Estimating The Effect Of Online Display Advertising On Browser Conversion

Ori Stitelman
media6degrees
37 East 18th Street
New York, NY, 10003
ori@media6degrees.com

Brian Dalessandro
media6degrees
37 East 18th Street
New York, NY, 10003
briand@media6degrees.com

Claudia Perlich
media6degrees
37 East 18th Street
New York, NY, 10003
claudia@media6degrees.com

Foster Provost
NYU Stern School of Business
New York, NY, 10012
fprovost@stern.nyu.edu

ABSTRACT

This paper examines ways to estimate the causal effect of display advertising on browser post-view conversion (i.e. visiting the site after viewing the ad rather than clicking on the ad to get to the site). The effectiveness of online display ads beyond simple click-through evaluation is not well established in the literature. Are the high conversion rates seen for subsets of browsers the result of choosing to display ads to a group that has a naturally higher tendency to convert or does the advertisement itself cause an additional lift? How does showing an ad to different segments of the population affect their tendencies to take a specific action, or convert? We present an approach for assessing the effect of display advertising on customer conversion that does not require the cumbersome and expensive setup of a controlled experiment, but rather uses the observed events in a regular campaign setting. Our general approach can be applied to many additional types of causal questions in display advertising. In this paper we show in-depth the results for one particular campaign (a major fast food chain) of interest and measure the effect of advertising to particular sub-populations. We show that advertising to individuals that were identified (using machine learning methods) as good prospective new customers resulted in an additional 280 browsers visiting the site per 100,000 advertisements shown. This result was shown to be extremely significant. Whereas, displaying ads to the general population, not including those that visited the site in the past, resulted in an additional 200 more browsers visiting the site per 100,000 advertisements shown (not significant at the ten percent level). We also show that advertising to past converters resulted in a borderline significant increase of an additional 400 browsers visiting the site for every 100,000 online display ads shown.

ADKDD '11, August 21 2011, San Diego, California USA

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; I.2.6 [**Artificial Intelligence**]: Learning—*induction*; I.5.1 [**Pattern Recognition**]: Models—*statistics*; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Algorithms, Design, Experimentation

Keywords

online display advertising, causal effects, predictive modeling, social networks, user-generated content, privacy

1. INTRODUCTION

Controlled experiments, often called randomized tests or A/B tests, are commonly used online to assess the effects of any kind of changes to the browsing experience (formally interventions) on browser behavior (see e.g.[6],[8],[7]). Some of those efforts have been devoted to extending these experiments to evaluating the causal effect of online display advertising on browser conversion ([8],[7]). However, the cost and difficulty of implementing A/B testing successfully in the display ad environment are very high, as will be explained below. This makes analytical approaches that can estimate the effect of ads while running the campaign regularly (observational setting; without creating any special testing controls) appealing. Despite the appeal of estimating the effects based on observational data there are many practical considerations that make this a difficult task. A number of analytical methods have been developed in a wide range of fields to estimate the causal effects of a binary treatment on a binary outcome of interest. Much of the relevant causal literature has been developed in the field of epidemiology and biostatistics (see e.g. [15] and [21]). Though the reasons for not using randomized tests in the medical literature are often different than in the advertising community, the lessons learned from the use of causal methods there are directly applicable to our current setting. For our purposes the treatment of interest is an advertisement and the outcome of interest is browser conversion such as visiting a

webpage of interest, providing an email address, or making an online purchase. Ultimately, we are concerned with answering the question: "What is the causal effect of online display advertising?"

For the remainder of this article we will refer to A/B tests as randomized tests. We will also interchangeably refer to treatment as the act of showing an advertisement and refer to a conversion or outcome of interest as taking a specific action on the brand's website. Furthermore, when we refer to display advertising we mean online display advertising. We will primarily focus in our experiments on whether or not a browser visits the website (site visit) of the advertising brand. We will also use the common convention that capital letters represent random variables and lower case letters represent realizations of a level of those variables.

There are many difficulties of implementing A/B tests online in general[5]. Kohavi, 2010, discusses the difficulties of implementing a randomized test in the online setting and provides examples of how the implementation of the randomization test can result in unforeseen artifacts that make estimating the intended effects difficult or impossible. In fact, they suggest using a form of testing called A/A/B, where there are three possible treatment scenarios–one for the current treatment, another for the current treatment using the A/B test implementation and a third for the new treatment using the A/B test implementation. The A/A/B test allows one to test if the observed effect is due to the new treatment or due to the implementation of the test. This type of set-up to assess the implementation of the test, though sensible, begins to add new costs to the implementation that must be considered. A paper presented by Google at the last KDD briefly addresses another major drawback of implementing A/B tests for assessing the effect of display advertising on browser conversion[1]. The major concern expressed was that advertisers would not want to pay to present advertisements, such as public service announcements (PSA), that did not promote their product. The fact that A/B tests are prone to unforeseen error and that marketers don't want to pay to present PSAs, coupled with the fact that there are additional overhead costs associated with A/B testing, suggests that alternative methods for estimating causal effects without intervening on the observed system would be preferable.

A common misconception is that randomized tests are the only study design in which causal effects may be estimated (see e.g.[6], [9]). In fact, there is a long history of literature devoted to estimating causal effects in observational, or non-randomized, settings (see e.g. [17],[21]). Rubin [17] established a *counterfactual* framework that defines the effect of all levels of possible treatment for each observed subject, and allows for the consideration and estimation of what would have happened when individuals received a specific level of treatment, possibly contrary to what is actually observed. The outcomes for the unobserved levels of treatment are referred to as potential outcomes in the literature. This framework also allows one to define summary measures that quantify the effect of the treatment on average for the population or sub-population of interest. Two commonly reported summary measures of interest include the difference in the outcome probabilities and the ratio of the outcome probabilities when treated versus not treated.[1] We will refer to the difference in conversion probabilities as the additive impact of the advertisement, and the ratio as the relative impact of the advertisement.

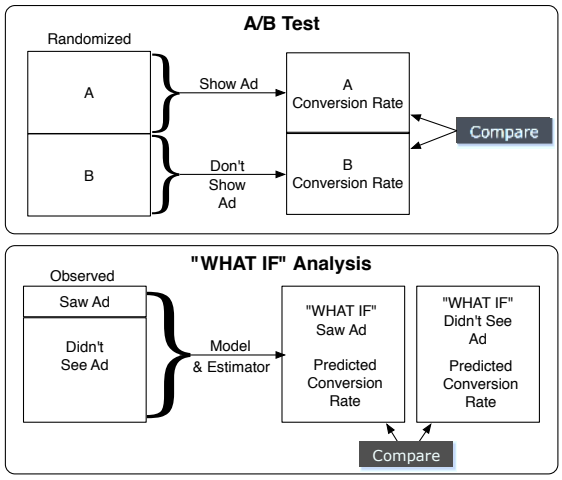

Figure 1: Diagram Of A/B Test And "What If" Analysis.

Chan et al., presented at last year's KDD, proposed the use of several related methods that are able to estimate the effect of advertising in observational data[1]. In particular, their paper focused on estimating the causal effect of advertising among those shown the advertisement, and highlighted the benefits of applying their methods in pipeline.[2] Several variants of the methods they surveyed as well as others will be discussed below. In addition to implementing variants of the methods they proposed, we will explore the benefits of estimating several other parameters of interest, explore another method for estimating causal effects (targeted maximum likelihood estimation (TMLE))[22]), and discuss some other advantages of estimating causal effects beyond its implementation in pipeline. Furthermore, we display the advantages of estimating causal effects within different segments of the population.

Consider what is the purpose of implementing an A/B, or randomized, test. The entire point of running the test is to easily identify the effect of the treatment, or in our case the display advertisement, on the outcome of interest. The first step in a randomized study is to randomly assign each subject, or browser, to one of two groups, A or B. The top box in figure 1 shows this approach. The motivation for random sampling is to ensure that both groups are similar with respect to the distribution of all relevant variables that can potentially affect the probability of taking the desired action (e.g., gender, browsing activity, past purchase activity, etc.). If the two groups were not similar, the difference in the observed effect might be due to those variables and not to the treatment. Variables that can affect both the probability of treatment and the probability of conversion would

[1]These summary measures are typically referred to as additive or attributable risk and relative risk, respectively, in the epidemiology and biostatistics literature because they were commonly used to quantify the increase in the risk of a disease or death when exposed to a possibly harmful substance.

[2]They define a pipeline as "an automated pipeline that retrieves data, computes estimates, and decides whether to release results, suppress results, or send them to an expert data analyst for review."

make it difficult to estimate the effect of the treatment and are known as *confounders*. Group A is then shown the advertisement and group B is not shown the advertisement. Next the conversion rate is calculated for both groups and is compared to assess the effect of showing the advertisement. If the randomization is successful it allows one to directly compare the outcomes of the two groups and the practice of relying on it to make the conclusion about the effect is known as the randomization assumption.

The bottom box in figure 1 outlines an approach for a counterfactual analysis that allows the estimation of the effect directly in the observed data even in the presence of confounders. In observational data one cannot directly compare the group that is shown the advertisement to the group that is not shown the advertisement because of the confounding of targeting. The group that was shown the ad was selected specifically *because* they are assumed to have a higher conversion rate. Therefore adjustment must be made for all variables that affect the conversion rate. This is done by analytically conducting the counterfactual or "what if" analysis. In general, the approach is to use a model/estimator to estimate the conversion rate had the entire population of interest been shown the advertisement. Subsequently, the conversion rate is estimated had the entire population not been shown the advertisement and the two conversion rates are then compared. A number of observational methods exist that provide a way to adjust for the fact that the treated and untreated groups are not the same with respect to the confounders in the observed data. Several of these methods are discussed and implemented in the subsequent sections.

In this paper we will describe a practical approach for estimating the causal effect of advertising. We will follow a unified approach that may be extended to other interesting causal questions of interest in the display advertising environment. This approach relies on (1) posing the question of interest (2) making assumptions about the observed data (3) clearly identifying a parameter that answers the question given the assumptions and (4) estimating the parameter. This approach loosely follows the roadmap for constructing a TMLE presented in [19]. However, rather than just presenting a TMLE we will discuss several ways that the parameter of interest may be estimated and expound on the advantages and shortcomings of each method. Finally, we will present an analysis of the effect of advertising for a major fast food chain and use the presented methods to assess the effect within different sub-populations of interest. The analysis presented is just an example that illustrates the method, many alternative questions of interest may be answered using the methods presented here (e.g. one could investigate the effect of different creatives). In summary our results show that advertising to past converters, "re-targeting", results in an additional 400 browsers visiting the site for every 100,000 online display ads shown (1.1 times more site visitors when shown the ad). Whereas, advertising to individuals that were identified (using machine learning methods discussed in [14]) as good prospective new customers results in a 1.5 times increase in conversions, which equates to an additional 280 browsers visiting the site per 100,000 advertisements shown (p-value $< 10^{-16}$). Finally, displaying ads to the general population, not including those that visited the site in the past, resulted in 2.4 times more site visits, or an additional 200 more browsers visiting the site per 100,000 advertisements shown (p-value = .13).

2. POSING THE QUESTION OF INTEREST

Primarily we will focus on answering the question: "What is the effect of display advertising on customer conversion?" In particular, we are not just interested in the immediate response of clicking. Increasingly, clicks are perceived as notoriously random and unreliable measure of effectiveness (see e.g. [2]). The much more relevant question is whether seeing an ad (without necessarily clicking on it) affects the probability of conversion (e.g., visiting the brand's website) within a reasonable timeframe (this is also known as a "post-view" conversion). Answering this question for the universe of all browsers may not be of great interest because that population includes a large portion of people who are highly unlikely to convert whether they are shown a particular advertisement or not. Moreover, the estimation problem is potentially much more difficult when looking at the entire population, rather than segments of the population, and may require larger data sets to answer the question of interest because of the low overall conversion rates. Fortunately, the questions that have the most financial relevance from the perspective of an advertiser relate to the effect of advertising on appropriate populations that have higher baseline conversion rates (even absent advertising) than the overall population. We will focus on estimating how advertising affects conversion, where conversion is measured in terms of future site visits. We will focus on answering this question for particular sub-populations of browsers. In particular, we will focus on answering the following three questions:

1. What is the effect of display advertising on conversions for individuals that have visited the site in the past?
2. What is the effect of display advertising on conversions for potential new customers (browsers with no past site-visits) that were targeted based on their natural expected tendency to convert at a higher rate than the general population?
3. What is the effect of display advertising on conversions for potential new customers in the general population?

For illustrative purposes, we have chosen a campaign with high conversion rates relative to other campaigns we have analyzed. By doing this we are able to estimate parameters that answer all three of these questions of interest reliably. In cases where the conversion rates are low it may be difficult to answer the third question of interest regardless of the estimation method employed.

3. THE OBSERVED DATA STRUCTURE AND LIKELIHOOD

In this section we will define the data structure we use, and introduce some notation as well as some overall causal assumptions about the observed system. We will then use those causal assumptions to define parameters of interest in the following section.

Our data structure is a common one in the causal literature. For each subject i (i.e., browser) we observe $O_i = (W_i, A_i, Y_i)$, where O_i is an observation from the true, and unknown, data generating distribution, P_0.[3] A is the binary random variable of intervention/treatment (i.e., showing an ad), where A equals one if an individual is treated and zero otherwise. W is a vector of baseline covariates that records

[3]The subscript 0 denotes truth, whereas a subscript n — will denote an estimate of the distribution.

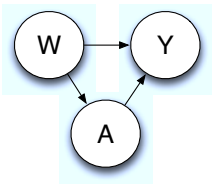

Figure 2: Causal Graph

information specific to a browser prior to intervention. W should in our case include the browser's past web-activity, past actions taken, past advertisements viewed, as well as any other relevant information that affects the outcome of interest and the treatment level the browser is exposed to. The random variable Y is the binary outcome of interest (i.e., is equal to one when a browser takes the action of interest and zero otherwise).

Now we can define a time ordering of the observed variables that allows us to factorize the data into an observed data likelihood. The time ordering of the observed random variables in this simple data structure corresponds to a causal graph that implies a particular factorization of the likelihood. The causal graph is shown in figure 2. More complicated observed data structures require careful defining of the time ordering of the variable and both time cues and subject matter knowledge should be used in constructing a causal graph. This causal graph lays out the assumptions one is making that allow for the construction of a parameter of interest that directly answers the scientific or business question of interest. Specifically, this causal structure states that the baseline covariates are not caused by targeting or conversion, and that targeting is not caused by conversion. Furthermore, it states that there are no other unobserved variables that cause both A and Y, or unobserved confounders. Unobserved variables that cause any of the nodes in the graph individually are okay. The factorization of the likelihood that corresponds to our causal assumptions is:

$$P_0(O) = \overbrace{P(W)}^{Q_{W_0}}\overbrace{P(A \mid W)}^{g_{A_0}(A,W)}\overbrace{P(Y \mid A, W)}^{Q_{Y_0}(A,W)}. \quad (1)$$

Thus, the likelihood factorizes with a part associated with the non-intervention nodes $Q_0 = (Q_{W_0}, Q_{Y_0})$ and a factor associated with the intervention node, g_{A_0}. We refer to the node for A as an intervention node because we are interested in what happens to the outcome when we intervene on A, showing an ad, for each browser.

4. DEFINING A PARAMETER OF INTEREST

The parameter of interest is specifically chosen to answer our primary question, "What is the effect of display advertising?" It is common in the machine learning community to refer to "tuning parameters" such as "k" in the k-nearest neighbor algorithm and the number of leafs in a regression tree in general as parameters. This is not how we use the term parameter here. In fact, when we define a parameter of interest, we define a quantity of interest that directly answers our question and thus we would like to obtain an estimate for it. This is a common use of the terms "parameter" and "parameter estimation" in the statistics community.

Now that the likelihood of the observed data is factorized according to a causal graph we can consider the interventions (showing an ad) on the observed system and how one may "observe" the outcome in the case of intervention $A = a$. The counterfactual distribution of the data structure under intervention is known by the causal inference community as the G-computation formula[16]. The G-computation formula is very similar to the do-calculus proposed by Pearl for causal analysis[11]. The A node which we are intervening upon in the causal graph is set to the intervention level, a, in the likelihood and the conditional distribution of A given W is removed from the likelihood since it is no longer a random variable (it is now deterministically set by the intervention). The following is the resulting G-computation formula, or distribution of the data under intervention $A = a$:

$$P_{0,A=a}(O) = P(W)P(Y \mid A = a, W). \quad (2)$$

The G-computation formula now may be used to guide the choice of causal parameter that will be useful for answering a particular business question of interest.

We are interested in the size of the effect of display advertising. If we knew the true distributions Q_{W_0} and Q_{Y_0} how would we answer this question? One straightforward approach would be to evaluate the conditional distribution of $Y \mid A, W$ for $A = 1$ and then again for $A = 0$ at all possible levels of baseline variables W. Then take the mean weighted by $P(W)$ for each group. This would result in the two quantities $E_W[Y_{A=1}]$ and $E_W[Y_{A=0}]$. Where $E_W[Y_{A=a}]$ is the mean of Y assuming everyone in a population is treated at level a. We can now combine $E_W[Y_{A=1}]$ and $E_W[Y_{A=0}]$ in useful ways to assess the effect of different levels of the treatment variable A. Two commonly used parameters of this type are the following:

$$\text{Additive Impact} = \Psi^{AI}(P_0) = E_W[Y_{A=1}] - E_W[Y_{A=0}] \quad (3)$$

$$\text{Relative Impact} = \Psi^{RI}(P_0) = E_W[Y_{A=1}]/E_W[Y_{A=0}] \quad (4)$$

The additive impact quantifies the additive effect of showing the advertisement to everyone in the population versus not showing the ad to anybody in the population. This value could be interpreted as the average number of additional conversions per 100 people had everyone been shown the ad versus had everyone not been shown the ad. Thus, if the additive impact were 3 percent the following statement would be appropriate: "Showing the ad versus not showing the ad results in 3 additional conversions per 100 browsers." The relative impact quantifies the multiplicative effect of the advertisement. This quantity is the ratio of the probability of the outcome had everyone been shown the ad divided by the probability of the outcome had nobody been shown the ad. Thus, a relative impact of 3 would correspond with the following statement: "Showing the ad makes browsers on average three times more likely to convert." The choice of parameter to estimate should be driven by the business question one is trying to answer and in many situations it may be useful to estimate both parameters. The additive impact directly addresses the return on investment while the relative impact is highly influenced by the level of the untreated conversion rate. If the untreated conversion rate is low a small additive impact will manifest itself as a large relative impact even-though the advertisment may have little affect on the number of additional customers converting. One particular

advantage of using these parameters to answer our question, rather than say log odds in a logistic regression, is that they are numbers that may be interpreted by statisticians and non-statisticians alike. Thus, the estimates may be handed off to individuals with business knowledge to make actionable decisions based on them.

Note that the parameters described above quantify the effect of the advertisement over the entire population of interest. Thus, any conclusions, or inferences, made from the estimates generalize to the entire population being examined. For example, as we do below, we estimate the additive impact of advertising among past site visitors for which a medium size ad serving company received a bid request[4]. Any inferences made based on an estimate are generalizable to this group. This is a different parameter than the ones estimated by Google in their paper where they estimated the effect only among the treated. Inferences made by the estimators and methodology proposed there are generalizable only to a population that is treated[1]. Thus, the parameters we propose to estimate here are generally valid for estimating the effect of advertising on a group of people to which one could potentially advertise and who may or may not have been advertised to in the past. This type of parameter is directly relevant to answer business questions that regard assessing the effect of potential interventions on the entire group of browsers that may be intervened upon.

5. ESTIMATION METHODS

The process described in the preceding sections has been primarily concerned with defining a parameter of interest that directly answers the question: "What and how big is the effect of display advertising?" The presented approach loosely follows the first few steps of the unified approach presented in [19] for estimating causal effects. In this section we explore alternative estimators that might be considered for estimating the parameters defined in the previous sections.

Each of the estimators we will present is a function of an estimate of the Q factor of the likelihood, the g factor of the likelihood, or both. Up until now we have not presented a model, or collection of possible data-generating distributions, for Q_{W_0}, Q_{Y_0}, and g_{A_0}. For Q_{W_0} we will use the empirical distribution, as this is the efficient non-parametric maximum likelihood estimator for this distribution (It will be explained later how this is done in practice). For the two conditional distributions, Q_{Y_0}, and g_{A_0}, ideally a non-parametric model which imposes no rigid assumptions on the functional form of the relationship would be used. However, practical concerns given the size of the data and dimensionality of the problem make this a computationally difficult task. So for the time being we employed a logistic regression model that performed cross-validated variable selection for both main terms and polynomials of order 2 to estimate the conditional distributions, Q_{Y_n} and g_{A_n}. Note however, that we are not interested in the estimated parameters of this logistic model, but use it only to generally estimate a functional dependence. One advantage of estimating the functional dependence of the underlying distributions rather than one specific beta of a linear model is that as computation power increases, better methods with less bias can come to bear that allow for searching over larger spaces. Those methods can be incorporated directly into our process for estimating causal effects. Thus, the better we get at estimating conditional distributions, the better we get at estimating causal effects. This is not the case for approaches that pre-specify the causal effect as a specific parameter of a linear regression or logistic regression model.

We will now present the different estimators $\psi_{n,*}$ of the parameters of interest Ψ shown above. We refer the reader to external sources for more in-depth understanding of each estimator, as that is outside the scope of this paper.[5] For each of them, we will define the estimator as well as describe some characteristics and develop intuition about when the estimator behaves well and when it may break down in practice. The estimators under consideration are unadjusted (UNADJ), a maximum-likelihood based evaluation of the g-computation parameter (MLE), inverse treatment weighted (IPTW), augmented-IPTW (AIPTW), and targeted maximum likelihood (TMLE). For a more in-depth treatment of these estimators see [4].

For each estimator, or method, we will estimate the average conversion rate for everyone as though they were shown the advertisement, $\psi^{a=1}_{n,METHOD}$, and for everyone as though they were all not shown the advertisement, $\psi^{a=0}_{n,METHOD}$. These estimates then may be combined to estimate the additive impact (AI) and relative impact (RI) in the following way:

$$\psi^{AI}_{n,METHOD} = \psi^{a=1}_{n,METHOD} - \psi^{a=0}_{n,METHOD} \quad (5)$$

$$\psi^{RI}_{n,METHOD} = \psi^{a=1}_{n,METHOD} / \psi^{a=0}_{n,METHOD} \quad (6)$$

We will now begin to describe the different methods of estimation. The unadjusted estimate (UNADJ) is a biased estimate of the causal effect because it does not account for the fact that individuals who are more likely to get advertised to are also more likely to convert. In other words, the estimator does not account for confounding. The UNADJ estimator for relative risk is the conversion rate of the treated divided by the conversion rate of the untreated. Similarly, the UNADJ estimator for the additive impact is the conversion rate of the treated minus the conversion rate of the untreated:

$$\psi^{a}_{n,UNADJ} = \frac{\sum_{i=1}^{n} I(A_i = a) Y_i}{\sum_{i=1}^{n} I(A_i = a)}. \quad (7)$$

An MLE based estimator is a substitution estimator that relies on a consistent estimate of the conditional distribution Q_{Y_0}. By a substitution estimator, we are specifically referring to the fact that the estimator follows the proper bounds of the model (i.e., estimates probabilities between 0 and 1) by evaluating at a particular P_n. It takes the following form:

$$\psi^{a}_{n,MLE} = \frac{1}{n} \sum_{i=1}^{n} Q_{Y_n}(a, W_i). \quad (8)$$

[4]A large percentage of the display advertising examined is flowing through ad-exchanges with real-time bidding. When a browser is visiting some site on the web, the site may send a request to such an exchange including relevant information on the browser. At that point, an auction is run in real-time and the highest bidder gets to show the browser an ad. A bid request is an event where the ad-exchange solicits potential bidders.

[5]The subscript n indicates this to be an estimate (rather than truth) based on n observations.

Some drawbacks of this method are that there is no available theory for the construction of the variance estimates and therefore confidence intervals. Furthermore, it is not robust to mis-specification of the outcome regression model.

The IPTW estimator is an estimating equation-based estimator that adjusts for confounding through g. Thus, it is a consistent estimate of the causal effect when g_n is a consistent estimate of g_0. Unlike a substitution estimator, the estimating equation estimators do not obey the bounds of the proper model (i.e. the may return estimates of probabilities not between 0 and 1). The estimator takes the following form:

$$\psi^a_{n,IPTW} = \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)Y_i}{g_{A_n}(A_i,W_i)}. \quad (9)$$

The A-IPTW estimator is an estimating equation-based estimator that is doubly robust. Double robustness means that it is a consistent estimate of the causal effect when either Q_n is a consistent estimate of Q_0 or g_n is a consistent estimate of g_0. The A-IPTW estimator is:

$$\begin{aligned}\psi^a_{n,A-IPTW} &= \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)}{g_{A_n}(A_i,W_i)}(Y_i - Q_{Y_n}(a,W_i)) \quad (10)\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(Q_{Y_n}(a,W_i)).\end{aligned}$$

Both the IPTW estimator and the A-IPTW estimator may blow up when $g_{A_n}(A_i,W_i)$ is not well-bounded. In situations where $g_{A_n}(A_i,W_i)$ is close to zero for some of the observed browsers these estimators may be unstable. A quick examination of these estimators reveals why they may be unstable. Since $g_{A_n}(A_i,W_i)$ is in the denominator, if it is close to zero it will result in a very large contribution to the estimating equation for the particular browser. This contribution is unbounded when $g_{A_n}(A_i,W_i)$ is not bounded and this can end up resulting in a very poor estimate. The lack of boundedness of $g_{A_n}(A_i,W_i)$ has been called a violation of the positivity assumptions or a violation of ETA (experimental treatment assumption). For a more thorough discussion of these types of violations see [12],[15].

The last estimator we will explore is the Targeted Maximum Likelihood Estimator (TMLE). TMLE is a substitution estimator, like MLE above. Moreover, the TMLE is double robust and locally efficient [10], like A-IPTW. The TMLE, $\Psi(Q^*_{Y_n})$, is of the following form:

$$\psi^a_{n,TMLE} = \frac{1}{n}\sum_{i=1}^{n}Q^*_{Y_n}(a,W_i), \quad (11)$$

where $Q^*_{Y_n}(a,W_i)$ is an update of $Q_{Y_n}(a,W_i)$ specifically chosen to target the parameter of interest. This update is done by fluctuating $Q_{Y_n}(a,W_i)$ with a parametric sub-model of the following form:

$$logit(Q^*_{Y_n}(a,W_i)) = logit(Q_{Y_n}(a,W_i) + \epsilon h_{a,n}(A_i,W_i)), \quad (12)$$

where $h_{a,n}(A_i,W_i) = I(A_i=a)/g_{A_n}(a,W_i)$. Using the logit here is just a computational trick to arrive at the TMLE and is not based on the fact that logistic regression was used to estimate the initial Q_{Y_0} or g_{A_0}. Implementing this sub-model fluctuation can easily be done using the standard glm function in most statistical packages with an offset equal to the logit of $Q_{Y_n}(a,W_i)$. The theoretical basis for the choice of $h(A,W)$ is explained in van der Laan's seminal paper on TMLE [23]. A more in-depth explanation of the implementation, as well as code for implementing this TMLE may be found in Gruber and van der Laan's gentle introduction to TMLE[3]. The TMLE, like the A-IPTW estimator will not return consistent estimates of the parameter of interest when both Q_n and g_n are mis-specified, as is the case when there are unmeasured confounders. However, the TMLE is not as sensitive to violations in the positivity assumption as the A-IPTW and IPTW estimators presented above because the contribution of each browser to the estimator is bounded between 0 and 1. Thus, the estimator follows a proper model and the final estimates are guaranteed to produce a proper probability that falls in the expected range. However, under more extreme violations of the positivity assumption, the TMLE may also lose some stability and extensions of TMLE that are more robust in these situations have been proposed and implemented [20, 4, 18]. These methods are outside the scope of the current paper.

Confidence intervals and p-values for the A-IPTW, and TMLE may be constructed using the variance of the influence curve as described in [3]. They can be similarly constructed for the IPTW estimator; however, they should be conservative for the IPTW estimator. Alternatively, bootstrap methods may be used to construct these estimates and have been shown to construct better estimates of confidence intervals in finite samples (see e.g [18]).

6. ANALYSIS

In this paper we focused on estimating the effect of display advertising for one marketer of interest. The marketing campaign analyzed was for a major fast food chain. Each of the above methods was implemented to estimate the effect of advertising in each of the following three sub-populations:

1. Individuals who have visited the chain's website in the past, or Action Takers (AT).
2. Individuals who have not visited the site in the past but were targeted based on their natural tendency to convert at a higher rate than the general population, or Network Neighbors (NN). These individuals were targeted using machine learning algorithms.
3. Individuals who have not visited the site, or Run-of-Network (RON).

In each case the data was sampled in the following way:

1. A day t_0 was defined
2. On t_0 all individuals within the specified sub-population for whom a medium-sized ad serving company received a bid request were sampled.
3. W, the vector of potential confounders at baseline, was recorded. These potential confounders included past browsing content, past browsing intensity, IP type (.com, .org, .gov, etc.), Internet connection type, browser used, number of times an ad network has seen the browser, and days since the browser was first seen, as well as if an individual visited the site in the past two days.
4. On the following day, t_0+1, it was recorded whether each sampled browser saw an advertisement for the

marketer of interest. $A = 1$ for individuals who saw an advertisement for the marketer. It was also recorded whether or not the individual took the desired action between t_0 and the time of the impression. If an individual took an action prior to seeing the impression, this action was recorded in their vector of baseline variables, W.

5. The action window was determined to be five days.
6. For those shown the advertisement, the action window begins the second the first advertisement is shown on $t_0 + 1$. For those not shown the advertisement, the action window starts at the same time of day on $t_0 + 1$ that the first bid request is observed on t_0.
7. Each browser was observed for the following five days, to the second, and once the window closed it was recorded if the browser converted in the action window. If an action was observed $Y = 1$ and if no action was observed $Y = 0$.

Since the data was sampled in this way the causal effects estimated may be used to make inferences about how advertising affects conversions in the sub-population (e.g. past site visitors) for which the ad company received bid requests. Choosing the time to start the action window for the untreated, $A = 0$, requires some assumptions since there is no action, such as an advertisement, for which to start the action window. Fortunately, since this is for the unadvertised group there is no reason to believe that right after the start of the window there should be a jump in the probability of converting right at that time. Thus, we chose to use time of day on $t_0 + 1$ that the first bid request was observed on t_0. We did a sensitivity analysis and chose other points in time to start the action window and no difference was seen in the results. It should be made completely clear that the action window was still exactly 5 days to the millisecond for both those shown the advertisement and those not shown the advertisement.

7. RESULTS

Table 1 shows the results of our experiments using the approaches described in the previous sections. In particular, for each of the 5 methods we show the estimates of the conversion rates (C-Rates) with and without the ad as well as the measures of impact. The p-values of 0.000 in the table indicate that the p-value was less than 10^{-16}. The p-value is not shown for the UNADJ estimates since those estimates are known to be biased, and thus the p-values do not provide relevant information. The p-values for the MLE estimate are not provided since there is no theoretical basis for their construction as discussed above. The p-values are displayed for the additive impact and the p-values for relative impact are similar.

In order to obtain reasonable results for the IPTW and AIPTW estimators, the estimated treatment probabilities had to be bounded at .98. Thus, all $g_{A_n}(a = 1, W_i)$ that were greater than 0.98 were set to 0.98. The results without this truncation are shown in Table 2. The violation in the positivity assumption has a drastic effect on the estimate of the conversion probability for the untreated. This is because there are levels of baseline variables that are almost perfectly predictive of being shown an ad, and the IPTW and AIPTW estimator contributions for those browsers are extremely high, as discussed above, causing the estimate to

AT					
	UNADJ	MLE	IPTW	A-IPTW	TMLE
No Ad C-Rate	3.6%	3.6%	15.4%	9.6%	3.7%
Ad C-Rate	4.4%	4.1%	4.1%	4.2%	4.1%
Relative Impact	1.2	1.1	0.3	0.4	1.1
Additive Impact	0.8%	0.5%	-11.4%	-5.5%	0.4%
p-value			0.17	0.48	0.05
NN					
	UNADJ	MLE	IPTW	A-IPTW	TMLE
No Ad C-Rate	0.51%	0.52%	0.52%	0.51%	0.52%
Ad C-Rate	1.03%	0.73%	0.81%	0.83%	0.80%
Relative Impact	2.0	1.4	1.5	1.6	1.5
Additive Impact	0.52%	0.21%	0.29%	0.33%	0.28%
p-value			0.000	0.000	0.000
RON					
	UNADJ	MLE	IPTW	A-IPTW	TMLE
No Ad C-Rate	0.15%	0.15%	0.15%	0.15%	0.15%
Ad C-Rate	0.37%	0.37%	0.35%	0.37%	0.35%
Relative Impact	2.5	2.5	2.4	2.5	2.4
Additive Impact	0.23%	0.23%	0.20%	0.22%	0.20%
p-value			0.126	0.097	0.125

Table 1: Conversion Rates and Impact Of Advertising for the three different subpopulations.

fall way outside the range of a probability, between 0% and 100%. While these observations have sufficient influence in the situation presented in Table 2 to drive the estimate out of the proper range, it is also possible that less severe violations can bias the results even if the estimates are within the proper range. For this reason the TMLE is preferable. Situations where the IPTW and A-IPTW estimators blow up and the TMLEs are stable are fairly common and not specific to the situation observed here. In fact, there are many articles that address this issue (see e.g., [13],[20],[18]). Notice in Table 1 even after bounding the denominator g_{A_n}, although the resulting IPTW and A-IPTW point estimates lie in the appropriate range for the ATs, they are still returning unreasonable (biased) results. In fact the point estimate for IPTW suggests that displaying the advertisement results in 11,400 fewer conversions per every 100,000 times the display advertisement is shown, and the A-IPTW estimate suggests 5,500 fewer conversions per 100,000. The fact that certain individuals are targeted for ads based on their characteristics suggests that violations in the positivity assumption are common in the display advertising environment, making the instability of the IPTW and A-IPTW a valid concern.

	IPTW	A-IPTW	TMLE
No Ad C-Rate	136.3%	-119901.8%	3.6%
Ad C-Rate	4.1%	4.2%	4.2%

Table 2: Impact Of Advertising To Past Converters With Unbounded g_{A_n} Causes IPTW and A-IPTW Estimates To Blow Up

Now we will make some general observations based on Table 1 presented above. For these observations, we will focus on the results based on the TMLE. For past action takers, AT, the advertisement results in 400 extra conversions for every 100,000 times the advertisement was displayed. However, this difference is only borderline significant, suggesting that there may or may not be an effect of showing the advertisement to past action takers. It is not particularly surprising that the effect is borderline significant considering that the act of retargeting, or displaying the advertisement

to people who have visited the site in the past, is a common practice in display advertising. The untreated, or those not exposed to the advertisement in the AT segment of the population, most probably have been shown the display advertisement several times by other firms that perform display advertising. For network neighbors, NN, if all of them were shown an advertisement 800 people out of 100,000 would have converted; whereas, if they were all not shown an advertisement 520 would have converted. The advertisement results in an extra 280 conversions for every 100,000 times the advertisement was displayed or a 1.5 times greater conversion rate. This difference in the conversion rates is extremely significant with a p-value of less than 10^{-16}. For run-of-network, RON, if all of them were shown an advertisement 350 people out of 100,000 would have converted; whereas, if they were all not shown an advertisement 150 would have converted. The advertisement results, within RON, in an extra 200 conversions for every 100,000 times the advertisement was displayed or a 2.4 times greater conversion rate. (Note that ratios may be different than just dividing treated by untreated probabilities that are displayed because of rounding.) The effect of the advertisement within RON is not significant at the 10 percent level.

Now for some general observations comparing the effects of advertising between sub-groups. The base conversion rates for ATs are much larger than for NNs. Unexposed ATs provide 3,180 more conversions than unexposed NNs per 100,000 browsers. In addition, the base rate (unexposed) for RON is 150 conversions per 100,000 browsers while the base rate for NNs is 520. This suggests that the machine learning algorithm mechanism used for choosing individuals to target is successfully choosing individuals that have higher base conversion rates than RON. In fact, those people grouped into the NN segments are 3.5 times more likely than RON to convert even when not shown the advertisement. Furthermore, the effect of advertising to the NNs is larger than the effect of advertising to RON: 280 extra conversions per 100,000 advertisements shown versus 200.[6] This suggests that the machine learning algorithms used to choose individuals to target for display advertisements based on their propensity to convert are successfully choosing individuals that are more likely to be influenced by the display advertisement.

8. INTERNAL METHOD VALIDATION

In this section we will present some evidence that the methods presented above are actually performing as expected. There are several simulations studies for TMLE that display how it performs under different scenarios and compare its performance to the other methods presented above (see e.g. [13], [20],[10]). We will present here some additional analyses that verify that the methods are working in our current setting.

First a negative test was performed. In this test we analyzed whether the advertisement for a different marketer, a telecommunications company, had any effect on the conversions for the fast food chain we were analyzing above. By performing this test, we can see if the methods we have implemented are returning spurious results when, in fact, there is no effect of the advertisement. In running this test we would expect the telecommunication company's advertisement to have no effect on the fast food conversions. Table 3 presents the results of this test for the marketer of interest's ATs. The TMLE estimates that browsers shown the advertisement will convert at a rate of 3.79 percent and those not shown the advertisement convert at a rate of 3.84 percent for an additive difference of -0.06 percent (p-value 0.89). Thus no effect of the telecommunication company's ad was observed.

	UNADJ	MLE	IPTW	A-IPTW	TMLE
No Ad C-Rate	3.84%	3.85%	3.89%	3.84%	3.84%
Ad C-Rate	4.07%	3.59%	3.89%	3.79%	3.79%
Relative Impact	1.06	0.93	1.00	0.99	0.98
Additive Impact	0.23%	-0.26%	-0.00%	-0.05%	-0.06%
p-value			.99	0.91	0.89

Table 3: Impact Of Telecommunication Company's Advertisement On Fast Food Conversion

Another test we ran to assess the validity of our proposed approach is that we compared the results to an A/B test that was recently run. The test was run for a different marketing campaign of a clothing retailer. The A/B test revealed a conversion rate for showing the advertisement of 3.26 conversions per 1,000 people compared to a TMLE estimate of 3.19 conversions per 1,000 people. For the untreated, the A/B test estimate was 9.9 conversions per 10,000 compared to 8.2 conversions per 10,000. Again, the IPTW and EE equation-based methods did not perform as well as the TMLE. For the untreated estimates, they were close; however, both methods estimated 2.2 conversions per 1,000 people for showing the advertisement.

9. CONCLUSION AND FURTHER WORK

The results displayed above show that by using methods developed in other fields for estimating causal effects, we can estimate the effect of advertising in observational data and reduce the demand for implementing an A/B, or randomized, test. These methods may also be used to estimate the effect of the advertisement within particular sub-populations of interest. We showed that for a particular fast food marketing campaign the display advertisement resulted in an additional 280 extra conversions per 100,000 non-past action takers that were targeted for advertising (NNs), and 200 additional conversions for run-of-network (RON). The effect estimated within the NNs is extremely significant while the effect within RON is not significant at the 10 percent level. We also showed that advertising to past action takers results in borderline significant increase of 400 extra conversions per every 100,000 times the advertisement is displayed. Despite the fact that the estimated additive impact for ATs is higher than for NNs (400 vs. 280), those conversions for non past action takers (NNs and RON) are potentially more valuable from the company's perspective because they represent a new stream of potential income, and once they convert they become ATs.

The above results also showcased the stability and robustness of using a particular double robust substitution estimator, targeted maximum likelihood estimation (TMLE), to estimate the causal effect of advertising. Inverse probability weighted estimators (IPTW) and estimating-equation

[6]For comparison between sub-groups that have different untreated conversion rates we prefer comparing based on additive impact since relative impact is highly influenced by the level of the untreated conversion rate as discussed above.

based double robust estimators (A-IPTW) tend to be unstable when estimating the causal effect of advertising in situations where there are levels of baseline variables that are highly predictive of browsers seeing a particular display ad in the sub-population of interest. The stability of TMLE relative to these other methods, in these situations where the positivity assumption is violated, makes it particularly appealing for estimating the causal effect of online display advertising.

The analysis presented here is just one example of how causal effect estimation methods may be implemented in the display advertising environment. Extensions of the approach presented here may be used to answer other causal business questions of interest. For example, the approach presented may be used to estimate the effect of the intensity of display advertising, the timing of display advertising, or the characteristics of the creative being displayed.

10. ACKNOWLEDGEMENTS

We would like to thank Tom Phillips, Rod Hook, Brian May, and Andre Comeau for their valuable insights and suggestions. We would also like to thank Edward Capriolo, whose patience and support in using the Hadoop distributed file system and HIVE query language was critical, and without whom this analysis would not have been possible.

11. REFERENCES

[1] 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 KDD*, KDD '10, pages 7–16, New York, NY, USA, 2010. ACM.

[2] J. Ebbert. The ctr means nothing says hp researchers Leighton and Satiroglu. http://www.adexchanger.com/research/clickthrough-rate-rethink11/, Mar. 2011.

[3] S. Gruber and M. van der Laan. Targeted maximum likelihood estimation: A gentle introduction. *UC Berkeley Division of Biostatistics Working Paper Series*, page 252, 2009.

[4] S. Gruber and M. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6(1):18, 2010.

[5] R. Kohavi and R. Longbotham. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter*, 12(2):31–35, 2010.

[6] R. Kohavi, R. Longbotham, D. Sommerfield, and R. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18:140–181, 2009. 10.1007/s10618-008-0114-1.

[7] R. Lewis and D. Reiley. Does retail advertising work: Measuring the effects of advertising on sales via a controlled experiment on yahoo. Technical report, Working paper, 2010.

[8] R. Lewis, D. Reiley, and T. Schreiner. Can online display advertising attract new customers? measuring an advertiser's new accounts with a large-scale experiment on Yahoo! Technical report, Working paper, 2010.

[9] S. Lewis. Mendelian randomization as applied to coronary heart disease, including recent advances incorporating new technology. *Circulation: Cardiovascular Genetics*, 3(1):109, 2010.

[10] K. Moore and M. van der Laan. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in medicine*, 28(1):39–64, 2009.

[11] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2008.

[12] M. Petersen, K. Porter, S. Gruber, Y. Wang, and M. van der Laan. Diagnosing and responding to violations in the positivity assumption. *UC Berkeley Division of Biostatistics Working Paper Series*, page 269, 2010.

[13] K. Porter, S. Gruber, M. van der Laan, and J. Sekhon. The relative performance of targeted maximum likelihood estimators. *UC Berkeley Division of Biostatistics Working Paper Series*, page 279, 2011.

[14] F. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In *Proceedings of KDD*, pages 707–716. ACM, 2009.

[15] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[16] J. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[17] D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688 – 701, 1974.

[18] O. Stitelman and M. van der Laan. Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1):21, 2010.

[19] M. van der Laan. Targeted maximum likelihood based causal inference: Part 1. *The International Journal of Biostatistics*, 6, 2010.

[20] M. van der Laan and S. Gruber. Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, 6(1):17, 2010.

[21] M. van der Laan and J. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.

[22] M. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company, 2011.

[23] M. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.