An Efficient Framework for Online Advertising Effectiveness Measurement and Comparison

Pengyuan Wang
Yahoo Labs
701 1st Ave, Sunnyvale
California 94089
pengyuan@yahoo-inc.com

Yechao Liu
Yahoo Inc.
11 West 19th Street
New York, NY 10011
yechao@yahoo-inc.com

Marsha Meytlis
Yahoo Inc.
11 West 19th Street
New York, NY 10011
mmeytlis@yahoo-inc.com

Han-Yun Tsao
Yahoo Labs
701 1st Ave, Sunnyvale
California 94089
hanyun@yahoo-inc.com

Jian Yang
Yahoo Labs
701 1st Ave, Sunnyvale
California 94089
jianyang@yahoo-inc.com

Pei Huang
Yahoo Inc.
11 West 19th Street
New York, NY 10011
peih@yahoo-inc.com

ABSTRACT

In online advertising market it is crucial to provide advertisers with a reliable measurement of advertising effectiveness to make better marketing campaign planning. The basic idea for ad effectiveness measurement is to compare the performance (e.g., success rate) among users who were and who were not exposed to a certain treatment of ads. When a randomized experiment is not available, a naive comparison can be biased because exposed and unexposed populations typically have different features. One solid methodology for a fair comparison is to apply inverse propensity weighting with doubly robust estimation to the observational data. However the existing methods were not designed for the online advertising campaign, which usually suffers from huge volume of users, high dimensionality, high sparsity and imbalance. We propose an efficient framework to address these challenges in a real campaign circumstance. We utilize gradient boosting stumps for feature selection and gradient boosting trees for model fitting, and propose a subsampling-and-backscaling procedure that enables analysis on extremely sparse conversion data. The choice of features, models and feature selection scheme are validated with irrelevant conversion test. We further propose a parallel computing strategy, combined with the subsampling-and-backscaling procedure to reach computational efficiency. Our framework is applied to an online campaign involving millions of unique users, which shows substantially better model fitting and efficiency. Our framework can be further generalized to comparison of multiple treatments and more general treatment regimes, as sketched in the paper. Our framework is not limited to online advertising, but also applicable to other circumstances (e.g., social science) where a 'fair' comparison is needed with observational data.

WSDM'14, February 24–28, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2556195.2556235.

Categories and Subject Descriptors

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

Keywords

Advertising, Causal Inference, Propensity Score, Gradient Boosting Trees, Feature Selection, Subsampling, Parallel Computing

1. INTRODUCTION

As the online advertising industry has evolved into an age of diverse ad formats and delivery channels, the market increasingly demands a reliable measurement and a sound comparison of ad effectiveness. A metric is needed to show change in brand interest independent of variables that characterize online users that enable us to isolate the effect of the campaign from the effect of other variables. The use cases for such a metric could be: 1) performance assessment of a single campaign, e.g., a website takeover, and 2) synergy effect of multiple campaigns, e.g., the effect of a website takeover on top of a targeted display campaign; 3) performance comparison of different campaigns, e.g., a website takeover versus a targeted display campaign. [1]

The basic idea for ad effectiveness measurement is to compare the performance (e.g. success rate[2]) of users who were and who were not exposed to a certain treatment of ads. In a randomized experiment, i.e., an A/B test, the direct comparison between the success rates of the two groups is unbiased. However in an advertising campaign, we rarely have opportunities to assign random treatments to the users. In such cases, the direct comparison may pro-

[1]Targeted campaigns are those campaigns that only serve ads to a specific group of users, e.g. users who visited the advertiser's website in past 30 days. Website takeover is an ad format that serves mainly for branding purposes. It's usually not targeted in a sense that users visiting the website during the flight will all get to see the takeover. Synergy effect refers to the fact that users exposed to multiple channels of campaigns will have a higher conversion rate than users exposed to a single channel.

[2]A success is an action favored by the campaign, such as click, search or site visitation. Success rate is the percentage of unique users who take a success action. In this paper we use success and conversion interchangeably.

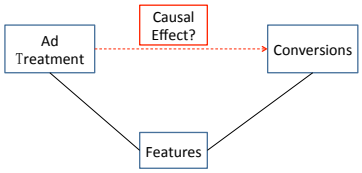


Figure 1: Confounding Effect of Features

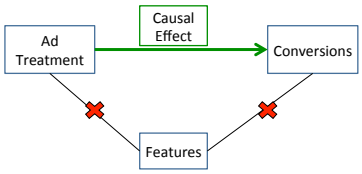


Figure 2: Causal Inference

duce selection biases if each of the groups has a different probability to be exposed to an ad. For example, imagine an auto campaign where all of exposed users are males and all of the non-exposed users are females. If the males generally have a larger conversion rate than females, the effectiveness of the campaign could be overestimated because of the confounding effect of the user features, in this case, gender: It might just be that males are more likely to be exposed and convert. Hence the relationship between the ad treatments and conversions is not causal without eliminating the selection bias. The intuition behind this argument can be seen from Figure 1, where the ad effect on conversion is confounded by the features (which is gender in this example). The causal inference is aiming to establish a causal connection between the ads treatment and the campaign performance, eliminate the impact of features as in Figure 2, and hence estimate the real impact of ads on conversions. We introduce causal inference in Section 2.1 in details.

A straightforward approach attempting to eliminate the impact of user features on the outcome (conversion or non-conversion) is to adjust the outcomes with a set of user-level covariates using regressions, for example, including them as independent variables in a regression as well as indicators of treatments to fit the individual outcomes. However the features of the users receiving each treatments may have complex correlations with the outcomes, and hence the causal effect of the treatment is difficult to estimate by adjusting the outcome with the covariates directly.

To address this problem, Rosenbaum and Rubin [27] established the causal inference in observational studies with propensity score as the major instrument for binary treatments. Defining the subjects of two treatments as 'control' and 'exposed' subjects respectively, the propensity score is defined as the estimated probability for a subject to be exposed, given a set of observed pre-treatment covariates. It is aiming to control for the difference between the groups of subjects receiving different treatments. Linear logistic regression [28, 29] was proposed and then widely used as a standard approach for the propensity score estimation. Other estimation approaches were also proposed, for example semiparametric regression (semiparametric single index and semiparametric binary quantile regression models [20]), and non-parametereic regression (kernal-based matching and local linear regression in [12], boosted regression in [22]). More robust result is reached via doubly robust (DR) estimator [25, 30], which is proven to have a smaller asymptotic variance.

The method rapidly becomes popular in various fields, including economics ([13]), epidemiology [10], longitudinal data [2], health care [4], social science [18], politics [15], online behavior study [8] and advertising [5, 21, 31, 7, 3]. Specifically in the online advertising area, [21] showed that observational data may lead to incorrect estimates, [31] explored the benefits of estimating several other parameters of interest and another method targeted maximum likelihood estimation (TMLE)),[7] used causal inference for a multi-attribution problem, and[3] used it in an experimental circumstance.

The above mentioned approaches were not specifically designed for industrial advertising campaign data with huge volume, high dimensionality and extremely sparse conversion. To the best of our knowledge, we are the first to construct an efficient framework to handle these challenges, which often exist for typical online advertising dataset. In this paper, we propose a novel approach that takes into account the characteristics of real, live campaigns. Our contribution can be summarized as follows:

1. We propose to employ gradient boosting stumps [14] for feature selection, and then gradient boosting trees (GBT) [9, 24] to model ad exposure probability and conversion probability, which selects a small yet sufficient set of features for causal inference and achieves substantially better out-of-sample performance than logistic regressions.

2. We develop a novel subsampling-and-backscaling approach to deal with extremely sparse (few successes) dataset, which significantly improves the out-of-sample performance of the probability models.

3. We devise a new parallel computation scheme to deal with large industrial data sets and combine it with the subsampling-and-backscaling approach to reach computation efficiency and robustness.

Chan et al. [5] also considered the industrial advertising data, but with moderate size. Our framework applies to a much larger dataset with sparser conversion, which is typical for industrial banner ads. The efficient calculation is enabled by the above contributions.

The rest of the paper is organized as follows. In Section 2, we briefly illustrate the intuition behind causal inference, and review the inverse propensity weighting (IPW) with DR estimation. In Section 3, we introduce gradient boosting stumps for feature selections and then GBT for propensity score and success rate modeling. The method is combined with a novel subsampling-and-backscaling approach that deals with extremely sparse conversion data in Section 4. To address the high volume of data, we propose a parallel computing algorithm in Section 5 for computation efficiency. In Section 6 we describe the marketing campaign data of a major auto insurance company used in this study and the procedure for data collection. In Section 7 we apply the proposed method to the dataset. We assess the model performance in Section 7.1, and further validate the causal inference with irrelevant conversion test in Section 7.2. Results in Section 7.2 verifies several aspects of our framework, including the features collection, feature selection scheme and choice of GBT model. In Section 8 we report the results on the whole dataset. In Section 9, we outline how to generalize the framework to multi-treatment cases and more general treatment regimes, where the treatment can be continuous and/or multi-dimensional. We conclude in Section 10.

As a roadmap, the algorithms are summarized in Table 1 to 4. Table 4 is the parallel computation (Section 5) algorithm, within which the subsampling-and-backscaling (Section 4) algorithm (Table 2 and 3) is embedded. The algorithm 2 further contains the IPW (Section 2.2) and DR (Section 2.3) estimation with GBT and feature selection (Section 3) algorithm as in Table 1.

2. A PRIMER ON CAUSAL INFERENCE

2.1 Intuitive Behind Causal Inference

Imagine the outcomes under the alternative treatments that may not have been observed. For example to evaluate the performance of a single campaign, one may consider, if no advertising campaign is running, what would happen? Would there be less conversions? If the campaign reaches every user, would there be more conversions? If the conversion rates in the two cases are the same, this campaign is not effective, i.e., it does not cause conversions. Only if there is a change in conversions under different treatments, should we be able to claim the campaign as effective. The same idea also works for causal inference on synergy effect of multiple campaigns. Imagine that in one situation all the subjects[3] see impressions from both of the two placements (takeover and targeted display) while in another situation the subjects only see impressions from the targeted display placement. The difference or ratio of the conversion rates under the two situations indicates the causal effect of the additional placement. In this section we focus on binary treatment, but the framework is readily generalizable to multiple or more treatment regimes as sketched in Section 9.

Usually the causal effect is measured by comparing a specific treatment (exposure) toward another (control) [4]. For example in the case where the performance of a single campaign needs to be evaluated, 'exposure' treatment is campaign impressions while 'control' treatment is no campaign impression. In the case where we measure the effect of a website takeover on top of a targeted display campaign, 'exposure' treatment is impressions from both placements while 'control' treatment is impressions only from the targeted display placement. In our case, the ad effectiveness is defined as the treatment effect of the exposure. Formally, we define conversions as the desired outcomes of advertising campaign. Each subject has two potential outcomes Y_c and Y_e under control and exposure treatments, respectively. The difference or ratio of the expectations ($E(Y_e) - E(Y_c)$ or $E(Y_e)/E(Y_c)$, respectively) can be used to evaluate the effectiveness of exposure. In reality, either Y_c or Y_e is observed for each subject and hence $E(Y_e)$ and $E(Y_c)$ need to be estimated.

Suppose that the control (e.g. no ad impression) and exposure (e.g. ad impressions) treatments are applied to two groups of subjects with no overlaps, indicted by $z_i = 0$ or 1 for subject $i = 1, 2, ..., N$, and hence the subjects are divided into exposed group ($z_i = 1$) and control group ($z_i = 0$). The success metric (such as conversion), is indicated by $y_i = 0$ or 1. A naive way to estimate $E(Y_e)$ and $E(Y_c)$ is to calculate the average success rates of the two groups, respectively, as in Equations 1 and 2.

$$r_{\mathrm{naive},exposed} = \frac{1}{\sum_i z_i} \sum_i z_i y_i; \tag{1}$$

$$r_{\mathrm{naive},control} = \frac{1}{\sum_i (1 - z_i)} \sum_i (1 - z_i) y_i. \tag{2}$$

Hence naive estimators (difference and ratio[5]) of the exposure effectiveness are defined as in Equations 3 and 4.

$$D_{\mathrm{naive}} = r_{\mathrm{naive},exposed} - r_{\mathrm{naive},control}; \tag{3}$$

$$R_{\mathrm{naive}} = r_{\mathrm{naive},exposed} / r_{\mathrm{naive},control}. \tag{4}$$

[3]In this advertising study, the subjects are online users.

[4]Another method is the difference-in-difference approach [26, 32].

[5]In the rest of the paper we will use amplifier to indicate the ratio of the conversion rates of the exposed and control groups.

The naive estimators are unbiased if the control and exposed groups of users are randomly sampled from the population. However in an advertising campaign, we rarely have opportunities to assign random treatments to the users. In many of the cases, the ad treatments are highly related to user features, such as network activity, website visitation, demographics, etc. In such cases, the estimated $r_{\mathrm{naive},exposed}$ and $r_{\mathrm{naive},control}$ cannot represent the whole population. In other words, it does not ensure the comparison of the two groups is on equal footing. The test in Section 7.2 verifies that for online advertising, the naive estimator can be severely biased.

Another straightforward approach to eliminate the impact of user features on the outcome, as mentioned in Section 1, is to adjust the outcomes with a set of user-level covariates X_i as well as treatment indicators, i.e. fit a model $y_i \sim f(X_i, z_i)$. However the estimated effect of z_i may not necessarily imply causal effect, since the model may not correctly address the relationship between z_i and X_i. We test such an approach in Section 7.2, and the result shows that, it may mistakenly 'detect' ad effectiveness while actually no effect should exist.

The basic idea to address the differences of the control and exposed groups is to consider the treatment z_i as a random variable depending on a set of pre-treatment covariates X_i for each subject i. Causal inference tries to balance the features of different treatment groups, and the methods introduced in this section rely on two assumptions.

Assumption 1: Stable unit treatment value assumption. "The (potential outcome) observation on one unit should be unaffected by the particular assignment of treatments to the other units" [6]. 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")[27]. Given the covariates X, the distribution of treatment assignments is independent of the potential outcomes. This assumption allows us to model the treatment with respect to the covariates, independent of the outcome. It means all the features that are related to both the treatment assignment and the success have been included in the model.

Ignoring the above assumptions may result in misspecification of the causality model, and hence result in biases in the estimation. In the methodology elaboration part we assume these two assumptions are satisfied. In the application, we check the causality model misspecification with an irrelevant response test as in Section 7.2.

In the rest of this section, we briefly review the IPW approach that addresses the selection bias, and DR estimator, which further improves the robustness of the result. More details can be found in [27], [25] and [30].

2.2 IPW

Assuming a feature vector X_i is available for each subject i with features collected before each subject's exposure, propensity score is defined as the probability $\hat{p}_i = P(z_i = 1|X_i), \forall i$. Using X to indicate features and z to indicate treatment without specifying a particular user, ideally, the exposed and control groups will be balanced in terms of propensity score, i.e. $P(z = 1|X, exposed) = P(z = 1|X, control)$. This can be reached in a randomized experiment, in which case the naive estimators in Equations 3 and 4 are unbiased. However in observational data, this assumption is usually violated.

Usually $\hat{p}_i$ is estimated by fitting a model $\hat{P}(X)$ to estimate probability to be exposed with respect to the covariate X. Specifically we model $\hat{p}_i \sim \hat{P}(X_i)$ where $z_i = 1$ with probability $\hat{p}_i$. The basic idea is to use the estimated $\hat{p}_i$ to match the two groups, rather

than to match the multi-dimensional X. In this paper we choose to use GBT with feature selection as in Section 3 to fit the model, with features chosen as in Section 6.

The IPW method proposes that each control subject is weighted by $1/(1-\hat{p}_i)$ and the exposed subjects is weighted by $1/\hat{p}_i$. [6] Hence the weighted success rates of the control and exposed groups are defined as in Equations 5 and 6.

$$r_{\mathrm{ipw},exposed} = \frac{1}{N}\sum_i z_i y_i/\hat{p}_i; \tag{5}$$

$$r_{\mathrm{ipw},control} = \frac{1}{N}\sum_i (1-z_i) y_i/(1-\hat{p}_i). \tag{6}$$

The IPW estimation of ad effectiveness (difference and amplifier) are then defined in Equations 7 and 8.

$$D_{\mathrm{ipw}} = r_{\mathrm{ipw},exposed} - r_{\mathrm{ipw},exposed}; \tag{7}$$

$$R_{\mathrm{ipw}} = r_{\mathrm{ipw},exposed}/r_{\mathrm{ipw},exposed}. \tag{8}$$

The above weighting strategy is to estimate the average exposure effect over the whole population. We are also interested in the average exposure effect on the subpopulation of subjects who actually got exposed. This is called the treatment on treated effect (TTE). For this estimation, the control subjects are weighted by $\hat{p}_i/(1-\hat{p}_i)$ and the exposed subjects are not weighted. Hence the TTE with IPW is calculated as in Equations 9 to 12.

$$r_{\mathrm{ipw},tte,exposed} = \frac{1}{\sum_i z_i}\sum_i z_i y_i; \tag{9}$$

$$r_{\mathrm{ipw},tte,control} = \frac{1}{\sum_{control} \hat{p}_i/(1-\hat{p}_i)}\sum_i (1-z_i) y_i \hat{p}_i/(1-\hat{p}_i). \tag{10}$$

$$D_{\mathrm{ipw},tte} = r_{\mathrm{ipw},tte,exposed} - r_{\mathrm{ipw},tte,control}; \tag{11}$$

$$R_{\mathrm{ipw},tte} = r_{\mathrm{ipw},tte,exposed}/r_{\mathrm{ipw},tte,control}. \tag{12}$$

2.3 DR Estimation

Based on IPW, [25] suggested further improvement of the estimator for robustness, i.e. smaller variance, which requires estimation of the probability to succeed under exposure and control treatments respectively.

Fit success model $\hat{M}_0(X)$ and$\hat{M}_1(X)$ to estimate probability to convert with respect to the covariate X under control and exposed treatments respectively, and we can estimate each subject's success probabilities under exposed and control conditions as $\hat{m}_{1i}$ and $\hat{m}_{0i}$. Specifically $\hat{M}_1$ is fitted with the observed exposed subjects and their features, where $\hat{m}_{i1} \sim \hat{M}_1(X_i)$, and $y_i = 1$ with probability $\hat{m}_{i1}$. The model $\hat{M}_0$ is fitted similarly with control user data. In this paper we again choose to use GBT with feature selection as in Section 3 to fit the model, with features chosen as in Section 6. Define $\delta_{\mathrm{i},exposed}$ and $\delta_{\mathrm{i},control}$ to be the adjusted observations "augmented" with cross terms $-\hat{m}_{1i}(z_i-\hat{p}_i)$ and $\hat{m}_{0i}(z_i-\hat{p}_i)$ and $r_{\mathrm{dr},exposed}$ and $r_{\mathrm{dr},control}$ to be the adjusted estimation of the success rates of exposed and control groups respectively as below.

$$\delta_{\mathrm{i},exposed} = \frac{z_i y_i - \hat{m}_{1i}(z_i-\hat{p}_i)}{\hat{p}_i}; \tag{13}$$

$$\delta_{\mathrm{i},control} = \frac{(1-z_i) y_i + \hat{m}_{0i}(z_i-\hat{p}_i)}{1-\hat{p}_i}, \tag{14}$$

$$r_{\mathrm{dr},exposed} = \frac{1}{N}\sum_i \delta_{\mathrm{i},exposed}; \tag{15}$$

$$r_{\mathrm{dr},control} = \frac{1}{N}\sum_i \delta_{\mathrm{i},control}. \tag{16}$$

Hence the DR estimators of the advertising effectiveness is defined as

$$D_{\mathrm{dr}} = r_{\mathrm{dr},exposed} - r_{\mathrm{dr},control}; \tag{17}$$

$$R_{\mathrm{dr}} = r_{\mathrm{dr},exposed}/r_{\mathrm{dr},control}. \tag{18}$$

The TTE is estimated similarly:

$$r_{exposed,tte} = \frac{1}{\sum_i \hat{p}_i}\sum_i \delta_{\mathrm{i},exposed}\hat{p}_i; \tag{19}$$

$$r_{control,tte} = \frac{1}{\sum_i \hat{p}_i}\sum_i \delta_{\mathrm{i},control}\hat{p}_i. \tag{20}$$

$$D_{\mathrm{dr},tte} = r_{exposed,tte} - r_{control,tte}; \tag{21}$$

$$R_{\mathrm{dr},tte} = r_{exposed,tte}/r_{control,tte}. \tag{22}$$

The idea in DR estimator is to "augment" the observed outcomes with a correction $-\hat{m}_{1i}(z_i-\hat{p}_i)$ (or $\hat{m}_{0i}(z_i-\hat{p}_i)$). The correction terms are products of the estimators from success model and the propensity score model. Intuitively if the error of either estimator is 0, the product of the two errors is 'zeroed out'. In fact [25] proved that, the estimation is unbiased if either the propensity model or the success model is correctly specified. The IPW with DR estimator algorithm is illustrated as in Table 1.

3. GBT WITH FEATURE SELECTION

The methodology described in section 2.2 and 2.3 requires to model the propensity score and success probability under control and exposure treatments with covariate X. In our dataset each internet user has thousands of features. To account for the large number of features, for the propensity model, we perform feature selection by fitting gradient boosting stumps [14] and choose the features with non-zero influence. We then feed the chosen features to GBT (with logistic loss) for propensity score model $\hat{P}$ [22]. Similar feature selection scheme and GBT model are applied to obtain the success model $\hat{M}_1$ and $\hat{M}_0$.

A comparison of the GBT and the standard logistic regression is presented in Section 7.1.2, which shows uniform superiority of GBT over logistic regression. The test in Section 7.2 verifies that the feature selection scheme captures sufficient features for causal inference while greatly reduces the number of features, and the estimated $\hat{p}_i$ by GBT can represent the characteristics of subjects.

The computation is conducted based on R ([23]) with gbm package ([11]) for GBT fitting.

4. SUBSAMPLING AND BACK-SCALING

The banner ad dataset typically has extremely sparse conversions. In order to better capture the pattern of the data, we propose a two-stage strategy for propensity score and success model fitting, consisting of a subsampling stage and a back-scaling stage. In the subsampling stage, we sample the converters with a higher sampling rate than the non-converters. The success rates of the two

[6] The basic intuition is that, a control subject belongs to its group with probability $1-\hat{p}_i$, and hence it is weighted by the inverse of this probability to infer the situation of the population. Similarly for the exposed subjects. See [25] for proofs.

Table 1: Algorithm of IPW with DR estimator

Input:	y_i, X_i, z_i for $i = 1, 2, ...N$.
Output:	$r_{\mathrm{dr},exposed}$, $r_{\mathrm{dr},control}$, D_{dr}, R_{dr}.
Step 1:	Fit the propensity score model. With users in the dataset, fit a model $\hat{P}(X)$ to estimate probability to be exposed with respect to the covariates X.
Step 2:	Estimate propensity score $\hat{p}_i = \hat{P}(X_i)$ for each user i.
Step 3:	Fit the success model for exposed users. With users in exposed group, fit model $\hat{M}_1(X)$ to estimate probability to succeed with respect to the covariate X. Estimate the success probability of each user in the population under exposed treatment $\hat{m}_{1i} = \hat{M}_1(X_i)$.
Step 4:	Fit the success model for control users. With users in control group, fit model $\hat{M}_0(X)$ to estimate probability to succeed with respect to the covariate X. Estimate the success probability of each user in the population under control treatment $\hat{m}_{0i} = \hat{M}_0(X_i)$.
	Whole Population Effect:
Step 7:	Calculate the adjusted success rates as in Equations 13 to 16.
Step 8:	Calculate the difference or amplifier as in Equations 17 or 18.
	TTE:
Step 7':	Calculatethe adjusted success rates as in Equations 19 and 20.
Step 8':	Calculate the TTE difference or amplifier as in Equations 21 or 22.

groups within the subsample are estimated as in Table 1. The subsample success rates are then back-scaled according to the sampling rates to estimate the population-level success rates in the back-scaling stage. The two-stage procedure is laid out as in Table 2 and 3.

The procedure applies to both population level effect and TTE. Numerical results are provided in section 7.1 to show substantial improvement of out-of-sample predictions by utilizing this strategy.

5. PARALLEL COMPUTING

The online ad dataset usually contains large volumes of users, and the computation time is substantially shortened by utilizing parallel computing. We propose to divide the whole dataset into subsamples, and conduct the subsampling-and-backscaling strategy as in Section 4. The computation on each of the subsamples yields an estimation of ad effectiveness, and the point estimation and variation of the population-level ad effectiveness are summarized from the collected subsample estimations. The algorithm is summarized in Table 4.

The parallel computation strategy is implemented on Hadoop (Apache^{TM} Hadoop® project) with an in-house package.

6. DATA

We apply our framework to a marketing campaign of a major auto insurance company, which consists of a specific website takeover and a direct response banner. The study is to measure the effectiveness of the website takeover on top of the direct response banner. The exposed group is defined as the users who were exposed to both the website takeover and the banner ads during a 10-day analysis period, while the control group subjects are only exposed to the banner ads. The control subjects must also have visited at least one of the websites visited by the exposed group,

Table 2: Algorithm of Subsamping

Input:	y_i, X_i, z_i for $i = 1, 2, ...N$.
Output:	$r_{e,sub}$, $r_{c,sub}$, $SP_{e,s}$, $SP_{e,ns}$, $SP_{c,s}$, $SP_{c,ns}$.
Step 1:	From the whole dataset, obtain a subsample such that the number of success users and non-success users are balanced in the subsample.
Step 2:	Count the number of subjects in each of the four groups (exposed & success, exposed & non-success, control & success, control & non-success) in the subsample as $n_{e,s}$, $n_{e,ns}$, $n_{c,s}$ and $n_{c,ns}$ respectively. (Subscripts "e" and "c" are short for "exposed" and "control", and "s" and "ns" are for "success" and "non-success", respectively.)
Step 3:	Count the number of subjects in each of the four groups in the population as $N_{e,s}$, $N_{e,ns}$, $N_{c,s}$ and $N_{c,ns}$ respectively..
Step 4:	Calculate the sampling probability of the four groups: $SP_{e,s} = n_{e,s}/N_{e,s}$, $SP_{e,ns} = n_{e,ns}/N_{e,ns}$, $SP_{c,s} = n_{c,s}/N_{c,s}$, $SP_{c,ns} = n_{c,ns}/N_{c,ns}$.
Step 5:	Calculate the estimated success rates of the exposed and control groups $r_{e,sub}$ and $r_{c,sub}$ with the sampled dataset, as in Table 1.

Table 3: Algorithm of Back-Scaling

Input:	$r_{e,sub}$, $r_{c,sub}$, $SP_{e,s}$, $SP_{e,ns}$, $SP_{c,s}$, $SP_{c,ns}$ from algorithm in Table 2.
Output:	r_{d}, r_{c}, D, R.
Step 1:	Estimate the population-level success rates. $r_{\mathrm{e}} = \frac{r_{e,sub}/SP_{e,s}}{r_{e,sub}/SP_{e,s}+(1-r_{e,sub})/SP_{e,ns}}$, $r_{\mathrm{c}} = \frac{r_{c,sub}/SP_{c,s}}{r_{c,sub}/SP_{c,s}+(1-r_{c,sub})/SP_{c,ns}}$.
Step 2:	Calculate the population-level ad effectiveness $D = r_{\mathrm{e}} - r_{\mathrm{c}}$ and $R = r_{\mathrm{e}}/r_{\mathrm{c}}$ (difference and amplifier respectively).

Table 4: Algorithm of Parallel Computing

Input:	y_i, X_i, z_i for $i = 1, 2, ...N$.
Output:	The estimated ad effectiveness and its variation.
Step 1:	Extract all the converters ($y_i = 1$) within the dataset.
Step 2:	Split the rest of the dataset into K chunks.
Step 3:	For each chunk $k = 1, 2, ..., K$, do steps 3.1 to 3.5:
Step 3.1:	Combine the chunk with all the converted users or a sample of them, such that the number of success users and non-success users are balanced in the sample dataset..
Step 3.3:	Compute $r_{e,sub}$ and $r_{c,sub}$ within this chunk, as in Table 2.
Step 3.4:	Scale $r_{e,sub}$ and $r_{c,sub}$ back to population level to obtain r_{e} and r_{c} and compute the ad effectiveness (difference and amplifier), as in Table 3.
Step 4:	Gather the results from each chunk together and observe the mean estimation and variation (SD).

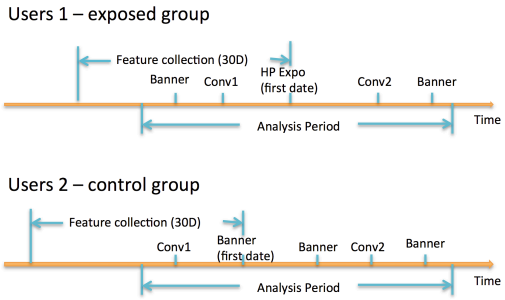


Figure 3: Timeline of Data Collection

and must have at least one feature that are available for the exposed group. These restrictions are to insure that the control and exposed subjects have common website visits and features.

The features are captured in a fixed length of time period (30 days) before the ad treatment. For each user of the exposed group, the features are captured in a 30-day time period before his/her first exposure to the website takeover in the analysis time period. For each user of the control group, the features are captured before his/her first exposure to the banner ads in the analysis time period. The timeline is illustrated in Figure 3.

The collected dataset[7] contains about 2.8 million users (2.0 million control and 0.8 exposed) with 11.7 thousand converters (8.4 thousand control and 3.3 thousand exposed). The dataset is extremely imbalanced since only 0.4% of the users converted. Each user is labeled with an exposure flag and a success flag (defined as a price quote of auto insurance on the advertiser's website). The features of each user, including website visitation, ad exposure, demographic information, market interest, etc., are collected repeatedly on a daily basis. They are then summarized by a weighted summation with a exponential time-decay factor, i.e. for a specific feature, the value observed d days before the advertising treatment is weighted by α^d where $0 < \alpha < 1$. We choose 1000 features which are most populated (i.e. least missing data) for the analysis. A sample of the features and corresponding values are shown in Table 5 for illustration. The features of the two groups show significant differences, and as an example, the network activity[8] of the exposed and control groups are shown in Figure 6(a) and (b).

According to the 'unconfoundedness' assumption as in Section 2.1, one must include all the important features that are related to both the treatment assignment and the success in the model, and one way to validate the choice of features is to conduct an irrelevant conversion test as in Section 7.2. The result shows that the features we suggest to collect in this section capture user's characteristics related to ad impression and success.

In this marketing campaign project, we focus on the TTE amplifier estimator.

[7]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.

[8]Network activity is a general indicator of how active across the web a user is and in this paper is defined as the summation of the feature scores for all the online activities, e.g. website visitation, banner impressions. etc.

Table 5: Sample Features

Feature	Value
Demographic | Gender | Male	0
Demographic | Gender | Female	1
Demographic| Age	27
Demographic| Family Size	1
...	
Interest | Celebrities	0.01
Interest | Auto | New	0.23
Interest | Auto | Used	0.65
...	
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
...	

7. MODEL PERFORMANCE AND VALIDATION

We evaluate and validate the model from two aspects. Section 7.1 focuses on the performance of the propensity score model and success model fitting. However these performance assessment does not provide information for the causality model misspecification which might be because of important feature missing. We validate the causal inference with an irrelevant conversion test in Section 7.2, which verifies various aspects of the framework.

7.1 Model Performance

The proposed procedure is applied to a sample (about 0.5% data) from the marketing campaign dataset. When fitting the propensity score model and the success models with GBT, we always use 50% of the data as training dataset and the rest as test dataset to choose the best number of trees.

7.1.1 Balance of the Two Groups

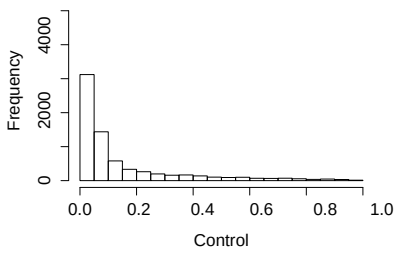
The IPW successfully balances the propensity score and hence the features of the users. The propensity scores of the control and exposed groups in the sample dataset before and after the weighting are summarized in Figure 4. Since the propensity score is the probability to be exposed, the figures suggest that after the weighting, the adjusted groups have similar probabilities to be exposed.

The QQ-plot of the propensity scores before and after the weighing (Figure 5) further shows that the weighing balances the propensity scores of the two groups.

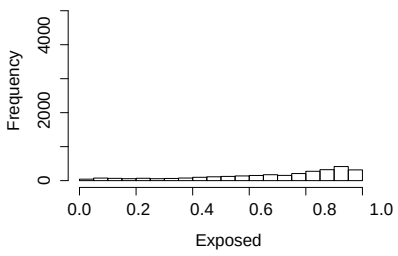
As an example of the user features, the network activities before and after the weighting are summarized in Figure 6 and QQ-plot Figure 7. Both of the figures show significant improvement in the network activity balance. Similar phenomena are observed for other important features, such as auto purchase intention, which implies balance between the adjusted control and exposed groups.

7.1.2 GBT with Feature Selection Performance

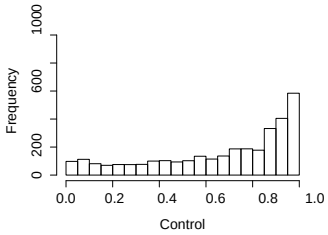
Through the feature selection with gradient boosting stumps , the percentage of chosen features varies for the propensity score and success rate models, ranging from 10% to 30%. The selection greatly reduces the number of features to be fed into GBT and hence the computation time. Yet the performance is similar to that of GBT on the full set of features (less than 1% difference in out-of-sample mean squared error (MSE).)



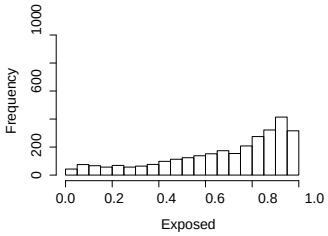
(a) Control, Before



(b) Exposed, Before

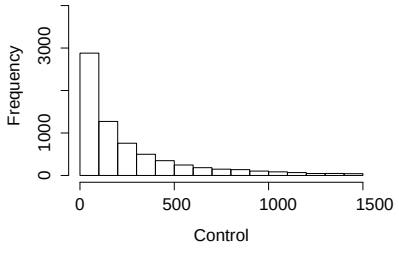


(c) Control, After

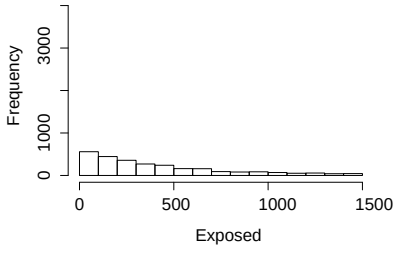


(d) Exposed, After

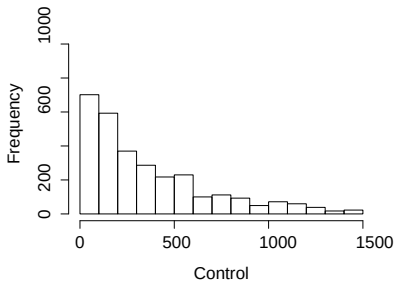
Figure 4: Propensity Score Before (a, b for control and exposed groups respectively) and After (c, d for control and exposed groups respectively) the Weighting



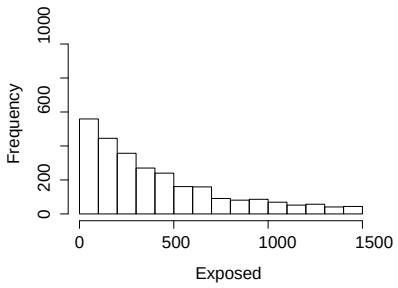
(a) Control, Before



(b) Exposed, Before



(c) Control, After



(d) Exposed, After

Figure 6: Network Activities Before (a, b for control and exposed groups respectively) and After (c, d for control and exposed groups respectively) the Weighting

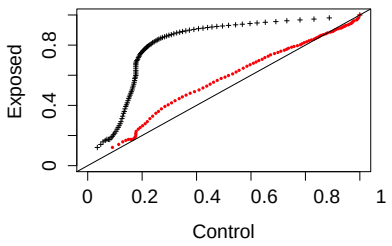


Figure 5: QQ Plot of Propensity Scores Before (cross mark) and After (dot mark) Weighting

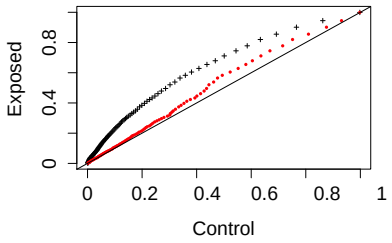


Figure 7: QQ Plot of Network Activity Before (cross mark) and After (dot mark) Weighting

We compare the performance of GBT with feature selection and the standard logistic regression by considering out-of-sample predictions. The MSE of GBT is 47% lower than the logistic regression. The out-of-sample ROC curve in Figure 8 further shows the advantage of the GBT approach.

In causal inference, the model has to not only fit the data well, but also represent and hence balance the features of the control and exposed groups well. One way to validate this point is to conduct an irrelevant conversion test as in Section 7.2. The test result suggests that our feature selection scheme chooses a small fraction of features, but the chosen ones are still sufficient for the causal inference. It also suggests that the GBT model constructs the propensity scores that can represent the high-dimensional user features.

7.1.3 Subsampling Performance

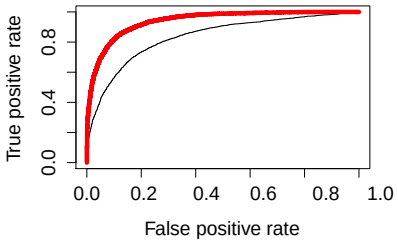


Figure 8: ROC of GBT (thick line) and Logit Regression (thin line)

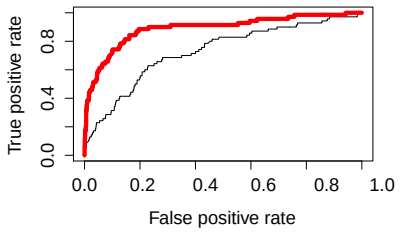


Figure 9: ROC with (thick line) and without (thin line) Subsampling

The two-stage strategy of subsampling and back-scaling in Section 4 improves the out-of-sample model prediction for the success models. As an example, we display the ROC curve of the success model within control group in Figure 9, which shows uniform superiority of the subsampling strategy.

7.2 Model Validation and Irrelevant Conversion Test

In the calculation, we observe less than 2% of the subjects with weights larger than 5, which suggests that the calculation is not sensitive to a small fraction of subjects and hence the robustness of the result.

We also calculate the effective sample sizes $((\sum_j w_j)^2 / \sum_j w_j^2$, where w_j is the weight for subject j in the sample, as suggested in [17]) of the two groups. The effective sample size of the control group is about 47% of the actual control sample size, and 60% for the exposed group. It results in about 2.0 controls per exposed user, and shows sufficient control sample size.

The above methods suggest validation and robustness of the result, but are not sufficient to provide information about causality model misspecification. As suggested in [26], one way to test the misspecification of the causality model is to use irrelevant outcomes, i.e. outcome not related to the treatment. Specifically, one may compose a dataset where the success flag is from a completely different business from the ad treatment, and hence is irrelevant to this particular advertising campaign. If the model successfully matches the control and exposed groups, the estimated ad effectiveness on the irrelevant conversion should be null. To be formal, it is a test with $H_0 : R = 1$ and $H_a : R \neq 1$, where R is the estimated amplifier for irrelevant conversion.

In this irrelevant conversion test, we flag the treatment according to an auto insurance ad impression while the conversion is flagged according to purchases of a consumer electronics brand. The naive amplifier is 0.714 with a bootstrap standard error 0.013, while the adjusted amplifier is 0.974 with a bootstrap standard error 0.075. Another irrelevant conversion test where the conversion is flagged according to purchase of an IT brand reveals similar results.

Before the adjustment, the estimated amplifier is different from 1 statistically significantly, which shows the invalidation of the direct comparison of the success rates. After the adjustment with the causal inference, we accept the null hypothesis that the ad has no effect on an irrelevant conversion, which means our method successfully captures and balances the differences between observed control and exposed groups, and hence conducts a fair comparison.

The result provides validation to multiple aspects in our framework: 1) The effect that the model 'drags' the amplifier toward 1 verifies that GBT captures the differences of the control and ex-

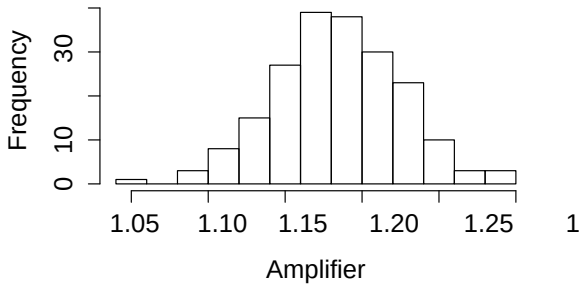


Figure 10: Histogram of TTE Amplifiers

posed groups, and the propensity scores estimated by GBT is a good representation of the user features. 2) The feature selection scheme chooses the features that are sufficient for causal inference while reducing the number of features substantially. 3) The features we suggest in Section 6 cover the major characteristics of online users that are related to ad exposures and conversions.

To show that directly fitting a model with features X_i and exposure indicator z_i may not necessarily work, we fit a logistic regression model and a GBT model for the irrelevant conversion with X_i and z_i. In the logistic regression model, the coefficient of z_i has a p-value nearly 0. In the GBT model, the exposure indicator z_i shows substantial influence. Both of the models mistakenly 'detect' the ad effect on irrelevant conversion, which should not exist.

8. MARKETING CAMPAIGN EFFECTIVENESS RESULTS

The naive amplifier (as in Equation 4) of the campaign summarized from the whole dataset is 0.94. According to the test in Section 7.2, the naive amplifier may be biased by the different characteristics of the control and exposed groups. With our proposed framework , we reach a population level TTE amplifier 1.184, i.e. the webpage takeover lift the conversion rate by 18.4% on top of the direct response banner ad. The collected amplifier estimations from each chunk have standard deviation 0.041, which suggest small variation in the results for different sub-datasets. The histogram (Figure 10) of the subsample amplifiers shows good robustness of the results. It also shows symmetry and uni-mode, which suggests that the average of the amplifiers from each chunk is a good representation of the amplifier of the population.

9. GENERALIZATION

In many of the cases, there are multiple advertising treatments that require fair comparison. For example ads with multiple designs (e.g. text, video and figure), or from multiple serving pipelines (e.g. banner, search and email). In such cases, it is straightforward to generalize this framework to multiple treatment situation, where the treatment indicator $z_i = t$ for user i where $t = 1, 2, ..., T$ indicates the treatment the user received. The key step is to modify the formula to estimated the success rate of each treatment group. Specifically, Equations 1 , 2, 5, 6, 15 and 16 need to be changed.

In the IPW approach, Equations 5 and 6 need to be changed to

$$r_{\mathrm{ipw},t} = \frac{1}{N}\sum_i I_{z_i=t} y_i / \hat{p}_i^t, \tag{23}$$

Table 6: Basic Algorithm of Propensity Function with Subclassification

Input:	y_i, X_i, treatment Z_i that might be continuous and/or multi-dimensional for $i = 1, 2, ...N$.
Output:	Estimated treatment effect.
Step 1:	Instead of propensity score which is probability of one treatment (exposure), one fits propensity functions $\hat{P}(X)$ with the pre-treatment covariates. It is a sufficient statistic for the distribution of treatments, namely $Z_i \perp X_i \vert P(X_i)$ and could be supported on the real line and be multi-dimensional.
Step 2:	Sub-classify the subjects with similar propensity functions together and forms S subclasses.
Step 3:	Within each subclass s, calculate the number of subjects n_s and estimate the treatment effect D_s or R_s.
Step 4:	Gather the results from each subclass and calculate the population treatment effect as a weighted average of D_s or R_s, where the weight is proportional to n_s.

where $\hat{p}_i^t$ is the estimated probability for user i to be exposed to treatment t, and hence $r_{\mathrm{ipw},t}$ is the estimated success rate for users of treatment t. A popular way to estimate $\hat{p}_i^t$ for multiple treatments is Multinomial Logistic Regression (MLR) [19]. Other approaches, such as GBT may also be applied.

For the DR approach, Equations 13 to 16 need to be changed to

$$\delta_{\mathrm{i},t} = \frac{I_{z_i=t} y_i - \hat{m}_i^t (I_{z_i=t} - \hat{p}_i^t)}{\hat{p}_i^t}; \tag{24}$$

$$r_{\mathrm{dr},t} = \frac{1}{N}\sum_i \delta_{\mathrm{i},t}, \tag{25}$$

where $\hat{m}_i^t$ is the estimated conversion probability for user i if he/she receives treatment t, and $r_{\mathrm{dr},t}$ is the estimated success rate under treatment t.

Besides weighting, another set of causal inference approaches is based on classifications, where one subclassifies similar subjects together, estimates the treatment effect within each subclass, and calculates the final treatment effect as a weighted average of the treatment effects within each subclass. By utilizing this approach, this framework can be further generalized to general treatment regimes through the concept propensity functions [16], where the treatments could be continuous and/or multi-dimensional. The basic algorithm is summarized in Table 6.

An example where the treatment could be multi-dimensional and non-binary is that, for a specific banner ad, considering the number of impressions per day and number of days with impression separately. Then Z_i = (# impressions/day, # days with impression), where Z_i is multi-dimensional and the elements of z_i are integer numbers rather than binary indicators. By the algorithm specified in Table 6, one can draw causal effect of two aspects of banner ads: the frequency of ads per day, and the frequency of days with impressions. These two aspects are usually considered together as the total number of impressions in a time period, but they may have different impacts if modeled separately.

10. CONCLUSIONS

In this paper, we propose a novel causal inference framework employing IPW and DR estimator. We advocate to fit the models

(propensity score model and success models) with GBT with a feature selection by gradient boosting stumps, and embed the approach in a subsampling-and-backscaling strategy to deal with extremely sparse conversions, thus better capture the pattern in the data. To deal with large volume of data, we devise a new parallel computing plan that provides the estimation of the mean ad effectiveness and the variation. We apply our framework to a real-world advertising dataset, which shows efficiency and robustness of the method.

Through the generalization discussions in Section 9, our framework is readily generalizable to multi-treatment cases and general treatment regimes, which can be used to compare multiple advertising treatments, multiple aspects of an ad simultaneously, and other general ad treatments. Also, this paper focuses on measuring the effectiveness of online ads, but the framework is readily applicable to other cases (e.g. social science) when a fair comparison needs to be conducted from observational data.

11. REFERENCES

[1] Apache™ hadoop® project. *http://hadoop.apache.org*.

[2] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[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] 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 the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–16. ACM, 2010.

[6] D. R. Cox. Planning of experiments. 1958.

[7] 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.

[8] 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.

[9] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[10] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

[11] Greg Ridgeway with contributions from others. *gbm: Generalized Boosted Regression Models*, 2013. R package version 2.0-8.

[12] S. Guo and M. W. Fraser. Propensity score analysis. *Statistical methods and applications*, 2010.

[13] J. J. Heckman, H. Ichimura, and P. Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

[14] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90, 1993.

[15] 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.

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

[17] L. Kish. Survey sampling. new york: J. *Wiley & Sons*, 643:16, 1965.

[18] 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.

[19] J. Ledolter. Multinomial logistic regression. *Data Mining and Business Analytics with R*, pages 132–149.

[20] S. F. Lehrer and G. Kordas. Matching using semiparametric propensity scores. *Empirical Economics*, 44(1):13–45, 2013.

[21] 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 the 20th international conference on World wide web*, pages 157–166. ACM, 2011.

[22] D. F. McCaffrey, G. Ridgeway, A. R. Morral, et al. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403–425, 2004.

[23] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[24] G. Ridgeway. Generalized boosted models: A guide to the gbm package. *Update*, 1:1.

[25] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

[26] P. R. Rosenbaum. *Observational studies*. Springer, 2002.

[27] 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.

[28] 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.

[29] P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[30] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[31] 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 (ADKDD 2011)*, 8, 2011.

[32] 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.