Behavioral Targeting, Machine Learning and Regression Discontinuity Designs

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Machine Learning

Outline

Abstract

- Behavioral targeting is a marketing application of ML, involving steps:
  1. Train a machine learning algorithm on a training dataset.
  2. Use that algorithm to score current or potential customers.
  3. When the score crosses a threshold, a treatment (e.g., offer, advertisement, recommendation) is assigned.
- The above steps give rise to opportunities for causal measurement of the effects of such targeted treatments using RDD.
- Investigating ML in a RD framework leads to several insights:
  - Under some conditions, RDD can be used to measure not only LATE, but also ATE.
  - In some situations, RD can be used to find bounds on the ATE even if point estimates are unable to be found.
- Apply this to the ML based targeting contexts:
  - Intercept-based scoring.
  - Slope-based scoring.

Experiment: Gold Standard for Causal Effects

- Causal effects estimation is crucial for marketers.

- Randomized experiment is the gold standard for obtaining causal effects.
  - Randomizing customers into treatment and control groups allows us to compare outcomes for those treated and those who are not.
  - This addresses self-selection and endogeneity in treatment.
  - But in behavioral targeting, such treatment policies naturally lead to **self-selection**.
  - Past behavior, correlated with outcomes, can bias treatment effects.
  - Gordon et al. (2019 & 2021): estimates of advertising effects on observational data fails to obtain causal effects.
  - Eckles and Bakshy (2017): experiments may be flawed in some situations, and high-dimensional datasets can yield comparable estimates.

Experiment & Quasi-Experiment

- Causal effects estimation is crucial for marketers, but conducting randomized experiments on a continuous basis can be expensive and slow.
- Exploring alternatives providing causal measurement at a lower cost and on a continuous basis (Sharma et al. 2015, Gomez-Uribe & Hunt 2015).

- **Quasi-experiments** exploit **naturally occurring randomness** in the DGP to study the effects of marketing treatments.
  - They allow for causal estimation but at lower costs and time.
  - Advertising: Narayanan & Kalyanam (2015).
  - Television advertising: Liaukonyte et al. (2015), Hartmann & Klapper (2018).
  - Promotional offers: Nair et al. (2011, 2017).

LATE → ATE I

- Nair et al. (2011) use RDD to estimate LATE in the context of behavioral targeting.
  - The study focuses on situations where targeting is **based on measures or summaries of past behaviors**.
  - It emphasizes the importance of carefully examining the validity of RDD in these contexts.

- This paper examines the validity and utility of RDD in contexts where a large number of such variables are **used through a ML framework to score customers**.
  - Not based directly on variables summarizing past behavior.
  - Leading to a natural application of ML algorithms for causal measurement.
  - The large number of variables underlying ML algorithms usually ensure that the score is continuous.
  - A continuous score, combined with a threshold rule for treatment, meets the conditions for RDD validity (Hahn et al. 2001, Lee & Lemeiux 2010, Imbens & Lemieux 2008, Nair et al. 2011).
  - **RDD allows to estimate LATE at the treatment threshold**.

LATE → ATE II

- Under some conditions, RDD can be used to measure ATE [novel finding]:
  - When **the score and the treatment effects are uncorrelated**, RD obtains **ATE** and not merely LATE.
  - Under some conditions, the firm can obtain ATE even when the score and treatment effects are correlated.
- In some situations, RD can be used to find bounds on the ATE even if point estimates are unable to be found.
  - We can derive bounds for ATE as a function of LATE estimates obtained using RDD if we have prior knowledge of treatment effect variance or there are naturally bounds on the variance.
- This paper proposes an approach with practical utility for obtaining treatment effects of interest in a variety of contexts.

Application I

ML based scoring systems in behavioral targeting can be applied using RDD:

1. **Intercept-based scoring**: customers are scored based on their likelihood of a positive outcome (e.g., purchase or churn).
   - Results on ATE discussed earlier can be useful in intercept-based scoring contexts under certain conditions.
2. **Slope-based scoring**: customers are scored based on their incrementality from treatment.
   - This is harder to do, as it requires the firm to have a way to measure this incrementality at the individual level.
   - RDD provides a simple and low-cost way to assess the validity of slope-based scoring algorithms.

Application II

Apply approach to an empirical setting involving the retargeting of display ad.

- The firm scored customers on their likelihood of purchase using a ML algorithm [intercept-based scoring].
  - Scoring was based on consumers' browsing and transaction activity.
  - Individuals exceeding a threshold were selected for display ad. retargeting.
  - Estimating causal effects of retargeted display ad. using RD approach.

- A field experiment with consumers selected based on the ML score.
  - Randomization into treatment and control groups was performed.
  - Only the treatment group received the retargeted ad. campaign.
  - This experimental design allows us to obtain causal effects of retargeting and validate the RD based estimates.
  - The experiment provides us placebo tests for verifying that we are not obtaining spurious estimates of the effects of retargeting using RDD.
  - We then go beyond LATE to obtain bounds on the ATE.

Behavioral Targeting

Behavioral targeting has a long history in marketing, utilizing customer response behaviors for personalized marketing actions.

Early Examples:

- Catalog marketers scored customers based on recency, frequency, and monetary value of their response behaviors (Shepard 1990).
- Point-of-sale scanners in retail stores generated individual-level purchase behavior data, leading to increased use of behavioral targeting in the grocery channel (Blattberg 1988).
- E.g., Catalina marketing issued customized coupons based on observed checkout behavior.

Click Stream and Path to Purchase Data

Internet Era:

- Internet growth led to a significant increase in behavioral data collection, particularly **web site browsing data**.
- Combining time-stamped logs with a visitor's session information (id or cookie) tie together the different pages that were visited during a session.
- This type of behavioral data is called click stream data, which enables **tracking of visitor paths on web** (Chatterjee et al. 2003, Montgomery et al. 2004, Bucklin et al. 2002), e.g., retail websites.
- Also, the inbound traffic on a web site has the referring URL, which is the source of the traffic, has become useful in building path to purchase models (Kannan et al. 2016).
- ML models predict future behavior based on past behavior and a vast number of predictors [methodological development is not our focus].

Online and Offline Advertising I

One of the most common application areas for behavioral targeting is online and offline advertising.

Retargeting in Display Advertising:

- Advertisers target display advertising to visitors whose on-site behaviors exceed a certain threshold based on machine learning scores.
- Customers can be grouped into separate buckets and scored, and those with scores above a cutoff value receive advertising.
- Sahni et al. (2019) retargeting display advertising to product viewers and cart creators.

Online and Offline Advertising II

Retargeting in Search Advertising:

- Retargeting List for Search Advertising (Google 2020) allows advertisers to score customers and retarget them for search advertising campaigns.
- Advertisers can leverage online browsing behaviors to target offline advertising, such as programmatic direct mail.

Behavioral Targeting on Ad Platforms:

- Ad platforms like Facebook can target advertising to users based on their behavioral data (He et al. 2014).
- User demographic information, interests, and past behavior are considered in the ad targeting process.

Recommendation Systems I

ML models are extensively used to build recommendation systems in e-commerce, content and entertainment platforms.

Types of Recommendation Systems (Adomavicius & 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 collaborative and content-based systems.

Recommendation Systems II

Example: Collaborative Recommendation System in Amazon

- Amazon's recommendation system suggests items based on the behavior of people with similar likes.
- In other words, the recommendation systems generate a relatedness score based on **past behavior of people with similar likes**
- ML is used to incorporate various factors such as liked recommendations, clicked items, compatibility, substitutes versus complements, sequential purchases, and the impact of time (Smith & Linden 2017).
- Reports indicate that a significant fraction of clicks on recommended products come from recommendation links (Smith & Linden 2017).
- Using experiments to obtain causal estimates of such systems on incremental clicks can inconvenience users (Sharma et al. 2015).
- Therefore, natural experiments have been proposed as an alternative. E.g., exogeneous shocks to traffic.

Recommendation Systems III

Example: Collaborative Recommendation System in YouTube

- YouTube uses collaborative recommendation systems to help users navigate their vast content libraries.
- Recommendation systems on YouTube have various objectives. E.g., users' specific interests, broad interests, or entertainment preferences.
- Scoring system on YouTube is a top N recommender (Davidson et al. 2010).
- ML techniques, such as association rule mining or co-visitation counts, are utilized to build these recommendation systems (Agrawal et al. 1993).

Recommendation Systems IV

Example: Collaborative Recommendation System in Netflix

- Netflix, a pioneer in online content streaming, uses 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).
- Netflix provides multiple recommendation categories such as "Top Picks", "Trending Now", "Continue Watching", and "Because You Watched".
- Algorithms blend popularity and personalization signals to generate scores and rankings for constructing pages and rows on Netflix.
- Netflix tracks user behavior, including the extent of catalog utilization and the take rate (fraction of recommendations resulting in a play).

Recommendation Systems V

Composite Score and Outcome Metrics

- In addition to providing recommendations, recommendation platforms observe outcome metrics such as user acceptance of recommendations and increased site usage.
- These outcome metrics, along with the relatedness score, can be incorporated into machine learning models to generate a composite score.
- The composite score considers factors like past user behavior and can generate a top-N set of recommendations, subject to a threshold value.

Pricing and Personal Selling I

Behavioral Targeting for Pricing and Promotional Offers

- Online behavior is used to target price and promotional offers, as discussed in Dube & Misra (2020) using the case of Ziprecruiter.
- Dube & Misra (2020):
  1. An experiment is conducted to understand the causal effects of pricing. Customers sign up for a trial period and provide background information.
  2. A high-dimensional demand model is trained on this data.
  3. Based on the pricing scheme developed from the first experiment, customized offers are implemented out of sample.
- This type of behaviorally targeted pricing is relevant for firms that offer content subscriptions or are in the software as a service (SAAS) sector.
- Firms in this sector practice a freemium model (Kumar 2014, 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 updgrades from the free version in freemium models.

Pricing and Personal Selling II

Personal Selling

- Personal selling efforts are a costly part of the marketing mix, and optimizing the allocation of sales effort has received huge attention.
- Behavior-based lead scoring is important in sales effort allocation. E.g., in pharmaceuticals, doctors are scored based on their past prescription behavior, and sales effort is assigned differentially across deciles (Narayanan & Manchanda 2009).
- Online behavior of prospects or customers can be used to score the quality of leads before passing them on for follow-up by the sales force. E.g., landing pages visited, emails opened, product pages visited, etc.
- The scores can be used to generate a marketing qualified lead (MQL) or a sales qualified lead (SQL).
- Leads crossing a threshold are prioritized for follow-up by the sales team.

Summary and Evaluation of Behavioral Targeting

- Behavioral targeting has extensive applications across the marketing mix.
  1. There is a marketing action.
  2. Individuals or treatments are scored by a ML algorithm using prior behavior data.
  3. The individual or treatment is assigned based on the score crossing a threshold or a rank ordering based on the score.
  4. Outcomes are observed both for customers with and without treatment.

- Incrementality is a key evaluation of behavioral targeting using ML. But:
  - Credible IV is likely to be context specific and in general hard to be found.
  - A/B testing and field experimentation are costly, may disadvantage users (Sharma et al. 2015), and are a bottleneck for rapid innovation.

- RDD-based methods are more frequent, causal, and less costly estimates.

- But, can RDD be an alternative to A/B test outcomes or field experiments to yield ATE estimate?

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

ML based scoring algorithms can be broadly classified into **intercept-based scoring** and **slope-based scoring**.

Intercept-based Scoring

- It attempts to estimate the heterogeneous intercepts across customers and target them based on this estimate crossing a threshold.
- It is common as it relies on simpler algorithms that relate outcomes of interest to a set of observable variables for consumers, potentially with exogenous variation coming through an experiment.
- E.g., estimating the propensity of a customer to convert when deciding on who to target with retargeted ads.
- E.g., finding the likelihood of churn for a subscription product in order to target consumers with proactive retention programs (Ascarza et al. 2018).

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

Slope-based Scoring

- It attempts to target customers based on an estimate of how they are likely to respond to the marketing intervention itself.
- It is harder to implement as they typically require estimation of heterogeneous treatment effects, using experiments to exogenously determine the treatment, and algorithms to obtain CATE.
- These CATEs can be used for generating optimal targeting policies (Hitsch & Misra 2018).
- Less commonly applied in practice because of higher data requirements, newer algorithmic approaches, and more sophisticated technique for implementation.

Background on Regression Discontinuity

Regression Discontinuity Designs (RDD)

- It 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, applying only to the obs. 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).

Formalization

Formally, let:
i index the obs.; y_i be the outcome of interest; x_i be the treatment; z_i be the forcing variable (score); $\tilde{z}$ be the threshold above which treatment is applied.

Treatment is defined by

$$x_i = \begin{cases} 1, & \text{if } z_i \geq \tilde{z} \\ 0, & \text{if } z_i < \tilde{z} \end{cases}.$$

RD estimate of the treatment effect β is given by

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

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

Behavioral Targeting and Regression Discontinuity

Behavioral targeting raises concern about validity of RDD: **self-selection**.

- E.g., 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 $z > \tilde{z}$.
- This makes RD invalid since 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.

This paper considers contexts where the behavioral targeting policy uses an underlying ML algorithm to generate scores.

- Combine behavioral data of consumers (e.g., history of browsing activity or purchases) into a score through a complex algorithm.
- Since the score is **continuous**, consumers are **unable to anticipate how specific actions they take affect the score**, and they are **uncertain about score and threshold**, RDD is valid (Nair et al., 2011).
- This makes RDD a useful candidate to measure LATE.

Beyond LATE I

The paper then analyze conditions under which RDD can be used to go beyond LATE towards measuring ATE.

Consider the DGP:

$$y_i = \alpha_i + \beta_i \cdot x_i + \epsilon_i, \text{ where}$$

α_i is the intercept; β_i is the slope (treatment effect); ϵ_i is the idiosyncratic shock. They are allowed to be heterogeneous across individuals.

- The linear specification is reasonable: continuous and differentiable function can be locally approximated by a linear specification.

Using $\hat{\beta}_{RD} = \lim_{\lambda \to 0} E\left[y_i | z_i = \tilde{z} + \lambda\right] - \lim_{\lambda \to 0} E\left[y_i | z_i = \tilde{z} - \lambda\right]$, we have

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

Substituting x_i as 0 or 1 based on the treatment conditions, and using the condition of continuity, we have

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

Beyond LATE II

The LATE estimate

- is not affected by how α_i and ϵ_i are correlated with z_i, as long as the condition of continuity of everything other than the treatment at the threshold is maintained.
- relies on the correlation between heterogeneous treatment effect β_i and z_i.

If the score is orthogonal to the slope ($z_i \perp \beta_i$)

$$z_i \perp \beta_i \implies \hat{\beta}_{RD} = E[\beta_i].$$

This expected value of β_i is the ATE.
$\rightarrow$ When **score is uncorrelated with slope**, ATE can be obtained via RDD.

If the score and slope are correlated, assume

$\beta_i = \gamma_0 + \gamma_1 z_i + \eta_i$, where $z_i \perp \eta_i$ (prototypical assumption) and $E[\eta_i] = 0$,

$$\hat{\beta}_{RD} = \lim_{\lambda \to 0} E\left[\gamma_0 + \gamma_1 z_i + \eta_i | z_i = \tilde{z} + \lambda\right] = \gamma_0 + \gamma_1 \tilde{z}.$$

Beyond LATE III

Whereas the ATE is

$$E[\beta_i] = \gamma_0 + \gamma_1 \bar{z}.$$

Therefore, the difference between the LATE and ATE is given

$$\hat{\beta}_{RD} - E[\beta_i] = \gamma_1(\tilde{z} - \bar{z}) = \frac{\text{cov}(\beta_i, z_i)}{\text{var}(z_i)}(\tilde{z} - \bar{z}) = \text{corr}(\beta_i, z_i)\sqrt{\frac{\text{var}(\beta_i)}{\text{var}(z_i)}}(\tilde{z} - \bar{z}).$$

If the correlation between the slope and score is between 1 and -1, then

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

- The only unknown is the variance of the treatment effect $\text{var}(\beta_i)$. Firms may have prior knowledge of it or its bounds.
- If threshold for treatment = mean score ($\tilde{z} = \bar{z}$), then LATE = ATE.
  $\rightarrow$ Firms can simply set $\tilde{z} = \bar{z}$ to find ATE.
  $\rightarrow$ RDD provides relatively low-cost estimation of ATE.

Scoring

Intercept-based Scoring: estimate the intercepts for customers, and set thresholds on these scores to determine which customers to target. That is,

$$z_i = \hat{\alpha}_i.$$

Slope-based Scoring: estimate the treatment effect of interest for each customer, and uses this estimate as a score in a targeting policy. That is,

$$z_i = \hat{\beta}_i. \text{ Therefore,}$$

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

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

$$\hat{\beta}_{RD} = \tilde{z}.$$

When the estimator is consistent and score is continuous, RD estimate of the LATE has to equal the threshold for treatment itself.

- RDD provides a way to continuously evaluate the validity of the underlying ML algorithm itself.

Context I

A advertiser of cellphone services to conduct a retargeted advtg. campaign.

- Consumers who visit any product page but depart the website without completing a purchase are eligible for a retargeted advertising campaign.
- With retargeting campaigns, these customers might be persuaded to return to the advertiser's website and complete their purchase.
- Issue: Retargeting campaigns involve customers who are **self-selected**, as evidenced by their prior interest in the advertiser's products.
  - Field experiment is costly: Sahni et al. (2019) spent 1 year to collect data.
- Issue: Retargeting campaigns may target consumers who are not interested in the advertiser's product.
  - This wastes money and could irritate these consumer.

Context II

This paper uses ML methodology to select customers.

1. Use historical data (e.g., pages visited, dwell times) to train a proprietary algorithm.
2. Generate a purchase propensity score using such data as imputs to the algorithm.
3. Consumers are eligible for retargeting if their scores are above the threshold.

Not measuring how responsive the customers would be to the retargeting campaign, but their baseline likelihood of making a purchase (intercept-based).

Context III

Key points of this empirical application

- This advertiser did not affect its advertising policy.
- The purchase propensity score (PPS) is normalized with zero threshold.
- Two sets of customers are ineligible for the retargeting campaign.
  - PPS < 0, and
  - PPS > 0, but randomized by experimental design into control group.
  - The second set is for placebo test: becasue both sets are not targeted with advtg., RDD-based estimate of treatment effects of advtg. should be zero.

Data Description

1.9 million consumers are assigned PPS through proprietary ML algorithm.

- About 1.3 million PPS < 0: ineligible for the retargeting campaign.
- About 570,000 PPS > 0. Experimental design randomized them into treatment and control groups (about 285,000 customers in each group).

Every customer is tracked both before and after the experiment using tracking cookies placed by the advtg. network on consumers' devices.

- This tracking is across devices.

The key outcome variables are "whether consumer make purchase" and "number of such purchase occasions".

Offline + online purchases, as well as online purchases alone are examined.

Summary Statistics I

Table 2: Summary Statistics - Customers with Score Lower than Threshold (i.e. PPS Score < 0)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Conversion	1,314,511	0.001	0.027	0	0	0	1
Conversion - Online	1,314,511	0.001	0.023	0	0	0	1
Conversion - Offline	1,314,511	0.0002	0.014	0	0	0	1
# Purchases	1,314,511	0.001	0.033	0	0	0	5
# Purchases - Online	1,314,511	0.001	0.029	0	0	0	5
# Purchases - Offline	1,314,511	0.0002	0.017	0	0	0	5
Prior Conversion	1,314,511	0.005	0.073	0	0	0	1
Prior Conversion - Online	1,314,511	0.003	0.053	0	0	0	1
Prior Conversion - Offline	1,314,511	0.003	0.052	0	0	0	1
# Prior Purchases	1,314,511	0.009	0.259	0	0	0	145
# Prior Purchases - Online	1,314,511	0.005	0.146	0	0	0	39
# Prior Purchases - Offline	1,314,511	0.005	0.209	0	0	0	145
PPS Score	1,314,511	−0.845	0.529	−14.499	−0.906	−0.721	0.000

Conversion: ≥ 1 instances of purchase after the triggering event of retargeting camapaign (a product page view but without completion of a purchase).

Summary Statistics II

Table 3: Summary Statistics - Customers with Score Higher than Threshold (i.e. PPS Score > 0)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Conversion	569,079	0.015	0.121	0	0	0	1
Conversion - Online	569,079	0.011	0.102	0	0	0	1
Conversion - Offline	569,079	0.004	0.066	0	0	0	1
# Purchases	569,079	0.017	0.153	0	0	0	17
# Purchases - Online	569,079	0.012	0.124	0	0	0	11
# Purchases - Offline	569,079	0.005	0.089	0	0	0	17
Prior Conversion	569,079	0.022	0.146	0	0	0	1
Prior Conversion - Online	569,079	0.011	0.104	0	0	0	1
Prior Conversion - Offline	569,079	0.011	0.107	0	0	0	1
# Prior Purchases	569,079	0.029	0.275	0	0	0	58
# Prior purchases - Online	569,079	0.013	0.141	0	0	0	27
# Prior purchases - Offline	569,079	0.016	0.232	0	0	0	58
PPS Score	569,079	0.918	0.704	0.008	0.335	1.509	20.143

From tables (2) & (3): higher PPS is associated with, on average, higher levels of past purchase and conversion.

Randomization Checks

Table 4: Randomization Checks

Variable	Mean - Treatment	Mean - Control	p-value
Prior Conversion	0.022	0.022	0.911
Prior Conversion - Online	0.011	0.011	0.473
Prior Conversion - Offline	0.012	0.011	0.649
# Prior Purchases	0.028	0.029	0.093
# Prior Purchases - Online	0.012	0.013	0.197
# Prior Purchases - Offline	0.016	0.016	0.225
PPS Score	0.920	0.917	0.154

Treatment and control groups are matched on all these variables, giving us confidence in the experiment.

Control Group vs PPS<0

Table 5: Comparison of Control Group Customers with those below the threshold (PPS Score <0)

	Mean - Below Threshold	Mean - Control	p-value
Conversion	0.001	0.015	0
Conversion - Online	0.001	0.011	0
Conversion - Offline	0.0002	0.004	0
# Purchases	0.001	0.017	0
# Purchases - Online	0.001	0.012	0
# Purchases - Offline	0.0002	0.005	0

Since both consumers in control group and with PPS<0 are not targetted, differences in the outcome variables between these groups represent the kind of selection that the PPS score is achieving.

Specification

- Use RDD-based approach to measure treatment effects of retargeting campaign.
- 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, use

$$y_i = \theta_0 + \theta_1 z_i + \theta_2 x_i + \theta_3 z_i x_i + v_i$$

  to conduct on the subset of the data with scores lying with the bandwidth intervals around the treatment threshold of 0.
- y_i is the outcome of interest, z_i is the score, x_i is the binary treatment variable and v_i is a random error.
- 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.

Empirical Strategy & Placebo Tests

Empirical Strategy

- measure the treatment effect for: overall conversion, conversion in both online and offline channels, and the number of purchase events overall as well as in the two channels.
- In all these instances, compare the treatment group (which by definition have score above threshold) with observations that had scores below threshold and which were not eligible for the retargeting campaign.

Placebo Tests (two sets of them)

1. Find RD estimate of treatment effect: control group vs obs. with PPS<0.
   $\rightarrow$ Both have no eligibility for retargeting campaign, so treatment effect should be zero.
2. Treatment group vs obs.with PPS<0, on the set of pre-experimental purchase and conversion variables.
   $\rightarrow$ These occurred before the treatment, so RDD-based estimates for these treatment effects should be zero.

Results: LATE & Placebo Tests

Table 6: Treatment Effects and Placebo Tests - Conversion

	Coefficient	Std. Err.	p-value	N
Treatment Effect	0.131	0.044	0.003	473
Placebo Test vs. Control	0.051	0.042	0.226	458
Placebo Test - Prior Conversion	0.037	0.055	0.506	473

Table 9: Treatment Effects and Placebo Tests - Number of Purchases

	Coefficient	Std. Err.	p-value	N
Treatment Effect	0.133	0.047	0.005	473
Placebo Test vs. Control	0.053	0.046	0.252	458
Placebo Test vs. Prior Purchases	−0.152	0.128	0.238	473

Retargeted advtg. increases conversions significantly & by a large magnitude. For placebo tests, there are no significant effects.

Results: Bounds on ATE I

Recall that

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

- We can obtain these bounds if we knew $var(\beta_i)$.
- The treatment effects can be reasonably bounded between 0 and 1, since the outcome itself is binary.
- Assume that the treatment effects $\sim U(0, 1)$, then $var(\beta_i) = 0.083$.

Results: Bounds on ATE II

Table 12: Bounds for ATE - Conversion

	coefficients	stderr	pvalues	deg.freedom
LATE	0.131	0.044	0.003	473
ATE - Lower Bound $[\beta_i \sim U(0,1)]$	0.041	0.044	0.177	473
ATE - Upper Bound $[\beta_i \sim U(0,1)]$	0.221	0.044	0.00000	473
ATE - Lower Bound $[var(\beta_i) = 0.03]$	0.077	0.044	0.040	473
ATE - Upper Bound $[var(\beta_i) = 0.03]$	0.185	0.044	0.00001	473
ATE - Lower Bound $[var(\beta_i) = 0.06]$	0.054	0.044	0.108	473
ATE - Upper Bound $[var(\beta_i) = 0.06]$	0.207	0.044	0.00000	473

When $\beta_i \sim U(0,1)$, upper bound of ATE is statistically significant; when $var(\beta_i) < 0.03$, both bounds are statistically significant.
$\rightarrow$ Given an estimate or guess of the variance of the treatment effect, we can obtain bounds on the ATE using the RDD.

As for choosing bandwidth:
$\uparrow$: increases potential bias in the RD estimate.
$\downarrow$: reduces degrees of freedom & leads to the estimates being insignificant.

Conclusion

In section 1-3:

- This paper proposes using RDD to estimate treatment effects in marketing, where customers are targeted based on ML scores above a threshold. RDD is suitable for estimating LATE in such contexts.
- Under some conditions, we can use RDD to find ATE or bounds on the ATE when point estimates are not feasible.

In section 4 (empirical application):

- There are significant effects of retargeted advertising on conversions.
- The placebo tests show that these effects are not spurious.
- Bounds on ATE are found.

Limitations of the proposed approach:

- There should be no other source of discontinuity at the treatment threshold other than the treatment itself.
- To find bounds on ATE, the score need to be uncorrelated with other factors that affect the treatment effect.
- This approach is not a substitute but a complement for experimentation.