More Bang for Their Bucks: Assessing New Features for Online Advertisers

Diane Lambert
Google, Inc.
76 Ninth Ave
New York, NY 10011
dlambert@google.com

Daryl Pregibon
Google, Inc.
76 Ninth Ave
New York, NY 10011
daryl@google.com

ABSTRACT

Online search systems that display ads continually offer new features that advertisers can use to fine-tune and enhance their ad campaigns. An important question is whether a new feature actually helps advertisers. In an ideal world for statisticians, we would answer this question by running a statistically designed experiment. But that would require randomly choosing a set of advertisers and forcing them to use the feature, which is not realistic. Accordingly, in the real world, new features for advertisers are seldom evaluated with a traditional experimental protocol. Instead, customer service representatives select advertisers who are invited to be among the first to test a new feature (i.e., white-listed), and then each white-listed advertiser chooses whether or not to use the new feature. Neither the customer service representative nor the advertiser chooses at random.

This paper addresses the problem of drawing valid inferences from whitelist trials about the effects of new features on advertiser happiness. We are guided by three principles. First, statistical procedures for whitelist trials are likely to be applied in an automated way, so they should be robust to violations of modeling assumptions. Second, standard analysis tools should be preferred over custom-built ones, both for clarity and for robustness. Standard tools have withstood the test of time and have been thoroughly debugged. Finally, it should be easy to compute reliable confidence intervals for the estimator. We review an estimator that has all these attributes, allowing us to make valid inferences about the effects of a new feature on advertiser happiness. In the example in this paper, the new feature was introduced during the holiday shopping season, thereby further complicating the analysis.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Experimental Design*

ADKD '07, August 12-15, 2007, San Jose, California, USA.

General Terms

Statistical Inference, Biased Sampling, Propensity Scores, Causal Modeling

1. INTRODUCTION

Randomized experiments are commonplace in the search engine (SE) industry. They are used to evaluate new ranking functions, changes to the user interface, and new algorithms for ad placement. These changes are typically tested on a sample of users that are chosen by randomly directing each query or cookie (depending on the study design) to the new conditions or the standard operating conditions.

Advertisers provide the revenue that allows search engines to provide free services to their users. SEs do their best to ensure that advertisers, like users, are happy by providing tools to manage and tune ads campaigns. These tools are continually improved and expanded according to changing business needs and advertiser feedback. When a new feature is introduced in the ads system front-end, it is often tested on a selected (i.e., white-listed) subset of advertisers before it is introduced to the entire advertiser base. Developers have limited control over which advertisers are white-listed, as this largely depends on customer service representatives (CSRs). CSRs whitelist advertisers for a variety of reasons, including the need to test functionality for both large and small advertisers and the desire to target advertisers that have requested the feature in the past. Indeed, these latter advertisers are especially important as they are likely to exercise the new feature because they are interested in it. The purpose of the whitelist trial is largely to ensure that the feature is bug-free and meets the desired performance criteria. But it is also of interest to learn if the feature increases customer satisfaction. Will advertisers be happier if they use the feature than if not?

We could ask advertisers directly if they are happy with the new feature, but self-reported satisfaction is often unreliable. Advertisers often tell you one thing, but their behavior indicates otherwise. Conventional thinking is that advertiser happiness is better reflected by their retention and spending. If advertisers feel that their return on investment is high, they will direct more money to their ad campaigns. Otherwise they will continue to spend at their current level, or, even worse, decrease their spending. We propose to address advertiser happiness in whitelist studies through metrics like retention and comparisons of pre-feature and post-feature spending behavior, correcting for the biases introduced by the whitelist selection process.

This paper is organized as follows. Section 2 introduces

the application and some of the complexities due to its non-random nature. Since the feature is not yet launched, we are not able to identify its exact nature, but the details are not germane to understanding the methodology. Section 3 then describes the basic statistical model for estimating effects in observational (non-randomized) studies. Section 4 reviews the notion of propensity scores, which measure selection bias. Incorporating propensity scores into the analysis leads to unbiased estimates of the effects of a new feature. In Section 5 we combine propensity scoring with outcome modeling to obtain better, *doubly robust*, estimates, so-called because inferences are valid even if only the propensity model or only the outcome model is correctly specified. (Lunceford and Davidian [7] give an excellent introduction to doubly robust estimators.) We argue that double robustness is extremely important because the analysis of non-random advertiser studies within a large SE company is likely to be automated. In Section 6, we apply the doubly robust estimator to our example, highlighting the data analysis steps along the way. Section 7 gives a high-level view of the literature that guided our thinking. Finally, in Section 8 we discuss other applications of the methods in the SE business.

2. WHITELIST TRIALS

The format of whitelist trials is similar across all new features. The CSRs choose a set of advertisers who are first offered the new feature in the ads front-end. Some advertisers are chosen for their willingness to test new features, some because they have asked for the new feature, some because the CSR believes that the new feature will benefit the advertiser, some because the advertiser is not entirely happy and the CSR is hoping to change that, and some for reasons that are perhaps not so obvious. Accordingly, the first users of a new feature are "selected" in two steps.

1. A CSR selects a whitelist of advertisers that will have access to the feature.
2. White-listed advertisers choose to use the new feature.

New features can require significant engineering resources, so whitelist testing is vital before a new feature is launched. A new feature could both enlarge the advertiser base and the spend of current advertisers, so any early indications that bear on these questions are important to detect, despite the complications that ensue from the nature of advertiser selection.

The set of advertisers on the whitelist is not a random subset of all advertisers, and the set of advertisers on the whitelist that choose to use a new feature is not a random subset of the white-listed advertisers. Advertisers who believe that they may benefit greatly from the new feature may be more likely to participate, or those who would benefit more may not want to be an early adopter, for example. So, neither the CSR selection nor the decision of a white-listed advertiser to use the feature can be considered to lead to random sampling.

The two step selection process results in a 3-way partition of advertisers.

1. Advertisers on the whitelist that use the feature.
2. Advertisers on the whitelist who do not use the feature.
3. Advertisers not on the whitelist.

To make the main ideas clearer, this paper considers only the first group of white-listed advertisers, which we call *users*, and a random sample of the third group of advertisers, which we call *controls*. (Note that the third group is not the same as a random sample of all advertisers, because all advertisers in the first two groups are excluded from controls.) The second group of advertisers would be needed to estimate the adoption rate of the feature or to estimate the effect of merely offering a new feature to advertisers, even if they do not use it. Since we do not consider these estimation problems in this paper, we ignore this group.

The challenge of assessing whether a new feature makes advertisers happier is exacerbated by irregularities in advertiser behavior that largely depend on business conditions outside an SEs control and occur whether or not a new feature is introduced. Figure 1 shows the variability in advertiser spending over an 18 week period for a small sample advertisers. Each panel of the plot corresponds to a single randomly chosen advertiser, and each point is the amount the advertiser spent on that day relative to its maximum daily spend over the 18 weeks. The curve is a smooth through the points and the light grey vertical line delineates the introduction of the new feature. The advertisers shown may or may not have used the feature.

Figure 1 shows that there is no "canonical" advertiser, and for many advertisers there is no canonical spending amount. Advertisers have nearly constant spend (like H), increasing spend (like J), decreasing spend (like G), and cyclical ups and downs (like B). While seasonality is apparent (and expected for this time of year), each advertiser has its own seasonal spending pattern. Teasing out effects in these data would be challenging even with random samples!

We also note that a whitelist is often not one fixed list, but a list that grows through time as more and more advertisers are included in the trial, perhaps growing over several months. The trial we analyze spans the winter holiday shopping season with advertisers added to the whitelist in waves, so that our methods of inferring advertiser happiness have to accommodate unusual traffic patterns. We entertained choosing a different ads front-end feature that avoided the holiday season for this paper, but felt that exceptional WWW traffic is rather the rule than the exception.

Finally, we acknowledge possible confusion in the term "feature" that has one interpretation in software engineering and another in machine learning. We try to avoid this confusion in the sequel by using "module" in reference to the new feature in the ads front-end system and "advertiser characteristic" in reference to features of advertisers that are related to usage and outcome.

3. THE GENERAL FRAMEWORK

We are interested in estimating the effect of a new ads front-end module on an outcome Y (e.g., advertiser spend) for a population of advertisers. We use the term *user* to denote white-listed advertisers that used the module and the term *control* for the comparison advertisers randomly chosen from those not on the whitelist. It is convenient to think of both users and controls as being part of the trial, although the controls are unaware of the trial while it is ongoing and unaffected by it. Usually there are also variables X that consist of both static characteristics of the advertiser, like tenure and country, and summaries of daily

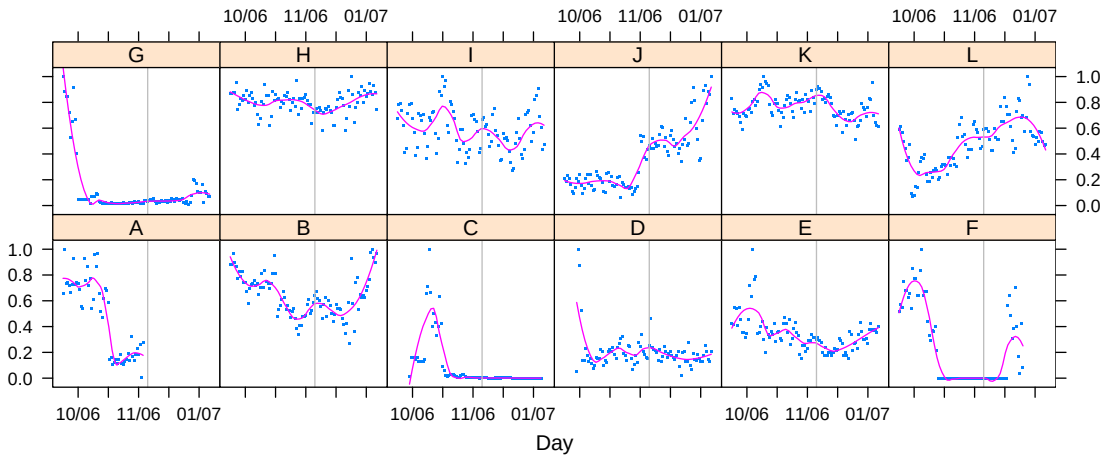

Figure 1: Daily spend as a fraction of maximum spend over the period for a random set of advertisers.

activity, like pre-trial spend. The only restriction on the variables X is that they should depend only on information that could be collected before the trial starts. It is also convenient to introduce a binary variable Z that denotes whether an advertiser was on the whitelist and used the module ($Z = 1$) or an advertiser was not on the whitelist ($Z = 0$). Thus, for each advertiser in the trial, the observed data consists of (Y, Z, X).

The problem with observational studies is that characteristics of advertisers that might affect the outcome Y might also affect whether the advertisers were on the whitelist and used the module. For example, advertisers with long tenure might be both more willing to experiment with new modules and have the financial resources to use them effectively. Differences in outcome then capture not only the effect of using the new model but also the uninteresting difference in tenure between users and controls. Such *confounding* of outcome and selection implies that the effects of using the new module on advertiser spend and retention cannot be estimated correctly (e.g., without bias) unless the sampling bias is taken into account.

The methodology for removing selection bias is best understood through the concept of counterfactuals, which are responses under conditions different from those used in the experiment. Here each advertiser has a pair of potential outcomes

Y_1**:** the outcome (e.g. spend) we would observe if it used the module

Y_0**:** the outcome we would observe if it did not use the module).

Of course, we cannot observe both outcomes for an advertiser. We can observe Y_1 for an advertiser that used the module, but we cannot observe Y_0 for a user, and we can observe Y_0 (but not Y_1) for a control. The unobservable outcomes are termed *counterfactuals* because we did not in fact observe them. (Holland [4] provides an insightful discussion of counterfactual reasoning in statistics.) Using the binary indicator variable Z we introduced earlier allows us to express our observed outcome as

$$Y = ZY_1 + (1 - Z)Y_0. \tag{1}$$

Equation (1) hints that there is a connection between making inferences in observational studies (whitelist trials, in our case) and missing data problems because Z indicates which potential outcome is observed and $1 - Z$ indicates which potential outcome is missing.

The difference $Y_1 - Y_0$ for any advertiser is using that module on that advertiser. Moreover, the distribution of $Y_1 - Y_0$ over all advertisers describes the distribution of the effect of module usage across advertisers. The mean difference over the population of advertisers

$$\Delta = E(Y_1 - Y_0) = E(Y_1) - E(Y_0). \tag{2}$$

is then the *average effect of using the module*. Note that we cannot observe Δ for any advertiser because we can never observe more than one of (Y_0, Y_1), yet the difference in $Y_1 - Y_0$ is an obvious way to describe the effect on an advertiser. The difficulty facing observational studies concerns the extent to which the observed data can be used to estimate Δ. We shall see that this depends on the relationship between the variables (Y, Z, X).

The mean outcome for the whitelisted advertisers that used the module is $E(Y|Z = 1) = E(Y_1|Z = 1)$, which is not the same as $E(Y_1)$ in Equation (2) unless Y_1 and Z are independent. Similarly, the mean outcome of the controls is $E(Y|Z = 0) = E(Y_0|Z = 0)$, which is not necessarily the same as $E(Y_0)$ when the controls are chosen from the set of advertisers that are not on the whitelist rather than from the set of all advertisers. In general, we cannot estimate Δ unless we make assumptions that allow $E(Y_1|Z = 1)$=$E(Y_1)$ and $E(Y_0|Z = 0) = E(Y_0)$.

One case is simple. Random assignment of subjects to user and control groups ensures that selection (Z) is independent of potential outcome (Y_0, Y_1). This justifies comparing the observed average differences of user and control groups, as is typically done in analyses of randomized experiments. That is, if outcome is independent of usage, then $E(Y|Z = 1) = E(Y_1)$ and $E(Y|Z = 0) = E(Y_0)$ and the average effect of using the module can be expressed by the

difference of the average of users and controls, so

$$\Delta = E(Y|Z=1) - E(Y|Z=0). \quad (3)$$

In non-randomized experiments like the whitelist trials of interest to us, progress can be made by exploiting the advertiser characteristics X that are associated with both selection Z and outcome Y. These characteristics are called *confounders*. If all the relevant confounders are known, then conditional on X we have the necessary independence of hypothetical outcomes and selection into the user or control group. Precisely, X includes all confounders if

$$(Y_0, Y_1) \perp Z \mid X \quad (4)$$

where the notation $\perp$ indicates independence, here conditional on X.

The notions of potential outcomes and confounders are powerful enough to provide consistent (i.e., asymptotically unbiased) estimates of the effect Δ. We next review some methods for removing confounding that build on these ideas.

4. PROPENSITY SCORE MATCHING

Knowledge of all the variables that affect both selection and the outcome, as in Equation (4), is not quite enough to get valid estimates of the average effect of using the module. We must also assume that every advertiser has a non-zero probability of being a user and a nonzero probability of being a control. If so, we can obtain valid estimates of the effect of using the module by partitioning X into level sets such that the values of all characteristics in the set are fixed, and then computing a difference Δ_k in the mean of user and control groups within each partition k. The separate estimates Δ_k can be combined to yield a consistent estimate of Δ.

In an important paper, Rosenbaum and Rubin [12] define the propensity score $p(x)$ as the conditional probability that an advertiser is in the group of white-listed users given it has characteristics x:

$$p(x) = P(Z = 1|X = x).$$

They prove that if (Y_0, Y_1, Z, X) satisfies equation (4) and if

$$0 < p(x) < 1 \text{ for all } x, \quad (5)$$

so that every advertiser has a nonzero chance of being a control or a user, then partitioning on $p(x)$ is as good as partitioning on x in the sense that

$$(Y_0, Y_1) \perp Z \mid p(X). \quad (6)$$

Partitioning on $p(X)$ instead of X itself can dramatically reduce the number of partitions in which separate estimates, Δ_k, need to be computed, especially because $p(X)$ is one dimensional regardless of the dimension of X. Rosenbaum and Rubin argue that the method can be applied using an estimate of the propensity score and much of the following work in the area has concerned diagnostics that suggest whether the estimated propensity score model should be trusted. Parenthetically, subsequent work, reviewed in [6], concludes that estimators based on an estimated propensity $\hat{p}(x)$ perform better than those based on the true (unknown) $p(x)$. The intuition is that $\hat{p}(x)$ captures some dependence of the outcome Y on X that does not affect selection bias and hence is not reflected in $p(x)$ itself.

The work of Rosenbaum and Rubin has led to numerous variations on the theme of matching. User and control advertisers can be paired by matching their propensity scores, or subclasses of users and control advertisers can be formed based on quantiles of estimated propensity scores and an average computed for each subclass, or outcome models that estimate Δ_k as a function of X can be fit within each subclass to capture any residual dependence that the propensity score model missed. Propensity score matching is used extensively in medical and social science applications, and until recently it was our preferred method of analysis for whitelist trials. Our main reservation about all variants of matching is the degree of care required in building the propensity score model and the degree to which the matched sets must balance the advertiser characteristics. If analysis is automated, then the care needed may not be taken. In our idealized view, we want our cake and we want to eat it too; specifically, we require an estimator that has good performance *and* that can be applied routinely by non-statisticians.

5. ALTERNATIVE METHODS

The challenge of making causal inferences from observational data is well-studied in statistics and there are many ways to proceed. The established methods all rely on assumptions (like no confounders beyond X) that are difficult to validate with sample diagnostics. Instead of giving a laundry list of methods for estimation for whitelist trials, we discuss two of the most commonly used methods and then a variant that combines both methods into a new method that is more attractive than either of the basic two methods alone. The hybrid estimator, called the "doubly robust" estimator, has certain advantages in our application, inasmuch as it protects against the constituent models being incorrectly specified. Since the doubly robust estimator can be built from simple logistic and ordinary regression models, for example, it can be applied without specialized software. Finally the doubly robust estimator has an estimated standard error that is both easy to compute and accurate [7].

5.1 Direct Outcome Models

Suppose that we knew the true relationship between the outcome Y and the pre-experiment variables X, and that it could be represented as $E(Y|X) = f(X, \beta)$ for an unknown β, and that the effect Δ is the same for all advertisers. We can then estimate the average effect of using the module in an asymptotically unbiased way by fitting the model

$$E(Y \mid Z, X) = f(X, \beta) + Z\Delta, \quad (7)$$

where β represents the effects of advertiser characteristics that are known before the start of the trial (including pre-trial outcomes, like spend), Δ is the effect of using the module, and ϵ is zero mean random noise. We can relax the condition of a constant effect Δ in equation (7) to also depend on X, in which case the mean effect to be estimated is $E\{\Delta(X) \mid X\}$.

Note that Δ in the direct outcome model (7) is exactly

equal to the average effect $E(Y_1) - E(Y_0)$. That is,

$$\begin{aligned} E(Y_1 - Y_0) &= E\{E(Y_1|X) - E(Y_0|X)\} \\ &= E\{E(Y|Z=1,X) - E(Y|Z=0,X)\} \\ &= E\{f(X,\beta) + 1 \times \Delta - f(X,\beta) - 0 \times \Delta\} \\ &= \Delta \end{aligned}$$

where the outer expectation in the first line, for example, averages over the distribution of X. Thus if Δ is constant in X, an unbiased estimate of the regression coefficient Δ is an unbiased estimate of the average effect. An hypothesis test $H : \Delta = 0$ captures whether using the module has influenced advertiser spending.

Estimating Δ by the direct outcomes method is both simple and dangerous. If we misspecify the outcome model (7), our estimate and test statistic do not capture the effect of using the module.

5.2 Inverse Propensity Weighting

If X contains enough information to remove selection bias (as assumption (6) requires), then the observed outcomes ZY_1 for the users on the whitelist satisfy

$$\begin{aligned} E\{E(ZY_1|X)\} &= E\{E(Y_1|\,X, Z=1)P(Z=1|X)\} \\ &= E\{E(Y_1|X)p(X)\}. \end{aligned}$$

Similarly,

$$E\{E((1-Z)Y_0|X)\} = E\{E(Y_0|X)P(Z=0|X)\}.$$

Together, these last two equations lead to the inverse propensity weighted estimator

$$\hat{\Delta}_{IPW} = n^{-1}\sum_{i=1}^{n}\left\{\frac{Z_iY_i}{\hat{p}(x_i)} - \frac{(1-Z_i)Y_i}{1-\hat{p}(x_i)}\right\},$$

where n is the total number of users and control advertisers in the whitelist study and $\hat{p}(x)$ is an estimate of the propensity score $P(Z = 1\,|\,X = x)$. Note that $\hat{\Delta}_{IPW}$ is asymptotically unbiased if $\hat{p}(x)$ is asymptotically unbiased; e.g.., if the correct propensity model is fit by logistic regression. In fact, [11] shows that any estimator of Δ that is asymptotically unbiased must involve inverse propensity weighting no matter what the distribution of (Y, Z, X) is.

5.3 Doubly Robust Estimation

Direct outcome estimates of Δ are valid (asymptotically unbiased) if the fitted outcome model for Y is correct. In practice this is often addressed by fitting separate outcome models, $m_1(x), m_0(x)$, to user ($Z = 1$) and control ($Z = 0$) advertisers. IPW estimates of Δ are valid if the fitted propensity model is correct. Surprisingly, there is a simple combination of the two methods of estimation that is asymptotically unbiased even if either the form of the assumed outcome models or the form of the assumed propensity model (but not both) is wrong. Because the estimate is robust to misspecification of either the outcome models or the propensity model, it is called doubly robust.

The doubly robust estimate $\hat{\Delta}_{DR}$ can be written in terms of the estimated propensities $\hat{p}(x_i)$ and the predictions $\hat{m}_1(x_i)$ and $\hat{m}_0(x_i)$ under the direct outcome mean models for the users and controls respectively. Note that there is a prediction for each advertiser in the trial under both the model $\hat{m}_1$ for the users and the model $\hat{m}_0$ for the controls.

There are two expressions for Δ_{DR} that we find convenient. First,

$$\begin{aligned} \hat{\Delta}_{DR} = n^{-1}\sum_{i=1}^{n}\frac{Z_iY_i - (Z_i - \hat{p}(x_i))\,\hat{m}_1(x_i)}{\hat{p}(x_i)} \\ - n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)\,Y_i + (Z_i - \hat{p}(x_i))\,\hat{m}_0(x_i)}{1-\hat{p}(x_i)}, \end{aligned}$$

which shows that $\hat{\Delta}_{DR}$ adjusts the inverse propensity weighted estimate $\hat{\Delta}_{IPW}$ with residuals of the Z_i from their fitted values $\hat{p}(x)$. Second,

$$\begin{aligned} \hat{\Delta}_{DR} = n^{-1}\sum_{i=1}^{n}\{\hat{m}_1(x_i) - \hat{m}_0(x_i)\} \\ + n^{-1}\sum_{i=1}^{n}\frac{Z_i(Y_i - \hat{m}_1(x_i))}{\hat{p}(x_i)} \\ - n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)(Y_i - \hat{m}_0(x_i))}{1-\hat{p}(x_i)}, \end{aligned}$$

which shows that $\hat{\Delta}_{DR}$ also adjusts the direct outcome estimates with their residuals. Robins, Rotnitzky and Zhao [11] show that $\hat{\Delta}_{DR}$ has the smallest asymptotic variance among all asymptotically unbiased estimates of Δ that are based on $\hat{\Delta}_{IPW}$.

Lunceford and Davidian [7] suggest a simple estimate of the standard error of $\hat{\Delta}_{DR}$ that can be used to give confidence intervals for Δ. We can rewrite the above equation for $\hat{\Delta}_{DR}$ as

$$\hat{\Delta}_{DR} = n^{-1}\sum_{i=1}^{n}\delta_i, \text{ where}$$

$$\begin{aligned} \delta_i =& \hat{m}_1(x_i) - \hat{m}_0(x_i) \\ &+ \frac{Z_i(Y_i - \hat{m}_1(x_i))}{\hat{p}_1(x_i)} - \frac{(1-Z_i)(Y_i - \hat{m}_0(x_i))}{1-\hat{p}(x_i)}. \end{aligned}$$

Then the variance of $\hat{\Delta}_{DR}$ can be estimated by

$$var(\hat{\Delta}_{DR}) = n^{-2}\sum_{i=1}^{n}(\delta_i - \hat{\Delta}_{DR})^2$$

Using simulation studies, Lunceford and Davidian [7] show that this variance estimate is remarkably accurate, yielding confidence intervals of correct size.

6. APPLICATION TO WHITELIST TRIALS IN ADS FRONT END SYSTEMS

6.1 Data Selection

The module we study was made available to 600 advertisers over a period of 11 weeks. Advertisers were added to the study in "waves" roughly every week. For each wave we sampled advertisers not on the whitelist at a 6:1 ratio to form a control group that shared the same wave start date. Although there is no single start date, we use the term "pretrial" to mean the period before an advertiser was added to the study.

The advertisers that the CSRs put on the whitelist are far from a random sample of all advertisers. For example, each advertiser has an assigned "tier" that governs the level of

Table 1: Advertiser characteristics considered in the propensity and outcome models.

Characteristic	Description
Tenure	the length of time the advertiser has been a customer
Tier	the (internally assigned) service level for the advertiser
Channel	the way by which the advertiser became a customer
Country	the country associated with the advertiser's billing address
ConversionTrk	a feature that advertisers can employ to track campaign performance
Reports	the average number of weekly reports requested by the advertiser
Impressions	mean number of ad impressions shown per day
CTR	the mean daily ratio of clicks to impressions (i.e., click through ratio)
AveDailySpend	the mean daily spend of the advertiser
SpendVariation	the variance of local trend in smooth daily spend
SpendPer1000	mean of daily spend per 1000 impressions

customer service that it receives. The highest tier is over-represented on the whitelist and the lowest tiers are under-represented. Our allocation of controls to weekly cohorts maintained the whitelist tier distribution, at least approximately. The main reason for doing this is to help ensure that the overlap (in advertiser characteristics) between user and control advertisers is adequate for propensity models to be applied (as required by 4). In the analysis and conclusions that follow, we sometimes distinguish between the top tier and others.

Of the 600 whitelisted advertisers, 284 used the new module; in what follows they constitute the users. We omit the 316 other white-listed advertisers who did not use the module from the analysis in this paper. As stated in Section 2, the white-listed non-users provide information about the adoption rate of the module, but that is not the focus of the present analyses.

We ran database queries to extract advertiser variables for each advertiser. The variables can be broken into two categories: static advertiser demographics and time-varying metrics like daily click through rate that describe the performance of advertisers' ad campaigns. The demographics are pre-trial conditions and as such are suitable predictors (and hence possible confounders) for our propensity and outcome models. The time-varying metrics were limited to eight weeks pre-trial and eight weeks post-trial. The pre-experiment metrics are suitable predictors for the propensity and outcome models, while the post-trial metrics can be outcomes. For example, the daily spend before the trial captures how advertisers manage their campaigns before they used the new module. Once the trial starts for an advertiser, daily spend is an outcome that we want to track. It would be possible to model the daily time series of performance directly but rather define functions, $f_k(x_t), k = 1, ...K$, that capture salient features relevant to either the decision to use the module or the value of the post-trial outcome.

Table 1 displays the advertiser characteristics that we use in our models. The first five are static demographic characteristics that are thought to influence use of the module and associated spend. The remaining advertiser characteristics are functions of time series that capture the daily performance of an ad campaign.

6.2 Outcomes

We consider two outcomes, *Retention* and *LogSpendRatio*, for each advertiser.

Retention is a binary variable that is one if an advertiser that was active in the eight weeks prior to its trial start remained active in the eight weeks following and zero otherwise. By active we mean that an advertiser had at least one ad impression shown in that period.

LogSpendRatio is based on the relative change in average daily spend (AveDailySpend) for an advertiser. Define S_{post} as the average daily spend for an advertiser in the eight weeks following trial start and S_{pre} as its average daily spend in the prior eight weeks. LogSpendRatio is the logarithm of the ratio of post-trial spend to pre-trial spend for an advertiser:

$$\begin{aligned} LogSpendRatio =& log_2(S_{post}/S_{pre}) \\ =& log_2(S_{post}) - log_2(S_{pre}). \end{aligned}$$

The doubly robust estimator estimates the mean (causal) effect of using the new module. That is, it estimates the mean change in the outcome if an advertiser previously not using the new module starts to use it where the mean averages over all advertisers. Define $\hat{\Delta}_{DR}(Retention)$ to be the doubly robust estimate of the mean change in Retention if an advertiser starts using the new module. Similarly define $\hat{\Delta}_{DR}(Spend)$ to be the doubly robust estimate of the mean difference in LogSpendRatio when an advertiser starts using the new module. Note that by construction the Spend outcome compares the change pre-and post-trial for the users to the controls.

The doubly robust method combines three separate models for each outcome under consideration. In our experience, good performance is obtained by using the same large set of advertiser characteristics as predictors in both the propensity score and the outcome models and using variable selection techniques (or L1-regularization) to mitigate overfitting. We note that even though the theory of the doubly robust estimator indicates that we need not get these models exactly right, it behooves us to make the best possible attempt.

6.3 The Propensity Score Model

The propensity score model captures how advertiser characteristics relate to the probability that an advertiser uses the module. It is possible the propensity model changes over waves but we found negligible evidence of this in our data. [This is easily assessed by including wave as a variable in the model.] All variables in Table 1 were included in a standard logistic regression model and Table 2 summarizes their importance in the models. A centered dot indicates that the advertiser characteristic was not important at the 1%

level. An * indicates statistical significance beyond the 1% level. The column headed "Prop Score" indicates which of these were important in distinguishing between white-listed users of the module and the controls. The fact that so many variables are statistically significant in the propensity score model highlights the fact that there is much selection bias in this study.

Table 2: Roles of advertiser characteristics in the propensity and outcome models. Variables were transformed as appropriate (e.g. log odds of CTR rather than raw CTR were used). A * indicates that the variable was statistically significant at the 1% level.

Advertiser	Prop	Retention		SpendRatio	
Characteristic	Score	User	Cntl	User	Cntl
Tenure	*	*	*	.	*
Tier	*	.	.	.	*
Channel	*	.	.	.	*
Country	*	.	.	.	*
ConversionTrk	*	.	.	.	.
Reports	*	.	.	.	.
Impressions	.	.	.	*	*
CTR	.	.	*	.	.
AveDailySpend	*	.	*	*	*
SpendVariation	*	.	.	.	.
SpendPer1000	*	.	.	*	*

Figure 2 describes the quality of the fitted propensity models. The ROC curve on the left-hand side plot shows the trade-off between true positive (a true user predicted to be a user) and false positive probabilities. The area under the curve (AUC) is a measure that captures the ability of the model to rank users and non-users appropriately; for our propensity score model we observe AUC=0.88.

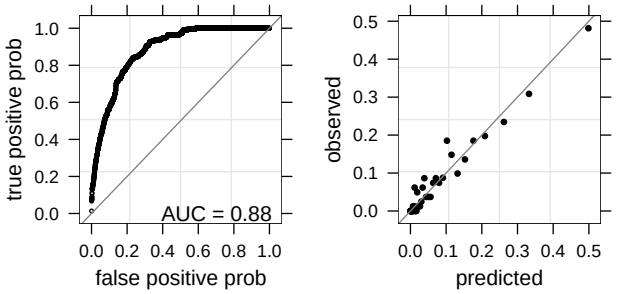

Figure 2: Plot illustrating the quality of the propensity score model.

The right-hand side panel of Figure 2 shows the agreement between the observed data and the fitted propensity model by tabulating the fraction of users within bands defined by the quantiles of fitted propensities. Perfect agreement would be obtained if all points fell on the 45° line through the origin. Our fitted propensity model is far from perfect, but it fits reasonably well across the whole range of estimated propensities.

6.4 The Direct Outcome Models

We base our outcome model for Retention on a logistic regression of the observed fraction of retained advertisers using the pre-trial advertiser characteristics listed in Table 1 as predictors. The columns in Table 2 headed by "Retention" indicate which of these characteristics demonstrated significant association with retention in the user and control groups. In contrast to self-selection, only a few variables affect retention when controls and users are directly modeled.

The outcome model for LogSpendRatio is a linear regression on the logarithm of SpendRatio. Mean daily number of impressions, mean daily spend, and mean daily spend per 1000 impressions are important in both the user and control models, but advertiser demographics are also important in the controls model. This is not surprising; the controls may be more heterogeneous than the users and the control sample is much larger than the user sample.

To assess the quality of the models, recall that the expression for the doubly robust estimator requires an estimate of the outcome for the user group of advertisers had they not used the module, and an estimate of the outcome for the control group of advertisers had they used the module. Thus, we need to use models fitted to one group of advertisers to predict outcomes for the complementary group of advertisers. Our trust in these models is guided by the extent to which prediction is interpolation rather than extrapolation.

Figure 3 conveys the degree to which the Retention models fitted to users and controls are appropriate for prediction. Each subpanel contains a kernel density estimate of the Mahalanobis distance (based on the in-model covariance) of the out-of-model instances from the in-model mean. In the left hand panel we show the distribution of distances for the model fitted to the controls. The solid curve corresponds to the controls (i.e., the in-model data) and the dashed curve corresponds to the users (i.e., the out-of-model data). The shapes of the distributions are quite different and we observe a bimodality in the distances for the in-model control set. In the right hand panel we show the distribution of distances for the model fitted to the users. The solid curve corresponds to the users (i.e., the in-model data) and the dashed curve corresponds to the controls (i.e., the out-of-model data). Here we observe less pronounced bimodality in the distances for the in-model (users) data and an excess of small distances for the out-of-model control data. In both cases, it appears that there is sufficient overlap in the distribution of distances to mitigate the concern with extrapolation.

The associated plots for LogSpendRatio are shown in Figure 4. For the model fitted to the controls we see extraordinary agreement in the distribution of distances. The model fitted to the users is worrisome since we have a long tail of distances for the out-of-model (controls) data. Most of the mass in the long tail overlaps with the in-model data so we are cautiously optimistic that the predicted LogSpendRatio's for the controls from the user model are valid at least for the vast majority of advertisers.

6.5 Doubly Robust Application Estimates

Given our satisfaction with the components of the doubly robust estimator, we are ready to compute $\hat{\Delta}_{DR}(Retention)$ and $\hat{\Delta}_{DR}(Spend)$ from the component models $\hat{p}(x)$, $\hat{m}_1(x)$, and $\hat{m}_0(x)$. The effect of tier is pronounced in the propensity score model, so we compute a separate mean effect (and

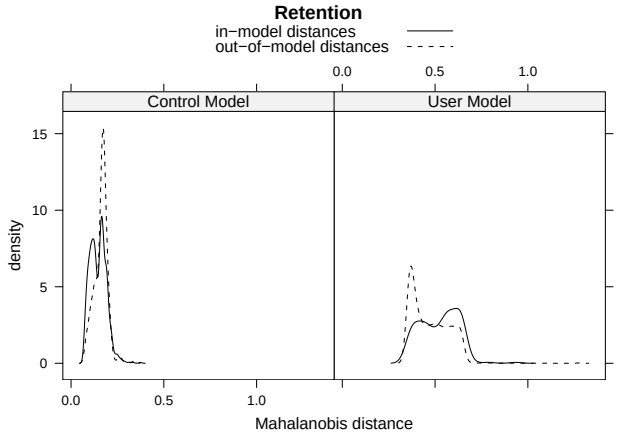

Figure 3: Plot illustrating the degree of extrapolation of the Retention outcome model.

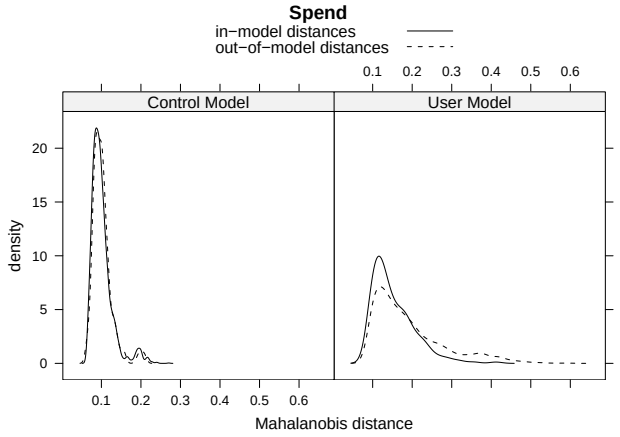

Figure 4: Plot illustrating the degree of extrapolation of the LogSpendRatio outcome model.

associated standard error) for tier 1 for both Retention and SpendRatio. For Retention, we obtain

$$\hat{\Delta}_{DR}(Retention) = \begin{cases} 0.044 \ (0.019), & tier = 1 \\ 0.094 \ (0.043), & tier \neq 1 \end{cases}.$$

In words this says that, on average, there is a +4.4% difference in retention for tier 1 advertisers who use the new module, while for non-tier 1 advertisers who use the module the difference in retention is +9.5%. Both estimated differences are statistically significant at the 0.05 level.

For the log of SpendRatio, we find

$$\hat{\Delta}_{DR}(Spend) = \begin{cases} 0.50 \ (0.40), & tier = 1 \\ 1.09 \ (0.39), & tier \neq 1 \end{cases}$$

In words, the module does not have a statistically significant effect on LogSpendRatio for tier 1 users. For non-tier 1 advertisers, the difference is 1.09 and it is statistically significant. Moving from the differences of log spend to the ratio of spend, the estimated effect corresponds to an increase in relative spend ratio of 2.13 ($2^{1.09}$) for users. Thus, users have a post-trial to pre-trial spend ratio that is more than doubled if they use the new module. Evidently, in addition to the retention gain we saw earlier, the new module has an upside in spend for non-tier 1 advertisers.

7. RELATION TO OTHER WORK

The field of statistics has long been associated with methods and models for the analysis of data from randomized and biased experiments. The papers cited in this section are meant as exemplars as each of these author has multiple papers (even books) in the area thereby demonstrating their longterm contributions to the area.

Fisher [3] is credited with the foundation of experimental design and the central role of randomization. Rubin [13] credits Fisher with the device of counterfactuals in developing the framework for causal inference for non-randomized studies, though Holland [4] calls the counterfactual framework "Rubin's model." Horvitz and Thompson [5] were pioneers in the area of sample surveys where the notion of correcting for non-representative samples first arose.

Our exposure to causal inference from observational data is through the work of Rosenbaum and Rubin [12] who introduced propensity score matching as a means of succinctly capturing differences in the selected users group and the controls. Robins and Rotnitzky [10] chart a more mathematical course and are responsible for the theoretical basis that underlies the doubly robust estimator. Imbens [6] provides a review of the field up to roughly five years ago. He mentions the doubly robust estimator in passing but concentrates on propensity score and outcome modeling separately, largely from the viewpoint of econometric studies. Our appreciation of the doubly robust estimator is due to the work of Lunceford and Davidian [7]. They complement the theory introduced by Robins and co-workers with empirical studies that convincingly demonstrate that the asymptotic properties of the doubly robust method apply to samples that occur in practice. They also introduce the sandwich variance estimator and demonstrate its accuracy.

Our analysis of the Spend outcome focused on the difference in log SpendRatios, or a difference-in-differences in LogSpend. A related approach is the difference-in-differences linear model (see Ashenfelter[1] and Ashenfelter and Card[2]) for LogSpendRatio that results in an estimate of $\Delta(Spend)$. We prefer the double robustness properties of $\hat{\Delta}_{DR}(Spend)$ whereby we fit separate models for users and controls and accommodate selection bias with the propensity score weighting.

McCaffrey, Ridgeway, and Morral [8] provide a link between the statistics literature and machine learning as they apply boosting to the estimation of the propensity score model. Smith and Elkan [14] provide another bridge between machine learning and statistics focusing on Bayesian networks as a means to formalize the conditional independence relationships that occur when training a model on a different population than the model will be applied to. Of course, Pearl [9] is the father of the foundation of causal inference in AI.

8. OTHER APPLICATIONS

Advertisers are an important component of the eco-system of modern search engines: advertisers come to a SE because it has users, users come because the SE provides high-quality search results, and publishers come because the SE has a

large inventory of ads that allows publishers to monetize their content. Improving the "user-experience" for all three of these groups is key to the viability and growth of a SE.

In this paper we introduced a new method for analyzing a common form of advertiser trial and illustrated the ideas on a specific study. Propensity score matching and the doubly robust estimator are broadly applicable within the enterprise, specifically to all three participants in the eco-system. Consider the application to studying whether a new self-selected service leads end-users to search more. In this scenario we have exactly the same problem — are there confounders that affect both the probability of self-selection and search frequency? The methods we propose can be applied once pre-experiment characteristics of users are extracted from logs. Some likely candidates are browser type, geo-location, use of other services, visit frequency, average number of searches per week, etc.

We are currently involved with exploring these new applications as well as productionalizing the application to our ads front-end testing environment.

9. REFERENCES

[1] O. Ashenfelter. Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics*, 67:648–660, 1985.

[2] O. Ashenfelter and D. Card. Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, 60(1):47–57, 1978.

[3] R. Fisher. *The Design of Experiments*. Hafner Publishing Company, 1935.

[4] P. Holland. Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81:945–970, 1986.

[5] D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

[6] G. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

[7] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23:2937–2960, 2007.

[8] D. McCaffrey, G. Ridgeway, and A. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425, 2004.

[9] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[10] J. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129, 1995.

[11] J. Robins, A. Rotnitzky, and L. Zhao. Analysis of semiparametric regression models with missing data. *Journal of the American Statistical Association*, 90:106–121, 1995.

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

[13] D. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[14] A. Smith and C. Elkan. A bayesian network framework for reject inference. In *Proceedings ACM SIGKDD*, pages 286–295, 2004.