Machine Learning, Behavioral Targeting and Regression Discontinuity Designs

Sridhar Narayanan Kirthi Kalyanam

Stanford University Santa Clara University[*]

October 2021

Abstract

The availability of behavioral data on customers and advances in machine learning methods have enabled scoring and targeting of customers in a variety of domains, including pricing, advertising, recommendation and personal selling. Typically, such targeting involves first training a machine learning algorithm on a training dataset, using that algorithm to score current or potential customers, and when the score crosses a threshold, a treatment such as an offer, an advertisement or a recommendation is assigned. In this paper, we highlight regression discontinuity designs (RDD) as a low-cost alternative to obtaining causal estimates in settings where machine learning is used for behavioral targeting. Our investigation leads to several new insights. Under appropriate conditions, RDD recovers the local average treatment effect (LATE). Further, we show that RDD recovers the average treatment effect (ATE) when: (1) The score is orthogonal to the slope of the treatment and (2) When the selection threshold is equal to the mean value of the score. We also show that RDD can estimate the bounds on the ATE even if we are unable to get point estimates of the ATE. That RDD can estimate ATE or bounds on ATE is a novel perspective that has been understudied in the literature. We also distinguish between two types of scoring: Intercept versus slope based and highlight the practical value of RDD in each context. Finally, we apply RDD in an empirical context where a machine learning based score was used to select consumers for retargeted display advertising. We obtain LATE estimates of the impact of the retargeted advertising program on both online and offline purchases, and also estimate bounds on the ATE. Our LATE estimates and ATE bounds add to the understanding

[*] The authors can be reached at sridhar.narayanan@stanford.edu (Narayanan) and kkalyanam@gmail.com (Kalyanam). They would like to thank participants of the 2019 and 2021 Marketing Science Conferences, the 2020 AI/ML Conference and seminar participants at Washington University St. Louis and Santa Clara University for their comments.

of the effectiveness of retargeting programs in particular on offline purchases which has received less attention.

Keywords: Regression Discontinuity, Causal Effects, Treatment Effects, Machine Learning

1 Introduction

Behavioral targeting is one of the earliest and most popular marketing applications of machine learning. The recent explosion of highly granular first party consumer data, have allowed firms to utilize machine learning algorithms to score consumers and target them based on these scores (Hitsch and Misra 2018). For example, an advertiser might observe shopper browsing behavior and conversion rates on a website and then use a machine learning model to score conversion rates based on browsing behavior. Targeting rules can then be set up on these scores - for example, consumers above a certain threshold on the score might be targeted with promotions aimed at accelerating conversion. Such targeting policies have been discussed and employed in a variety of contexts including recommendation systems (Adomavicius and Tuzhilin 2005, Davidson et al. 2010, Gomez-Uribe and Hunt 2015, Smith and Linden 2017), online advertising (He et al. 2014), customized pricing (Dube and Misra 2020) and in the context of personal selling (Syam and Sharma 2018).

In this paper, we highlight regression discontinuity designs (RDD) as a low-cost alternative to obtaining causal estimates in settings where machine learning is used for behavioral targeting. We show that under appropriate conditions, RDD can provide local average treatment effects (LATE). However, machine learning contexts have an interesting feature in that the score that they generate can be based on the slope of the treatment (slope score), the purchase propensity (intercept score) or some combination of these . Incorporating the ideas of slope versus intercept scoring into RDD yields several new insights/results: (1) When the slope is orthogonal to the score, RDD yields the Average Treatment effect (ATE), (2) When the selection threshold, which is under the control of the firm, is equal to the mean value of the score, RDD yields the ATE, (3) Even when point estimates of the ATE are not feasible, RDD can generate bounds on the ATE.

RDD's ability to obtain local average treatment effects (LATE) has not only been a primary focus of the literature but has also been the source of criticism. For example Angrist and Pischke [2010] note the criticism of external validity and the related charge that "experimentalists are playing small ball while big questions go unanswered". Consequently there has been some interest in going beyond LATE, and obtaining average treatment effects (ATE). Angrist and Rokkanen [2015] obtain causal estimates away from the cutoff in a regression discontinuity design using dependent variable predictors other than the running variable or score. Conditional on these predictors the running variable is assumed to be ignorable. Eckles et al. [2020] propose a new approach to identification, estimation, and inference in regression discontinuity designs that exploits exogenous measurement error in the score variable. The approach proposed in this paper relies on the relationship between the score and the slope of the treatment in a machine context to either obtain the ATE or bounds on the ATE. Thus, we offer a novel perspective about going from LATE to ATE that is also very relevant to a variety of machine learning contexts.

The estimation of causal effects of marketing treatments is of central interest to marketers. The typical gold standard for obtaining causal effects is an experiment, where randomization of customers into treatment and control groups allows us to compare outcomes for those treated and those who are not while keeping everything else fixed between the two groups. This addresses concerns that treated customers are *systematically* differ from untreated customers due to reasons such as self-selection and endogeneity in treatment. This is of particular concern in the behavioral targeting context, where customers are targeted with treatment based on their past behavior. To the extent that past behavior is correlated with the outcomes of interest, such treatment policies naturally lead to self-selection. Gordon et al. [2019] show that estimates of advertising effects using even state of the art methods on observational data fails to obtain causal effects - in magnitude and often even in terms of the signs of the effects. Gordon et al. [2021] discuss these issues as well. A more general discussion of the challenges of inferring causality in machine learning contexts is provided by Judea Pearl in various books including "The Book of Why" (Pearl and Mackenzie 2018). An alternative perspective to Gordon et al. [2021] is provided by Eckles and Bakshy [2017] who argue that experiments might be flawed in some situations and that having very high dimensional data sets can produce estimates that are close to those obtained from experimentation.

A number of studies in the past have applied a variety of experimental designs to obtain causal effects of marketing interventions. One of the issues with experimentation is that it is expensive and slow. Therefore, firms often experiment in an episodic rather than continuous basis. This has led to calls for alternatives to experimentation (Eckles and Bakshy 2017), that allow for causal measurement, but at low cost and on a continuous, rather than episodic basis (Sharma et al. 2015, Gomez-Uribe and Hunt 2015). Researchers have used a variety of quasi-experimental approaches that exploit naturally occurring randomness in the data generating process to study the effects of marketing treatments including search advertising (Narayanan and Kalyanam 2015), television advertising (Liaukonyte et al. 2015, Hartmann and Klapper 2018) and promotional offers (Nair et al. 2011, 2017). These allow for causal estimation but without the costs and time associated with experimentation.

The context of behavioral targeting leads to particular concerns around the use of quasi-experimental approaches for causal estimation of treatment effects. Nair et al. [2011] study the contexts in which regression discontinuity designs can be used for finding causal estimates of local average treatment effects in contexts of behavioral targeting. The study looks at situations where targeting is based directly on measures of past behaviors or summaries of past behaviors, and makes the case that the researcher needs to carefully examine the validity of RDD in these contexts. We build on this study by specifically examining the validity and utility of RDD in contexts where targeting is based not on the variables summarizing past behavior directly, but where a large number of such variables are used through a machine learning framework to score customers. We document that this gives rise to a natural application of machine learning algorithms for causal measurement. Importantly, we document that under some conditions, RDD can be used to obtain not just local average treatment effects (LATE), but also average treatment effects (ATE), a novel finding for this literature. In some other contexts, it can be used to obtain bounds for the ATE even if point estimates are not feasible. Thus, we propose an approach with practical utility for obtaining treatment effects of interest in a variety of contexts.

We show that machine learning based targeting policies present natural opportunities for the use of regression discontinuity designs. This is because the large number of variables underlying machine learning algorithms usually ensure that the score is continuous. A continuous score, combined with a threshold rule for treatment, where consumers above a given threshold are treated and those below are not, meet the conditions for validity of regression discontinuity designs (Hahn et al. 2001, Lee and Lemeiux 2010, Imbens and Lemieux 2008, Nair et al. 2011). Thus, at the threshold at which there is treatment, we can obtain local average treatment effects using RDD. We additionally examine conditions under which we can use RDD to go beyond LATE. With heterogeneous treatment effects, we show that the degree to which the consumer-level score is correlated with these treatment effects matters. When the score and the treatment effects are uncorrelated, RD obtains ATE and not merely LATE. Further, under some conditions, when the threshold for treatment is chosen carefully, the firm can obtain ATE even when the score and treatment effects are correlated. Finally, we derive the bounds for ATE as a function of the LATE estimates obtained using the regression discontinuity design. With some prior knowledge regarding the variance of the distribution of treatment effects, or where there are naturally bounds on the variance, we can obtain the bounds for the ATE.

We discuss the application of these results in two contexts in which RDD can be used with behavioral targeting using machine learning based scoring systems. Machine learning algorithms have been used in two ways to score customers. In the first, which we term *intercept-based scoring*, customers are scored on their likelihood of having a positive outcome. For instance, if the outcome is purchase, consumers might be scored on the likelihood of making a purchase. If the outcome of interest is churn, then the likelihood of churning constitutes the score. By contrast, consumers could be scored on their *incrementality* from treatment - we term this *slope-based scoring.* The algorithm attempts to predict the incremental effects of the marketing treatment and bases the score on it. This is inherently harder to do, as it requires the firm to have a way to measure this incrementality at the individual customer level. We show that RDD provides a simple relatively low-cost way to assess the validity of the slope-based scoring algorithm. In intercept-based scoring contexts, we find that the results on ATE discussed earlier can be of use under some conditions. Thus, we discuss practical ways in which RDD can be employed in machine learning based targeting contexts.

We then apply our approach to an empirical setting involving the retargeting of display advertising. In this application, we partnered with a firm that used a machine learning score to select customers for retargeted advertising. This score was obtained using a machine learning algorithm that took as its input the consumers' browsing and transaction activity. Individuals whose scores exceeded a threshold were selected for the retargeting of display advertising. We obtain estimates of the causal effects of retargeted display advertising using our regression discontinuity approach. We have data on both online and offline purchases and hence are able to investigate the cross channel impact of retargeted display advertising. We also go beyond local average treatment effects to obtain bounds on the average treatment effect.

In the next section, we discuss various applications of behavioral targeting and machine learning. We then discuss regression discontinuity designs, specifically looking at conditions under which they can be used to obtain local average treatment effects, average treatment effects and bounds on the latter. We discuss applications of RDD to machine learning based targeting contexts, specifically discussing intercept-based and slope-based scoring policies. Next, we present our empirical application to retargeted display advertising. Finally, we conclude with implications of our work, and limitations.

2 Behavioral Targeting and Machine Learning

2.1 Behavioral Targeting

Behavioral targeting has a long history in marketing. Some of the early documented examples of behavioral targeting are from catalog marketing (Shepard 1990). Catalog marketers observe their customers' response to catalog mailings. They then score customers based on recency, frequency and monetary value of their response behaviors and select customers who exceed a threshold for follow-up mailings or other marketing actions. The installation of point of sale scanners in retail stores led to a dramatic increase in data on individual-level purchase behaviors collected in the grocery channel (Blattberg 1988). This led to an increase in the use of behavioral targeting in the grocery channel. For example, Catalina marketing, a commercial coupon marketing company issues coupons at the checkout of grocery stores based on observing a consumer's checkout ticket. Some implementations of the Catalina system were based on observing a single purchase of a shopper. Rossi et al. [1996] analyzed the value of purchase data for behavioral targeting in a setting that emulated the Catalina implementation. They examined the benefits of a customization strategy where the face value of a coupon is customized to an individual based on their purchase behavior.

2.2 Click Stream and Path to Purchase Data

The growth of the internet lead to a dramatic increase in the collection of behavioral data, to an even greater extent than the arrival of scanners in stores. One of the most common types of behavioral data available to firms is web site browsing data. Web servers create a time-stamped log of each page accessed by a visitor. These logs, combined with a visitor's session id or session cookie can tie together the different pages that were visited during a session. In the context of a retail web site, these web logs can provide information on whether a visitor added an item to a shopping cart, what items were added, how many pages were browsed on a session and the dwell time on specific pages. By combining these data, the visitor's path through the web site can be traced. This type of behavioral data is often referred to as click stream data and there is a growing literature on modeling these kinds of data (Montgomery et al. 2004, Bucklin et al. 2002). The focus of these studies is to use past behavior to predict future behavior. In addition to these onsite browsing behaviors, the inbound traffic on a web site also has a reference to the website-the referring URL-that was the source of the traffic. For example the referring URL would indicate if the visitor came from an ad campaign on a search engine, or a price comparison engine or by directly typing in the URL into the browser. Information on the inbound channel has also become useful in building path to purchase models (Kannan et al. 2016). This combination of on site browsing behavior and the referring channel creates a rich source of behavioral data for different types of targeting in many application areas in marketing.

The richness of the click stream and path to purchase data has created opportunities to predict future behavior based on past behavior using machine learning models that typically take as inputs a very large number of predictor variables, and aim to relate them to some outcome of interest. Various methods have been developed to improve the predictive accuracy (Breiman 2001) and the computational efficiency of these models (Sparapani et al. 2019). The methodological developments in this field are beyond the scope of this paper. But what is of interest is that these machine learning applications focus on predicting future behavior and do so by generating a score for each customer or visitor based on their past behavior. Customers whose score crosses a certain threshold are then targeted for marketing actions. The choice of the threshold can be driven by available budgets or by an economic criteria such as return on investment or break-even analysis.

2.3 Online and Offline Advertising

One of the most common application areas for behavior targeting is online or offline advertising. For example, in a practice called retargeting, an advertiser can target display advertising to visitors whose on site behaviors exceed a certain threshold in terms of the machine learning score. Advertisers can group customers into separate buckets and then score them. Customers whose scores are above a cutoff value receive advertising. For example Sahni et al. [2019] demonstrate an example of the retargeting of display advertising to two different groups of customers - product viewers and cart creators. In the context of search advertising, Google offers a feature called Retargeting List for Search Advertising (Google [2020]) that allows advertisers to score customers and retarget them for search advertising campaigns. Online browsing behaviors can also be used to target offline advertising.

In addition to advertisers targeting users, ad platforms such as Facebook can also target advertising to users using behavioral data. He et al. [2014] describe the use of machine learning to target ads based on predicted clicks at Facebook. On the Facebook platform, ads are not associated with a query but instead with user demographic information, interests *and* past behavior. Machine learning can be used to score the set of candidate ads and the ads that are potentially above a threshold or score the highest can be selected to be shown to the user. This is also a behavioral targeting context since the score is based on prior user behavior.

2.4 Recommendation Systems in eCommerce, Content and Entertainment Platforms

Machine learning models are also extensively used to build recommendation systems in electronic commerce contexts. These systems can be broadly classified into collaborative systems, content systems and hybrid systems (Adomavicius and Tuzhilin 2005). Collaborative systems provide recommendations based on items that users with similar tastes and preferences have liked in the past. Content based systems provide recommendations primarily based on the content or products that the user has shown a preference for in the past. Hybrid systems combine elements of both content based systems and collaborative systems.

The canonical example of collaborative recommendation systems is the system of recommendations on Amazon. These systems as they started out were loosely defined as "people are likely to buy items that others like them bought". In other words the recommendation systems generate a relatedness score based on past behavior of people with similar likes. While these recommendation systems started out as a simple relatedness count system, they now use machine learning to incorporate what recommendations are liked, clicked on, what items are compatible, substitutes versus complements, sequential purchases and the impact of time (Smith and Linden 2017). Reports suggest that a fairly large fraction of the clicks received by a recommended product comes from a recommendation link (Smith and Linden 2017). However obtaining causal estimates of *incremental* clicks due to recommendation systems is challenging since they require experimentation that can inconvenience users (Sharma et al. 2015). Natural experiments such as an those involving exogeneous shocks to traffic have been proposed as an alternative (Sharma et al. 2015).

Another popular and highly visible example of collaborative recommendations are systems on video sites such as YouTube and Netflix. Users navigate the vast amounts of information on content platforms such as YouTube using a combination of search, browsing and recommendations. Recommendation systems on sites such as YouTube have many objectives. Some users have specific interests and some users have broad interests. Some users want to be entertained. The scoring system on YouTube is a *top N* recommender (Davidson et al. 2010). A well known technique to build recommendation systems is to use association rule mining or co-visitation counts (Agrawal et al. 1993).

In addition to providing recommendations, the platform also observes many outcome metrics such as whether the user accepted recommendations and other measures such as whether the user increased the usage of the site. These outcomes metrics can be incorporated into a machine learning model along with the relatedness score to generate a more composite score that incorporates other factors such as past user behavior. Such a composite score could generate a top N set of recommendations, where N is determined by the space available on the display. The score could also be subject to a threshold value.

Collaborative recommendations are also extensively used in the online content streaming industry. Netflix, the pioneer in this category initially started predicting user ratings for video rentals but has now evolved to use an ensemble of prediction algorithms to help users select content. According to Netflix, the typical consumer loses interest after 60 to 90 seconds, reviewing 10-20 titles across one or two screens (Gomez-Uribe and Hunt 2015). So it is important to provide effective recommendations. Netflix has multiple recommendations categories such as 'Top Picks', 'Trending Now', 'Continue Watching', and 'Because You Watched'. Netflix also uses algorithms to construct pages and rows. These algorithms use different signals blending popularity and personalization to generate scores and rankings. Netflix tracks many outcome metrics including the extent of catalog utilization by the user and the take rate, the fraction of recommendations that resulted in a play.

2.5 Pricing and Personal Selling

Online behavior is also used to target price and promotional offers. For example consider the case of Ziprecruiter as discussed in Dube and Misra [2020]. The authors first run an experiment to understand the causal effects of pricing. Customers in this experiment sign up for a trial period and also provide background information. A high dimensional demand model is then trained on this data. In a second experiment, the proposed pricing scheme, i.e. customized offers, are implemented out of sample. In the language of machine learning, customers are scored regarding their price sensitivity using the high dimensional data that was obtained in the first experiment. This higher dimensional data includes usage behavior. This type of behaviorally targeted pricing has a lot of appeal to firms that offer content subscriptions or are in the software as a service (SAAS) sector. Firms in this sector practice a freemium model (Pujol 2010, Seufert 2013), where some basic features of a product are offered free with access to premium features for a fee. Behavioral targeting can be applied to customize and target pricing for upgrades from the free version in freemium models.

Personal selling efforts are perhaps the most costly part of the marketing mix. The optimal allocation of sales effort, territory design, balancing effort across the product life cycle and balancing the reliance on incentives has received considerable attention (Zoltners et al. 2008). Behavior-based lead scoring has also been an important practice in the allocation of sales effort. For example in the pharmaceutical industry, doctors are scored into deciles based on their past prescription behavior and sales effort is then assigned differentially for the deciles (Narayanan and Manchanda 2009). In the situation where leads are generated online, the behavior of the prospect or customer on the web site can be used to score the quality of the lead before passing them on for follow up by the sales force. Marketing automation platforms such as Salesforce.com and Marketo have started providing guidelines for scoring different behaviors of leads. For example some of the online behaviors include the landing pages visited, emails opened, data or spec sheets downloaded, product and service pages visited, the use of a part finder, the use of inventory lookup, visiting urgent delivery segments of the web site.[1]. The scores are assigned by the marketing function within the organization to generate a marketing qualified lead (MQL)[2] and by the sales function to generate a sales qualified lead (SQL)[3]. The leads can be scored on different criteria by assigning points to different behaviors or using an automated algorithm. Lead scores that cross a threshold are typically assigned for follow up by the sales team.

2.6 Summary and Evaluation of Behavioral Targeting

As the discussion above shows, behavioral targeting has been extensively applied across the entire marketing mix. A common theme across all of these applications is that there is a

[1] https://www.salesforce.com/products/marketing-cloud/best-practices/basic-science-behind-lead-scoring/

[2] https://blog.hubspot.com/marketing/definition-marketing-qualified-lead-mql-under-100-sr

[3] https://blog.marketo.com/2018/07/is-your-lead-sales-qualified-how-to-tell.html

treatment (a catalog mailing, a customized coupon, an offline or an online ad, a recommendation, a personalized price, a sales call). Individuals or treatments (a personalized price, a recommendation) are *scored* by a machine learning algorithm using prior behavior data. The individual or treatment is assigned based on the score crossing a threshold or a rank ordering based on the score. Outcomes are observed both for customers with and without treatment. For instance, in the case of targeting a specific ad for display advertising, the outcomes would be subsequent visits to the website of the advertiser or purchasing the advertiser's products, and these would be tracked for customers who were targeted with ads as well as those who were not.[4]

The common key evaluative question is the causal effect (incrementality) of behavioral targeting using machine learning. For example if the conversion rate (or other outcomes) for the customer who was selected using the machine learning score and was shown an ad is statistically difference from the customer who was not selected for the ad using the machine learning score, and the treatment was experimentally assigned using randomization, then the measured effects of the ad would be considered causal. However, in spite of the widespread use of machine learning in behavioral targeting, evaluation is an ongoing challenge. Evaluation approaches include prediction testing of scoring models on holdout samples. There is some awareness and recognition that the results from these models have to be carefully evaluated for incrementality. For example in the context of recommendation engines, the volume of clicks from the recommendation engine are often viewed as a measure of its success. Using a data set in the context of recommendations on Amazon, Sharma et al. [2015] point out that although recommendation clicks account for a large fraction of the traffic for recommended products, at least 75% of this activity would likely occur in the absence of recommendations (i.e. they are not incremental). They propose an instrumental variable strategy that relies on demand shocks to estimate the incrementality of behavioral targeting of recommendations. We note that finding credible instrumental variables is likely to be

[4] Advertising platforms such as Google and Facebook can track these "conversion" events on the retailer platforms using what is called a conversion pixel.

context specific and in general challenging (Angrist and Pischke 2010).

A/B testing and field experimentation have also been proposed in the context of behavioral targeting using machine learning. For example, researchers at Netflix report using A/B tests to evaluate recommendation algorithms (Gomez-Uribe and Hunt 2015). However a recurring theme is that field experiments are costly, that they may disadvantage some users, and that they inconvenience users (Sharma et al. 2015) and are a bottle neck for rapid innovation. In their evaluation of recommendation systems at Netflix, Gomez-Uribe and Hunt [2015] also state that behavior-based recommendation systems are subject to strong feedback loops and hence require careful validation. Narayanan and Kalyanam [2015] also note the existence of strong feedback loops but in the context of search advertising. To continue with the Netflix example, Gomez-Uribe and Hunt [2015] voice concerns that experimentation is a bottleneck for rapid innovation and explore offline experimentation (backcasting models and other forms of holdout prediction) as an alternative. However they conclude that

> “We need to have a better alternative to offline experimentation that allows us to iterate as quickly, but that is more predictive of A/B test outcomes.”

A similar perspective is echoed by Adomavicius and Tuzhilin [2005] in their assessment of recommendation systems. They note that

> “Understandably, it is expensive and time consuming to conduct controlled experiments with users ..therefore experiments that test recommendation quality on an unbiased random sample are rare”

These perspectives make the case for the development of methods that provide more frequent estimates, are causal but require less effort and costs than a field experiment. Our perspective is that RDD-based methods meet this criteria of causal, more frequent and less costly (in terms of time and money) estimates. But an important question is whether RDD can indeed be an alternative to A/B test outcomes or field experiments that typically yield an average treatment effect (ATE) estimate. We examine this important issue in this paper.

2.7 Types of Scoring - Intercept-scoring and Slope-scoring

Machine-learning based scoring algorithms can be broadly classified into two buckets - intercept-based scoring and slope-based scoring. In intercept-based scoring, as the name suggests, the firm is attempting to estimate heterogeneous intercepts across customers and target them based on this score crossing a threshold. Examples include finding the likelihood of churn for a subscription product in order to target consumers with proactive retention programs (Ascarza et al. 2018). Such types of scoring is quite common, as it relies on simpler algorithms that relate outcomes of interest to a set of observable variables for consumers.

Slope-based scoring, on the other hand, attempts to target customers based on an estimate of how they are likely to respond to the marketing intervention itself (Athey et al. 2017, Hitsch and Misra 2018, Wager and Athey 2018. These are harder to implement as they typically require estimation of heterogeneous treatment effects, using experiments to exogenously determine the treatment, and algorithms to obtain conditional average treatment effects (CATE). These CATEs in turn can be used for generating optimal targeting policies (Hitsch and Misra 2018). This is less commonly applied in practice because of the requirements it places on the data, the relative recency of the algorithmic approaches to this problem and the technical sophistication required on the part of firms to implement such policies. In this paper, we will discuss the estimation of causal effects using RDD in both of these situations, where consumers are scored on their intercepts and when they are scored on their slopes.

3 Regression Discontinuity and Behavioral Targeting with Machine Learning

3.1 Background on Regression Discontinuity

Regression discontinuity designs (RDD) can be employed to measure treatment effects when treatment is based on whether an underlying continuous forcing variable or score crosses a threshold. Under the condition that there is no other source of discontinuity, the treatment, which applies only to the observations with score above the threshold, induces a discontinuity in the outcome of interest at the threshold. Thus, the limiting values of the outcome on the two sides of the threshold are unequal and the difference between these two directional limits measures the treatment effect. A necessary condition for the validity of the RD design is that the forcing variable itself is continuous at the threshold (Hahn et al. 2001).

Formally, let i index the observation, let y_i denote the outcome of interest, x_i the treatment and z_i the forcing variable (henceforth referred to as score), with $\tilde{z}$ being the threshold above which there is treatment. Thus, treatment is defined by

$$x_i = 1 \iff z_i \geq \tilde{z} \tag{1}$$

$$x_i = 0 \iff z_i < \tilde{z} \tag{2}$$

Then the RD estimate of the treatment effect β is given by

$$\hat{\beta}_{RD} = \lim_{\lambda \to 0} \mathbb{E}\left[y_i | z_i = \tilde{z} + \lambda\right] - \lim_{\lambda \to 0} \mathbb{E}\left[y_i | z_i = \tilde{z} - \lambda\right], \lambda > 0 \tag{3}$$

Practical implementation involves finding these limiting values non-parametrically using a local regression, often simply a local linear regression (Fan and Gijbels 1996) within a prespecified bandwidth λ of the threshold $\tilde{z}$ and then assessing sensitivity to the bandwidth.

More details on estimating causal effects using RD designs, including the difference between sharp and fuzzy RD designs, the selection of non-parametric estimators for the limits on the two sides, the choice of bandwidth λ and the computation of standard errors can be found in Hahn et al. [2001], Imbens and Lemieux [2008] and Lee and Lemeiux [2010].

3.2 Behavioral Targeting and Regression Discontinuity

While the basic scoring and selection aspects of machine learning seem to fit the requirements of a RDD, the behavioral targeting raises particular concerns about validity of RDD. The fact that past behaviors are used to target customers with marketing interventions raises questions about self-selection by consumers who might undertake some actions with anticipation of future marketing interventions that might benefit them. For instance, consider a loyalty program that has benefits for consumers with score crossing a particular threshold. Consumers who are aware of the policy and their own scores might be induced to undertake actions that makes their score z cross the threshold $\tilde{z}$, and thereby receive the benefits. This would make RD invalid, because the customers who chose to undertake actions to cross the threshold would not be otherwise comparable with customers who did not, even at the limit. Nair et al. [2011] discuss the conditions under which RDD can obtain causal effects in such contexts, pointing out that uncertainty about the score or threshold might allow for measurement of local average treatment effects (LATE) using RDD.

In this paper, we consider contexts where the behavioral targeting policy uses an underlying machine learning algorithm to generate scores. The prototypical machine learning context involves a large number of variables - that is often the reason a machine learning algorithm is needed in the first place. Most of these variables are based on behavioral data of consumers - for instance, the history of browsing activity or purchases. These are combined into a score through a complex algorithm. The very complexity of the algorithm, the fact that it is based on a large number of behavioral variables and that the scores and thresholds for treatment are unobserved support the validity of RDD in these contexts. The continuous nature of the score, the fact that consumers are unable to anticipate how specific actions they take affect the score, and uncertainty of the score and threshold satisfy the conditions laid out in Nair et al. [2011] for validity of RDD. This makes RDD a useful candidate to measure LATE.

In marketing contexts these local effects are often interesting in and of themselves, since they measure the effect of treatment at a relevant margin - the threshold at which treatment takes place. The threshold is typically selected based on some underlying objective. The objective might be to target customers who have positive expected profits and measuring the treatment effect for the marginal customers who have expected profits at the zero threshold is often critically important for firms. However, randomized controlled experiments allow the marketer to obtain average treatment effects. These help evaluate the policy as a whole. For instance, average treatment effects help a marketer evaluate whether there are positive returns on average to advertising. RDD typically fails to address this need to obtain average treatment effects. In the next section, we examine whether and under what conditions RDD can go beyond LATE and obtain average treatment effects.

3.3 Beyond local average treatment effects

For the purpose of this analysis, consider the following data generating process.

$$y_i = \alpha_i + \beta_i \cdot x_i + \varepsilon_i \tag{4}$$

In this data generating process, the intercept α_i, slope or treatment effect β_i and idiosyncratic shock ε_i are allowed to be heterogeneous. This linear specification allows us to analyze the estimation of treatment effects using RDD in a tractable manner. Given that the RDD relies on finding limiting values of the outcome on the two sides of the threshold, the linear specification is a reasonable one, as any continuous and differentiable function can be locally approximated by a linear specification. The conditions of continuity and differentiability are typical requirements of the RDD itself, and therefore, these are not additional conditions we are imposing.

The treatment effect is β_i but since we do not observe the same unit of observation in both treated and untreated (i.e. control) conditions, the inferential aim is average treatment effects. RDD specifically measures the local average treatment effects in equation (3) where we substitute the expression for y_i in the data generating process in equation (4) to obtain

$$\hat{\beta}_{RD} = \lim_{\lambda \to 0} \mathbb{E}\left[\alpha_i + \beta_i \cdot x_i + \varepsilon_i | z_i = \tilde{z} + \lambda\right] - \lim_{\lambda \to 0} \mathbb{E}\left[\alpha_i + \beta_i \cdot x_i + \varepsilon_i | z_i = \tilde{z} - \lambda\right] \tag{5}$$

Separating terms in these expectations and substituting x_i as 0 or 1 based on the treatment conditions in equations (1) and (2), we get

$$\begin{aligned} \hat{\beta}_{RD} = &\lim_{\lambda \to 0} \mathbb{E}\left[\alpha_i | z_i = \tilde{z} + \lambda\right] - \lim_{\lambda \to 0} \mathbb{E}\left[\alpha_i | z_i = \tilde{z} - \lambda\right] + \\ &\lim_{\lambda \to 0} \mathbb{E}\left[\beta_i \cdot 1 | z_i = \tilde{z} + \lambda\right] - \lim_{\lambda \to 0} \mathbb{E}\left[\beta_i \cdot 0 | z_i = \tilde{z} - \lambda\right] + \\ &\lim_{\lambda \to 0} \mathbb{E}\left[\varepsilon_i | z_i = \tilde{z} + \lambda\right] - \lim_{\lambda \to 0} \mathbb{E}\left[\varepsilon_i | z_i = \tilde{z} - \lambda\right] \end{aligned} \tag{6}$$

In this equation, the continuity of the score z_i, the intercept α_i and the shock ε_i implies that the limits of the expectations on the two sides of the threshold are equal. Thus,

$$\lim_{\lambda \to 0} \mathbb{E}\left[\alpha_i | z_i = \tilde{z} + \lambda\right] = \lim_{\lambda \to 0} \mathbb{E}\left[\alpha_i | z_i = \tilde{z} - \lambda\right] \tag{7}$$

and

$$\lim_{\lambda \to 0} \mathbb{E}\left[\varepsilon_i | z_i = \tilde{z} + \lambda\right] - \lim_{\lambda \to 0} \mathbb{E}\left[\varepsilon_i | z_i = \tilde{z} - \lambda\right] \tag{8}$$

Thus, we obtain

$$\hat{\beta}_{RD} = \lim_{\lambda \to 0} \mathbb{E}\left[\beta_i | z_i = \tilde{z} + \lambda\right] \tag{9}$$

Although these are not new results to the literature, there are some important points to reiterate based on the above formalization. First, the regression discontinuity estimate is the local average treatment effect at the treatment threshold, i.e. it is the local average treatment effect for those observations for which the score is the threshold. Second, this estimate is not affected by how the intercept α_i and idiosyncratic shock ε_i are correlated with the score, as long as the condition of continuity of everything other than the treatment at the threshold is maintained. But the local average treatment effect does rely on the correlation between the heterogeneous treatment effect or slope β_i and the score z_i. We will explore this next.

3.3.1 Score is Orthogonal to Slope

Let's take the case when the score is orthogonal to the slope. Thus, $z_i \perp \beta_i$. Under this condition, the expectation in equation (9) becomes an unconditional expectation. And since the expectation is unconditional, the limit is the same at any value of z_i. Thus,

$$z_i \perp \beta_i \implies \hat{\beta}_{RD} = \mathbb{E}\left[\beta_i\right] \tag{10}$$

This expected value of β_i is the *average treatment effect (ATE)*. Thus, we can see that when the score is uncorrelated with the slope, we can obtain the ATE using a regression discontinuity design. While this result may seem obvious, we note that to the best of our knowledge, this result has not been highlighted or discussed in the literature.

3.3.2 Score and Slope are Correlated, and $\tilde{z} = \bar{z}$

We next explore what happens when the score and slope are correlated. To do this, consider the relationship between the score and the slope to be given by the following relationship

$$\beta_i = \gamma_0 + \gamma_1 z_i + \eta_i \tag{11}$$

Consider the situation where the score and the shock in this expression are uncorrelated, i.e. $z_i \perp \eta_i$. This may in general be difficult to justify, but in contexts where there is a large volume of data, typical to machine learning based targeting situations, this may not be an unreasonable assumption.

Then, substituting equation (11) into the LATE expression in equation (9), we get

$$\begin{aligned} \hat{\beta}_{RD} &= \lim_{\lambda \to 0} \mathbb{E}\left[\gamma_0 + \gamma_1 z_i + \eta_i | z_i = \tilde{z} + \lambda\right] \\ &= \lim_{\lambda \to 0} \mathbb{E}\left[\gamma_0 | z_i = \tilde{z} + \lambda\right] + \lim_{\lambda \to 0} \mathbb{E}\left[\gamma_1 z_i | z_i = \tilde{z} + \lambda\right] + \lim_{\lambda \to 0} \mathbb{E}\left[\eta_i | z_i = \tilde{z} + \lambda\right] \end{aligned} \tag{12}$$

With a zero mean, uncorrelated error η_i, the third term in the expression goes to 0. We therefore get

$$\begin{aligned} \hat{\beta}_{RD} &= \gamma_0 + \gamma_1 \lim_{\lambda \to 0} \mathbb{E}\left[z_i | z_i = \tilde{z} + \lambda\right] \\ &= \gamma_0 + \gamma_1 z \end{aligned} \tag{13}$$

The ATE is the unconditional expectation of β_i, which when substituting in the expression of β_i in equation (11) is

$$\mathbb{E}\left[\beta_i\right] = \mathbb{E}\left[\gamma_0 + \gamma_1 z_i + \eta_i\right] = \gamma_0 + \gamma_1 \bar{z} \tag{14}$$

where $\bar{z} = \mathbb{E}\left[z_i\right]$ is the mean value of the score. Therefore, the difference between the LATE and ATE is given by

$$\hat{\beta}_{RD} - \mathbb{E}\left[\beta_i\right] = \gamma_1 \left(\tilde{z} - \bar{z}\right) \tag{15}$$

Equation (15) provides us with another important result. When the threshold for treatment $\tilde{z}$ is the same as the mean of the score $\bar{z}$, the difference between the LATE and ATE is zero. In other words when $\tilde{z} = \bar{z}$, the RD estimate is equal to the ATE. Note that we are assuming that the slope and score are correlated as per Equation (11) and that this result does not require any additional assumptions. This is an important result because the firm has the threshold $\tilde{z}$ under its control in many marketing contexts. The firm chooses this threshold, and by setting it to be the mean score in even a subset of the observations or for a subset of time periods, it can find average treatment effects without having to experiment with the entire set of users.

Thus, RDD provides opportunities for continuous, and relatively low-cost estimation of *average treatment effects.* Firms are often concerned about the opportunity costs involved in setting aside a group of consumers in a control group who do not receive the treatment concerned and the costs involved in targeting a random set of users in the treatment group, even though some of them might be poor targets. Adjusting an already established threshold-based targeting policy might be easier to implement and be seen as less costly or risky. And such measurement using the RDD approach provides a continuous evaluation of the average treatment effects rather than the episodic evaluation feasible using an experimental approach. We now turn our attention to obtaining bounds for the ATE.

3.3.3 Bounds for the Average Treatment Effect

Given that γ_1 is the slope coefficient in the regression of z_i on β_i and substituting in the expression for γ_1 as the ratio of the covariance of β_i and z_i and the variance of z_i

$$\gamma_1 = \frac{cov\left(\beta_i, z_i\right)}{var(z_i)} \tag{16}$$

we get

$$\begin{aligned} \hat{\beta}_{RD} - \mathbb{E}\left[\beta_i\right] &= \frac{cov\left(\beta_i, z_i\right)}{var(z_i)}\left(\tilde{z} - \bar{z}\right) \\ &= correlation\left(\beta_i, z_i\right)\sqrt{\frac{var\left(\beta_i\right)}{var(z_i)}}\left(\tilde{z} - \bar{z}\right) \end{aligned} \tag{17}$$

where we have used the relationship between covariances, correlations and variances.

Now, let us consider the case when the correlation between the slope and score is 1. In this situation

$$\hat{\beta}_{RD} - \mathbb{E}\left[\beta_i\right] = \sqrt{\frac{var\left(\beta_i\right)}{var(z_i)}}\left(\tilde{z} - \bar{z}\right) \tag{18}$$

Similarly, when the correlation is -1, we get

$$\hat{\beta}_{RD} - \mathbb{E}\left[\beta_i\right] = -\sqrt{\frac{var\left(\beta_i\right)}{var(z_i)}}\left(\tilde{z} - \bar{z}\right) \tag{19}$$

For values of this correlation between 1 and -1, it is easy to see that this difference between the LATE and ATE also has to lie between these two extremes. Thus, we get the result

$$\mathbb{E}\left[\beta_i\right] \subset \left(\hat{\beta}_{RD} - \sqrt{\frac{var\left(\beta_i\right)}{var(z_i)}}|\tilde{z} - \bar{z}|, \hat{\beta}_{RD} + \sqrt{\frac{var\left(\beta_i\right)}{var(z_i)}}|\tilde{z} - \bar{z}|\right) \tag{20}$$

The absolute value for the difference $(\tilde{z} - \bar{z})$ comes from the fact that which of the two bounds is greater than the other depends on whether the threshold $\tilde{z}$ is greater or lesser than the mean score $\bar{z}$.

Note that in the expression above, the bounds for the average treatment effect $\mathbb{E}\left[\beta_i\right]$ depend on the threshold, the mean score and RD estimate, which are all known. The only unknown is the variance of the treatment effect. In some situations, the researcher might have prior knowledge about this variance, or might be able to bound it (for instance, when the dependent variable is bounded, like in the case of a binary dependent variable).

3.4 Discussion of the Utility of RDD in Intercept-based scoring

Now we discuss the utility of RDD approaches to finding treatment effects in cases where the scoring approach involves estimating the intercepts for customers, and setting thresholds on these scores to determine which customers to target. In other words, the score is an estimate of the intercept α_i in equation (4).

$$z_i = \hat{\alpha}_i \tag{21}$$

We have seen that RDD can be used to obtain local average treatment effects under conditions of continuity of the slope at the threshold. This, in and of itself, is interesting in a variety of contexts. An understanding of effects of marketing treatments at the margin at which such treatments are undertaken can provide interesting insights to the marketer, as seen in the examples in Nair et al. [2011]. Estimates of treatment effects at the margins can also in some instances provide marketers an assessment of how far away from optimal their targeting policies are. For instance, the objective of a retargeting policy might be to generate positive expected profits. If the marginal customer has expected profits far away from zero, it might provide the marketer a signal about the departure from optimality of such a policy. Experimenting with the targeting threshold, and continuous evaluation of the local average treatment effects at the various experimental thresholds might provide firms with assessments of their targeting policies while retaining the simplicity of intercept-based scoring.

We have also seen that we can use RDD to obtain average treatment effects - ATE - in the special case where the score and the slope are orthogonal. While this is a theoretical argument, it may still be of relevance in contexts where it can be argued that the intercept (which is what the score is trying to estimate) and the slope are uncorrelated.

More practical than this is the result we have found earlier for bounds on the estimate of the ATE as a function of the RDD-based estimate. Thus, with either prior knowledge of the variance of the treatment effect $var\left(\beta_i\right)$ or with bounds placed on this, we can estimate bounds on the ATE. Further, if the firm sets the threshold $\tilde{z} = \bar{z}$, then RDD provides the ATE.

3.5 Discussion of the Utility of RDD in Slope-based scoring

In slope-based scoring, the firm attempts to estimate the treatment effect of interest for each customer, and uses this estimate as a score in a targeting policy. For instance, an advertising targeting policy might attempt to score customers on estimates of their response to the advertising campaign, estimate the return on investment for each customer, and then target customers with positive returns. Thus, the score is

$$z_i = \hat{\beta}_i \tag{22}$$

Substituting this in the expression for the RDD estimate in equation (9), we get

$$\hat{\beta}_{RD} = \lim_{\lambda \to 0} \mathbb{E}\left[\beta_i | \hat{\beta}_i = \tilde{z} + \lambda\right] \tag{23}$$

If $\hat{\beta}_i$ is a consistent estimate of the treatment effect β_i, it is true (at least aymptotically), that

$$\hat{\beta}_{RD} = \tilde{z} \tag{24}$$

This is an important result because under the conditions we have assumed (consistency of the estimator, continuity of the score), we see that the regression discontinuity estimate of the local average treatment effect has to equal the threshold for treatment itself. If it does not, the assumptions underlying this result have to be untrue. With continuity of $\hat{\beta}_i$ typically not in question in machine learning applications, it would imply that the estimate is not consistent.

Thus, in slope-based scoring contexts, RDD provides a way to continuously evaluate the validity of the underlying machine learning algorithm itself. If the underlying algorithm is correctly estimating the treatment effects, the RDD-based LATE estimate must be equal to the threshold itself. If not, it would call for re-evaluation of the underlying algorithm or suggest that the algorithm is not consistent with slope based scoring. This is an important result as it would allow an analyst to discern the nature of the scoring rule and validate institutional perspectives in this regard.

This provides an alternative to the experimental evaluation of the algorithm, which can be costly and time-consuming and episodic. By contrast, with RDD we can obtain an evaluation that is continuous. This is of particular relevance since even an experimental validated algorithm for estimating slope might need reevaluations as time passes, market conditions change, technologies change and new competitors enter the market. Continuous, low-cost evaluation of the algorithm can be of potentially great value to marketers employing slope-based scoring for behavioral targeting.

3.5.1 Summary of Key Theoretical Results

A summary of our key theoretical results (also summarized in Table 1) are as follows:

1. In intercept based scoring, when $z_i \perp \beta_i \implies \hat{\beta}_{RD} = \mathbb{E}[\beta_i]$ which is the ATE.

2. In Intercept based scoring, assuming a heterogeneous data generation process as given by Equation (15), then $\tilde{z} = \bar{z}, \implies \hat{\beta}_{RD} = \mathbb{E}[\beta_i]$, which is the ATE.

3. In intercept based scoring, the bounds for the ATE $\mathbb{E}[\beta_i]$ are given by

$$\left(\hat{\beta}_{RD} - \sqrt{\frac{var\left(\beta_i\right)}{var(z_i)}}|\tilde{z} - \bar{z}|, \hat{\beta}_{RD} + \sqrt{\frac{var\left(\beta_i\right)}{var(z_i)}}|\tilde{z} - \bar{z}|\right) \tag{25}$$

   .

   In all of the above three cases since the decision maker does not have a slope based score, the ATE or the bounds on the ATE obtained via RDD should be useful to the decision maker.

4. In slope based scoring $\hat{\beta}_{RD} = \mathbb{E}[\beta_i|\tilde{z}]$, which is the LATE. The availability of the LATE provides continuous causal evidence of the validity of the slope based scoring policy.

5. In slope based scoring $\hat{\beta}_{RD} = \tilde{z}$, which is the threshold. This provides a consistency check on the slope based scoring rule.

4 Empirical Application: The Targeting of Retargeted Display Advertising

4.1 Retargeted Display Advertising

In this subsection, we describe the context of our empirical application. We partnered with a major advertiser conducting a retargeted advertising campaign. The advertiser is a major provider of cellphone services, and also sells equipment to consumers, including cellphones, accessories, aside from cellular voice, text and data plans. Consumers who visit any product page, including for cellphone plans and equipment, but depart the website without completing a purchase are eligible for a retargeted display advertising campaign.

Retargeted display advertising campaigns reach customers who have, through their product page visits, shown that they are interested in the product. Targeting these customers with display advertising after they leave the advertiser's website without making a purchase, might persuade them to return to the advertiser and complete their purchase either online or offline. Retargeted display advertising often uses dynamic and personalized creatives, the effects of which have been studied by Lambrecht and Tucker [2013] and Bleier and Eisenbeiss [2015].

Retargeted display advertising is a controversial practice. The dynamic creatives are quite visible to consumers and policy makers who may react negatively to such overt tracking. There might be multiple mechanisms behind the response to retargeted advertising. Sahni et al. [2019] show that such the retargeting effect might be driven by a reminder mechanism and/or a competitive blocking mechanism (Burke and Srull 1988), whereby retargeting reduces the likelihood that the consumer subsequently visits a competing seller's website and is targeted in turn by their retargeting advertisements.

Industry reports tout the benefits of retargeting and advertisers report spending as much as 10% of their marketing budgets on retargeting (add cite to Adroll 2014. also cited in Sahni et al). However a fundamental issue with evaluating retargeting campaigns is that they involve customers who are highly self-selected, as evidenced by their prior interest in the advertiser's products. Thus, correlations between advertising and subsequent actions might reflect this selection, rather than the effect of the display retargeting campaign. To causally evaluate the effects of retargeting, experimental approaches have been proposed in the literature (Sahni et al. 2019, Johnson et al. 2017b). An issue with the use of these approaches is that they are potentially costly because they involve significant manpower to execute. If retargeting does work then there is the opportunity cost of lost sales for the control group. The field studies reported by Johnson et al. [2017b] and Sahni [2015] used control groups that were as high as 50% of the experimental observations.

Retargeting experiments can be time-consuming. Sahni et al. [2019] report that the retargeting field experiment executed by them required about 1 year of data collection. They were able to obtain causal estimates of the impact of retargeting on revisits but not on conversions. Johnson et al. [2017b] report causal estimates of the conversion effects of retargeting. *Low-cost* alternatives for *continuous* evaluation of the effects of retargeting would be a very useful alternate methodology to firms. The practical importance of this advertising practice suggests that additional causal estimates of the impact of retargeted display advertising on conversions, such as those reported in this study would benefit our overall understanding of this practice. Further the offline impact of retargeted display advertising has not been investigated and our study sheds light on offline effects.

4.2 Machine Learning Scoring and Selection

An additional issue with retargeting campaigns is that while they select users who have shown interest in the advertiser's products, they also potentially target consumers who are not interested in the advertiser's product, as evidenced by their departure from the advertiser's website without completing a purchase. Targeting such consumers with advertising might not just be wasteful, it might even lead to negative effects if the advertising campaigns irritate the consumer. They also come at the cost of advertising to other, potentially more promising consumers or using the money spent on those consumers in more positive ways. To deal with this issue, the focal advertiser in our empirical application partnered with a third-party AI-enabled marketing and advertising services firm and used a machine learning algorithm applied on consumer browsing and (when available) transaction data to select consumers for a retargeting campaign. While others (Sahni et al. 2019, Johnson et al. 2017b) have studied the retargeting of display advertising, the use of machine learning for audience selection is novel to this application. If retargeting involves consumers who are highly self-selected, as evidenced by their prior interest in the advertiser's products, then the addition of a machine learning score and selection policy adds another layer of selection further highlighting the importance of causal estimates.

Consumers were selected using a machine learning methodology that scored them on purchase propensity. Historical data were used to train a proprietary algorithm (details of which we are not privy to) to estimate the likelihood that a customer would be a purchaser. The data for this algorithm included the pages visited by the customer, their dwell times on various parts of the website and on specific pages, their signed-in status, and when available, their past transaction history. These data were then supplied as inputs to the algorithm to generate a score, which is correlated to their likelihood of making a prior purchase. The firm then arrived at a threshold score, based on its business objectives, above which consumers were eligible for retargeting. Consumers with score below this threshold were not eligible for the campaign, and no further action was taken in the case of such consumers. Outcomes including purchases (both online and offline) were tracked for all customers regardless of their score and whether they were targeted for the advertising campaign or not. The advertising campaign lasted for up to two weeks after being triggered, and outcomes for all consumers were tracked for a period of 6 weeks subsequently.

4.3 Data Description

We obtained data for a total of about 1.4 million customers who visited the advertiser's website during the period for which we have data, visited one or more product pages but left the website without completing a purchase. All of these customers were assigned scores through the proprietary machine learning algorithm described in the previous sub-section. Of these consumers, about 1.1 million had scores below the threshold of 0 and were ineligible for the retargeting campaign. There were a little over 0.3 million customers with scores above the threshold and thereby eligible for the retargeting campaign.

Every customer was tracked both before and after the experiment using tracking cookies placed by the advertising network on consumers' devices. Since this network has consumers signed in on to one of their applications or on their web browsers across devices, this tracking took place across their devices. Thus, a consumer might trigger an advertising campaign on their laptop computer, be served ads on the computer as well as their mobile devices, and complete a purchase on a different desktop computer. As long as they are signed in on their different devices at some point of time, their cookies across these devices are linked and the firm is able to link the triggering event, advertising campaign status and outcomes. Both past and prior transactions were tracked using this approach. Also, the firm tracked offline purchases (i.e. if the customer walked into a brick and mortar store of the advertiser) when feasible for all customers who had been in signed in state at the time the advertising campaign was triggered.

For the purpose of this paper, the pertinent data are the machine learning scores and their purchases before and after the start of the advertising campaign for them. The key outcome variables we study is whether they make any purchases (conversion), and separately in the online and offline channels separately. We also examined the total number of items purchased overall and in the two channels separately, but the results were very similar to those for the conversion variable, since most consumers who converted only purchased one item. Hence, we do not separately report estimates for this set of outcome variables.

Table (2) reports the summary statistics for these variables for all customers in the dataset. A few observations about the summary statistics. Conversion, which refers to one or more instances of purchase after the triggering event of the retargeting camapaign (a product page view but without completion of a purchase) is a relatively rare event, with about 0.4% of consumers converting. This is not atypical of retargeting campaigns in general. About 3/4ths of these conversions take place online, although when one looks at past conversions, it is about equally distributed between online and offline purchases. Given the category - mobile devices, accessories and cellphone plans - this is again not surprising, given that the shift from offline to online retail is more recent in the cellphone context than in many other categories such as books and personal electronics. Another possibility is that retargeting with digital ads shifts purchases online. Another observation is that the score has a mean of about 0.15. Since the score is normalized by the standard deviation, this implies that the threshold, which is at 0, is about 0.15 of a standard deviation below the mean. This is of relevance given that the bounds estimator for the ATE relies on the difference between the mean score and threshold.

The fact that the threshold, which is at 0, is about 0.15 of a standard deviation below the mean has another important implication. Recall our previous result from equation (15), that if $\tilde{z} = \bar{z}$, then the RDD estimator recovers the ATE. In this application since $\tilde{z} \neq \bar{z}$, we can be sure that the RDD estimator does not recover the ATE, a useful clarification that can be discerned by simply examining the summary statistics of the score and the value of the threshold.

Table (3) reports the same summary statistics for consumers whose score was below the zero threshold. These consumers were not eligible for the retargeting campaign after the triggering event. They had lower conversion than the average consumer in the sample and lower levels of conversion in the past. The same applied to number of purchase events as well. Table (4) reports the summary statistics for consumers who had scores above 0, and hence were eligible to be included in the retargeting campaign. Reading tables (3) and

(4) together, it is clear that higher scores are associated with, on average, higher levels of past purchase and conversion, with the consumers with scores greater than zero having past conversion incidences at about 5 times the rate of those with scores below zero, and the number of purchases at almost 4 times. The machine learning score and the corresponding threshold are indeed selecting consumers who have a high prior propensity to buy. Since prior purchase propensity is likely to be correlated with future purchase propensity, a case can be made that the individuals selected for advertising have a higher propensity to buy. This selection underscores the importance of a causal estimate.

4.4 Empirical Strategy

In this subsection, we use our RDD-based approach to measure treatment effects of the retargeting campaign. We use local linear regressions to find the limiting values of the outcomes on the two sides of the threshold, choosing bandwidths for the RDD manually, while assessing sensitivity to the bandwidth choice. Practically, this is achieved using the following regression conducted on the subset of the data with scores lying within the bandwidth intervals around the treatment threshold of 0.

$$y_i = \theta_0 + \theta_1 z_i + \theta_2 x_i + \theta_3 z_i x_i + \nu_i \tag{26}$$

where y_i is the outcome of interest, z_i is the score, x_i is the binary treatment variable and ν_i is a random error. It is standard practice in the RDD literature to incorporate z_i and the interaction of z_i and x_i as regressors. The parameters θ_1 and θ_3 reflect the different slopes of the outcomes with respect to the score on the two sides of the threshold. The treatment effect we aim to measure is given by θ_2.

Our empirical strategy is to measure the treatment effect for three outcomes of interest - overall conversion (a binary variable indicating whether the customer made one or more purchases), and conversion in the online and offline channels. In these three instances, we compare the treatment group (with score above the threshold) with observations that had scores below threshold and which were not eligible for the retargeting campaign.

In figure (1) we plot conversion rate from the current campaign against the score. Since conversion is a binary variable, it would not be meaningful or informative to plot it against score directly. Instead, we divide the data into equally sized intervals based on the score. We then compute the average conversion for all observations within the interval, and plot this mean conversion rate against the score. As can be seen from the figure, the mean conversion rate is close to zero at all points below the score of zero. There is a discontinuous jump in the mean conversion rate for all the score intervals above the threshold score of zero. This discontinuity is what we rely on to find the RDD estimates of the effect of retargeting.

We similarly plot the online and offline conversion rates in figures (2) and (3) respectively. We see similar disontinuities there as well, though in the case of offline conversion, the discontinuity is less sharp, in that the confidence intervals of the nonparametric fit lines on the two sides overlap to a greater degree than for the other variables.

4.5 Results: Local Average Treatment Effects

In this subsection, we report the results of our empirical analysis. Rather than reporting the full set of regression results for each outcome variable, we report the main parameter of interest, which is the treatment effect θ_2 in equation (26).

Table (5) reports the results of this analysis. The main findings is that the LATE shows that retargeted advertising increases total conversions significantly. The estimated coefficient is 0.034 and by a large magnitude (0.034 incremental conversions relative to a baseline of 0.001 from table (3)). The increase in total conversions are largely driven by significant increases in online conversion; there is no significant increase in offline conversions as a result of the retargeting advertising campaign. Thus, we have demonstrated that RDD can be used to estimate local average treatment effects of retargeting and that these are statistically and economically significant, particularly in terms of the effect on online conversions and purchases.

The RD estimates in Table (5) provide another important insight. Recall from Equation (24) that if the firm's targeting is based on slope based scoring, then $\tilde{\beta}_{RD} = \tilde{z}$. In other words the RDD estimate is equal to the threshold. In our application the threshold is equal to zero and since the RDD estimate is statistically different from zero (and estimated very precisely), it provides evidence that the firm is not engaged in slope based scoring. This result combined with correlation of the score with prior purchases in figure 4 increases our confidence that the firm is most likely engaged in some form of purchase propensity or intercept based scoring. We note that RDD allows the analyst to reach this conclusion and this ability to diagnose the nature of the scoring can be very useful when institutional knowledge is not available or weak.

4.6 Results: Bounds on Average Treatment Effects

Next, we examine if we would be able to obtain upper and lower bounds for the ATE. Recall from equation (20) that we can obtain these bounds if we knew the variance of the treatment effects, i.e. $var\left(\beta_i\right)$. This is not known to us, but we can see that in the case of the conversion variables, the treatment effects can be reasonably bounded between 0 and 1, since the outcome itself is binary. Assume that the treatment effects are uniformly distributed between 0 and 1. When this happens, it is easy to see that $var\left(\beta_i\right) = 0.083$, which is obtained from the expression for the variance of a uniform distribution.

With this level of the variance of the treatment effect, we obtain the lower and upper bounds of the ATE and report it in table (7). Also reported are the lower and upper bounds when we have smaller variances in the treatment effects given that the uniform distribution assumption is perhaps conservative. We arbitrarily choose the values for the variance at 0.0005, 0.01, 0.03 and 0.06 and assess what the bounds of the ATE are. Since the expression for the bounds in equation (20) is linear, the standard errors for the bounds estimates of ATE are the same as those for the LATE. We find that lower bound of ATE in the case of the conservative uniform distribution assumption for treatment effects is not statistically significant but the upper bound is (and always will be if the LATE is statistically significant). When the variance of the treatment effects is lower at 0.005, the lower bound is positive but statistically insignificant. For higher variances, the lower bound is typically not statistically significant in this application. However, in other applications, especially where the treatment effects are stronger, one might see significant results for both bounds. Overall the upper bounds in table 7 suggest that the ATE is economically meaningful. The upper bounds of 11% and 14.2% are similar to the ATE estimates reported by Johnson et al. [2017a].

We replicate this analysis for online conversion and report the results in table (8). We find that in this case, both the lower and upper bounds are significant and positive for a variance of the treatment effect set at 0.005. For higher variances, the lower bound is not statistically significant.

Thus, we can see that we can obtain bounds on the ATE using the RDD design given an estimate or guess of the variance of the treatment effect. This is a potentially powerful new result in the context of RDD, which has so far used only for estimation of local average treatment effects. We also note that the bounds suggest that the ATE can potentially be zero, but this might simply indicate the degree of response or slope heterogeneity and hence an opportunity for slope based targeting. Such bounds can be very useful to decision makers who are interested in the ROI of advertising campaigns. Further given that firms conduct many different advertising campaigns concurrently which might make it impractical to experimentally evaluate all campaigns, obtaining bounds in this manner for certain campaigns can be helpful.

4.7 Robustness checks

In this sub-section, we assess the power of the RD estimator, particularly given the null effects in some cases, and also assess the sensitivity of RD estimates to the choice of bandwidth, or the window around the treatment threshold within which observations are used for the estimation procedure. Using a significance level of 95% and power of 80%, we compute the minimum sample size needed to detect the effects we estimate for overall conversion, online conversion and offline conversion. We find that the minimal sample size are 556 observations (across treatment and control) for detecting the lift in conversion we estimate, 549 observations for online conversion, and 5866 observations for offline conversions. Given that our sample was 4886 observations within the bandwidth, we have adequate power in the sample for detecting both overall conversion and online conversion. The data are a little under=powered for detecting offline conversion, but is adequate when the power of the test is set at 70% instead of 80% (the minimum number of observations in this case is 4612, lower than the 4886 we have).

There are some placebo tests we conduct to assess robustness of our results. One set of placebo tests is to find RD esimates at the same threshold for a set of variables where we do not expect to find any treatment effect. One such variable is past conversions. We should not expect any effect of treatment on past conversions. Figures (4) through (6) plot the prior conversion rates against the score. We see that the confidence intervals for the non-parametric fit curves on the two sides of the threshold overlap greatly, to the extent that the confidence interval on the left of the threshold is almost entirely subsumed within the confidence interval on the right. This is true for all three outcome variables - overall conversion, online conversion and offline conversion.

Another placebo test we conduct is to compute the RD estimates at other randomly determined thresholds. We find that the treatment effects are insignificant at 92 out of 100 randomly determined thresholds on either side of the actual threshold. This can be seen visually in figures (1) through (3) as well, where there are no other obvious points of discontinuity like the one at the actual threshold.

Finally, we assess sensitivity of our estimates to the choice of bandwidth and report these for conversion as an outcome in table (9). As the bandwidth increases, the potential bias in the RD estimate increases due to the reliance on observations farther away from the threshold. Recall that the RDD-based estimate is based on the limiting values of the outcomes on the two sides of the threshold, and hence inclusion of more observations farther from that threshold can increase the bias of the estimates, especially with our use of local linear regressions on the two sides of the threshold to estimate the limiting values. On the other hand, reducing the bandwidth reduces our degrees of freedom and leads to the estimates being insignificant. Thus, our reported estimates are at a bandwidth of 0.02.

5 Conclusion

In this paper, we examine the use of regression discontinuity designs (RDD) for the estimation of treatment effects in contexts where the marketing treatment is based on scoring customers based on their past behaviors using a machine learning algorithm and assigning treatment to customers above a predetermined threshold. Such approaches to targeting are commonly founds across a variety of domains from advertising to recommendation systems to customized pricing. We draw attention to RDD as a natural application to finding local average treatment effects in these contexts.

We further show that under some conditions, we can use RDD to go beyond local average treatment effects to finding average treatment effects or bounds on the ATE when point estimates are not feasible. We apply these insights to two types of machine learning based targeting systems - involving intercept-based scores and slope-based scores. For intercept based scores, the RDD estimates provide a low cost and timely approach to obtain causal estimates. In the case of slope-based scoring systems, RDD can provide *continuous causal validation* of the underlying machine learning algorithm itself.

We present an empirical application in the context of the targeting of retargeted advertising. Machine learning scores were constructed from customers' past browsing and transaction history. In this application, we find that there are significant effects of retargeted advertising on conversions in the online channel, but not in the offline channel. We also find the bounds on the local average treatment effects and find statistically significant estimates of both the lower and upper bounds under certain assumptions on the variance of the treatment effects.

Finally we discuss some limitations of our proposed approach. We need some assumptions to be true for our approach to work. There should be no other source of discontinuity at the treatment threshold other than the treatment itself. This rules out relatively simple algorithms that violate the necessary continuity condition for the score. Our results for finding bounds on the average treatment effects require the score to be uncorrelated with other factors that affect the treatment effect. While this is a reasonable and commonly employed assumption for many of the contexts we propose our method for, it is not a testable assumption and may not be reasonable in some contexts. We do not make the case that our approach is a substitute for experimentation. It should be viewed as a complement to other approaches, particularly experiments, for obtaining estimates of the causal effects of the marketing treatment, and these can be used to periodically validate the continuous measurement that our approach facilitates.

Overall, we suggest that this is a useful approach to conduct continuous and low cost evaluation of marketing treatments in a variety of contexts with behavioral targeting based on underlying machine learning algorithms.

References

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

J. D. Angrist and J.-S. Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30, 2010.

J. D. Angrist and M. Rokkanen. Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344, 2015.

E. Ascarza, S. A. Neslin, O. Netzer, Z. Anderson, P. S. Fader, S. Gupta, B. G. S. Hardie, A. Lemmens, B. Libai, D. Neal, F. Provost, and R. Schrift. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*, 5(1-2):65–81, 2018.

S. Athey, G. Imbens, T. Pham, and S. Wager. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–281, 2017.

R. C. Blattberg. Assessing and capturing the soft benefits of scanning," a study conducted for the coca-cola retailing research council. *May*, 3:1–43, 1988.

A. Bleier and M. Eisenbeiss. Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, 34(5):669–688, 2015.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

R. E. Bucklin, J. M. Lattin, A. Ansari, S. Gupta, D. Bell, E. Coupey, J. D. Little, C. Mela, A. Montgomery, and J. Steckel. Choice and the internet: From clickstream to research stream. *Marketing Letters*, 13(3):245–258, 2002.

R. R. Burke and T. K. Srull. Competitive interference and consumer memory for advertising. *Journal of consumer research*, 15(1):55–68, 1988.

J. Davidson, B. Liebald, J. Liu, P. Nandy, T. V. Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingstone, and D. Sampath. The youtube recommendation system. *10: Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

J.-P. Dube and S. Misra. Personalized pricing and customer welfare. Working Paper, University of Chicago., 2020.

D. Eckles and E. Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. Working Paper, MIT Sloan School., 2017.

D. Eckles, N. Ignatiadis, S. Wager, and H. Wu. Noise-induced randomization in regression discontinuity designs. *arXiv preprint arXiv:2004.09458*, 2020.

J. Fan and I. Gijbels. *Local Polynomial Modeling and its Applications*. Chapman & Hall, London, 1996.

C. A. Gomez-Uribe and N. Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, 6(4): 13:1–19, 2015.

Google, April 2020. URL `https://support.google.com/google-ads/answer/2701222?hl=en`.

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

B. R. Gordon, K. Jerath, Z. Katona, S. Narayanan, J. Shin, and K. Wilbur. Inefficiencies in digital advertising markets. *Journal of Marketing*, 85(1):7–25, 2021.

J. Hahn, P. Todd, and W. van der Klaauw. Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica*, 69:201–209., 2001.

W. R. Hartmann and D. Klapper. Super bowl ads. *Marketing Science*, 37(1):78–96, 2018.

X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. Candela. Practical lessons from predicting clicks on ads at facebook. In *Eighth International Workshop on Data Mining for Online Advertising*, pages 1–9, 2014.

G. Hitsch and S. Misra. Heterogeneous treatment effects and optimal targeting policy evaluation. Working Paper, University of Chicago., 2018.

G. W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

G. A. Johnson, R. A. Lewis, and E. I. Nubbemeyer. Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research*, 54(6):867–884, 2017a.

G. A. Johnson, R. A. Lewis, and E. I. Nubbemeyer. Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research*, 44:867–884, 2017b.

P. Kannan, W. Reinartz, and P. C. Verhoef. The path to purchase and attribution modeling: Introduction to special section. *International Journal of Research in Marketing*, 2016.

A. Lambrecht and C. Tucker. When does retargeting work? information specificity in online advertising. *Journal of Marketing research*, 50(5):561–576, 2013.

D. S. Lee and T. Lemeiux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

J. Liaukonyte, T. Teixeira, and K. C. Wilbur. Television advertising and online shopping. *Marketing Science*, 34(3):311–330, 2015.

A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing science*, 23(4):579–595, 2004.

H. S. Nair, W. R. Hartmann, and S. Narayanan. Identifying causal marketing mix effects using a regression discontinuity design. *Marketing Science*, 30(6):1079–1097, 2011.

H. S. Nair, S. Misra, W. J. H. IV, R. Mishra, and A. Acharya. Big data and marketing analytics in gaming: Combining empirical models and field experimentation. *Marketing Science*, 36(5):699–725, 2017.

S. Narayanan and K. Kalyanam. Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science*, 34(3):388–407, 2015.

S. Narayanan and P. Manchanda. Heterogeneous learning and the targeting of marketing communication for new products. *Marketing Science*, 28(3):424–441, 2009.

J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., 2018.

N. Pujol. Freemium: attributes of an emerging business model. *Available at SSRN 1718663*, 2010.

P. E. Rossi, R. E. McCulloch, and G. M. Allenby. The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340, 1996.

N. S. Sahni. Effect of temporal spacing between advertising exposures: Evidence from online field experiments. *Quantitative Marketing and Economics*, 13(3):203–247, 2015. ISSN 1570-7156.

N. S. Sahni, S. Narayanan, and K. Kalyanam. An experimental investigation of the effects of retargeted advertising: The role of frequency and timing. *Journal of Marketing Research*, 56(3):401–418, 2019.

E. B. Seufert. *Freemium economics: Leveraging analytics and user segmentation to drive revenue*. Elsevier, 2013.

A. Sharma, J. M. Hofman, and D. J. Watts. Estimating the causal impact of recommendation systems from observational data. In *ACM Conference on Economics and Computation*, 2015.

D. Shepard. *The new direct marketing: how to implement a profit-driven database marketing-strategy*. Irwin, 1990.

B. Smith and G. Linden. Two decades of recommender systems at amazon.com. *IEEE Internet Computing*, 21(3):12–18, 2017.

R. Sparapani, C. Spanbauer, and R. McCulloch. Nonparametric machine learning and efficient computation with bayesian additive regression trees: the bart r package. *Journal of Statistical Software*, pages 1–71, 2019.

N. Syam and A. Sharma. Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. *Industrial Marketing Management*, 69:135–146, 2018.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

A. A. Zoltners, P. Sinha, and S. E. Lorimer. Sales force effectiveness: A framework for researchers and practitioners. *Journal of Personal Selling & Sales Management*, 28(2): 115–131, 2008.

Table 1: Summary of Key Theoretical Results

Scoring Method	Required Condition	Result	RDD Estimate Provides
Intercept	$z_i \perp \beta_i$	$\hat{\beta}_{RD} = \mathbb{E}\left[\beta_i\right]$	ATE
Intercept	$\tilde{z} = \bar{z}$	$\hat{\beta}_{RD} = \mathbb{E}\left[\beta_i\right]$	ATE
Intercept or Slope	None	$\mathbb{E}\left[\beta_i\right] \subset \hat{\beta}_{RD} \pm \sqrt{\frac{var(\beta_i)}{var(z_i)}}\lvert\tilde{z} - \bar{z}\rvert$	Bounds on ATE
Slope	None	$\hat{\beta}_{RD} = \mathbb{E}\left[\beta_i \lvert \tilde{z}\right]$	LATE
Slope	None	$\hat{\beta}_{RD} = \tilde{z}$	Consistency Check

Table 2: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Conversion	1,403,913	0.004	0.061	0	0	0	1
Conversion - Online	1,403,913	0.003	0.052	0	0	0	1
Conversion - Offline	1,403,913	0.001	0.033	0	0	0	1
Purchases	1,403,913	0.004	0.075	0	0	0	9
Purchases - Online	1,403,913	0.003	0.062	0	0	0	9
Purchases - Offlinbe	1,403,913	0.001	0.042	0	0	0	6
Prior Conversion	1,403,913	0.009	0.093	0	0	0	1
Prior Conversion - Online	1,403,913	0.005	0.068	0	0	0	1
Prior Conversion - Offline	1,403,913	0.004	0.066	0	0	0	1
Prior Purchases	1,403,913	0.013	0.252	0	0	0	145
Prior Purchases - Online	1,403,913	0.006	0.130	0	0	0	39
Prior Purchases - Offline	1,403,913	0.007	0.212	0	0	0	145
Score	1,403,913	0.147	1.108	−16.460	−0.180	−0.030	7.490

Table 3: Summary Statistics - Customers with Score Lower than Threshold (i.e. Score < 0)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Conversion	1,099,810	0.001	0.024	0	0	0	1
Conversion - Online	1,099,810	0.0004	0.021	0	0	0	1
Conversion - Offline	1,099,810	0.0002	0.013	0	0	0	1
Purchases	1,099,810	0.001	0.030	0	0	0	5
Purchases - Online	1,099,810	0.0005	0.026	0	0	0	5
Purchases - Offline	1,099,810	0.0002	0.015	0	0	0	5
Prior Conversion	1,099,810	0.005	0.068	0	0	0	1
Prior Conversion - Online	1,099,810	0.002	0.049	0	0	0	1
Prior Conversion - Offline	1,099,810	0.002	0.048	0	0	0	1
Prior Purchases	1,099,810	0.008	0.250	0	0	0	145
Prior Purchases - Online	1,099,810	0.004	0.123	0	0	0	39
Prior Purchases - Offline	1,099,810	0.004	0.214	0	0	0	145
Score	1,099,810	−0.319	0.591	−16.460	−0.250	−0.030	−0.010

Table 4: Summary Statistics - Customers with Score Higher than Threshold (i.e. Score > 0)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Conversion	304,103	0.015	0.122	0	0	0	1
Conversion - Online	304,103	0.011	0.104	0	0	0	1
Conversion - Offline	304,103	0.004	0.066	0	0	0	1
Purchases	304,103	0.017	0.151	0	0	0	9
Purchases - Online	304,103	0.012	0.123	0	0	0	9
Purchases - Offline	304,103	0.005	0.084	0	0	0	6
Prior Conversion	304,103	0.024	0.152	0	0	0	1
Prior Conversion - Online	304,103	0.013	0.111	0	0	0	1
Prior Conversion - Offline	304,103	0.012	0.107	0	0	0	1
Prior Purchases	304,103	0.030	0.258	0	0	0	27
Prior Purchases - Online	304,103	0.015	0.152	0	0	0	27
Prior Purchases - Offline	304,103	0.016	0.202	0	0	0	25
Score	304,103	1.832	0.885	0.010	1.180	2.480	7.490

Table 5: Treatment Effects - Regression Discontinuity Estimates

	Coefficient	Std. Err.	p-value	N
Conversion	0.034	0.009	0.0001	$4,886$
Conversion - Online	0.032	0.007	0.00001	$4,886$
Conversion - Offline	0.001	0.005	0.774	$4,886$

Table 6: Treatment Effects - Correlational Estimates

	Coefficient	Std. Err.	p-value	N
Conversion	0.014	0.0001	0.0000	$1,403,913$
Conversion - Online	0.010	0.0000	0.0000	$1,403,913$
Conversion - Offline	0.004	0.0000	0.0000	$1,403,913$

Table 7: Bounds for ATE - Conversion

	coefficients	stderr	pvalues	deg.freedom
LATE	0.034	0.009	0.0001	$4,886$
ATE - Lower Bound $[\beta_i \ \sim U(0,1)]$	-0.095	0.009	0	$4,886$
ATE - Upper Bound $[\beta_i \sim U(0,1)]$	0.163	0.009	0	$4,886$
ATE - Lower Bound $[var(\beta_i) = 0.005]$	0.002	0.009	0.406	$4,886$
ATE - Upper Bound $[var(\beta_i) = 0.005]$	0.065	0.009	0	$4,886$
ATE - Lower Bound $[var(\beta_i) = 0.01]$	-0.044	0.009	0.00000	$4,886$
ATE - Upper Bound $[var(\beta_i) = 0.01]$	0.111	0.009	0	$4,886$
ATE - Lower Bound $[var(\beta_i) = 0.03]$	-0.044	0.009	0.00000	$4,886$
ATE - Upper Bound $[var(\beta_i) = 0.03]$	0.111	0.009	0	$4,886$
ATE - Lower Bound $[var(\beta_i) = 0.06]$	-0.076	0.009	0	$4,886$
ATE - Upper Bound $[var(\beta_i) = 0.06]$	0.143	0.009	0	$4,886$

Table 8: Bounds for ATE - Online Conversion

	coefficients	stderr	pvalues	deg.freedom
LATE	0.032	0.007	0.00001	$4,886$
ATE - Lower Bound [$\beta_i \sim U(0,1)$]	-0.097	0.007	0	$4,886$
ATE - Upper Bound [$\beta_i \sim U(0,1)$]	0.161	0.007	0	$4,886$
ATE - Lower Bound [$var(\beta_i) = 0.005$]	0.001	0.007	0.463	$4,886$
ATE - Upper Bound [$var(\beta_i) = 0.005$]	0.064	0.007	0	$4,886$
ATE - Lower Bound [$var(\beta_i) = 0.01$]	-0.012	0.007	0.043	$4,886$
ATE - Upper Bound [$var(\beta_i) = 0.01$]	0.077	0.007	0	$4,886$
ATE - Lower Bound [$var(\beta_i) = 0.03$]	-0.045	0.007	0	$4,886$
ATE - Upper Bound [$var(\beta_i) = 0.03$]	0.110	0.007	0	$4,886$
ATE - Lower Bound [$var(\beta_i) = 0.06$]	-0.077	0.007	0	$4,886$
ATE - Upper Bound [$var(\beta_i) = 0.06$]	0.142	0.007	0	$4,886$

Table 9: Sensitivity of Treatment Effect Estimates to Bandwidth Choice - Conversion

	Coefficient	Std.Err.	p-value	N
Bandwidth = 0.02	0.034	0.009	0.0001	$4,886$
Bandwidth = 0.03	0.072	0.016	0.00001	$5,195$
Bandwidth = 0.01	0.026	0.007	0.0001	$4,362$

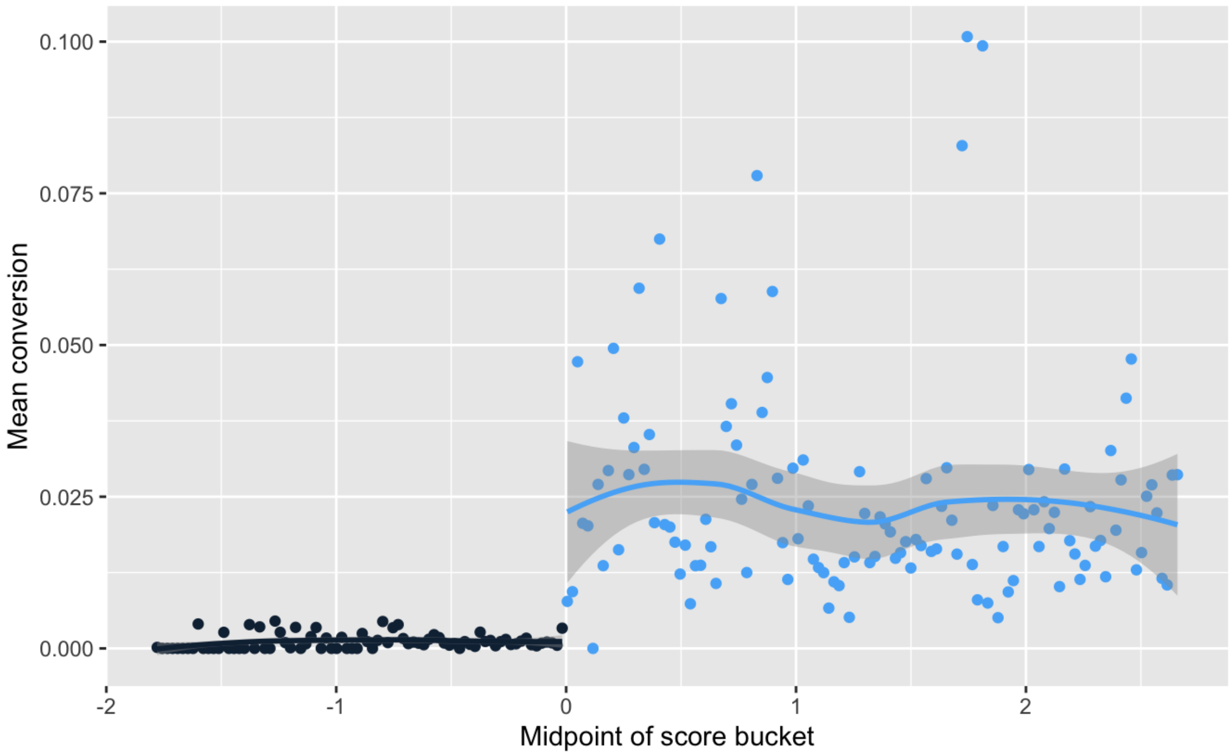

Figure 1: Conversion vs. Score

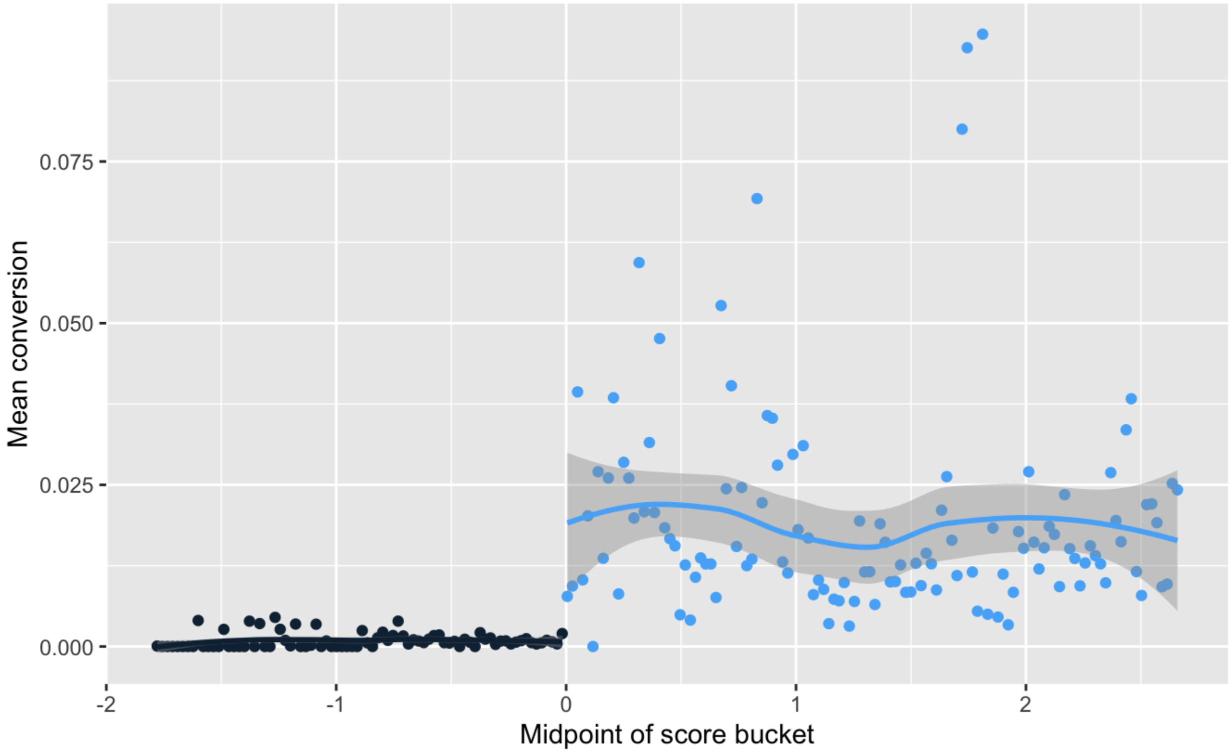

Figure 2: Online Conversion vs. Score

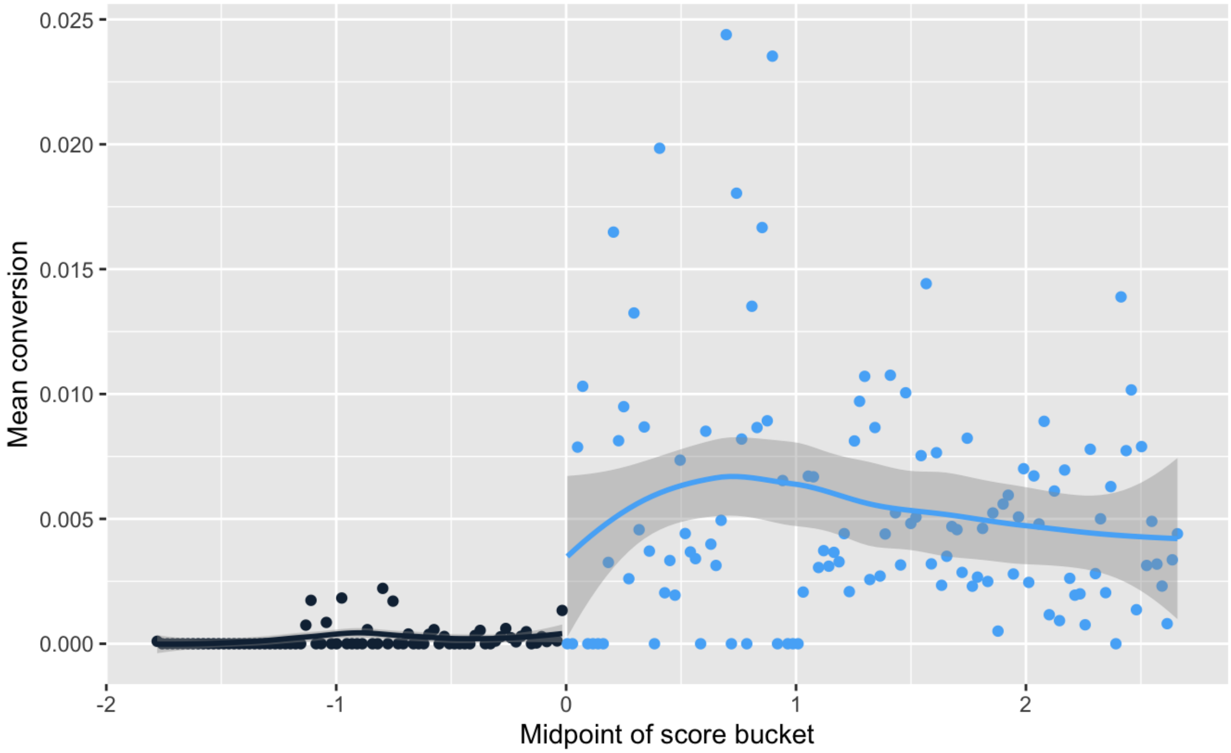

Figure 3: Offline Conversion vs. Score

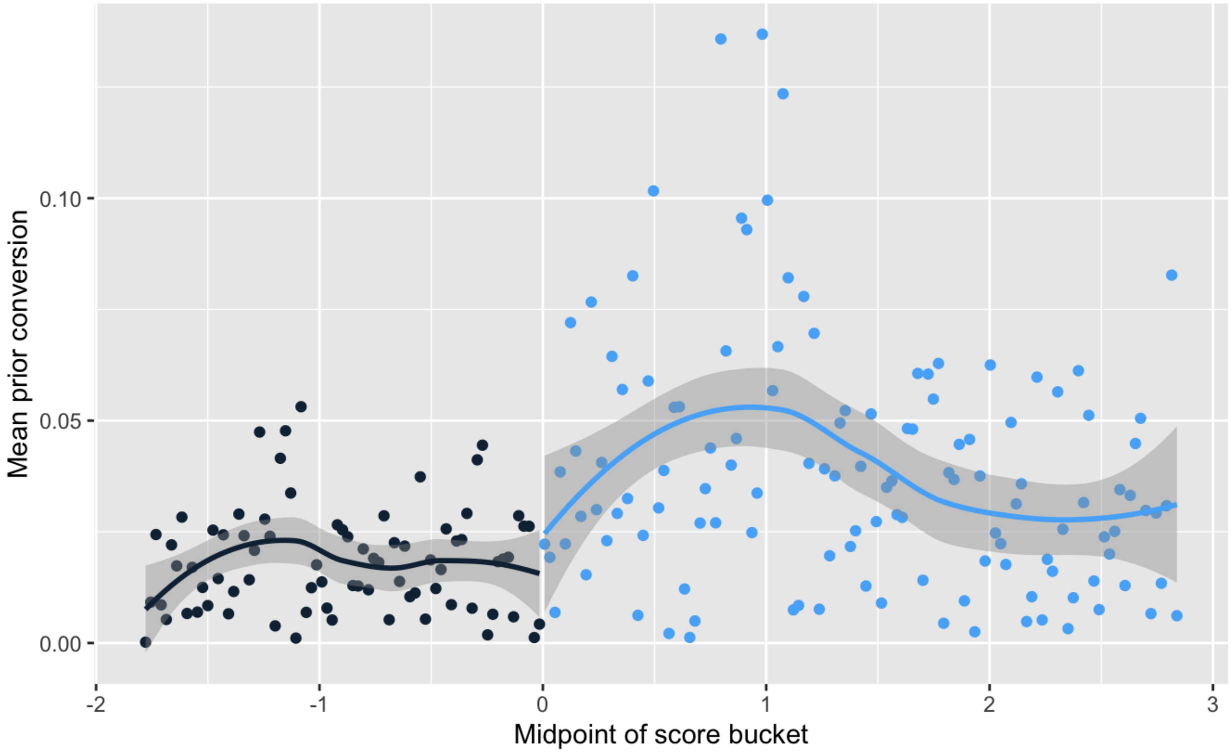

Figure 4: Prior Conversion vs. Score

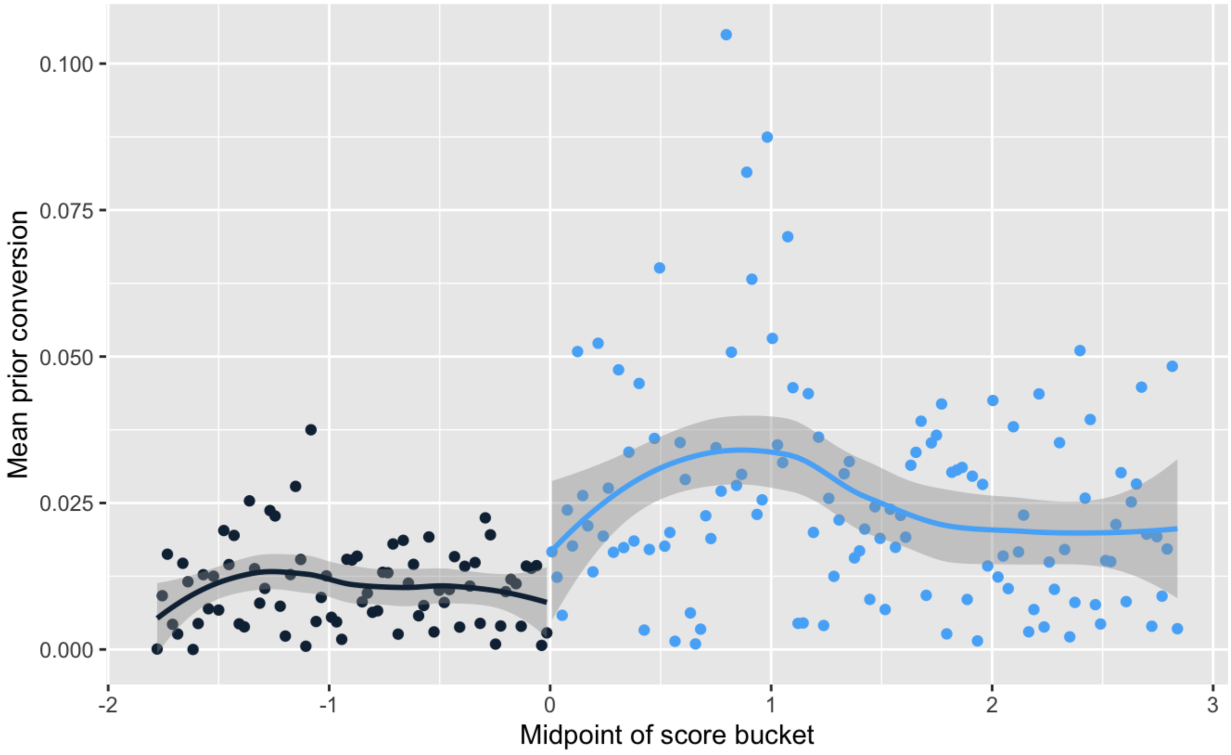

Figure 5: Prior Online Conversion vs. Score

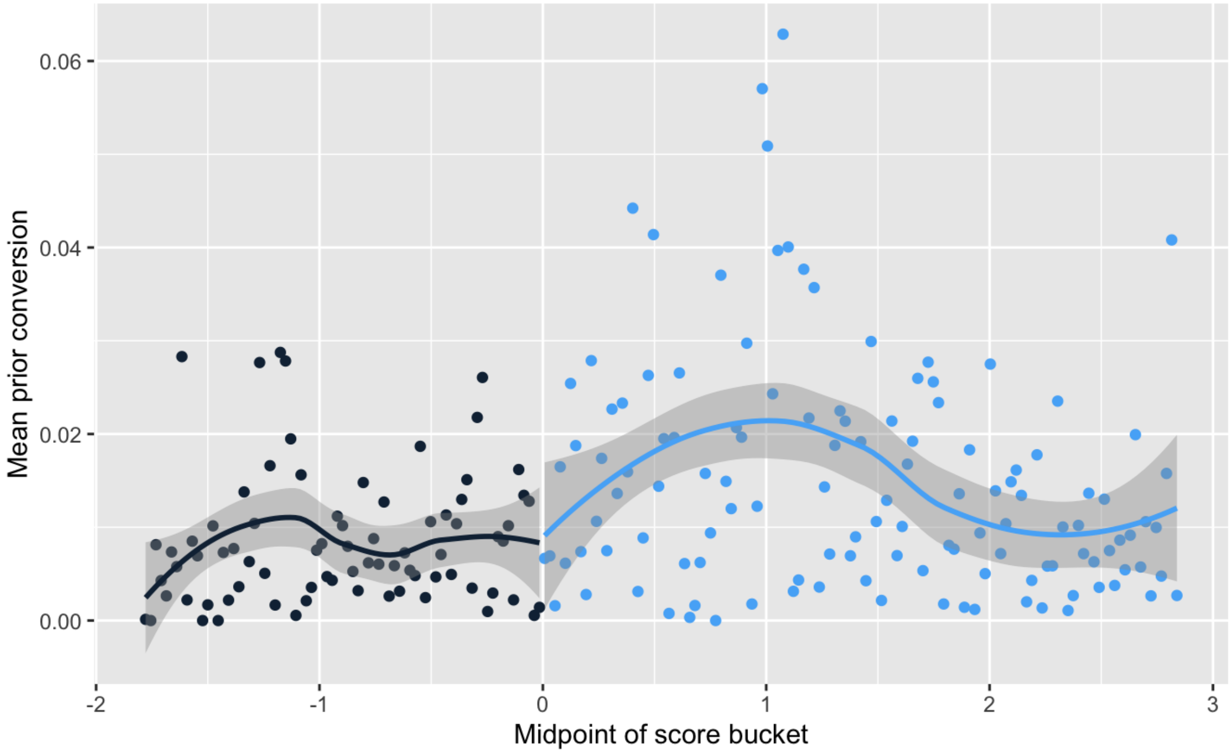

Figure 6: Prior Offline Conversion vs. Score