averaged treatment effect (ATE) as

$$\widehat{\text{ATE}} = \sum_s \frac{N_s}{N} \{R_s(\mathbf{t}) - R_s(\mathbf{t}_0)\}, \tag{4}$$

where $\mathbf{t}_0$ is the baseline treatment.

2.2.2 Bagged Tree-Based Causal Inference

A single tree suffers from the instability issue since a small perturbation to the samples may lead to tremendous changes in the constructed tree [7]. The bootstrap aggregating (bagging) introduced by Breiman [6] is frequently applied to enhance the performance of non-robust methods by reducing the variance of a predictor . We here adopt the bagging strategy to improve the robustness of our framework.

In the bagged tree-based causal inference, we repeatedly generate bootstrap samples (i.e., a set of random samples drawn with replacement from the dataset), estimate the treatment effect based on the samples, and calculate the final results by averaging the results from the bootstrap sample sets at the end. The detailed implementation is in Table 3. The theoretical justification of this bagged tree-based causal inference model is due to the unbiased estimation of single tree-based method in Theorem 1 and the consistency of the bagging method for CART shown by Bühlmann and Yu [8].

Table 3: Algorithm of Bagged Tree-Based Model

Input:	Y_i, $\mathbf{X}_i$, treatment $\mathbf{T}_i$ for $i = 1, 2, ...N$.
Output:	Estimated treatment effect for $\mathbf{t}$.
Step 1:	Construct a bootstrap sample $\mathcal{D}^*$ according to the empirical distribution of the observations.
Step 2:	Compute the bootstrapped treatment effect estimator $\widehat{\text{ATE}}^*$ based on $\mathcal{D}^*$ via Table 2.
Step 3:	Repeat Steps 1-2 B times and output the final estimator $\widehat{\text{ATE}}_{\text{B}}$ in (5) and SD in (6).

Note that our bagged causal inference model is able to establish the confidence interval of the estimated treatment effect. In Step 3 of Table 3, we can calculate the bootstrapped mean $\widehat{\text{ATE}}_{\text{B}}$ and standard deviation SD of $\widehat{\text{ATE}}_{\text{B}}$ according to B bootstrapped treatment effect estimators in Step 2. Specifically,

$$\widehat{\text{ATE}}_{\text{B}} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\text{ATE}}^{*(b)}, \tag{5}$$

$$\text{SD} = \frac{1}{\sqrt{B}}\left(\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\text{ATE}}^{*(b)} - \widehat{\text{ATE}}_{\text{B}}\right)^2\right)^{1/2}, \tag{6}$$

where b is the bootstrap sample set index. If the bootstrapped estimators follow a normal distribution, we can obtain the 95% confidence interval of ATE as $(\widehat{\text{ATE}}_{\text{B}} - 1.96 * \text{SD}, \widehat{\text{ATE}}_{\text{B}} + 1.96 * \text{SD})$. Otherwise, the middle 95% quantile of $\widehat{\text{ATE}}^{*(1)}, \ldots, \widehat{\text{ATE}}^{*(\text{B})}$ could be used.

Another advantage of the bagged model is that flexible subsampling strategies can be incorporated into robust causal inference framework, which is useful, even necessary when addressing some practical problems. For instance, modern online ad datasets typically suffer from the severely imbalanced outcome, e.g., most users are not exposed to any ad. Our robust causal inference framework is able to employ the subsampling and backscaling strategy as described in [38]. Specifically, in the subsampling step, we sample the exposed users with a higher probability than the non-exposed users and estimate the treatment effect via our framework based on the sample dataset. The success rates calculated from the subsamples are then back-scaled to the whole population level according to the sampling rates. This method has been examined carefully in [38] and shown to achieve substantial improvement of out-of-sample predictions.

2.2.3 Summary

The utilization of the tree-based structures fully automizes the causal inference pipeline in a nonparametric way without prespecified assumptions for outcome or treatment models, and the bagging further improves the robustness. To sum up, we visualize the approach with the following roadmap:

1. In order to obtain a valid causal inference of the treatment, we need to account for the confounding impact of user features (Figure 1).
2. Employ a tree model to estimate the propensity function (Figure 2). The tree model automatically classifies the users into groups within which the distribution of treatment is homogeneous.
3. Conditioned on each propensity function, the treatment and user features are independent. Therefore, within each tree leaf node, the impact of the treatment on the outcome is the real treatment effect (Figure 3). We then estimate the population-level treatment effect by a weighted average of the estimators from each leaf node.

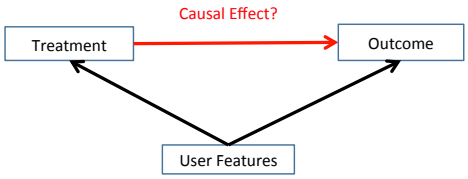

Figure 1: Confounding effect of user features.

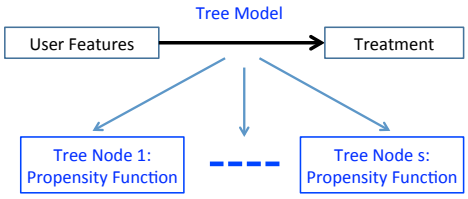

Figure 2: Estimate the propensity function via the proposed tree model.

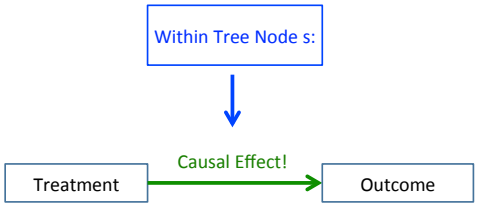

Figure 3: Obtain causal effect by conditioning on the propensity function estimated from each tree leaf node.

3. SIMULATION STUDY

To better understand our framework, we use two simulated examples to asses our tree-based causal inference framework and compare it with direct modeling methods. The simulation results show that, when there are confounding covariates contributed to both treatment assignments and outcome, which is common in observational studies, direct modeling methods without propensity function adjustment lead to severe biases, while our tree-based framework shows no significant biases and significantly improves the accuracy and robustness of the treatment impact measurement.

3.1 Without Treatment Effect

In the first simulation, the treatments have no causal effect on the outcome and the superficial correlation between treatments and outcome is due to the confounding effects of the covariates.

We set the 5-dimensional covariate $\mathbf{X}$ as the exponential values of five randomly generated standard normal variables. This generation mimics our real data where the covariates are all nonnegative. We then generate two binary treatments T_1, T_2 and a binary outcome Y. This process is summarized as follows.

$$\mathbf{Z} = (Z_1, \ldots, Z_5)^T \sim N_5(0, \mathrm{I});$$
$$\mathbf{X} = \exp(\mathbf{Z});$$
$$T_j \sim \text{Bernoulli}(\pi_j) \text{ with } \text{logit}(\pi_j) = \beta_j^T \mathbf{X} \text{ for } j = 1, 2;$$
$$Y \sim \text{Bernoulli}(\pi_y) \text{ with } \text{logit}(\pi_y) = \beta_y^T \mathbf{X}.$$

We set the coefficients $\beta_1 = (1, -1, -1, -1, 1)$, $\beta_2 = (1, 0, -1, -1, 1)$, and $\beta_y = (1, -1, -1, -1, 0)$. We choose these coefficients to ensure that the outcome is roughly balanced. In the above simulation, the treatment $\mathbf{T}$ is not involved in the generation of Y and there is no treatment effect. The sample size is fixed as $n = 1000$.

Table 4: Percentages of mistakenly considering the coefficients of treatments to be different than 0. "T1" and "T2" refer to the treatment T_1 and treatment T_2, respectively, and "either" means either the effect of treatment T_1 or that of treatment T_2 is overestimated.

Model A			Model B		
T1	T2	Either	T1	T2	Either
11%	9%	20 %	0	0	0

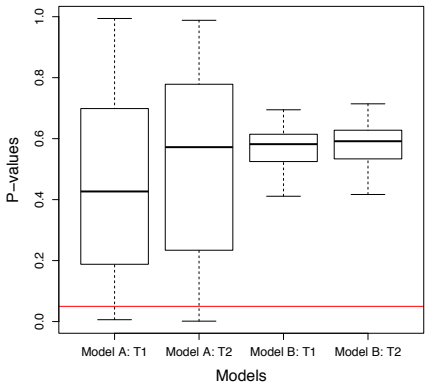

Figure 4: P-values for the coefficients of treatments. "Model A: T1" ("Model A: T2") corresponds to the effect of treatment T_1 (T_2) in Model A. Similar meanings are used for Model B. When the p-value is less than 0.05 (red line), the corresponding estimated treatment effect is significantly different from 0.

We compare the direct modeling method and the proposed tree-based method in fitting the simulation dataset.

1) **Model A: Direct Modeling.** A linear logistic regression with dependent variable Y and independent variables $(\mathbf{T}, \mathbf{X})$. The treatment impact is measured as the fitted coefficient of $\mathbf{T}$.

2) **Model B: Our Proposed Model.** The proposed tree-based method in Table 2. In step 3 of Table 2, within each mode s, we fit a linear logistic regression for Y with independent variables $(\mathbf{T}, \mathbf{X})$. The treatment impact is captured by the averaged coefficients of $\mathbf{T}$ over all nodes.

We repeat the above generation and inference process 100 times, and record the number of times that these two methods mistakenly consider the coefficients of $\mathbf{T}$ to be significantly different than 0 under the 95% confidence level. Recall that the true causal impacts of both treatments are 0. However, as shown in Table 4, the direct modeling approach (Model A) incorrectly discovers the causal impact of either treatment 20% of the times. On the other hand, our approach (Model A) always correctly finds that the treatment impact is not significantly different than 0. This illustrates the advantage of our model.

Next we plot the p-values of the fitted coefficients for treatments T_1 and T_2 in Figure 4. Note that the true coefficients for both treatments are not significantly different from 0 and ideally the p-values should be larger than 0.05. As shown in Figure 4, Model A overestimates the treatment effect as some of their p-values are very small. Furthermore, it produces a very unstable p-values over the 100 replicates, which hinders its applicability in practice. In contrast, Model B delivers very stable and accurate inference results.

Finally, we demonstrate the covariance balance performance of our tree-based algorithm. Figure 5 plots the absolute values of the correlations between each treatment and each of the covariates in Model A and Model B. Due to the propensity function adjustment, our Model B shows a considerable reduction in the correlation over Model A. This illustrates why our causal inference model improves the performance over the directly modeling methods.

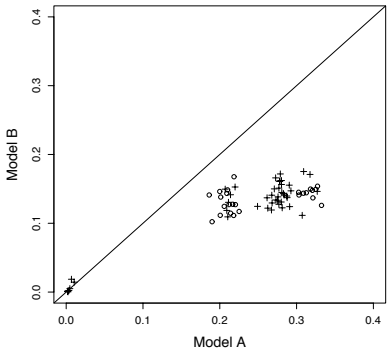

Figure 5: Correlations between two treatments and the covariates in Model A and Model B. The circle (cross) indicates the correlation with treatment T_1 (T_2). Results are from 5 replications.

3.2 With Treatment Effect

In the second simulation, both treatments have real causal effects on the outcome. Specifically, we generate data as follows.

$$
\begin{aligned}
&X_1, X_2, X_3, X_4, X_5 \overset{iid}{\sim} U[0,1];\\
&T_1 \sim \text{Bernoulli}(\pi_1) \text{ with } \text{logit}(\pi_1) = 1 - X_1 - X_2;\\
&T_2 \sim \text{Bernoulli}(\pi_2) \text{ with } \text{logit}(\pi_2) = 1 - 2X_2 - 2X_3;\\
&Y \sim \text{Bernoulli}(\pi_y) \text{ with } \text{logit}(\pi_y) =\\
&\quad -2 + 2T_1 - 2T_2 + 2X_1 - 2X_2 + 2X_3 - 2T_1X_3 + 2T_2X_4.
\end{aligned}
$$

In this example, covariates X_1, X_2, X_3 are confounders that affect both treatment $\mathbf{T}$ and outcome Y, covariate X_4 is an causal effect modifier, and X_5 is a totally irrelevant feature.

In this 2-dimensional treatment example, the true effect of treatment T_j is $\text{E}[Y(\mathbf{t}_j)] - \text{E}[Y(\mathbf{t}_0)]$ with $\mathbf{t}_1 = (1,0)^T$, $\mathbf{t}_2 = (0,1)^T$, and $\mathbf{t}_0 = (0,0)^T$. Simple calculation implies that the true treatment effects of T_1 and T_2 are 0.196 and -0.138, respectively.

We compare two direct modeling methods (Models 1-2 below) and the proposed tree-based methods (Models 3-4 below) as well as their bagged versions (Models 5-6 below).

1) **Model 1**: a naive approach which estimates the treatment effect of T_j as $\bar{Y}(\mathbf{t}_j) - \bar{Y}(\mathbf{t}_0)$ for $j = 1, 2$.

2) **Model 2**: a linear logistic regression with dependent variable Y and independent variables $(\mathbf{T}, \mathbf{X})$.

3) **Model 3**: the proposed tree-based method in Table 2. In step 3 of Table 2, within each leaf node s, we estimate the treatment effect of T_j as $\bar{Y}(\mathbf{t}_j) - \bar{Y}(\mathbf{t}_0)$ for $j = 1, 2$.

4) **Model 4**: the proposed tree-based method in Table 2. In step 3 of Table 2, within each leaf node s, we fit a linear logistic regression for Y with independent variables $(\mathbf{T}, \mathbf{X})$.

5) **Model 5**: the bagged version of **Model 3**.

6) **Model 6**: the bagged version of **Model 4**.

To measure the accuracy of the treatment effect estimators, we compute the absolute difference of the true treatment effect and the estimated treatment effect of above models. In Figure 6, we report the sum of errors of both treatments for each model over 100 replications. The standard deviation bar of the errors are also shown. As demonstrated in Figure 6, our Model 6 achieves the minimal error and our tree-based models outperform their counterparts without propensity function adjustment. Specifically, for the naive treatment estimation methods, Model 1 has the largest error. After tree-based propensity function adjustment, Model 3 improves the estimation accuracy and Model 5 further improves the accuracy via bagging. For the logistic modeling method, Model 6 outperforms

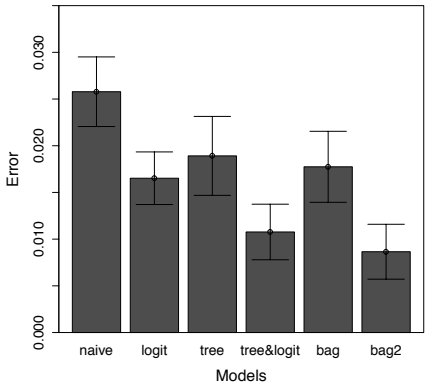

Figure 6: Errors of treatment effect estimation in Section 3.2. In the X-axis, "naive", "logit", "tree", "tree&logit", "bag", and "bag2" refer to Models 1-6, respectively.

the tree-based logistic method Model 4, while the latter greatly improves the direct logistic method Model 2. To sum up, the tree-based methods significantly reduce the errors in treatment effect estimation of the direct methods, and the bagged versions help to further enhance the performance.

4. REAL APPLICATIONS

In this section we apply the proposed framework on two real campaigns. The first campaign is from a major telecommunication company, where we measure the impact of ads with different frequencies. Hence the treatment is a one-dimensional scalar (Section 4.1). The second campaign is from a major auto insurance company and involves multiple ad exposure across TV and online platforms (Section 4.2). We measure the frequency impact of the ads from both platforms, as well as the synthetic impact of the two platforms. In both campaigns, the success action is defined as an online quote.

Note that our framework applies to general ad treatments. The two applications serve as examples to show the efficacy of the methodology, and there is not obvious obstacle to apply the framework to measure treatments including other factors of ad strategies.

4.1 Single Treatment Study

This dataset contains about 0.7 millions users (0.5 million non-exposed users and 0.2 million exposed users) collected during a 3-day campaign[3]. The overall success rate, i.e., the percentage of users making a success action, is about 2.4%. In this example, the ad treatment is the frequency of ad exposures, which is continuous and one-dimensional. The maximal number of ad frequency is 331, while 95% of the frequency are less than 17. This data contains 1010 user features, including the demographic information, online activities, and ad exposure information. Among these features, some are potential confounding variables related to both treatment and outcome. For example, a user with more active online behaviors tends to see more ads and simultaneously has a larger chance to make a success action. In this application, we aim to measure the causal impact of ad exposure frequency, removing the bias due to user characteristics.

To illustrate how our tree model builds a homogeneous treatment density within each node, we compare the variance of the treatment

[3]The reported dataset and results are deliberately incomplete and subject to anonymization, and thus do not necessarily reflect the real portfolio at any particular time.

before and after the tree adjustment. In the original dataset, the variance of the treatment is 110.14. After utilizing our tree model, the weighted variance reduces to 79.08, which has a 28% improvement over the original variance.

We next illustrate the frequency impact on the success rates. According to the algorithm in Section 2, within each leaf node, there could be various ways to estimate the treatment impact via controlling the confounding effect of the covariates on treatments. This choice of modeling method within each node is not the focus of this paper. In this section, for illustration purpose, we utilized a straight-forward estimation method. Specifically, we compare 1) a naive approach which computes the plain success rates corresponding to the various ad frequencies and 2) an adjusted approach via our tree-based causal inference model where the corresponding success rates are computed within each node and then averaged with weights proportional to the node sizes as in (4).

The naive and adjusted estimators of the treatment impact are shown in Figure 7. The general trend of the naive estimation indicates that success rate increases as the number of ad exposure increases. However, interestingly, our method suggests that the success rate increases at the beginning, then decreases, and finally stabilizes. Specifically, the maximal success rate (0.055) is obtained when the users are shown 12 ads, and then success rate stabilizes around 0.045. This agrees with the findings in [24] that ad frequency has nearly linear increasing effect at the beginning and has nearly constant effect for users eligible to see more ads. The one standard deviation bar shows that the maximal success rate at 12 ads is significant larger than others, where the standard deviations get larger as the ad frequency increases due to less samples. To sum up, our result advises that 12 ads are sufficient to maximize the success rate in this campaign and there is little demand for this telecommunication company to deliver more than 12 ads to the eligible users.

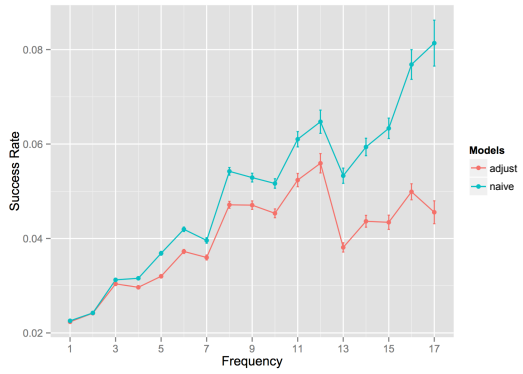

Figure 7: Success rates and their corresponding one standard deviation bars of the naive and adjusted models. X-axis is the frequency of ad exposure.

4.2 Cross-Platform Study

In the cross-platform study from an auto insurance company, the treatment is a two-dimensional vector, containing the numbers of ad exposures from TV and online platform, separately. We measure the impact of TV and online ads together, and hence address the synthetic impact of ad exposure from both platforms.

The dataset contains about 37 million users with 23 million non-exposed users and 14 million exposed users during a 30-day campaign. The original data are extremely imbalanced since the success rates are only 0.204% in the non-exposed group and 0.336% in the exposed group. To deal with this imbalance issue, we employ the subsampling and backscaling proposal in Section 2.2.2, based on which the success rates of non-exposed group and exposed group in the sample increase to 16.9% and 16.7%, respectively. The summary statistics of the original dataset and one sample are shown in Tables 5.

Table 5: Sizes of all online-tv data and one of its sample. "success" or "no success" refers to the size of users who make or do not make a success action, respectively. "M" refers to million.

	Non-exposed		Exposed		Total
	Success	No Success	Success	No Success	
All	46472	23 M	48082	14 M	37 M
Sample	1784	8757	1189	3270	15000

The features include the demographic information, personal interest, and online and TV activities. A sample of the features and their corresponding values are shown in Table 6 for illustration. Specifically, the demographic information consists of the user's gender, age, etc.; the personal interest measures how a user is interested in a specific category, e.g., auto; the online activity captures how often a user visits a particular website and the ad exposures to other companies; and the TV activity collects the TV watching information and the TV ad exposures. In this campaign, there are 2542 features in total.

Table 6: Sample features in the cross-platform study

Feature	Value
Demographic Info and Interest	
Demographic | Gender | Male	0
Demographic | Gender | Female	1
Demographic| Age	27
...	
Interest | Celebrities	0.01
Interest | Auto | New	0.23
Interest | Auto | Used	0.65
...	
Online Network Activities	
Site Visitation | Finance	67.4
Site Visitation | Movies	1.3
Site Visitation | Sports	0.0
...	
Ad Impression | Auto | Company 1	7.24
Ad Impression | Insurance | Company 2	9.43
...	
TV Activities	
TV Program Viewership | Movies	2.5
TV Program Viewership | Sports	53.1
...	
TV Ad Impression	132.7
...	

To demonstrate each step of our algorithm in Table 2, we show a single tree fitted by treating the two-dimensional treatment as the dependent variable and the covariates as the independent variables in Figure 8. In this single tree, nodes 4, 5, 8, 9, 10, and 11 are the leaf nodes.

Within each leaf node in Figure 8, we calculate the success rates of non-exposed group and the exposed group for a given treatment, and hence the treatment effect is estimated as the difference of the two success rates. Then the population level treatment effect is estimated as the weighted average of the results from each node with weight proportional to the node sizes. We take the treatment with 1 tv ad exposure and 2 online ad exposures as an example to illustrate the estimation process. Table 7 shows the results in estimating its treatment effect.

To compare the results from naive estimation without propensity adjustment and the causal inference estimation with the proposed framework, we first show the naive estimator for the ad frequency impact by simply computing the averaged outcomes corresponding

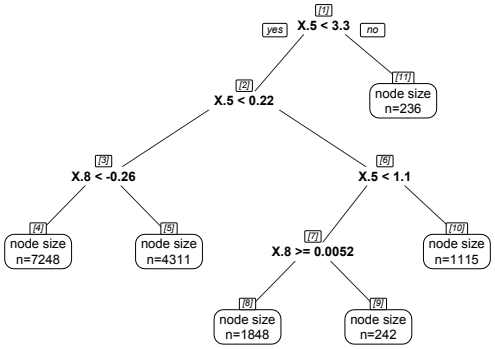

Figure 8: Tree of causal inference model in the cross-platform study.

Table 7: Treatment effect of the case with 1 tv ad exposure and 2 online ad exposure in the cross-platform study. The success rates, treatment effects (TE), and averaged treatment effect (ATE) are given in 10^{-3}.

Node Index	Size	Non-exposed Success Rate	Treatment Success Rate	TE	ATE
[4]	7248	1.14	3.84	2.70	
[5]	4311	0.85	1.45	0.60	
[8]	1848	0.56	0.66	0.10	1.86
[9]	242	0.42	0	−0.42	
[10]	1115	0.92	6.70	5.78	
[11]	236	3.32	0	−3.32	

to various treatments. We group both TV and online ad frequencies as 0, 1, 2, 3 ,4, 5, 6-10, and 11-15 buckets. We employ this grouping scheme since the frequency decreases sharply when it is larger than 5 and most of the frequency is less than 15. As shown in Figure 9, the naive estimator implies that the highest success rate is obtained when the users are shown 11-15 TV ads and 11-15 online ads. In addition, it shows that generally the ad effects get larger as the number of ad exposures increases for both TV and online platforms. However, we will show that this plausible conclusion is biased and the superficial treatment effect is affected by the confounding effect of the user features.

By controlling the confounding effects of the covariates, our tree-based casual inference estimator is able to generate an unbiased estimator. We employed our bagging tree-based algorithm with B=100 according to Table 3. As illustrated in Figure 10, the largest success rate is obtained when the users are shown 5 online ads and 5 TV ads. Furthermore, we find that the online ad effect is marginally larger than the TV ad by comparing the success rate of 0 TV ad exposure (first column in Figure 10) with that of 0 online ad exposure (first row in Figure 10). This suggests that users generally has a larger chance to conduct quotes on the insurance company website when they are shown online ads instead of the TV ads. Finally, similar to the discovery in Section 4.1, both the online and TV ad effects will increase to a maximal value and then decrease as the users are shown more ads. Therefore, it is crucial for the ad providers to make appropriate adjustment based on the number and type of the ads the users have been exposed to.

Furthermore, we employ the bootstrapping approach to estimate the standard deviation of the ATE estimator based on (6). Figure 11 shows the top five highest success rates as well as their corresponding one standard deviation bars. Clearly, the combination of 5 online ads and 5 TV ads is shown to achieve a significantly larger success rate than other combinations.

In order to illustrate the flexibility of our tree-based causal inference algorithm, in contrast to the above nonparametric method applied within each leaf node of the constructed tree, we fit a sparse logistic regression [18] with the success as the binary outcome, and the ad exposures from the two platforms and their interaction term as well as the user features as the independent variables. The tuning

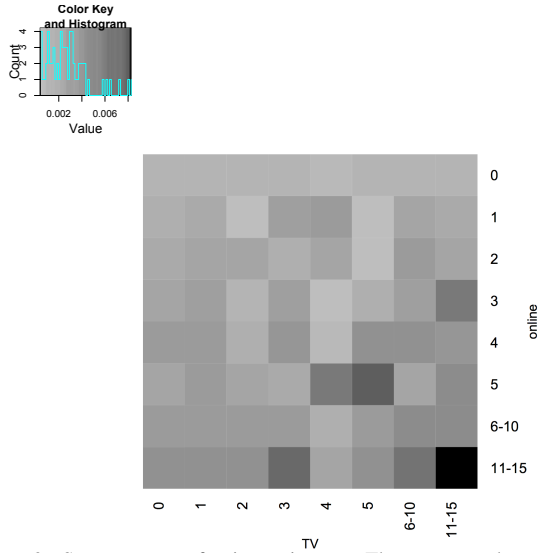

Figure 9: Success rates of naive estimator. The rows are the online ad frequency and the columns are the TV ad frequency

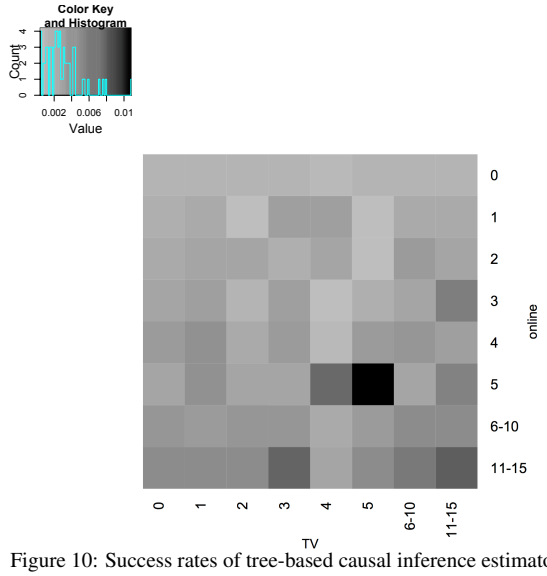

Figure 10: Success rates of tree-based causal inference estimator.

parameter λ in the sparse logistic regression model is selected via cross validation [16]. The causality coefficients of the ad exposure from online, TV and interaction are 0.066, −0.001, and −0.0001 with the standard deviations 0.0393, 0.0183, and 0.0005. This ensures that online ad exposure has relatively positive effect on the success rate while the TV ad exposure has no significant effect. Hence the treatment effect is dominated by the online ad exposures, which is consistent with our findings with the above nonparametric method.

4.3 Model Validation

We have proved that the proposed causal inference framework provides an unbiased estimation of the treatment effect in Section 2, and verified it with several simulations in Section 3. Here we further validate our tree-based causal inference model with the cross-platform analysis in Section 4.2 by showing the covariate balancing effect in real data. Following [21] we first normalized each covariate and then regress it on both the online and TV treatments via the Gaussian linear regression. We record the p-values corresponding to the t-statistics for the coefficients of the treatments in each regression. As shown in Figure 12, the lack of balance is evident in the original data since most of the coefficients of the TV

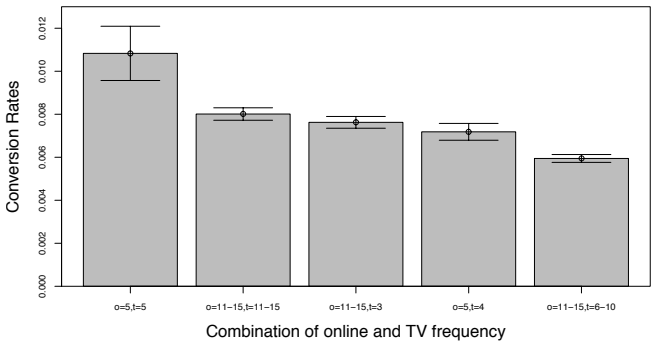

Figure 11: Top 5 success rates and their one standard deviation bars according to our tree-based causal inference estimator. In the X-axis, label "o=5,t=5" refers to the combination of 5 online ad exposures and 5 TV ad exposures. Similar meanings for other labels.

treatment (before: T_2) are significantly different from 0. Therefore, without controlling the balance of the covariates the direct modeling approaches may lead to severe biases in the treatment effect estimations [30]. After adjusting the propensity function via our tree-based method, the percentages of nonzero coefficients of the online and TV treatments reduce from 13.3%, 86.7% to 0 and 6.2%, respectively. Hence our approach successfully balances the covariates and leads to more accurate estimations.

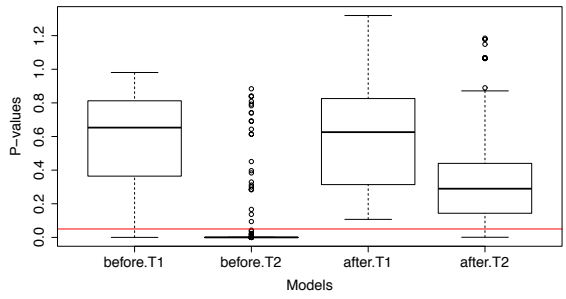

Figure 12: P-values of coefficients by regressing $\mathbf{X}$ on $\mathbf{T}$. The red line is the baseline 0.05. "before.T1" ("before.T2") is the box plot of the p-values for the coefficient of online (TV) treatment before adjustment; "after.T1" ("after.T2") is the corresponding p-values after applying our tree-based model.

5. RELATED WORK

In this section we briefly review three folds of related work: 1) the experiment-based measurement, 2) the propensity-based causal inference that focuses on binary treatments, and 3) general treatment causal inference. The first line of work serves as the gold standard, but meets challenges in practice. We link the second and third lines of literatures with our work and reveal the novelty of our framework.

The experimental approach is the gold standard to estimate treatment effect. When the randomized experiments are available, various regression approaches by regressing the outcome on the treatments can be employed to compute the ad effectiveness from causality perspective (i.e., how the ad treatment causes changes in the outcomes) [3, 27]. However, in reality, the cost and difficulty of successfully performing fully randomized experiments are very high [10, 35], and some experiments may not even be feasible for complex treatments. To avoid the difficulties in experimental approaches, direct comparison and regression-based approaches have been utilized in observational studies. However, direct comparison of the outcomes from different treatment groups in the observational data can lead to severe overestimation of the effects of advertising [1, 27], and the regression-based approaches rely on the pre-specified functional form of the model and are vulnerable to model misspecification [29, 32, 34]. In addition, when the outcome is binary, prior research suggested that at least 8 events should be observed for every covariate in the regression model [9], which prevents its application in practice [2]. Therefore, an unbiased and flexible model is in demand in the observational study.

The causal inference methodologies are able to offer an unbiased estimation of the ad treatment effect from observational data and hence fundamentally overcome the disadvantages of experiments and direct observational studies. Most of the previous causal inference methodologies focus on binary or categorial treatments. The causal inference framework for the observational studies with a binary treatment was original proposed by Rosenbaum and Rubin [30]. They first introduced the concept of propensity score and the inverse propensity score weighting estimation method. This method is also generalized to multi-category treatments [31] and ordinal treatments [23, 28], which have been widespread in various fields, including health care [4], social science [26], politics [20], online behavior study [14]. In online advertising market, causal inference methods have been developed to estimate the causal effects of a binary treatment. For example, [10] applied the causal analysis to industrial advertising data of moderate size, [35] explored the benefits of estimating several other parameters of interest via the targeted maximum likelihood estimation, and [13] used causal inference for a multi-attribution problem. Yet, these methods were typically applied to single treatment case with small to medium user group and moderate success rates. Recently, [38] applied the inverse propensity weighting causal inference method for the large scale online advertising data. Nevertheless, their work mainly focused on the univariate treatment scenarios and was not well suited for general and complex treatment measurements. Besides propensity-based causal inference methodologies, structural equation model [5], inference based on decomposition of the joint distribution of the treatment and response variables [22, 11], and before-and-after studies [39] are also used for causal inference from different points of view.

To expand the scope of causal inference from a binary treatment to a general treatment, a propensity function-based framework is proposed [21]. However, in practice, it is a non-trivial task to estimate this propensity function and there is few investigation on how to choose the number of sub-classes in grouping the users with similar propensity functions. In reality, the treatment effect estimation could be sensitive to the choice of the sub-class size. Another causal inference framework is the causal inference tree (CIT) in the machine learning community [36]. It models both the treatment and outcome simultaneously with respect to the user characteristics by imposing a parametric assumption on the joint density of treatment and outcome.

Our tree-based causal inference framework is substantially different with all the previous work. Our framework is a nonparametric approach that does not require specific assumptions of the joint or separated density functions of ad treatment and outcome as in [36]. Furthermore, it automatically groups subjects (i.e., users in online advertising cases) within the same leaf node of the tree, and hence avoid the arbitrary specification of number of classes as in [21]. It is fully general and flexible that the treatment can be multi-dimensional combining discrete and/or continuous treatment factors, and it is computationally more efficient than regression-based propensity score methods. Most importantly, we prove that our treatment effect estimation is unbiased under weak assumptions.

6. CONCLUSION AND DISCUSSION

This paper proposes a robust tree-based causal inference framework for complex treatment measurement. Our framework utilizes the tree-based structure embedded in a bagging procedure to achieve efficient computation, flexible modeling, unbiased estimation and robust inference. It is able to provide practical guideline to assess advertising strategy. To show the efficacy of our framework, we apply it to two real world applications—an online advertising campaign and a cross-platform campaign. Our framework successfully provides the causal impact of ads with different frequencies and further shows that the ad frequency has a treatment effect cap, which is usually over-estimated by naive estimation. Hence it is important for the ad providers to make appropriate adjustment for ad frequency to reach optimal results.

In this paper, the proposed methodology solves the problem of general treatment measurement. However, in some extreme cases that the treatments and confounding features are both high-dimensional and sparse, and direct application of the methodology is computationally infeasible. For the future work, we would like to design a more sophisticated way for modeling $p(\mathbf{T}|\mathbf{X})$. Another interesting direction is to use the causal inference framework to online media layout optimization problem [37] and effectiveness measurement of user engagement strategies [15].

7. REFERENCES

[1] M. Abraham. The off-line impact of online ads. *Harvard Business Review*, 86(4): 28, 2008.

[2] P. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46:399–424, 2011.

[3] J. Barajas, J. Kwon, R. Akella, A. Flores, M. Holtan, and V. Andrei. Marketing campaign evaluation in targeted display advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, page 5. ACM, 2012.

[4] A. Basu, D. Polsky, and W. G. Manning. Use of propensity scores in non-linear response models: the case for health care expenditures. Technical report, National Bureau of Economic Research, 2008.

[5] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[6] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[7] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman & Hall/CRC, 1984.

[8] P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[9] M. Cepeda, R. Boston, J. Farrar, and B. Strom. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158:280–287, 2003.

[10] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of SIGKDD*, pages 7–16. ACM, 2010.

[11] Z. Chen, K. Zhang, L. Chan, and B. Schölkopf. Causal discovery via reproducing kernel hilbert space embeddings. *Neural computation*, pages 1–34, 2014.

[12] D. R. Cox. *Planning of experiments.* Wiley, 1958.

[13] B. Dalessandro, C. Perlich, O. Stitelman, and F. Provost. Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, page 7. ACM, 2012.

[14] A. Dasgupta, K. Punera, J. M. Rao, X. Wang, J. Rao, and X.-J. Wang. Impact of spam exposure on user engagement. In *USENIX Security*, 2012.

[15] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *Proceedings of WSDM*. ACM, 2013.

[16] J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version 1.4.*, 2009.

[17] G. Harrison and J. A. List. Field experiments. *Journal of Economic Literature*, 27:1013–1059, 2004.

[18] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[19] K. Hullsiek and T. Louis. Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, 2:179–193, 2002.

[20] K. Imai. Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *American Political Science Review*, 99(2):283–300, 2005.

[21] K. Imai and D. A. Van Dyk. Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467), 2004.

[22] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[23] M. M. Joffe and P. R. Rosenbaum. Invited commentary: propensity scores. *American Journal of Epidemiology*, 150(4):327–333, 1999.

[24] G. A. Johnson, R. A. Lewis, and D. H. Reiley. Add more ads? experimentally measuring incremental purchases due to increased frequency of online display advertising. *Working Paper*, 2013.

[25] R. Kohav and R. Longbotham. Unexpected results in online controlled experiments. *SIGKDD Explorations*, 12(2), 2010.

[26] M. Lechner. Earnings and employment effects of continuous gff-the-job training in east germany after unification. *Journal of Business & Economic Statistics*, 17(1):74–90, 1999.

[27] R. A. Lewis, J. M. Rao, and D. H. Reiley. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of WWW*, pages 157–166. ACM, 2011.

[28] B. Lu, E. Zanutto, R. Hornik, and P. R. Rosenbaum. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96(456):1245–1253, 2001.

[29] S. Perkins, W. Tu, M. Underhill, X. Zhou, and M. Murray. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, 9:93–101, 2000.

[30] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[31] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[32] P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[33] G. Rosenkrans. The creativeness & effectiveness of online interactive rich media advertising. *Journal of Interactive Advertising*, 9(2):18–31, 2009.

[34] D. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127:757–763, 1997.

[35] O. Stitelman, B. Dalessandro, C. Perlich, and F. Provost. Estimating the effect of online display advertising on browser conversion. *Data Mining and Audience Intelligence for Advertising*, 8, 2011.

[36] X. Su, J. Kang, J. Fan, R. A. Levine, and X. Yan. Facilitating score and causal inference trees for large observational studies. *The Journal of Machine Learning Research*, 13(1):2955–2994, 2012.

[37] L. Tang, R. Rosales, A. Singh, and D. Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of CIKM*, pages 1587–1594. ACM, 2013.

[38] P. Wang, Y. Liu, M. Meytlis, H.-Y. Tsao, J. Yang, and P. Huang. An efficient framework for online advertising effectiveness measurement and comparison. *Proceedings of WSDM*, 2014.

[39] P. Wang, M. Traskin, and D. S. Small. Robust inferences from a before-and-after study with multiple unaffected control groups. *Journal of Causal Inference*, pages 1–26, 2013.