Marketing Science Institute Working Paper Series 2022

Report No. 22-106

Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement

Brett R. Gordon, Robert Moakler and Florian Zettelmeyer

"Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement" © 2022

Brett R. Gordon, Robert Moakler and Florian Zettelmeyer

MSI Working Papers are Distributed for the benefit of MSI corporate and academic members and the general public. Reports are not to be reproduced or published in any form or by any means, electronic or mechanical, without written permission.

Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement*

Brett R. Gordon
Kellogg School of Management
Northwestern University

Robert Moakler
Ads Research
Meta

Florian Zettelmeyer
Kellogg School of Management
Northwestern University and NBER

January 7, 2022

Abstract

Randomized controlled trials (RCTs) have become increasingly popular in both marketing practice and academia. However, RCTs are not always available as a solution for advertising measurement, necessitating the use of observational methods. We present the first large-scale exploration of two observational methods, double/debiased machine learning (DML) and stratified propensity score matching (SPSM). Specifically, we analyze 663 large-scale experiments at Facebook, each of which is described using over 5,000 user- and experiment-level features. Although DML performs better than SPSM, neither method performs well, despite using deep learning models to implement the propensity scores and outcome models. The median absolute percentage point difference in lift is 115%, 107%, and 62% for upper, mid, and lower funnel outcomes, respectively. These are large measurement errors, given that the median RCT lifts are 28%, 19%, and 6% for the funnel outcomes, respectively. We further leverage our large sample of experiments to characterize the circumstances under which each method performs comparatively better. However, broadly speaking, our results suggest that state-of-the-art observational methods are unable to recover the causal effect of online advertising at Facebook. We conclude that observational methods for estimating ad effectiveness may not work until advertising platforms log auction-specific features for modeling.

Keywords: Digital Advertising, Field Experiments, Causal Inference, Observational Methods, Advertising Measurement, Double ML.

* Data was de-identified such that consumers could not be identified and was analyzed in aggregate when possible. We thank Neha Bhargava, JP Dubé, Dean Eckles, Garrett Johnson, Randall Lewis, Oded Netzer, Kelly Paulson, Julian Runge, and seminar participants at Facebook, the Marketing Science Conference, the Conference on Digital Experimentation, the NYU-Temple-CMU Conference, and the Virtual Quantitative Marketing Seminar. To be allowed to access the data required for this paper, Gordon and Zettelmeyer were part-time employees of Facebook (Academic Researchers, 3 hours/week). E-mail addresses for correspondence: b-gordon@kellogg.northwestern.edu, f-zettelmeyer@kellogg.northwestern.edu, rmoakler@fb.com.

1 Introduction

In recent years, randomized controlled trials, or RCTs, have become increasingly popular in marketing practice and academia. This trend follows three important developments. First, many firms have invested heavily in their data analytics capabilities. These investments have resulted in experimentation (and other analytical) capabilities and a more wide-spread recognition among practitioners that RCTs are the gold standard of measurement (Kohavi et al., 2020). Second, several leading advertising platforms have created experimentation tools that allow advertisers to run RCTs at no additional cost.[1] Finally, marketing academics have increasingly used RCTs to execute their research agendas (Lewis et al., 2015).

For advertising measurement, however, RCTs are not always available as a solution. Advertisers may operate under internal pressure to forgo a control group to maximize a campaign's reach. RCTs can also be technically difficult or even impossible to implement on many ad platforms (Johnson, 2020). These are among the reasons why advertisers have looked to observational methods as alternative solutions for causal inference. We are motivated by the common practice in which an advertiser—instead of running an RCT—chooses a population to target with an ad campaign. In practice, only a subsample of these eligible users are eventually exposed. To estimate the treatment effect, the advertiser extracts various measures from the exposed group (and potentially the unexposed group) and uses these to estimate the causal or "incremental" effect of the ad campaign.

Recently, Gordon et al. (2019) analyzed the performance of certain observational methods from the program evaluation literature (Imbens and Rubin, 2015) using 15 large-scale RCTs at Facebook. To simulate a setting in which an advertiser had not implemented an RCT, they used data from only the RCT treatment groups and applied popular observational methods. Although the causal estimates obtained from the observational method in some cases came close to those from the RCT, more often they were substantially different.

However, these results do not necessarily imply that observational methods cannot measure the causal effect of advertising. This is for three reasons. First, Gordon et al. (2019) used a small number of experimental studies, and moreover, these studies were not chosen to be representative of the studies advertisers run. Second, the paper focused only on traditional methods from the program evaluation literature; today there are newer methods at the intersection of machine learning and causal inference (Chernozhukov et al., 2018; Athey et al., 2019; Athey and Wager, 2018; Hartford et al., 2017). Third, the data in Gordon et al. (2019) contained only a small set of observable features to match consumers in applying the program evaluation methods.

This paper addresses the above shortcomings. First, we analyze many more experimental studies

[1] Google: `https://www.thinkwithgoogle.com/intl/en-gb/marketing-resources/data-measurement/a-revolution-in-measuring-ad-effectiveness/`. Facebook: `https://www.facebook.com/business/help/552097218528551`. Microsoft: `https://help.ads.microsoft.com/#apex/3/en/56908/-1`.

– 663 studies run between November 2019 and March 2020 on the Facebook ad platform. These studies were chosen to be representative of the large-scale experiments advertisers run on Facebook in the United States. The median ad study ran for 30 days with about 7.3 million users across the test and control groups. Within the test group, 77% of users were exposed to at least one ad impression, and the median campaign accrued over 22 million impressions. These studies represent a range of industry verticals such as E-commerce, Retail, Travel and Entertainment/Media. Many of these studies measure several different conversion outcomes across a "purchase funnel," such as page views (upper funnel), adding an item to a digital shopping cart (mid funnel), and purchase (lower funnel).

Second, we estimate the causal effect of advertising using double/debiased machine learning (DML) (Chernozhukov et al., 2018). Traditional machine learning techniques, while appealing for their flexibility and efficiency, can produce biased estimates of causal effects due to regularization and overfitting. DML corrects for the bias introduced by regularization through orthogonalization and removes the bias from overfitting through cross-validation. This technique has become popular for estimating causal effects in a variety of settings, including both in academia and industry.[2] We also evaluate the preferred program evaluation method from Gordon et al. (2019), stratified propensity score matching (SPSM). For both models, we employ highly flexible and scaleable deep learning methods to estimate each model's underlying components. Adopting a scaleable method is important given the size and number of experiments we study.

Third, we use a much larger set of observable features. Having access to a rich set of features helps us address one of the most significant concerns with observational modeling: since ad exposure is non-random, the exposed and unexposed populations are not directly comparable. To obtain an unbiased estimate of the causal effect, observational models appeal to some version of the unconfoundedness assumption. Loosely speaking, this assumption requires that a user's potential outcomes (e.g., do they purchase or not) are independent of treatment status, conditional on the user's observable features. Having a rich set of features that are predictive of treatment and conversion is thus critical to help justify this assumption. We use four groups of variables: (1) a dense set of descriptive user features (e.g., age, gender, number of friends, number of ad impressions in the last 28 days); (2) a sparse set of user interest features (e.g., cooking interests? movie interests?); (3) estimated action rates (e.g., expected probability of a user converting given an ad); (4) prior campaign-related conversion activity (e.g., 30-day lagged outcomes). A number of these features vary over time within each user and likely reflect the intensity of their online browsing

[2] As of this writing (1/5/2022), Chernozhukov et al. (2018) has 1,046 citations in Google Scholar. See Athey and Imbens (2019) for a general review on the relevance of machine learning methods for empirical research. In industry, the technique is used at Uber (`https://medium.com/teconomics-blog/using-ml-to-resolve-experiments-faster-bd8053ff602e`) and Microsoft (`https://medium.com/data-science-at-microsoft/causal-inference-part-2-of-3-selecting-algorithms-a966f8228a2d`), and see the industry case studies presented in the tutorial in Syrgkanis et al. (2021).

behavior, helping us to address the issue of activity bias (Lewis et al., 2011). Other variables play significant roles in the ad delivery mechanism at Facebook.

Collectively, these improvements in the ad study sample, methodology, and features help us to provide a more confident and comprehensive answer to the question of *whether* observational approaches can recover the causal effects of advertising. Moreover, using a large and representative set of experiments allows us to characterize *when* observational methods come closer to recovering the causal effect of advertising. If there are types of ad campaigns where observational models perform well, some advertisers may be able to utilize these approaches to measure the impact of their ad campaigns. We examine these ad campaign segments through an exploratory analysis that attempts to understand the characteristics of studies for which observational models produce more accurate causal effect estimates.

In terms of results, we find that SPSM performs poorly, despite making use of an extensive set of user-level features and a sophisticated machine learning model to estimate the propensity score. DML is, on average, less upwardly biased than SPSM. However, the remaining bias is substantial. The median absolute percentage point difference ("absolute error," or AE) in lift is 115%, 107%, and 62% for upper, mid, and lower funnel outcomes, respectively. These are large measurement errors, given that the median RCT lifts are 28%, 19%, and 6% for the funnel outcomes, respectively.

We find that observational methods perform comparatively better for prospecting campaigns rather than for remarketing campaigns and when the counterfactual conversion rates are smaller. Prospecting campaigns tend to employ broader targeting rules, whereas remarketing campaigns restrict attention to narrower groups of users who probably already interacted with the advertiser. Both findings could be due to the fact that prospecting campaigns tend to have lower baseline conversion rates than remarketing campaigns, such that conversions in the test group are more likely incremental to the ad campaign.

In addition, we observe improved performance of SPSM and DML when studies have more users, when a smaller share of users in the test group are exposed, and when the propensity model performs better. These findings point towards the fact that an overall larger set of unexposed users provides a better "candidate pool" to help either model estimate counterfactual outcomes for the exposed users. However, even with a large set of users, the underlying predictive models need to achieve some level of accuracy in terms of differentiating between exposed and unexposed users. Nonetheless, even under the best of circumstances, observational methods do not reliably estimate an ad campaign's causal effect.

To the best of our knowledge, the quality of the data and sophistication of models we have used is close to the best available at the most sophisticated advertising platforms and far exceeds what individual advertisers have access to without partnering with an advertising platform. Nonetheless, these results suggest that observational approaches generally fail to measure the true effect of advertising accurately. Since we are using state-of-the-art observational approaches which have

worked successfully in other domains, we conclude that the data at our disposal, as extensive as it is, is inadequate to control for the selection induced by advertising platforms.

We speculate that, even though we use an extensive set of user features, the relevant variables that lead to selection are not recorded in our data. To see why this might be the case, consider that the selection of users into exposed versus unexposed groups occurs inside advertising auctions. When a bid request to show an ad to a user is triggered, advertising platforms use a ranking algorithm to determine which ad, among all ads for which advertisers have placed bids, will be shown. The exact nature of the ranking algorithm is unknown to bidders but usually takes into account the bid amount, the estimated click-through or conversion rate, and a relevance penalty to ensure that users are only shown ads that the advertising platform has deemed relevant to them. These features are continuously updated over time for each user and may be different for each new auction. Given the number of features and users in these systems, only the most recent values are stored for future auctions and features from old auctions are not retained. Additionally, advertising platforms regularly record ("log") the winning ad. However, the volume of bid requests makes it extremely costly to record the ads that failed to win the auction as well as the bid request-level user and platform features used in every auction.

Therefore, statistical models lack the information needed to determine why one user but not another user, both in the target group, would have been shown a specific ad. The reason may be the relevance penalty, the estimated click-through rate, the specific set of competing ads, or other features that are specific to the interaction between each user and the set of ads that were ranked. This discussion has two important implications. First, no advertiser or third-party measurement company has access to this type of data without the explicit consent of the advertising platform. Second, even if these data were logged by the advertising platform, estimating an observational model would be challenging because controlling for selection would require a selection model at the individual bid-request, not the user level. The unit of analysis would be extremely granular and require massive storage and computing power.

This paper makes three contributions. First, we are the first to characterize the performance of a large set of studies that are representative of the large-scale experimental studies advertisers run on Facebook in the United States. We describe the results in a way that is common in the ad industry: by industry vertical and by whether the measured outcome lies in the upper, middle, or lower part of the purchase funnel. The results we describe complement the results on TV ad effects from Shapiro et al. (2021) and can serve as prior distributions that digital advertisers can use for decision making. Second, we add to the literature on whether observational methods using comprehensive individual-level data are "good enough" for ad measurement, or whether they prove inadequate to yield reliable estimates of advertising effects. Our results support the latter. Third, we characterize the circumstances under which a common program evaluation approach, SPSM, and a newer method at the intersection of machine learning and causal inference, DML, perform better

or worse at recovering the causal effect of advertising. We conclude that observational approaches are unlikely to succeed unless advertising platforms fundamentally change what data they log and how they implement observational methods.

This paper follows a series of pioneering studies that evaluate the performance of observational methods in gauging digital advertising effectiveness. Lewis et al. (2011) is the first paper to compare RCT estimates with results obtained using observational methods (comparing exposed versus unexposed users and regression). They faced the challenge of finding a valid control group of unexposed users: their experiment exposed 95% of all US-based traffic to the focal ad, leading them to use a matched sample of unexposed international users. Blake et al. (2015) documents that non-experimental measurement can lead to highly sub-optimal spending decisions for online search ads. However, in contrast to our paper, Blake et al. (2015) use an aggregate difference-in-differences approach based on randomization at the level of 210 media markets as the experimental benchmark and therefore did not implement individual-level causal inference methods. The closest paper to ours is Gordon et al. (2019), which we have described as a starting point and motivation for this work.

A recent paper by Tunuguntla (2021) proposes a novel observational method to estimate ad effects using detailed bid request-level data. The approach circumvents some of the typical endogeneity problems by framing the treatment effect as the advertiser's bid and by the researcher knowing the exact form of the targeting rule. The paper shows the proposed method is able to accurately recover the ad effect when comparing it to the effect obtained from one RCT. A difficulty in our setting is that Facebook (and to our knowledge other advertising platforms like Google) do not log the information necessary to reconstruct the targeting rule. Tunuguntla (2021) was able to circumvent this problem by building and bidding using his own Demand Side Platform (DSP).

This paper proceeds as follows. We first briefly review how advertising works at Facebook, how Facebook implements RCTs, and what determines advertising exposure. In Section 3 we describe how we selected our studies, give an overview of the studies, and describe the user-level features we use to estimate our observational program evaluation model. In Section 4 we explain how we measure the causal effect of an advertising campaign and then present the results of our RCTs. In Section 5, we first introduce the two methods we use to estimate the causal effect of advertising and then show the results. Section 6 explores when observational models do better or worse. Section 7 offers concluding remarks.

2 Overview of Ad Experiments at Facebook

In this section, we describe how Facebook conducts advertising campaign experiments. We explain how these experiments work and describe the measurement challenge in advertising. This is an abbreviated version of the discussion in Section 2 of Gordon et al. (2019).

2.1 Ads at Facebook

Facebook provides advertisers a number of tools and choices for designing new ad campaigns. To launch an ad campaign, an advertiser needs to make three decisions. First, the advertiser needs to choose the primary objective of their campaign. The choices for an objective include increasing awareness of their brand, improving consideration through engagement with the campaign's media, or driving conversions such as sales. Given the chosen objective, Facebook's ad platform will aim to find a broad audience of users who are likely to take the intended action and are more likely to respond positively to the ad. The second choice advertisers need to make is to refine the potential audience defined by the earlier choice of objective. For example, if an advertiser selects conversions as their objective, Facebook will find users who are more likely to convert from among the entire population on Facebook. However, an advertiser may choose to refine this population by focusing on only a particular age range, geographic location, set of interests, or previously observed behaviors. The choice of objective and target audience determine which users may potentially be served an ad from a specific advertiser. Finally, after making these choices that describe the advertiser's target audience, the advertiser chooses the "creative" for their ad. This involves making selecting the ad's image or video, the dimensions of the ad, and the overall design including text and other visuals.

Like most other online advertiser platforms, ads on Facebook are delivered as the result of an auction. This auction is a modified version of a second-price auction where the winning bidder pays only the minimum amount necessary to have won the auction. To balance whether the winning ad from an auction maximizes value for both users and advertisers, the final bid considered in the auction is made up of three components: (1) the bid placed by the advertiser, (2) the probability that the user in the auction will take the action consistent with the advertiser's desired objective, and (3) the quality of the ad (derived from feedback about the ad, whether people hide the ad, and by identifying potential "low-quality" attributes of the ad).[3]

Throughout this paper, we focus on ad campaigns where the advertiser is looking to drive conversions, such as purchases, signing up for a mailing list, or viewing a specific web page. In practice, these conversion events are measured through a "conversion pixel" which is a small piece of code provided by Facebook that advertisers add to specific pages on their website to log specific outcomes. A conversion pixel "fires" when the page it is on is loaded by a user, reporting information about the event back to Facebook for measurement purposes. For example, to log a purchase, an advertiser may place a pixel on an order confirmation page which would only be served and loaded if a sale was finalized. Since pixels are only attached to web pages owned by advertisers, Facebook relies on advertisers to classify the type of outcome that is measured by a corresponding pixel.[4]

[3] `https://www.facebook.com/business/help/430291176997542?id=561906377587030`

[4] Advertisers focusing on measuring outcome events from within mobile apps may also consider building on their pixel setup through the use of the Facebook SDK (`https://www.facebook.com/business/help/1989760861301766?id=378777162599537`).

2.2 Conversion Lift

To measure the effectiveness of a conversion-focused ad campaign, advertisers can utilize Facebook's "Conversion Lift" product to setup an advertising experiment.[5] In Conversion Lift, the ad platform randomly assigns all users in the advertiser's target audience to either a test or control group according to the advertiser's preferred proportion. In the test group, users may receive an ad from the advertiser if that ad wins the auction (we discuss why a test group user may not be exposed in the next section).

In the control group, users are guaranteed not to see an ad from the advertiser's campaign even though their ads will still be involved in the auction process. The reason the ad still participates in the auction, even though it will never be served to the user, is to enable a fair comparison for the sake of ad measurement. When an ad from the advertiser running a Conversion Lift study wins the auction, Facebook will instead serve the user the ad that would have won if that advertiser's ad had not been running. The result of this process is that users in the control group may be served a variety of ads from many advertisers due to the number of competitors and the diverse set of users involved. Whatever ad they are shown, it corresponds to the correct counterfactual ad that would have been displayed in the absence of the ad campaign from the focal advertiser.

Given Facebook's use of a single-user login and identifier that persists across devices and is present for any ad exposure, ad experiments at Facebook are able to avoid contamination between test and control groups. The structure of ad experiments at Facebook results in an unbiased measure of the effect of an ad campaign. These results measure the average treatment effect for the ad media part of the campaign on Facebook and do not generalize to media being run on other channels and are not a measure of future ad effects at a different point in time.

2.3 Ad Exposure

While users in the control group are never shown an ad from a campaign running a Conversion Lift study, users in the test group may or may not be exposed to an ad. Exposure to an ad in the test group is not random—it is due to factors such as user behavior and activity, advertiser characteristics, and platform-level details about the auction.

Users must visit Facebook during a campaign to be exposed to an ad. However, users that are more likely to visit Facebook are also generally more active on the web, and are more likely to take the online action that meets the objective of the advertiser. Additionally, each time a user visits Facebook and an ad auction takes place, the diversity of competing advertisers can vary drastically depending on features such as the time of day and market conditions. While an advertiser that values a user highly will most likely be towards the top of the bid ranking, any one advertiser will not be guaranteed to win a specific user in a specific auction due to the choices of other advertisers

[5] `https://www.facebook.com/business/m/one-sheeters/conversion-lift`

outside of their control. Finally, modern ad delivery systems rely on a complex set of features and predictive models. After an advertiser makes audience targeting choices during campaign setup, the delivery system will continuously make updated predictions on whether a specific user is likely to take the action the advertiser's desired action. As a result, throughout the course of a campaign, the specific users that are more likely to be exposed to an ad may change.

This paper relies on this experimental setup at Facebook as it lets us measure the causal effects of an advertiser's ads through a comparison of the test and control groups, but it also enables us to leverage the one-sided compliance in the test group to mimic a setting where an advertiser chose not to run an experiment but to simply run an ad campaign on Facebook.

3 Data

This section describes how we selected our studies, gives an overview of the studies, and describes the user-level features we use to estimate the observational models.

3.1 Study Selection

The advertising studies analyzed in this paper were chosen to be representative of large-scale advertising experiments run in the United States on the Facebook ad platform. These studies cover a wide range of verticals, targeting choices, campaign objectives, conversion outcomes, sample sizes, and test/control splits. The studies we analyze are a random subset from the set of studies started between November 1, 2019, and March 1, 2020, and had at least one million users in the test group. For each study, we selected all outcomes with at least 5,000 conversions in the test group.[6]

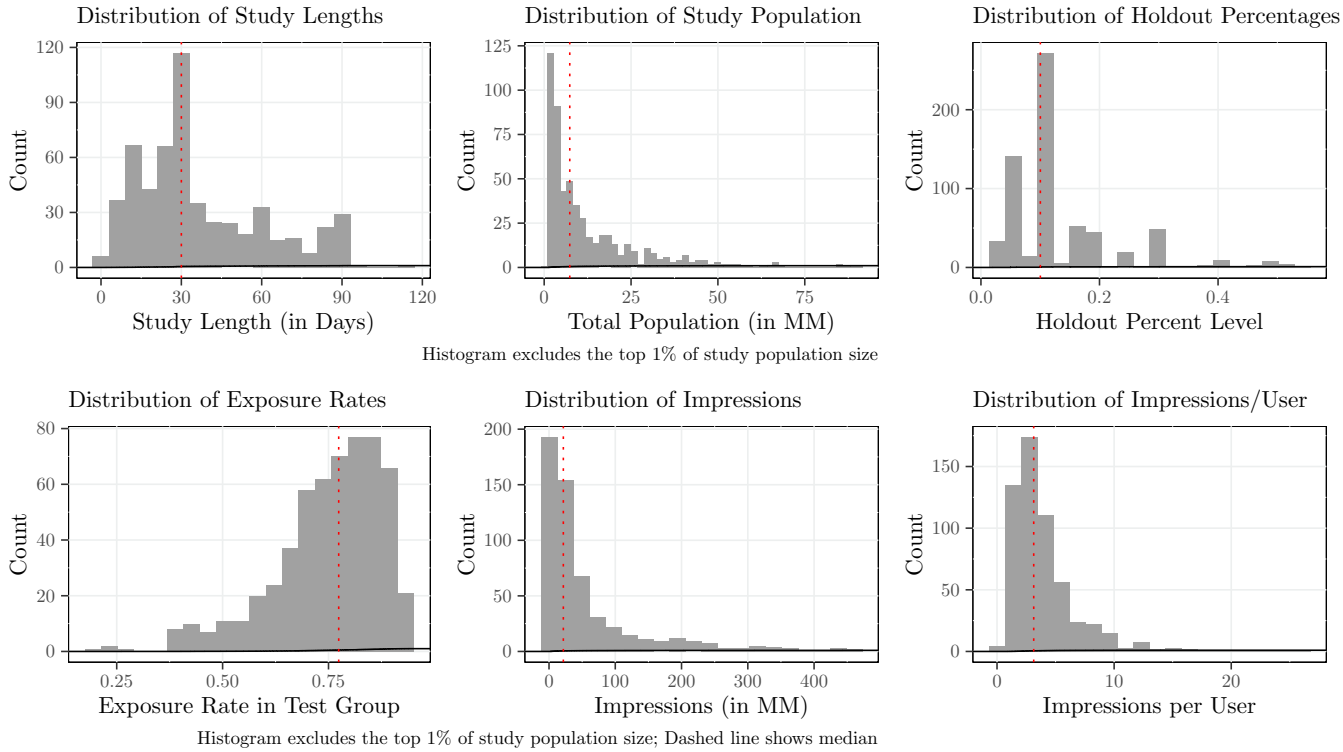
3.2 Overview of Studies, RCTs, and Conversion Events

Our dataset consists of the results of 563 experimental studies run on Facebook in the United States.[7] As Figure 1 shows, studies vary widely by length, by study population size, by the fraction of users in the holdout group, by the rate at which targeted consumers were exposed, and the number of impressions. The median of study length is 30 days and includes 7,372,103 users across test and control groups. The median holdout percentage places 90% of users in the test group and

[6]These minimums were selected by first starting with larger cutoffs during initial versions of these analyses in an effort to provide observational models the largest possible amount of data. We then continually lowered these thresholds to randomly add additional studies while balancing overall computational resources until timing become the main constraint and adding additional studies wasn't feasible.

[7]An earlier version of this paper, dated February 17, 2021, used 850 experimental studies that were selected using the same selection criteria. This earlier version only presented results using SPSM. Updating the analysis to include DML required us to drop some of the experimental studies because data retention policies at Facebook meant that the necessary individual-level data were no longer available. From an expositional perspective, we opted to present results from both models using the same sample of 563 studies. The results for SPSM are similar when using the superset of 850 studies.

Figure 1: Distribution of Study Characteristics



10% in the control group. For those in the test group, the median exposure percentage was 77%, while 23% of users were never exposed. The median of ad impressions per study is 22,115,390.

Table 1: Studies by Number of RCTs Conditions

# Treatment-Control Pairs	# Studies	Percent
1	488	86.7
2	54	9.6
3	17	3
4	4	0.7

Of these studies, 75 contained more than one treatment-control pair, giving us a total of 663 treatment-control pairs.[8] Table 1 shows the distribution of the number of treatment-control pairs conducted for each study. For simplicity, we refer to all these treatment-control pairs as "studies." Most studies measure several different conversion outcomes, such as purchases, page views, downloads, etc. Industry practitioners classify conversion outcomes by whether they occur earlier or later in a hypothetical "purchase funnel." For example, page views occur early in the purchase funnel, adding items to a cart occurs later, and purchase occurs last. Our 663 studies capture a

[8]At Facebook, one treatment-control pair is called a "cell."

total of 1,673 conversion events, measuring different conversion outcomes. Henceforth, we will refer to each conversion event experimental study as an "RCT." We classify RCTs into "Upper Funnel" (601), "Mid Funnel" (475), and "Lower Funnel" (597) experiments. As we describe in Section 2.1, outcomes are measured using "pixels" which advertisers choose to place on their online properties. Table 2 shows the distribution of RCTs (described by their pixel), grouped by when they occur in the purchase funnel.

Table 2: Distribution of Conversion Events

Pixel Name	Funnel Position	N	Percent
view_content	Upper	410	24.5
search	Upper	121	7.2
lead_referral	Upper	70	4.2
add_to_cart	Mid	266	15.9
initiate_checkout	Mid	138	8.2
add_to_wishlist	Mid	34	2
add_payment_info	Mid	21	1.3
tutorial_completion	Mid	16	1
purchase	Lower	409	24.4
app_activate_launch	Lower	97	5.8
complete_registration	Lower	91	5.4

Our advertisers come from many different industry verticals. Table 3 shows the distribution of RCTs by vertical.[9] E-commerce, Retail, Travel, and Entertainment/Media make up over 75% of the RCTs we observe.

Table 3: Conversion Events by Industry Vertical

Industry Vertical	N	Percent
E-Commerce	504	30.1
Retail	377	22.5
Financial Services/Travel	322	19.2
Entertainment/Media	145	8.7
Tech/Telecom	124	7.4
Consumer Packaged Goods	105	6.3
Other	96	5.7

Advertisers measure outcomes across the purchase funnel. Since we define RCTs as the individual conversion event experimental studies run by advertisers, we use RCTs as the unit of observation for the remainder of the paper and analyze outcomes by funnel.

[9]Some smaller verticals were combined to ensure all analyses were sufficiently sized to prevent identifying individual advertisers.

3.3 User Features

For each user in an RCT, we observe a large set of variables logged before users are potentially exposed to an ad in the campaign. We group these features as describing a dense set of descriptive user features, a sparse set of user interest features, estimated action rates, and prior campaign-related outcome activity. These features play a significant role in serving ads as they directly contribute to determining the winner of an ad auction and describing whether a user is likely to convert for any given study. Here is a description of each feature group:

1. **Dense features.** User characteristics such as gender, age, household size, as well as Facebook-specific attributes describing the age of the user's account, number of posts, friends, Likes, and comments, devices used to access Facebook, and measures of activity such as ad impressions, clicks and conversions across the Facebook family of apps. Although some of these variables are effectively time invariant (e.g., gender, age), measures of activity are logged in rolling windows of either 84 or 28 days, and so these vary across user-study combinations. We log these characteristics on the Sunday before the start of each advertising study.

2. **Sparse features.** Facebook allows users to express interests in a large number of different topics. For example, users can express interest in hobbies such as cooking, watching certain genres of movies, listening to certain kinds of music, playing different sports or video games, or being interested in technology, science, the outdoors, or traveling. We log sparse features on the Sunday before the start of each advertising study.

3. **Estimated action rates.** As described in Section 2.1, when a user is first considered as a candidate for exposure in an ad auction for a particular advertiser, Facebook estimates the probability that showing the ad to this user will lead to a desired advertiser outcome. We log these features the first time the ad auction considers a user for a given advertising study.

4. **Prior campaign outcomes.** To measure the results of advertising campaigns, advertisers use conversion pixels (see Section 2.1) to log user outcomes. We measure up to a month of prior conversion data for each outcome event, depending on when the advertiser installed a conversion pixel.

The second group consists of thousands of features, whereas the remaining feature groups comprise roughly 500 descriptive variables. Due to the differences in treated and untreated groups inherent in an observational analysis setup, we utilize these features to create a balanced set of exposed and unexposed users, specifically for understanding treatment status, to satisfy unconfoundedness as described earlier. These features are also used for purely predictive aspects of estimation, namely when describing conversion outcomes for users. While some of these features

are specific to the Facebook platform, many other digital services would have analogs that describe similar sets of features as those we describe.

4 Analysis of Experiments

In this section, we first explain how we measure the causal effect of an advertising campaign and then present the results at the conversion event experimental study level with our 1,673 RCTs.

4.1 Methods

To explain our measurement approach, we make use of the potential outcomes notation. In this subsection we summarize the exposition in Gordon et al. (2019), which in turn used material in Imbens (2004), Imbens and Wooldridge (2009), and Imbens and Rubin (2015). All variables below are specific to an RCT, and so we do not include such a subscript.

Each ad study contains N individuals who are randomly assigned to test or control conditions through $Z_i = \{0, 1\}$. Exposure to ads is given by $W_i(Z_i) = \{0, 1\}$. Users assigned to the control condition are never exposed to any ads from the study, $W_i(Z_i = 0) = 0$. However, exposure is endogenous outcome among users assigned to the test group, such that $W_i(Z_i = 1) = \{0, 1\}$ (i.e., there is one-sided non-compliance). We observe a set of features $X_i \in \mathbb{X} \subset \mathbb{R}^P$ for each user that are unaffected by the experiment. The potential outcomes are $Y_i(Z_i, W_i(Z_i)) = \{0, 1\}$. Given a realization of the assignment, and the subsequent realization of the endogenous exposure variable, we observe the triple Z_i, $W_i = W_i(Z_i)$, and $Y_i = Y_i(Z_i, W_i)$.

As a first step, the intent-to-treat (ITT) effect compares outcomes across random assignment status:

$$\text{ITT} = \mathbb{E}\left[Y(1, W(1)) - Y(0, W(0))\right] \ , \tag{1}$$

with the sample analog being

$$\widehat{\text{ITT}} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i(1, W_i) - Y_i(0, W_i)\right) \ . \tag{2}$$

This calculation rests on the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1978), which requires that a user only receive one version of the treatment and that a user's treatment assignment does not interfere with another user's potential outcomes. The ad experiments on Facebook should satisfy both conditions. The platform's single-user login design helps ensure it shows the right ad to the right user. Although the general form of interference is untestable, we would not expect significant "spillover" effects in the context of these online ad experiments.

However, the observational models we use do not produce ITT estimates because they lack experimental control groups. To compare our RCT estimates with those obtained from observational methods, we instead estimate the average treatment effect on the treated (ATT),

$$\tau \equiv \text{ATT} = \mathbb{E}\left[Y(1, W(1)) - Y(0, W(0)) | W(1) = 1\right] \ . \tag{3}$$

One way to estimate the ATT is to assume the following exclusion restriction: $Y_i(0, W) = Y_i(1, W)$, for any W, which requires that random assignment only affects a user's outcome through receipt of the treatment. This assumption allows us to estimate the ATT using two-stage least squares (2SLS) with assignment Z as an instrument for endogenous exposure W (Imbens and Angrist, 1994). The outcome equation could include heterogeneous effects such as

$$Y_i = \alpha_i + \tau_i W_i + \varepsilon_i \ , \tag{4}$$

in which case the estimate we obtain $\hat{\tau}$ can be interpreted as the average effect among those who were exposed.[10]

4.2 Results

We begin by reporting the estimated ATTs across all 1,673 RCTs. As Figure 2 shows, most ATTs are below 0.01, while some can be as high as 0.13. However, ATTs are difficult to interpret since they contain no information on whether the ATT is "small" or "large."

Hence, to more easily interpret outcomes across advertising studies, we report most results in terms of *lift*, the incremental conversion rate among treated users expressed as a percentage,

$$\begin{aligned} \ell &= \frac{\text{Conversion rate due to ads in the treated group}}{\text{Conversion rate of the treated group if they had } \textit{not} \text{ been treated}} \\ &= \frac{\tau}{\mathbb{E}[Y|Z=1, W=1] - \tau}. \end{aligned} \tag{5}$$

The denominator is the estimated conversion rate of the treated group if they had not been treated. Reporting the lift facilitates the comparison of advertising effects across studies because it normalizes the results according to the treated group's baseline conversion rate, which can vary significantly with study characteristics (e.g., advertiser's identity, the outcome of interest).

Figure 3 shows the distribution of lifts across all RCTs. The average lift is 52%, while the median lift is 9%. Johnson et al. (2017) report a similar median lift estimate in their analysis of 432 experiments on the Google Display Network. Figure 3, however, masks differences between

[10] When we interpret the ATT, it is always conditional on the entire treatment (e.g., a specific ad delivered on a particular day and time) and who is targeted with the treatment. In the context of online advertising, the "entire treatment" includes the advertising platform and its ad-optimization system.

Figure 2: ATTs across all RCTs

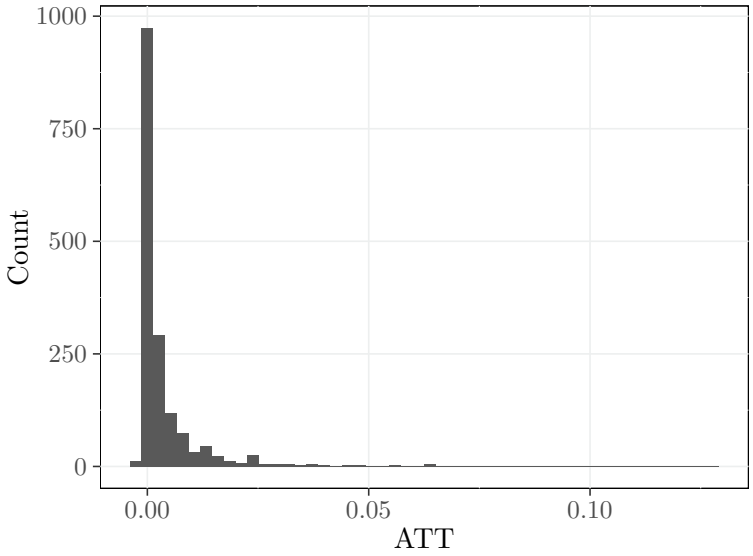
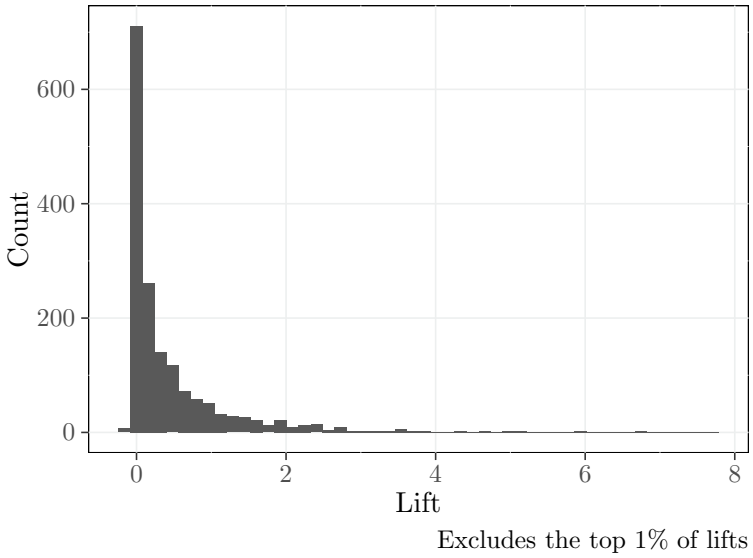
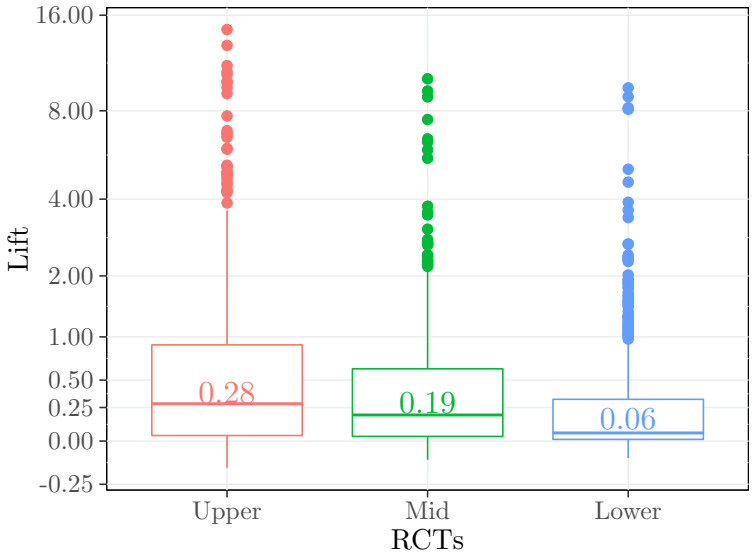


Figure 3: Lifts across all RCTs



Excludes the top 1% of lifts

Figure 4: Lifts by Purchase Funnel Position



event funnel outcomes. As Figure 4 shows, the median lower funnel outcome is smaller than the median lift of mid-funnel outcomes while upper funnel outcomes are higher still.

Variance across studies can vary substantially. Figure 5 displays lifts by the conversion outcome's position along the purchase funnel together with 95 percentile bootstrapped confidence intervals. Also, we indicate whether the lift is statistically different from zero at the 5% level.

We find that 75.8% of experiments with upper-funnel RCTs show lifts that are statistically different from zero. For mid-funnel RCTs, this number is 73.7%, while 59.6% of experiments with lower-funnel RCTs are statistically different from zero.

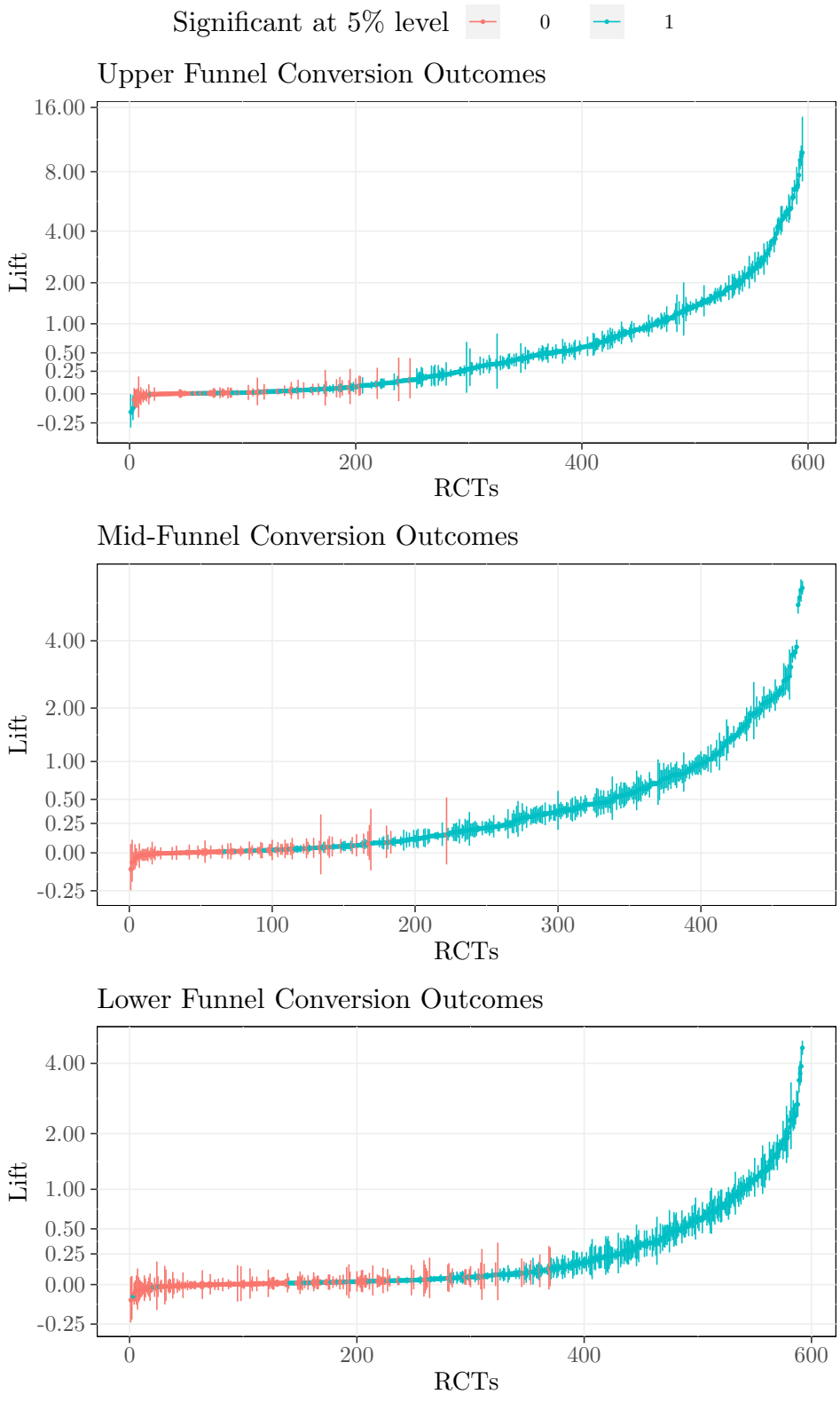
Next, to interpret the result regarding statistical significance, we want to know whether the experiments we observe were adequately powered. Figure 6 shows the proportion of studies that had a 50% ex-ante power to detect a given lift at the 5% significance level.[11]

Across purchase funnel categories, 75% of RCTs were powered to detect a lift of 10%. 85-90% of studies were powered to detect a lift of 20%. Given that the lift in mid and upper-funnel RCTs is typically higher than for lower-funnel RCTs, this explains why the fraction of insignificant lift estimates is higher for lower-funnel RCTs.

Lifts vary widely by industry vertical. Figure 7 shows the distribution of lifts for the seven industry vertical groupings we introduced in Table 3, separated by purchase funnel position.

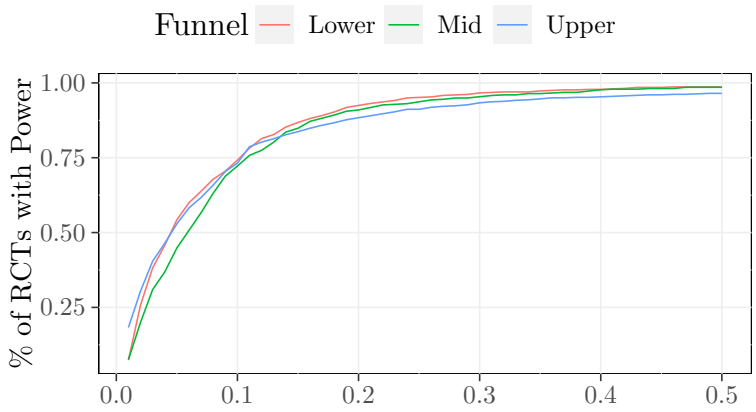
[11] We follow the procedure suggested by Gelman and Hill (2007) and implemented in Shapiro et al. (2021), namely to identify the proportion of studies for which the standard error of the lift estimate is less or equal to detectable lift divided by 1.96.

Figure 5: Lifts by Purchase Funnel Position



Figures excludes the top 1% lifts for each position in the purchase funnel

Figure 6: Detectable Lift



5 Observational approaches

In this section, we discuss how to apply stratified propensity score matching (SPSM) and double/debiased machine learning (DML) methods to estimate the ATT of an ad campaign. The following thought experiment motivates the analysis in this section. Rather than conducting an RCT, an advertiser implemented an ad campaign without an explicit no-ad control group, thus requiring the application of an observational method to measure the impact of the ad. In this setting, all users who satisfied the targeting criteria were made eligible to see the ads. Due to the endogeneity of ad exposure (see Section 2.3), a non-random subset of eligible users are exposed. Comparing outcomes between exposed and unexposed users is likely to produce a biased estimate of the treatment effect due to systematic differences in the users. To properly estimate the causal effect, SPSM and DML use a large set of observable features of users to help adjust for differences between the exposed and unexposed groups.

5.1 Methods

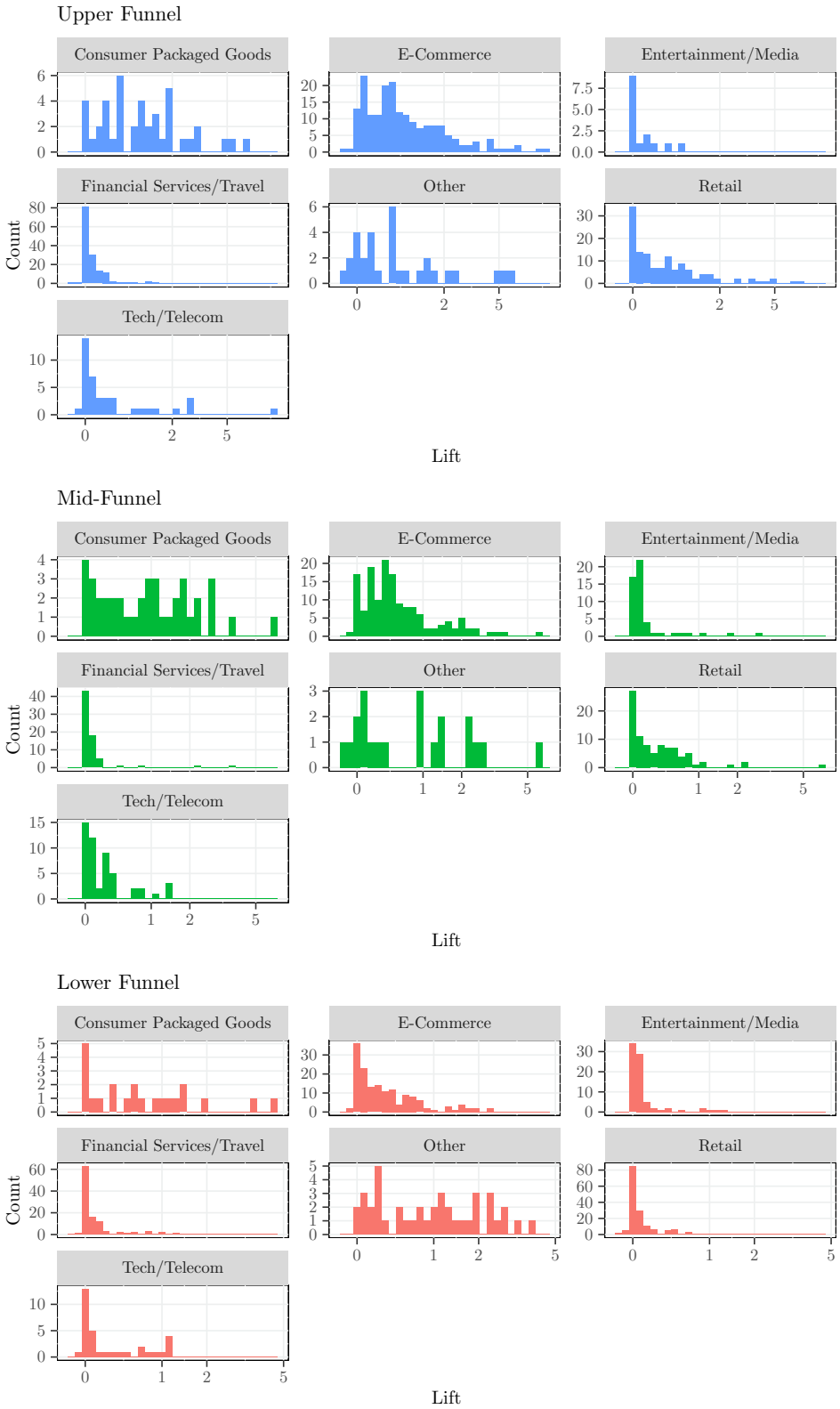
In the observational analysis that follows, we focus on the test group exclusively and ignore the control group. With a slight abuse of notation, we redefine potential outcomes as

$$Y_i(W_i) \equiv Y_i(Z_i = 1, W_i) \ . \tag{6}$$

Thus, in the observational data, for each user we observe the triple (Y_i, W_i, X_i). The ATT obtained using observational method m is

$$\tau^m = \mathbb{E}\left[Y(1) - Y(0) | W = 1\right] \tag{7}$$

Figure 7: Distribution of Lifts by Industry Vertical



Figures exclude the top 1% lifts for each position in the purchase funnel

and the lift is

$$\ell^m = \frac{\tau^m}{\mathbb{E}[Y|W=1] - \tau^m} \ . \tag{8}$$

If treatment status W_i were random and independent of X_i, comparing the conversion rates of exposed and unexposed users would provide an estimate of the campaign's causal effect. Going forward we will refer to this as the "exposed-unexposed" baseline. This ATT effect is

$$\tau^{\text{e-u}} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \ , \tag{9}$$

with a corresponding lift of

$$\ell^{\text{e-u}} = \frac{\tau^{\text{e-u}}}{\mathbb{E}[Y(0)]} \ .$$

This estimate serves as a useful baseline of comparison with SPSM and DML because it has not been corrected to reflect any self-selection into treatment.

However, treatment status W_i usually is *not* random and independent of X_i. Ignoring this fact will lead to bias in the resulting causal effects. In addition to SUTVA, observational models attempt to correct for this bias under several assumptions. One assumption is unconfoundedness,

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \ \mid \ X_i \ , \tag{10}$$

such that potential outcomes are independent of treatment status, conditional on X_i. This (untestable) assumption implies there are no unobserved individual characteristics that are associated with the treatment and potential outcomes. The second assumption is overlap,

$$0 < \Pr(W_i = 1|X_i) < 1, \quad \forall X_i \in \mathbb{X} \ .$$

which requires the probability of treatment to be bounded away from zero and one for all values of user characteristics. Both of these assumptions are necessary to apply SPSM and DML.[12]

5.1.1 Stratified Propensity Score Matching (SPSM)

The first method we use to address the non-randomness of treatment is propensity score matching (Dehejia and Wahba, 2002; Stuart, 2010). The propensity score, $e(X_i)$, is the conditional probability

[12] For DML, in Chernozhukov et al. (2018), see section 5.1 regarding unconfoundedness and Assumption 5.1(c) for overlap.

of treatment given features X_i,

$$e(X_i) \equiv \Pr(W_i = 1 | X_i = x) \ . \tag{11}$$

Under strong ignorability, Rosenbaum and Rubin (1983) establish that treatment assignment and the potential outcomes are independent, conditional on the propensity score,

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \ \mid \ e(X_i) \ . \tag{12}$$

This result shows that the bias from selection can be eliminated by adjusting for the propensity score.

In standard propensity score matching, we find the one (or more) unexposed users with the closest propensity score to each exposed user to estimate the treatment effect. Since this is very computationally burdensome, instead, we stratify on the propensity score: After estimating the propensity score, $\hat{e}(X_i)$, we divide the sample into strata such that within each stratum, the estimated propensity scores are approximately constant. This method, known as stratified propensity score matching (SPSM), scales well and achieves good feature balance without an over-reliance on extrapolation (Imbens and Rubin, 2015). Gordon et al. (2019) found that propensity score matching and stratification on the propensity score produced similar results.

We first partition the estimated propensity scores into J intervals of $[b_{j-1}, b_j)$, for $j = 1, \ldots, J$. Let B_{ij} be an indicator that user i is contained in stratum j,

$$B_{ij} = \mathbb{1} \cdot \{b_{j-1} < \hat{e}(X_i) \leq b_j\} \tag{13}$$

Each stratum contains $N_{wj} = \sum_{i=1}^{N} 1 \cdot \{W_i = w\} B_{ij}$ observations with treatment w. The ATT within a stratum is estimated as

$$\widehat{\tau}_j^{spsm} = \frac{1}{N_{1j}} \sum_{i=1}^{N} W_i B_{ij} Y_i - \frac{1}{N_{0j}} \sum_{i=1}^{N} (1 - W_i) B_{ij} Y_i \ . \tag{14}$$

The overall ATT is the weighted average of the within-strata estimates, with weights corresponding to the fraction of treated users in the stratum relative to all treated users,

$$\widehat{\tau}^{spsm} = \sum_{j=1}^{J} \frac{N_{1j}}{N_1} \cdot \widehat{\tau}_j^{spsm} \ , \tag{15}$$

where N_1 is the number of users in the treated group. We split the sample into 100 equally spaced strata of 1%.[13]

[13] We explored the approach proposed in Imbens and Rubin (2015), which uses the variation in the propensity scores to determine the number of strata and their boundaries, which was used in Gordon et al. (2019). In the present

The variance of the estimator is

$$\hat{V}_{wj} = \frac{S_{wj}^2}{N_{wj}}, \qquad \text{where} \quad S_{wj}^2 = \frac{1}{N_{wj}} \sum_{i:B_{ij}=1, W_i=w} \left(Y_i - \bar{Y}_{wj}\right)^2 \tag{16}$$

$$\hat{V}(\hat{\tau}^{spsm}) = \sum_{j=1}^{J} \left(\hat{V}_{0j} + \hat{V}_{1j}\right)^2 \cdot \left(\frac{N_{1j}}{N1}\right)^2 . \tag{17}$$

5.1.2 Double/Debiased Machine Learning (DML)

In the past few years, the machine learning community has made vast improvements to predictive modeling procedures with new statistical methods and advances in computational hardware. Given the focus of these models on making accurate predictions, they are trained on data sets for which the true answer is known for a set of records and are then applied to new, unseen data. However, in causal inference settings, where the goal is not simply predictive power and where we will never observe true outcomes for any individual record, a direct application of machine learning methods to estimate causal effects can lead to invalid, biased, results.

In recent years, new work had aimed to combine the advantages of machine learning with the causal inference goals of traditional econometrics. Specifically, new literature has addressed the main reasons why predictive models may struggle with causal inference, namely the bias that arises from regularization and overfitting. The double/debiased machine learning (DML) approach introduced by Chernozhukov et al. (2018) corrects for both of these sources of bias by using orthogonalization to account for the bias introduced by regularization and by implementing cross-fitting to remove bias introduced by overfitting. Double machine learning methods build on common econometric approaches by combining the benefits of cutting edge machine learning with causal inference methods such as propensity score matching.

Let outcomes and treatments be given by

$$Y_i = g(W_i, X_i) + u_i \tag{18}$$

$$W_i = e(X_i) + \nu_i \ , \tag{19}$$

where we assume $E[u|X, W] = 0$ and $E[\nu|X] = 0$. Since treatment W_i and user characteristics X_i enter into the nonlinear function $g(\cdot)$, we allow for possibly complex heterogeneous effects of treatment on outcomes. Thus, the ATT is

$$\theta = E[g(1, X) - g(0, X)|W = 1]. \tag{20}$$

To orthogonalize this setup, we can partial out the effect of X on W in three steps: (1) directly

sample of RCTs, we did not find that it made a meaningful difference in our estimates while being computationally much more demanding.

predicting W based on X, (2) directly predicting Y based on X, and (3) by regressing the residuals from step two on the residuals from step 1. In practice, we do this using the score function:

$$\psi(X;\theta,\eta) = \frac{W(Y - g(W,X))}{N_1} - \frac{e(X)(1-W)(Y - g(W,X))}{N_1(1-e(X))} - \frac{D\theta}{N_1} \tag{21}$$

where $\eta = (g, e)$ is a nuisance parameter and N_1 is again the proportion of treated users. If we set the expectation of this function to zero, $\sum_{i=1}^{N} E(\psi) = 0$, we can solve for θ.

We use cross-fitting to account for the bias introduced by overfitting. We start by randomly partitioning our data into K subsets. One partition, k, is held out and we fit models for W and Y on the remaining subsets. We use these models to estimate θ_k in the held-out partition. We continue for the remaining $K-1$ partitions, holding one out and modeling on the remaining subsets, until we have an estimate of θ_k for each partition. To get our final estimate of θ, we average across the partition estimates, $\frac{1}{K}\sum_{k\in K}\theta_k$.

An estimate of the variance can be calculated as

$$\sigma^2 = J^{-2}\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}[\psi(X_i;\theta,\eta)]^2 \tag{22}$$

$$J = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi(X_i;\eta). \tag{23}$$

5.1.3 Estimation

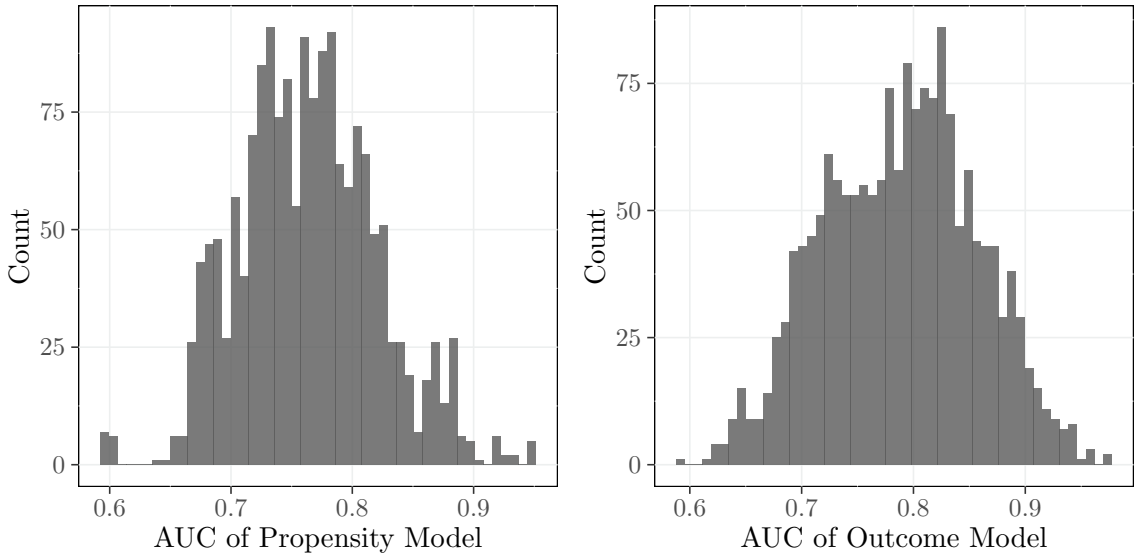
Both SPSM and DML require propensity score models. We estimate the propensity score using a deep learning model due to its ability to leverage both large amounts of dense and sparse features and its ability to parallelize model training across a cluster of computers (Naumov et al., 2019). Deep learning models are characterized by a set of hyperparameters that include the network topology, dropout rate, learning rate, and the number of passes through the training data.

We build a propensity score model using the four groups of features described in Section 3.3 separately for each of the 1,673 RCTs. Additionally, because hyperparameter sweeping requires training a large number of models, we determined a set of optimal hyperparameters by sweeping over a random subset of our propensity score models. We used these hyperparameters across all RCTs.

We obtain propensity scores for each user in our RCTs by training each propensity model using three-fold cross-validation. We use two folds to train a model for estimating propensity scores for the remaining fold. This process takes place three times so that each fold is used two times for training and once for estimation.

The DML model also requires an outcome model ($g(\cdot)$). Similar to our approach for the propen-

Figure 8: Distribution of AUCs in Propensity (SPSM and DML) and Outcome Model (DML only)



sity score models, we trained a deep learning model using three-fold cross-validation to predict whether an individual converts. For these outcome models we used the same set of features described in Section 3.3, as well as the treatment status for each user. We selected hyperparameters by sweeping over a random subset of models and selecting the best set to use across all RCTs. In total, we trained 5,019 propensity score models and 5,019 outcome models for the 1,673 RCTs in our data. Given the size of the RCTs and the number of features, training this number of models required significant cluster computing resources.

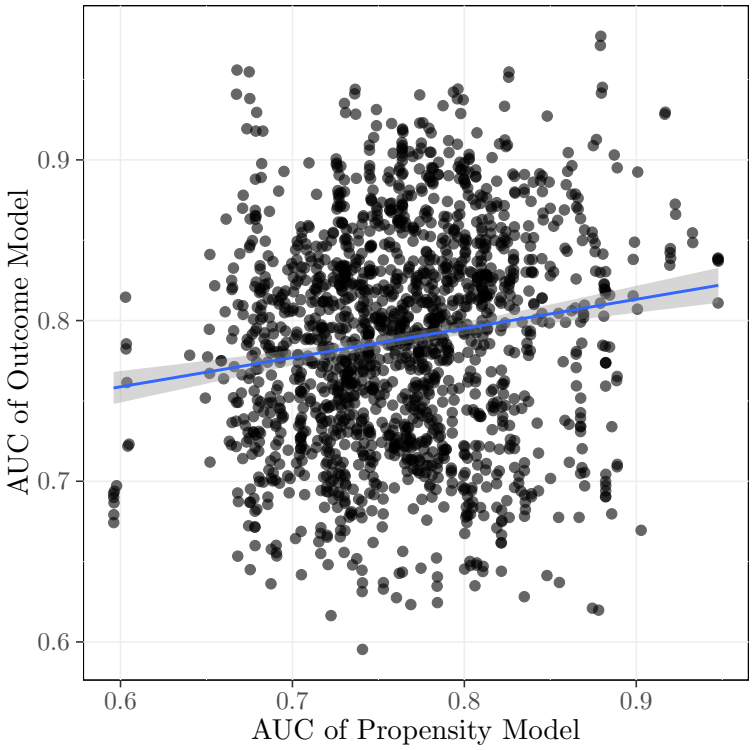
5.2 Model Fit

We summarize the fit of our propensity score and outcome models in Figure 8. We see a wide distribution of model performance with an "area under the ROC curve" (AUC) between 0.6 and 0.95 with a median of 0.76. This variation is expected given that the RCTs in our sample represent a number of different industry verticals, outcome funnel positions, population sizes, and ad campaign targeting and delivery settings.

The AUCs of the outcome model also show a wide distribution with AUCs between 0.6 and 0.98 with a median of 0.79.

The fit of the propensity score and outcome models are not highly correlated. Figure 9 shows a scatterplot of the AUCs of the two models for each RCT. This should not be surprising given that these models predict very different outcomes. The propensity model predicts treatment assignment and—under the identifying assumption for SPSM—should not perfectly predict. In contrast, the outcome model predicts conversions and performs better when the fit is higher.

Figure 9: Scatterplot of Propensity and Outcome Model AUCs



5.3 Results

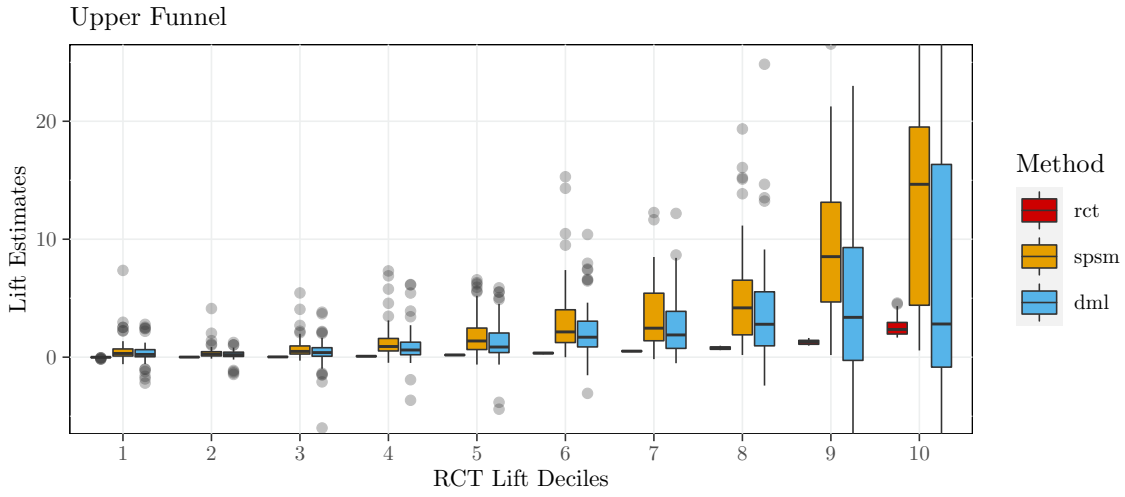
We begin reporting our results with a visual depiction of SPSM and DML lifts ("observational lifts") in Figure 10. To do so, for each purchase funnel position, we split all RCTs into lift deciles where decile 1 contains RCTs with the lowest positive estimated lift and decile 10 contains RCTs with the highest positive estimated lifts. For each decile we show three boxplots where the left-most boxplot summarizes the RCT lifts. Since we form deciles based on these RCT lifts the interquartile range is small. The middle and right boxplots for each decile summarize the lifts estimated by SPSM and DML, respectively. Because the boxplots are difficult to interpret for the lowest deciles, we also summarize the medians, 25th and 75th percentiles in Tables 4.

Scanning across upper, mid, and lower funnel results in Figure 10 reveals a few key results. First, SPSM overestimates the RCT lift by a large amount. To see this result more easily, we characterize each study by the absolute percentage error (APE) between a given method m's lift estimate and the RCT lift estimate as

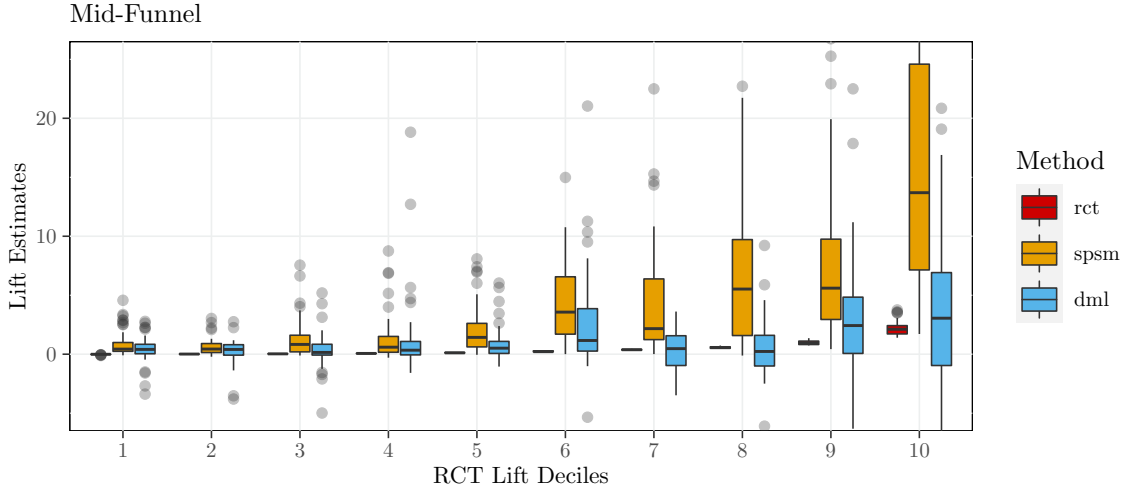
$$\text{APE}^m = \left| \frac{\ell^m - \ell^{rct}}{\ell^{rct}} \right| . \tag{24}$$

For example, suppose an RCT yields an RCT lift estimate of 10% and an observational lift estimate of 50%. We would say that the observational method overestimates by a factor if 5.

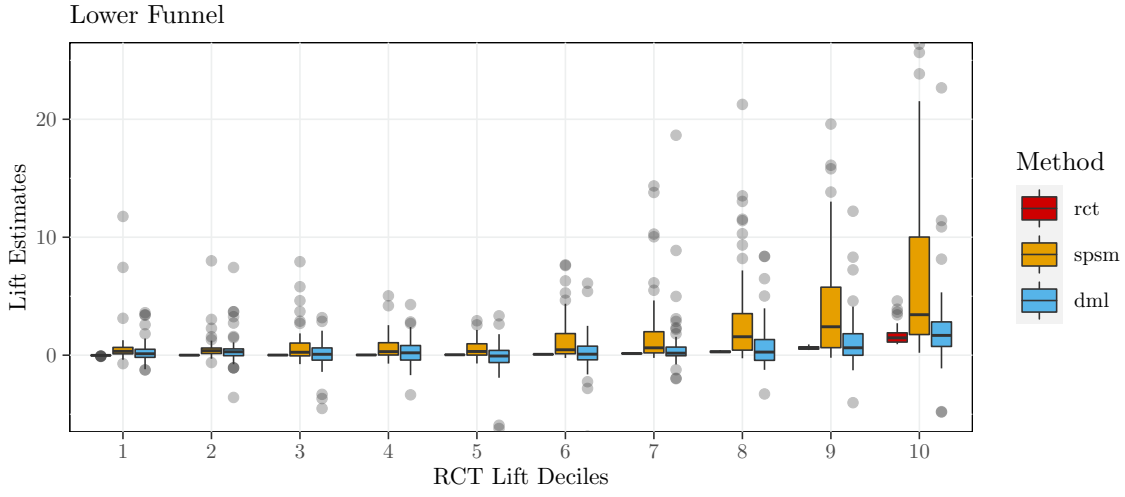
Figure 10: Comparison of RCT Lifts with Lifts Estimated using SPSM and DML



Figures exclude the top 1% lifts for each position in the purchase funnel



Figures exclude the top 1% lifts for each position in the purchase funnel



Figures exclude the top 1% lifts for each position in the purchase funnel

Table 4: Comparison of RCT Lifts with Lifts Estimated using SPSM and DML[†]

Funnel	RCT Decile	p50-RCT	p25-SPSM	p50-SPSM	p75-SPSM	p25-DML	p50-DML	p75-DML
Upper Funnel	1	-0.005	0.099	0.317	0.699	0.038	0.232	0.64
	2	0.007	0.104	0.266	0.466	0.062	0.22	0.428
	3	0.025	0.266	0.489	0.954	0.082	0.39	0.805
	4	0.081	0.529	0.906	1.589	0.207	0.609	1.274
	5	0.188	0.648	1.375	2.469	0.39	0.862	2.051
	6	0.347	1.24	2.144	4.02	0.87	1.694	3.053
	7	0.513	1.393	2.459	5.425	0.746	1.885	3.887
	8	0.768	1.896	4.18	6.528	0.962	2.791	5.544
	9	1.274	4.68	8.524	13.134	-0.271	3.378	9.297
	10	2.349	4.403	14.656	19.512	-0.843	2.814	16.34
Mid Funnel	1	-0.006	0.244	0.433	0.99	0.039	0.406	0.84
	2	0.01	0.14	0.438	0.902	-0.081	0.402	0.805
	3	0.03	0.199	0.829	1.607	-0.074	0.152	0.843
	4	0.065	0.164	0.595	1.507	-0.058	0.339	1.08
	5	0.115	0.62	1.426	2.615	0.066	0.511	1.085
	6	0.214	1.698	3.567	6.564	0.263	1.162	3.86
	7	0.379	1.236	2.166	6.382	-0.96	0.471	1.565
	8	0.558	1.58	5.523	9.715	-1.001	0.232	1.599
	9	0.945	2.942	5.599	9.751	0.065	2.427	4.835
	10	2.113	7.143	13.691	24.588	-0.961	3.052	6.919
Lower Funnel	1	-0.014	0.115	0.327	0.663	-0.181	0.123	0.503
	2	0.002	0.141	0.384	0.609	-0.041	0.289	0.536
	3	0.01	-0.058	0.249	1.031	-0.418	0.077	0.619
	4	0.021	0.016	0.302	1.059	-0.411	0.201	0.82
	5	0.037	0.01	0.316	0.969	-0.624	-0.069	0.407
	6	0.075	0.122	0.46	1.833	-0.392	0.087	0.766
	7	0.145	0.195	0.625	1.991	-0.046	0.173	0.688
	8	0.293	0.429	1.565	3.527	-0.437	0.265	1.328
	9	0.602	0.638	2.411	5.767	-0.006	0.629	1.822
	10	1.474	1.752	3.433	10.016	0.741	1.68	2.825

[†] Table excludes the top 1% lifts for each position in the purchase funnel.

Conversely, if the observational estimate was 5%, we would say that the observational model underestimates by a factor of 2. Table 5 summarizes the degree to which SPSM and DML over- or underestimate RCT lifts, expressed as the APE by decile of the RCT lift. Because APE is difficult to interpret for non-positive RCT lifts, the following table summarizes results only for positive RCT lifts (this excludes 10.04% of RCTs).

Table 5: Absolute percentage error (APE) between SPSM/DML and RCT lifts†

	Upper Funnel		Mid-Funnel		Lower Funnel	
RCT Decile	SPSM	DML	SPSM	DML	SPSM	DML
1	53.93	52.02	57.27	87.5	129.05	186.94
2	19.12	22.44	30.77	24.52	40.61	45.28
3	12.31	12.52	12.55	11.39	19.44	32.22
4	6.87	4.55	9.3	4.66	9.28	16.46
5	5.56	4.85	9.54	5.97	5.5	8.85
6	3.46	3.39	8.77	4.09	6.65	5.03
7	4.68	2.55	7.34	2.65	4.35	3.01
8	5.82	2.94	9.72	2.48	3.84	1.75
9	4.73	1.96	4.06	2.97	0.96	0.86
10	3.06	1.68	4.51	1.5	2	0.77
Median	6.96	5.57	9.48	4.88	7.64	6.72

† Table contains results only for positive RCT Lifts.

For example, for decile 5 in the upper funnel, the table lists a 5.56 for SPSM. This means that SPSM overestimates the RCT lift by a factor of 5.56 or 556%. The table shows that SPSM overestimates the RCT lift by about a factor of 54 to 129 in RCT lift decile 1 and by a factor of 3.06 to 4.51 in RCT lift decile 10. For the median study, SPSM overestimates the RCT lift by a favor of 6.96, 9.48, and 7.64 for upper, mid, and lower funnel outcomes, respectively.

The second key result is that DML overestimates the RCT lift somewhat less. For the median study, DML overestimates the RCT lift by a factor of 5.57, 4.88, and 6.72 for upper, mid, and lower funnel outcomes, respectively.

One problem in interpreting the summaries in Table 5 is that they are multiplicative and thus tend to be large for RCT's in the lower deciles with smaller lifts. Hence, Table 6 reports instead the absolute error (AE) of SPSM and DML, respectively. The absolute error of a given method, m, is

$$\text{AE}^m = \left|\ell^m - \ell^{rct}\right| \ . \tag{25}$$

For example, the "RCT Decile 5 entry" in this table (1.35) means that for decile 5, the average AE between the RCT lift and SPSM lift is 135 percentage points. This metric confirms that DML leads to less mis-estimation than SPSM for upper and mid-funnel outcomes. SPSM and DML are

Table 6: Absolute error (AE) between SPSM/DML and RCT lifts

	Upper Funnel		Mid-Funnel		Lower Funnel	
RCT Decile	SPSM	DML	SPSM	DML	SPSM	DML
1	0.3	0.4	0.45	0.59	0.37	0.35
2	0.27	0.28	0.44	0.5	0.44	0.4
3	0.49	0.52	0.78	0.53	0.37	0.49
4	0.78	0.69	0.55	0.57	0.37	0.63
5	1.35	0.94	1.37	0.65	0.38	0.55
6	1.82	1.41	3.15	1.22	0.34	0.55
7	2.62	1.48	2.16	1.34	0.6	0.38
8	4.99	2.58	5.03	1.58	1.31	0.85
9	6.77	2.96	4.06	2.98	1.73	0.79
10	14.41	8.66	14.99	4.92	2.67	1.29
Median	1.28	1.15	1.6	1.07	0.62	0.62

comparable for lower funnel outcomes according to the AE metric.

The third key result is that the interquartile range of DML estimates is smaller than that of SPSM estimates. This is true in particular for mid- and lower funnel outcomes.

The fourth key result is that DML, while generally superior to SPSM, still fails to reliably approximate the RCT estimates. The median absolute percentage point difference (AE) between RCT and DML lift estimates is 115%, 107%, and 62% for upper, mid, and lower funnel outcomes, respectively. These are very large measurement errors, given that the median RCT lifts are 28%, 19%, and 6% for the equivalent funnel outcomes, respectively.

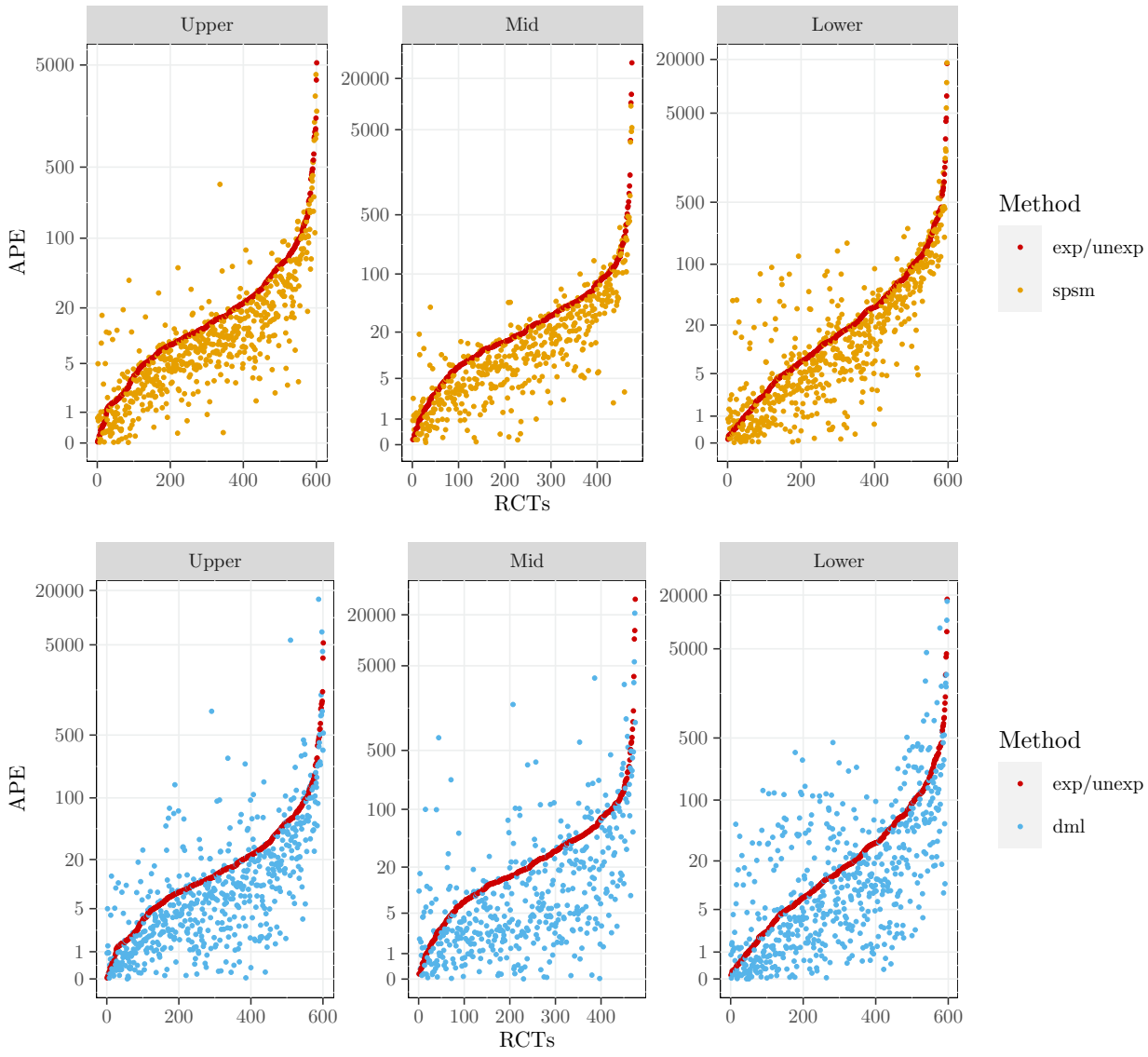
5.4 Reducing Selection Bias

The poor performance of observational methods raises the question of whether these methods are reducing the bias inherent from comparing exposed and unexposed users, i.e., whether the methods improve upon the "exposed-unexposed" estimate. If these methods and the large feature set being used do not lead to better performance, it may indicate that observational methods are mis-specified or that the features we are utilizing are not useful in explaining selection.

In Figure 11 we sort all RCTs from the lowest APE when using the exposed-unexposed lift estimate to the largest APE. We then plot the APE for the exposed-unexposed and for the SPSM estimates by purchase funnel position. The top panel in the figure shows that SPSM improves upon the exposed-unexposed estimate in the majority of cases. The APE of the SPSM estimate nearly always falls below the APE of the exposed-unexposed estimate. In other words, SPSM reduces the selection bias that arises from non-random assignment into the exposed and unexposed groups. However, it does not manage to reduce the bias enough to come close to the RCT estimates.

The bottom panel in Figure 11 shows that DML often reduces the APE of an RCT more than SPSM does, however, compared to SPSM there are more cases where DML increases the APE

Figure 11: Absolute Percentage Error by Event Funnel



compared to the exposed-unexposed estimate.

We can formalize the degree to which SPSM and DML manage to reduce the APE of the exposed-unexposed estimate by calculating the remaining percentage bias (RPB), i.e., the remaining selection effect, after each method has tried to reduce the total selection effect inherent in the exposed-unexposed estimate with

$$\text{RPB} = \left(1 - \frac{\text{APE}^{\text{e-u}} - \text{APE}^{m}}{\text{APE}^{\text{e-u}}}\right) * 100 \tag{26}$$

where m $\in$ {SPSM, DML}. Figure 12 reports the RPB estimates by funnel position. The mean, median, and mode of the remaining percentage bias distribution for all three purchase funnel positions are below 100%. This finding shows that SPSM and DML usually reduce the selection bias present in these RCTs. However, RPB is not bound by 100%. In some cases the predictive models underlying SPSM or DML estimates may be noisy or trying to estimate a small effect which can result in situations where exposed-unexposed estimates may actually provide a closer estimate of RCT results.

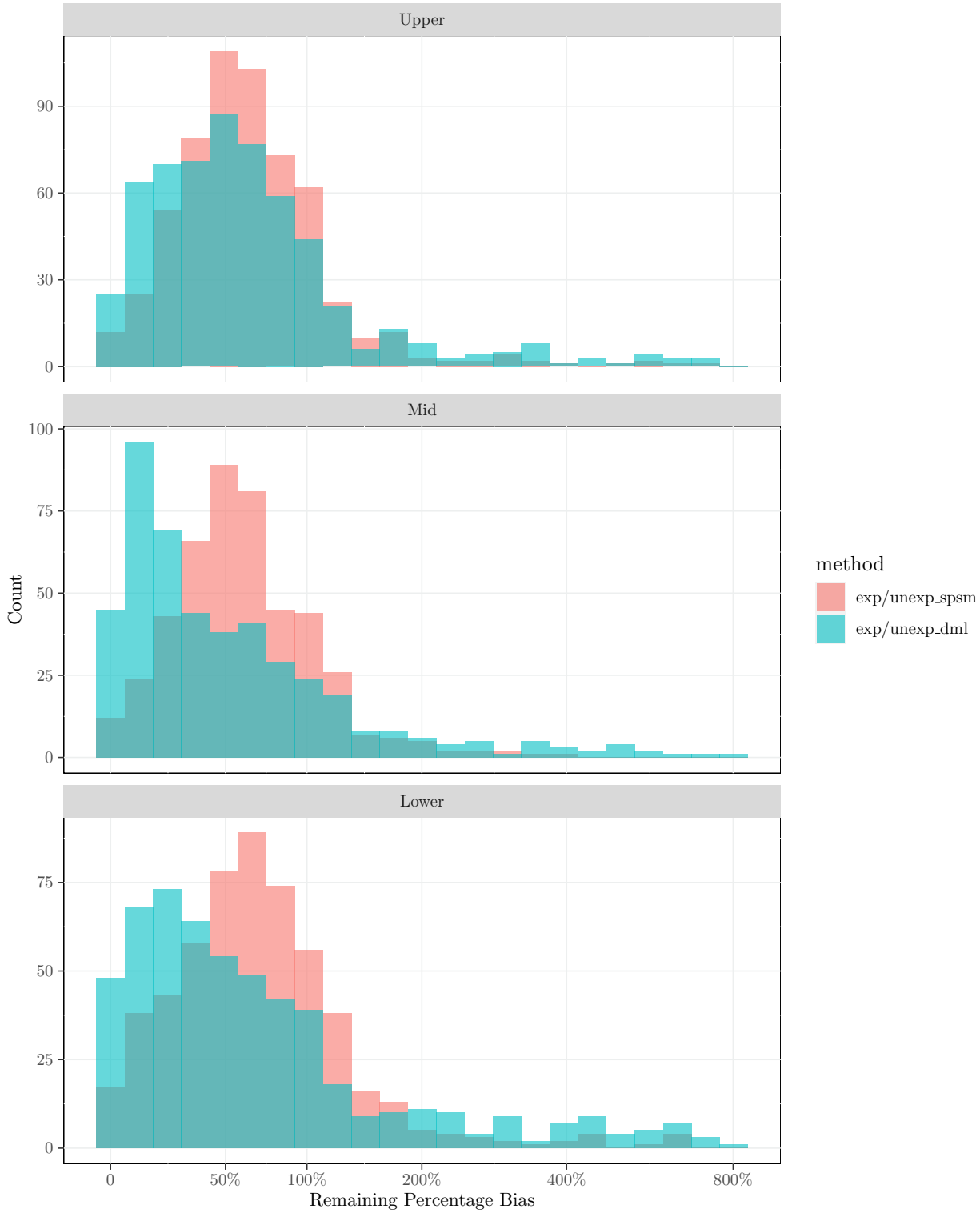
Table 7: Percent of Studies with RPB Below Threshold

Method	Funnel Position	RPB <100%	RPB <50%	RPB <20%
DML	Upper	82.2%	51.2%	19.3%
DML	Mid	83.9%	66.5%	42%
DML	Lower	77.9%	59.2%	29.3%
SPSM	Upper	86.7%	39.3%	8.2%
SPSM	Mid	85.1%	42.3%	10.4%
SPSM	Lower	79.7%	39.2%	13.6%

However, as Table 7 shows, the overall performance of SPSM and DML in reducing selection bias is relatively poor: SPSM improves on the exposed-unexposed estimate in 79.7% to 86.7% of RCTs, depending on the purchase funnel position. Consistent with the lower panel of Figure 11, DML does a bit worse: it improves on the exposed-unexposed estimate in 77.9% to 82.2% of RCTs, depending on the purchase funnel position. However, DML does better than SPSM at reducing the selection bias by 50% or more (51% to 59% vs. 39% to 42%) and at reducing the selection bias by 80% or more (19% to 42% vs. 8% to 13%).

These results lead to two main findings. First, observational models such as SPSM and DML are better at estimating RCT results when compared to a simple comparisons of exposed and unexposed users. In scenarios where treatment is not randomly assigned, methods that control for self-selection into treatment led to noticeable improvements in the accuracy of causal estimates. Second, we observe that even though there is variation in performance across the types of observational models and position in the purchase funnel, observational methods still regularly produce estimates that

Figure 12: Remaining Percentage Bias by Purchase Event Position



are unable to accurately estimate causal effects. Even in the best cases observational methods were unreliable.

6 Explaining Performance

The previous subsection shows considerable heterogeneity in how well SPSM and DML can measure the causal effect of ad campaigns. While SPSM and DML lift estimates are, on average, much higher than RCT lift estimates, this is not always true. Ad campaigns are described by a wide range of features that are either predefined or chosen by an individual advertiser such as the industry vertical an advertiser is part of, the baseline conversion rate that advertisers could expect without ads, the way they target and deliver ads, and the size of a campaign. If there are segments of ad campaigns where observational models perform well, it is possible that some advertisers could utilize these approaches to obtain causal estimates on their campaign's effectiveness. To understand what these segments look like, we now investigate whether the characteristics of the ad campaign explain when SPSM and DML perform comparatively better.

To see how characteristics of the ad campaign are correlated with SPSM and DML performance, we build a predictive model with the APE of an observational method as the target and core study characteristics as features. The unit of observation is an RCT. We include the following study characteristics: The length of the study (in days), the number of users in the test group, the conversion rate of the control group, the exposure rate (what percent of targeted users in the test group were exposed to the ad), the position of the conversion outcome in the purchase funnel, the out-of-sample predictive performance of the propensity scoring model (measured as "area under the ROC curve" or "AUC"), and the prospecting ratio (which is 1 if all targeted users were prospects, 0 if all targeted users were chosen for remarketing, and intermediate values for a mixture of prospecting and remarketing). For DML we also include the out-of-sample predictive performance of the outcome model (measured as AUC).

We exclude RCTs with the highest 5% APE in each purchase funnel position (for SPSM, this excluded RCTs with APEs above 122 for upper-funnel outcomes, 157 for mid-funnel outcomes, and 230 for lower-funnel outcomes; for DML this excludes RCTs with APEs above 149 for upper-funnel outcomes, 250 for mid-funnel outcomes, and 297 for lower-funnel outcomes). This leaves a final sample of 1,596 and 1,588 RCTs to investigate SPSM and DML, respectively.

To account for non-linearities and interactions, we use a random forest. Since this model allows for interactions, we train the model on all RCTs (for the two observational methods across the three purchase funnel positions). Tuning the model yields two randomly selected features at each cut in the tree and a minimum node size of one. We use permutation importance to evaluate the relative importance of each explanatory variable. The results are scaled from 100 (most important

feature) to 0 (least important feature) and shown in Table 8.[14]

Table 8: Variable Importance for SPSM and DML (Random Forest explaining APE)

Relative Variable Importance	SPSM	Relative Variable Importance	DML
Prospecting (vs. Remarketing) Ratio	100	Experiment Length (in days)	100
Exposure Rate	63.3	Number of Users in Test Group	82.3
Control Group Conversion Rate	18.9	Exposure Rate	71.3
Lower Funnel Dummy	11.9	AUC from Propensity Score Model	67.1
Experiment Length (in days)	10	Prospecting (vs. Remarketing) Ratio	65.4
Number of Users in Test Group	8.1	AUC from Outcome Model	41.2
AUC from Propensity Score Model	6.7	Control Group Conversion Rate	20.9
Mid-Funnel Dummy	0	Mid-Funnel Dummy	9.1
		Lower Funnel Dummy	0

Among the top three most important features for explaining absolute percentage error, DML and SPSM highly rank the prospecting ratio and control group conversion rate. Otherwise, the models differ in the order of variable importance. One of the most important features for SPSM, Lower Funnel Dummy, is towards the very bottom of the ranking for DML. In the feature importance ranking, the scaled feature importance scores drop significantly after the third feature for SPSM but tend to decrease more uniformly for DML.

To better evaluate the relationship between study characteristics and the performance of SPSM and DML, Figures 13 and 14 visualize the relationship between each study characteristic and APE^{SPSM} and APE^{DML} while accounting for the average effect of the other study characteristics in the model. These plots are referred to as partial dependence plots (PDP) (Greenwell, 2017). Note that each plot displays a "rug" along the x-axis, indicating deciles of the observations for each study characteristics. In these plots, the PDP for SPSM are graphed on the left while the plots for DML are graphed on the right. We sort the plots by decreasing variable importance of each variable in the SPSM model.

To interpret the PDP of prospecting ratio, notice the rug: half of the RCTs targeted a majority of users by prospecting only (no remarketing). The APE of these experiments is lower than those of experiments that target most users through remarketing, for both SPSM and DML. We speculate that remarketing introduces additional endogeneity, which the observational methods have difficulty correcting. Next, we consider the conversion rate of the control group. The smaller that conversion rate, the better SPSM and DML perform. We believe the reason to be straightforward: If untreated users do not convert, no selection effect needs to be corrected by the observational methods. Third, it seems that a larger number of users in the test group is associated with a lower APE, except

[14]We use "permutation importance," as proposed by Breiman and Cutler. The process consists of establishing a baseline R^2 (for a continuous outcome variable) by putting the test samples down the random forest. Next, for each study characteristic, in turn, permute its values, put the permuted data down the random forest, and recompute R^2. Finally, let the importance of a study characteristic be the drop in R^2 caused by having permuted that study characteristic. The more R^2 drops after permuting, the more important the study characteristic. See `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`

Figure 13: Partial Dependence Plots (Random Forest explaining APE)

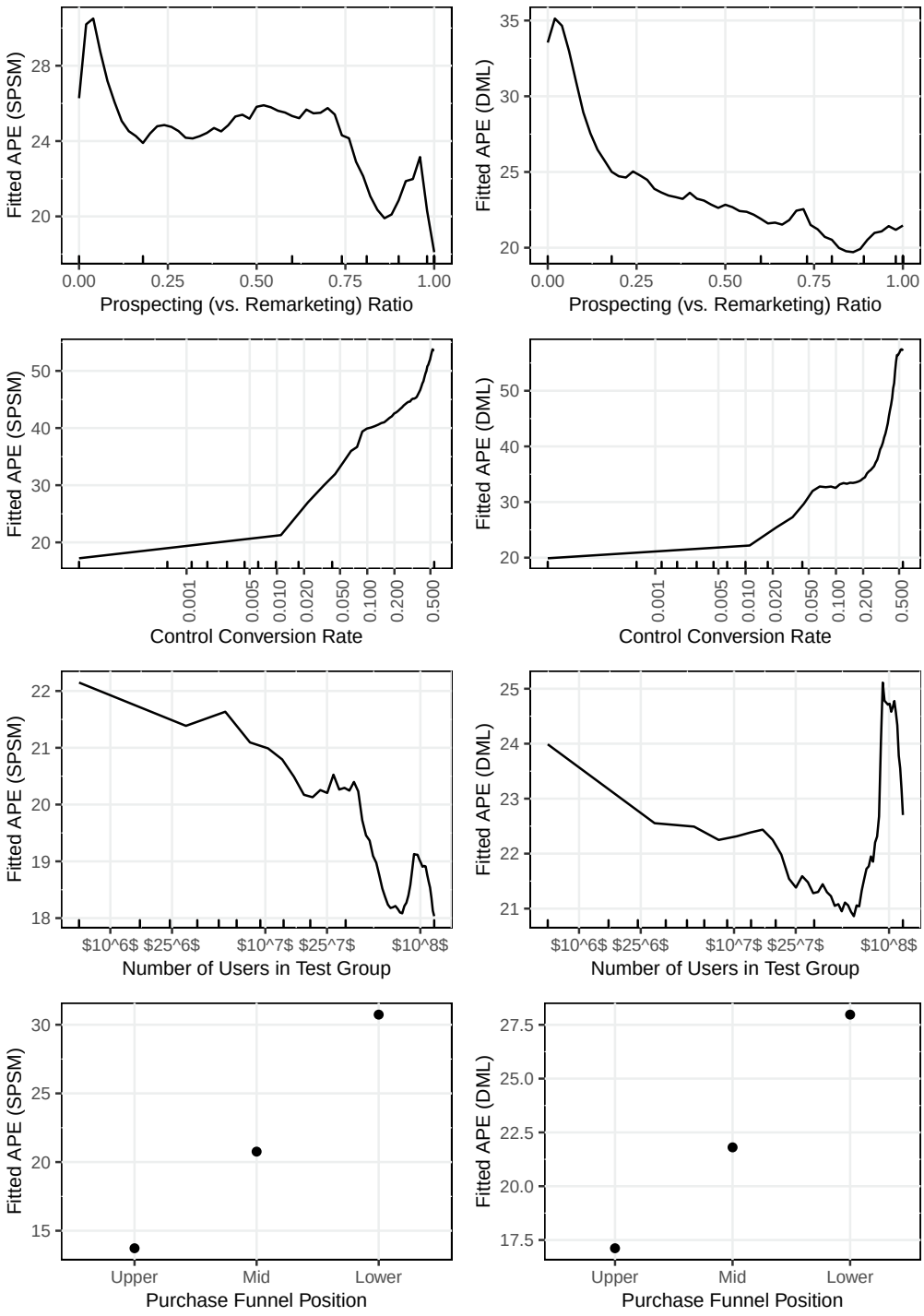
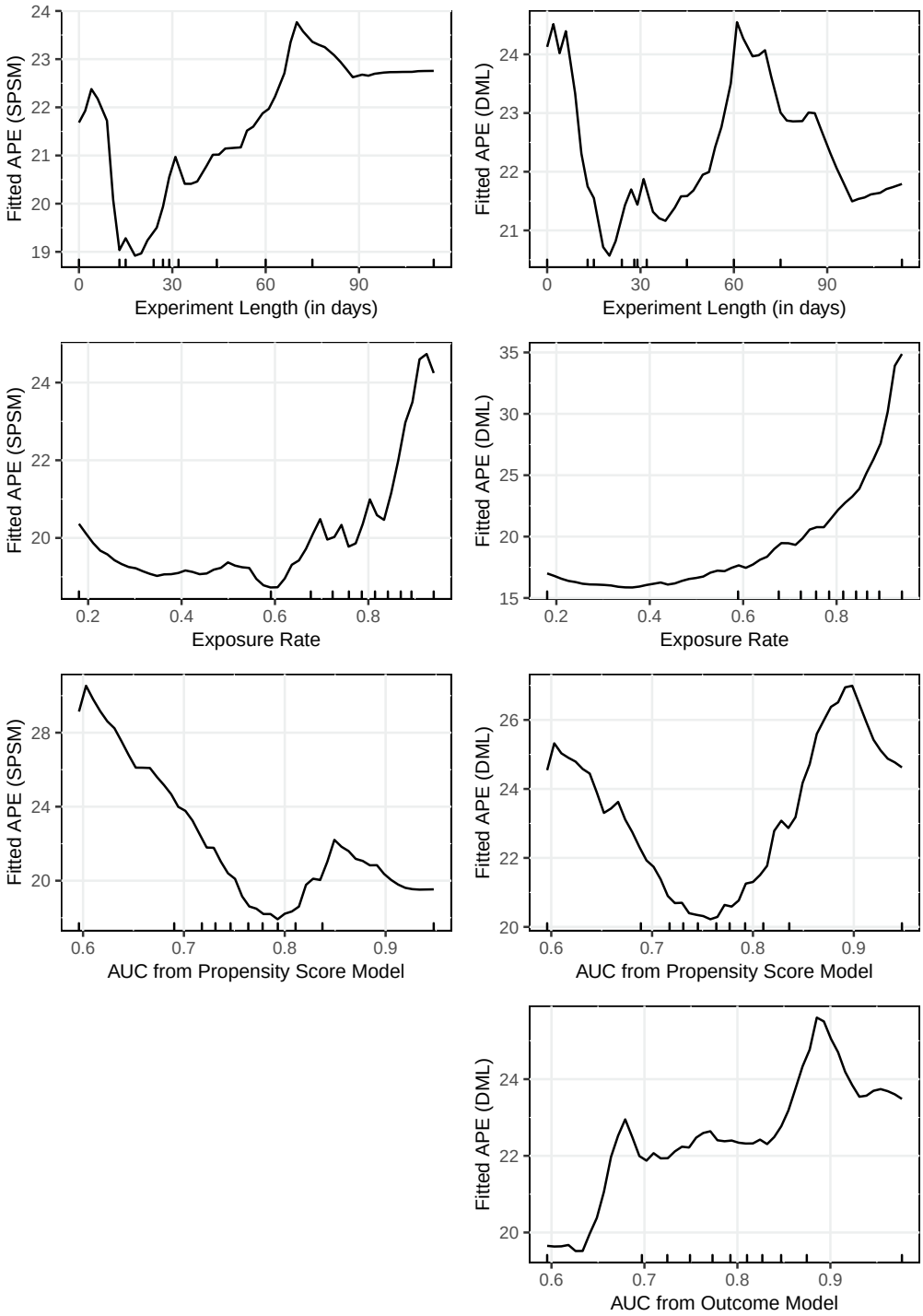


Figure 14: Partial Dependence Plots (Random Forest explaining APE)



for the very largest number of users for DML. Perhaps a larger sample helps the selection model. Fourth, we find that both observational methods perform better for upper-funnel than for mid- or lower funnel purchase outcomes. We think this is because there is more scope for selection for lower-funnel outcomes.

Moving to Figure 14, we do not find a clear relationship between study length and model performance. Next, we find that the highest exposure rates are associated with the worst APEs. We think this is because matching does best when there are many potential matches in the unexposed group to balance the sample. We also find that for SPSM a better-fitting propensity model (higher AUC) is associated with lower APEs. We do not have a prior about this relationship since a propensity score model should not perfectly predict. In fact, we do not see the same pattern for DML. Finally, surprisingly, we don't find a monotonic relationship between the the AUC of the outcome model and model performance.

Overall, SPSM and DML seem to work better in some settings. However, even in these settings, the average percentage error of both observational methods remains high.

7 Conclusion

This paper makes three contributions. First, we are the first to characterize the effectiveness of Facebook advertising using a large set of experimental studies that are representative of the large-scale RCTs advertisers run on Facebook in the United States. We showed that the median RCT lifts are 28%, 19%, and 6% for upper, mid, and lower funnel outcomes, respectively. We also found that 75.8%, 73.7%, and 59.6% of experiments are statistically different from zero for upper, mid, and lower funnel outcomes, respectively. The fraction of non-significant results can be partially explained by the fact that 25% of RCTs were not sufficiently powered to detect a lift of 10%. Moreover, we described ad effectiveness by industry vertical. These results could serve as useful prior distributions to aid firms' digital advertising decisions, complementing the results on TV ad effects in Shapiro et al. (2021).

Second, we add to the literature on whether observational methods using comprehensive individual-level data are "good enough" for ad measurement, or whether they prove inadequate to yield reliable estimates of advertising effects. Specifically, we used stratified propensity score matching and double/debiased machine learning to analyze 1,673 RCT at Facebook, each of which was described using over 5,000 user and experiment level features. We find that SPSM and DML both overestimate the RCT lift by a large amount. While the newer machine learning-based DML approach performs better than the traditional program evaluation SPSM approach, it still fails to reliably approximate the RCT estimates. The median absolute percentage point difference (AE) between RCT and DML lift estimates is 115%, 107%, and 62% for upper, mid, and lower funnel outcomes, respectively. These are very large measurement errors, given that the median RCT lifts are 28%,

19%, and 6% for the respective funnel outcomes.

Third, we characterize the circumstances under which SPSM and DML perform better or worse at recovering the causal effect of advertising. While both approaches seem to work better in some settings, even in those cases, the average percentage error of both observational methods remains high.

The data and models we have used surpass what individual advertisers are able to use for ad measurement and represent close to the peak of what large advertising platforms are currently able to leverage. However, even with this level of data and models, and when using observational approaches that have worked well in other domains, our results suggest that observational approaches generally are unable to measure the true causal effect of ads.

We conclude that observational approaches are unlikely to succeed unless advertising platforms pursue one of two paths. The first path is to fundamentally change what data advertising platforms log and how they implement observational methods. This path requires recording vastly more data than is currently done in practice—it would require logging the features that lead to ad rankings at the user-bid-request level. In addition, this path would require a selection model at the user-bid-request, as opposed to only at the user level. This unit of analysis is extremely granular and would require massive storage and computing power.

The second path frames the task of using non-experimental data to estimate an ad campaign's causal effect as a prediction problem. The premise of this approach is that an advertising platform has run a large set of RCTs and therefore has a collection of "ground truths" of advertising effects; in this approach the unit of observation is an RCT itself. This path would try to model the relationship between the RCT lift and a set of easily observed non-causal summary, or "proxy", metrics that correspond to non-experimental measures of campaign effectiveness. For example, the number of users who convert within a time window (e.g., 1 day, 7 days) after having clicked on the ad could be used to predict RCT results. Gordon et al. (2022) pursue this approach.

References

Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*, 685–725.

Athey, S., R. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Athey, S. and S. Wager (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1128–1242.

Blake, T., C. Nosko, and S. Tadelis (2015, January). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica 83*(1), 155–174.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*, C1–C68.

Dehejia, R. H. and S. Wahba (2002, February). Propensity score matching methods for non-experimental causal studies. *The Review of Economics and Statistics 84*(1), 151–161.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st ed.). Cambridge University Press.

Gordon, B., R. Moakler, and F. Zettelmeyer (2022). Predicting advertising incrementality using non-incremental metrics. *Working paper*.

Gordon, B., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science 38*(2), 193–225.

Greenwell, B. M. (2017). pdp: An r package for constructing partial dependence plots. *The R Journal 9*(1), 421–436.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017, 06–11 Aug). Deep IV: A flexible approach for counterfactual prediction. In D. Precup and Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning*, Volume 70 of *Proceedings of Machine Learning Research*, pp. 1414–1423. PMLR.

Imbens, G. and D. B. Rubin (2015, April). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (1st ed.). Cambridge University Press.

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

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Johnson, G. A. (2020). Inferno: A guide to field experiments in online display advertising. *Working paper, available at SSRN: http://ssrn.com/abstract= 3581396*.

Johnson, G. A., R. A. Lewis, and E. I. Nubbemeyer (2017). The online display ad effectiveness funnel & carryover: Lessons from 432 field experiments. Working paper, SSRN.

Kohavi, R., D. Tang, and Y. Xu (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (1st ed.). Cambridge University Press.

Lewis, R., J. Rao, and D. Reiley (2011). Here, there, and everywhere: Correlated online behaviors can lead to overestimtes of the effects of advertising. In *Proceedings of the 20th International Conference on World Wide Web*, pp. 157–66. Association for Computing Machines.

Lewis, R., J. Rao, and D. Reiley (2015). Measuring the effects of advertising: The digital frontier. In A. Goldfarb, S. Greenstein, and C. Tucker (Eds.), *Economic Analysis of the Digital Economy*. University of Chicago Press.

Naumov, M., D. Mudigere, H. M. Shi, J. Huang, N. Sundaraman, J. Park, X. Wang, U. Gupta, C. Wu, A. G. Azzolini, D. Dzhulgakov, A. Mallevich, I. Cherniavskii, Y. Lu, R. Krishnamoorthi, A. Yu, V. Kondratenko, S. Pereira, X. Chen, W. Chen, V. Rao, B. Jia, L. Xiong, and M. Smelyanskiy (2019). Deep learning recommendation model for personalization and recommendation systems. *CoRR abs/1906.00091*.

Rosenbaum, P. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica 70*, 41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics 6*, 34–58.

Shapiro, B., G. Hitsch, and A. Tuchman (2021). Tv advertising effectiveness and profitability: Generalizable results from 288 brands. *Econometrica 89*(4), 1855–1879.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science 25*(1), 1–21.

Syrgkanis, V., G. Lewis, M. Oprescu, M. Hei, K. Battocchi, E. Dillon, J. Pan, Y. Wu, P. Lo, H. Chen, T. Harinen, and J.-Y. Lee (2021, August). Causal inference and machine learning in practice with econml and causalml: Industrial use cases at microsoft, tripadvisor, uber. *KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 4072–4073.

Tunuguntla, S. (2021). Display ad measurement using observational data: A reinforcement learning approach. *Working paper*.