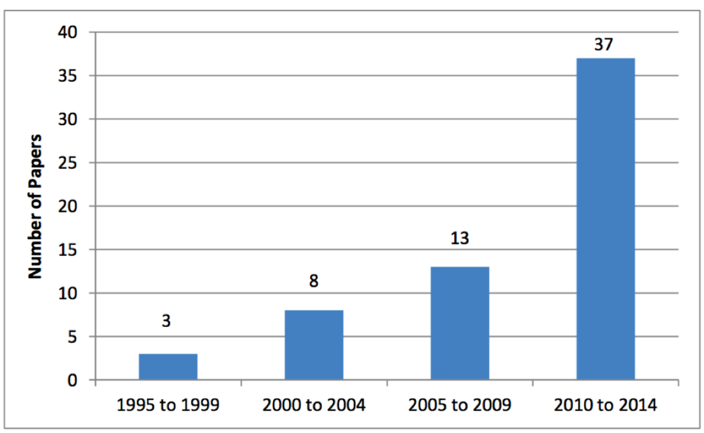
Field Experiments in Marketing

Oded Netzer

Empirical Models in Marketing

Papers in Marketing Using Field Experiments

Figure 1: Publications by Year

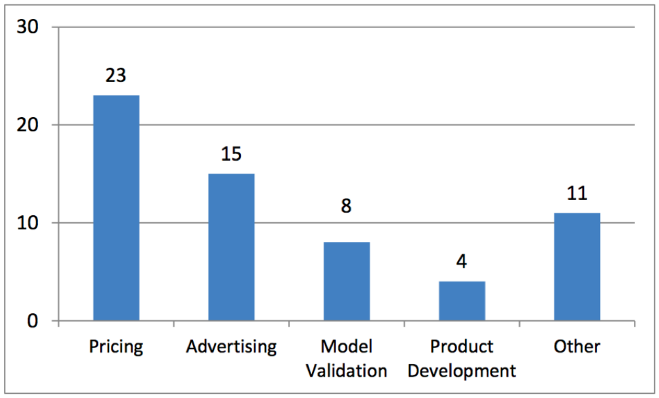


This figure reports the number of papers published by 5-year period. The sample size is 61 (papers).

Source: Field Experiments in Marketing, Duncan Simester 2015

Topics of Field Experiment in Marketing

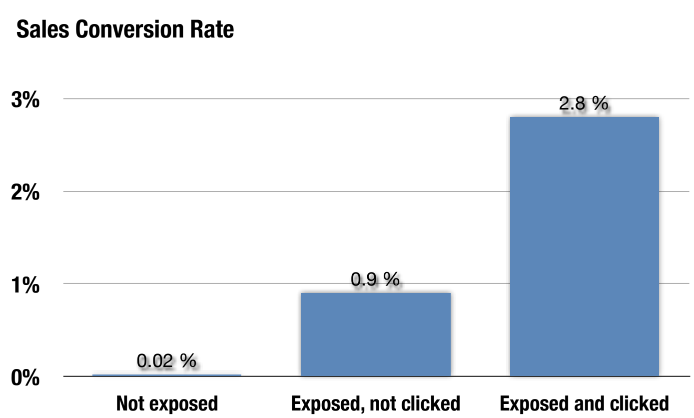
Figure 4: Choice of Topics



This figure reports the number of published papers by topic. The sample size is 61 papers.

Source: Field Experiments in Marketing, Duncan Simester 2015

Does advertising generate incremental sales?



Source: Gordon 2017

What may be going on?

Ad Exposure $\longrightarrow$ Conversion

Ideally, how would we measure causation?

How would a consumer behave in two worlds that are identical except for one difference:

1. In one world they experience our action
2. In the other world they do not

Of course, you can't experience an action and not experience an action at the same time

Experiments try to solve this problem by creating comparable groups

Steps to run an experiment

1. What is the business initiative you want to test?
2. What is the outcome of interest?
3. Who should be included in the control group (A)?
4. Who should be included in the test group (B)?

How to create test and control groups?

Assign people to groups completely randomly

Some Basic Notation

- We care about the treatment effect. $\Delta = E(y_i^T - y_i^C)$

- From data we can get

$$\begin{aligned}\hat{\Delta} &= E(y_i^T|T) - E(y_i^C|C) \\ &= E(y_i^T|T) - E(y_i^C|T) - E(y_i^C|C) + E(y_i^C|T) \\ &= \underbrace{E(y_i^T - y_i^C|T)}_{\text{Treatment efect on treated}} - \underbrace{(E(y_i^C|C) - E(y_i^C|T))}_{\text{Inherent difference}}\end{aligned}$$

Source: Navdeep Sahni, Stanford University

$\Delta = \hat{\Delta}$ if assumptions hold

$$\Delta = E(y_i^C - y_i^T)$$

$$\widehat{\Delta} = \underbrace{E(y_i^T - y_i^C | T)}_{\text{Treatment efect on treated}} - \underbrace{(E(y_i^C | C) - E(y_i^C | T))}_{\text{Inherent difference}}$$

$\Delta = \widehat{\Delta}$ if

- Inherent differences vanish because of randomization
- Treatment done on a representative sample
- SUTVA (Stable Unit Treatment Value Assumption) holds – Rubin (1980)
- Treatment effect is individualistic

Source: Navdeep Sahni, Stanford University

Effects are not the same for every unit

- Treatment effect depends on observed and unobserved factors
  - $\Delta(X, \varepsilon)$
    - X and ε are vectors of observed and unobserved characteristics
- Usefulness of the experiment depends on learning about $\Delta(X', \varepsilon')$
- Determining the mechanism is valuable: how does treatment cause the effect?
  - Consider this objective at experiment design stage
  - Describe the experiment with this objective in mind

Source: Navdeep Sahni, Stanford University

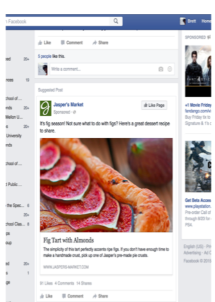
7

Challenges with running field experiments

- Too many possibilities, not clear which theory is most important to test
- Manipulating relevant constructs of a theory may be difficult (e.g., consumer expectations)
- Avoid Hawthorne Effects
  - Consumers might react in unreal ways if they get an unexpected stimulus
- Treatments move from equilibrium

Source: Navdeep Sahni, Stanford University

Sometimes choosing control is not obvious…

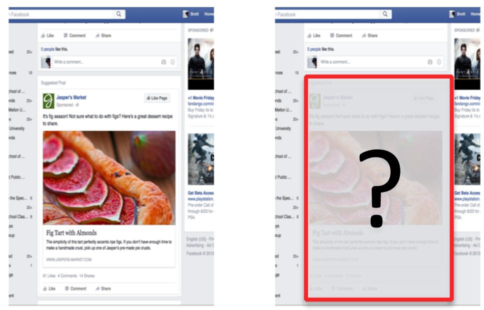


Test

Control

Source: Gordon et al. 2018

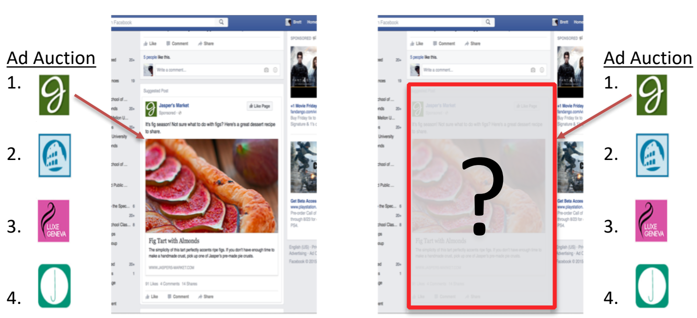
What ad should the control user see?



Test Control

Source: Gordon et al. 2018

Each ad is the result of an auction



Test Control

Source: Gordon et al. 2018

Serve the next ad in the auction queue

Source: Gordon et al. 2018

Unit of randomization vs. unit of analysis

Randomize individuals or aggregate units (markets)

- Ideally randomize at the individual level
  - More observations possible
  - Richer heterogeneity to explore
- But, sometimes feasibility is an issue
  - E.g., TV advertising can be varied across markets only
- Or the question requires it
  - Is one store format better than another?
- Implications for analysis

Want to make precise conclusions? Power calculations

- Power: probability of rejecting when the null is false
- How small of an effect do you want to detect?
- Pick a prior effect size
  - Business relevant quantitative effect: "I want to be able to detect a 5% change in sales"
  - From past literature
- What sample size will give precise enough (minimum detectable effect (MDE))?

$$MDE = (t_{(1-k)} + t_{\alpha})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

- σ: standard deviation of y
- N: sample size
- P: proportion of sample in treatment
- k: power; t_{1-k}=0.84 for 80% power
- α: desired statistical significance level

Source: Navdeep Sahni, Stanford University

Good company collaborators

- Collaborating business may not see value in testing a theory
  - The mechanism has to be consequential for the company
- Can run a high powered experiment
- Get the concerns our of the early → ask questions
- Question every step of implementation
- Randomization not alphabetization

Analysis - Are treatment and control groups balanced?

- Conduct randomization checks
- For any baseline characteristic Z, we expect:

$$E(Z_i \mid T) - E(Z_i \mid C) = 0$$

- Useful to check for lagged dependent variables

Analysis - Estimating treatment effects

- Compare unconditional means: $\widehat{\Delta} = E_i(y|T) - E_i(y|C)$
  - t-test or ANOVA
- Using linear regression

$$y_i = \alpha + \beta T_i + \epsilon_i$$

- $\hat{\beta}$ is an unbiased estimator of Δ

… but OLS gets wrong standard errors (Freedman 2008a)

  - Estimate heteroskedasticity robust standard errors

- Logit regression gets inconsistent results (Freedman 2008b)

Analysis - Standard errors clustered to unit of randomization

- When analysis is not at the randomization level

$$y_{im} = \alpha + \beta T_{im} + \epsilon_{im}$$

- Cannot assume observations iid
    - Individual behavior within units of m may be correlated
- Use cluster-robust standard errors to allow for errors to be correlated with m
    - See Cameron, Gelbach and Miller (2008) for adjustment when number of clusters is small

Analysis - Increasing power with covariates?

- Precision can be increased by reducing noise
- Adding covariates that explain noise in y can help

$$y_i = \alpha + \beta T_i + \gamma X + \epsilon_i$$

- $\hat{\beta}$ is an estimator of Δ
- Standard errors decrease because noise is lower

- X could include Baseline characteristics, lagged dependent variables
- However, $\hat{\beta}$ may be biased for small samples ("gold standard" of experiments lost; Freedman 2008; Deaton 2010).
- Be aware of p-hacking…

Analysis - Exploring heterogeneity in treatment effect

- One can interact the treatment dummy with baseline characteristics

$$y_i = \alpha + \beta_i T_i + \epsilon_i$$

Linear case

$$y_i = \alpha + \beta T_i + \theta T_i X_i + \gamma X_i + \epsilon_i$$

- θ represents how the causal effects of T varies with
  - It does not say how X causes the effect of T to change
- Concern about data-mining
  - 5% of the interaction terms will be statistically significant at 95%, by chance

Retention futility: Targeting high-risk customers might be ineffective

Eva Ascarza – JMR 2018

Research question

Should firms target customers at highest risk of churning?

- Are high-risk customers more sensitive to retention interventions?

How should firms target retention efforts?

Lift-based approach to proactive churn management
- Leverages experimentation to capture observed heterogeneity
- More **effective**
- **Feasible** and **generalizable**
- Identifies **segments** of customers who should be targeted

Ascarza 2018

Risk vs. Sensitivity

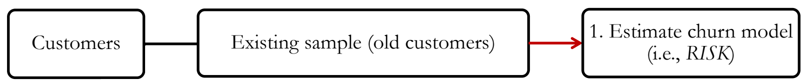
- Observed characteristics X_i
- Probability that a customer churns $P[Y_i|X_i]$
- Current approach estimates: $RISK_i = P[Y_i|X_i]$

(Targeted) proactive retention campaign
- Probability that a customer churns
  - $P[Y_i|X_i, T_i = 1]$ if i is targeted
  - $P[Y_i|X_i, T_i = 0]$ if i is not targeted
- We should estimate

$$P[Y_i|X_i, T_i = 0] - P[Y_i|X_i, T_i = 1]$$

Ascarza 2018

Most commonly used targeting method

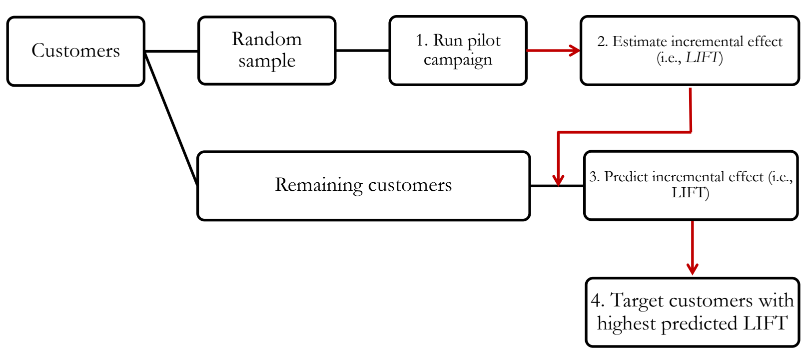


$$RISK_i = P[Y_i|X_i]$$

Logit/Probit, Lasso, Random forests, Bagging and boosting methods, Support Vector Machine… or a combination of them

Ascarza 2018

Lift-based targeting method



$$LIFT_i = P[Y_i|X_i, T_i = 0] - P[Y_i|X_i, T_i = 1]$$

Ascarza 2018

Ascarza 2018

Two approaches

Statistics, Political Sciences Economics, Biology …

Conditional Average Treatment Effect (CATE)

- Response function:
  - Interactions
  - GAM (additive models)
- Transform DV
- Causal Trees
- Causal Forest

(e.g., Heckman and Vytlacil 2005; Feller and Holmes 2009; Athey and Imbens 2016; Wager and Athey 2017; Athey, Tibshirani, and Wager 2017)

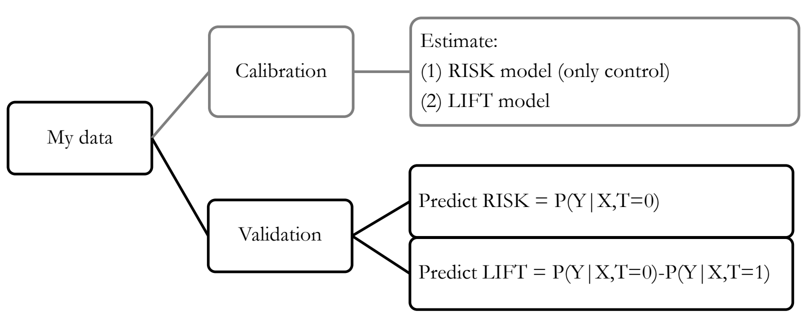
Data Mining, Practitioners, Machine Learning …

Uplift / True-lift / Incremental response / Net-modeling

- Two model approach
- Tree-based uplift
  - Additive
  - Multiplicative
- Causal K-nn
- Uplift Random Forest

(e.g., Radcliff and Surry 1999; Rzepakowski and Jaroszewicz 2012; Soltys, Jaroszewicz, and Rzepakowski 2015; Guelman et al. 2015)

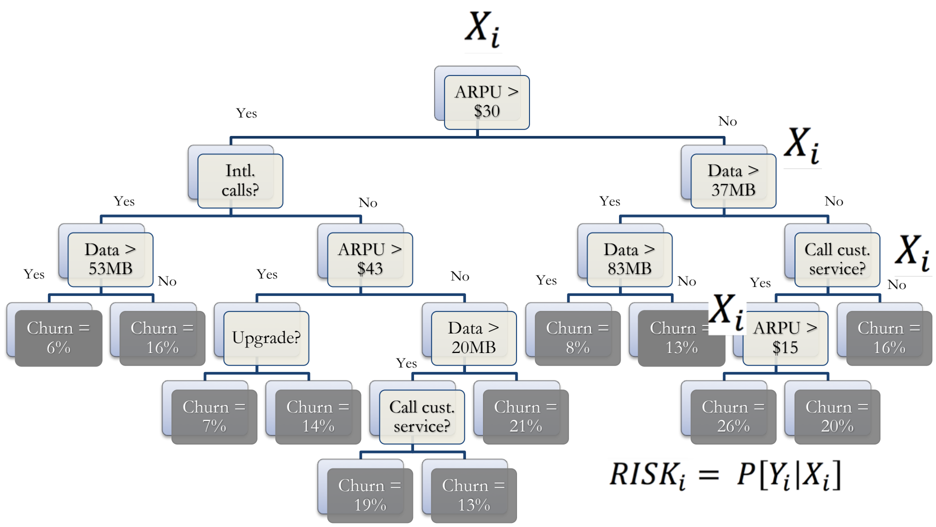
Empirical approach



(1) RISK models: GLM, LASSO, Random Forests, SVM
(2) LIFT models: GLM w/ interactions, Uplift Random Forest

Ascarza 2018

Decision tree for churn prediction

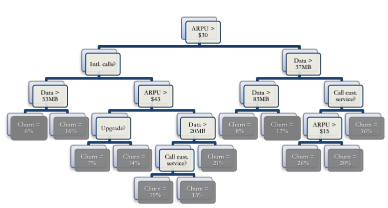


$$RISK_i = P[Y_i|X_i]$$

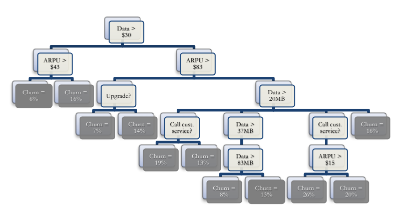
Ascarza 2018

From one tree to a (random) forest

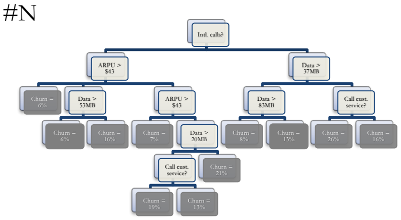
Tree #1



Tree #2



Tree #N



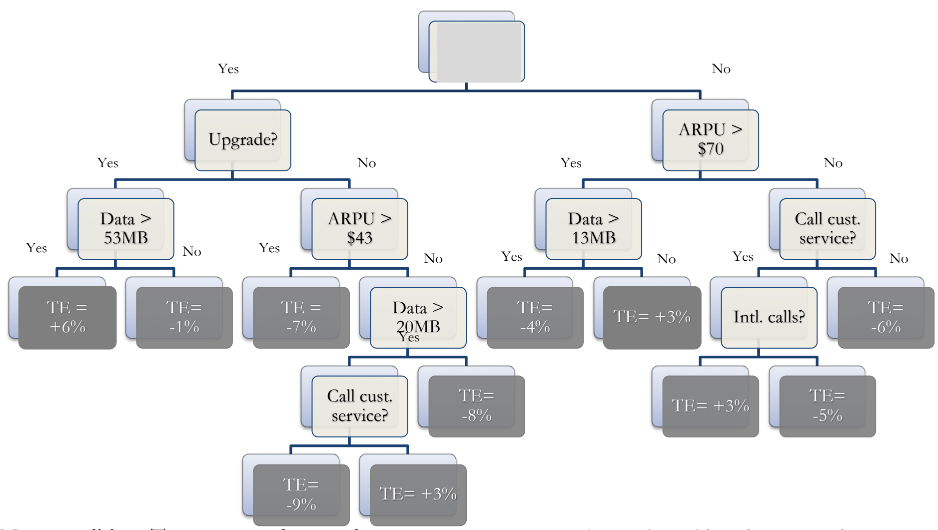
. . .

More stable (outliers "cancel out"), non-linear relationships …

(Breiman 2001)

From RISK to LIFT

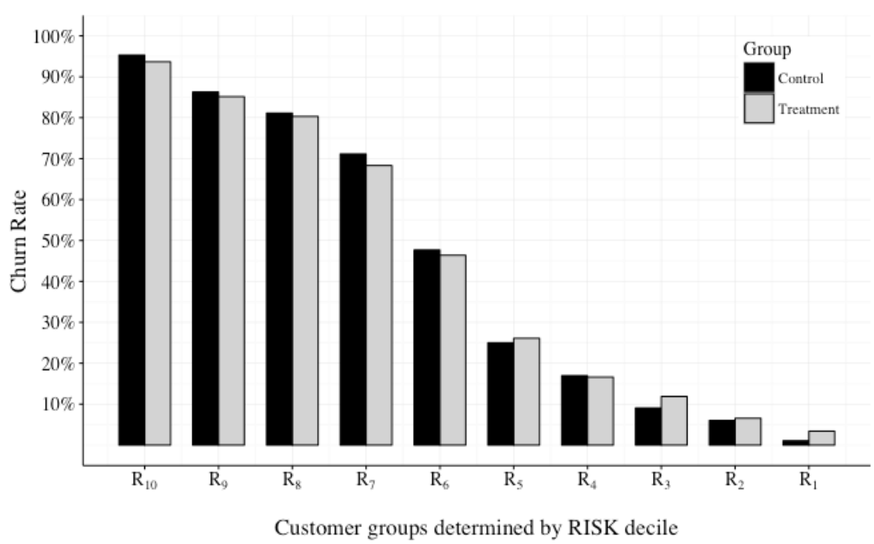
Ascarza 2018



New condition: Treatment and control observations in all end nodes.

(Rzepakowski and Jaroszewicz 2012; Guelman, Guillén and Pérez-Marín 2015)

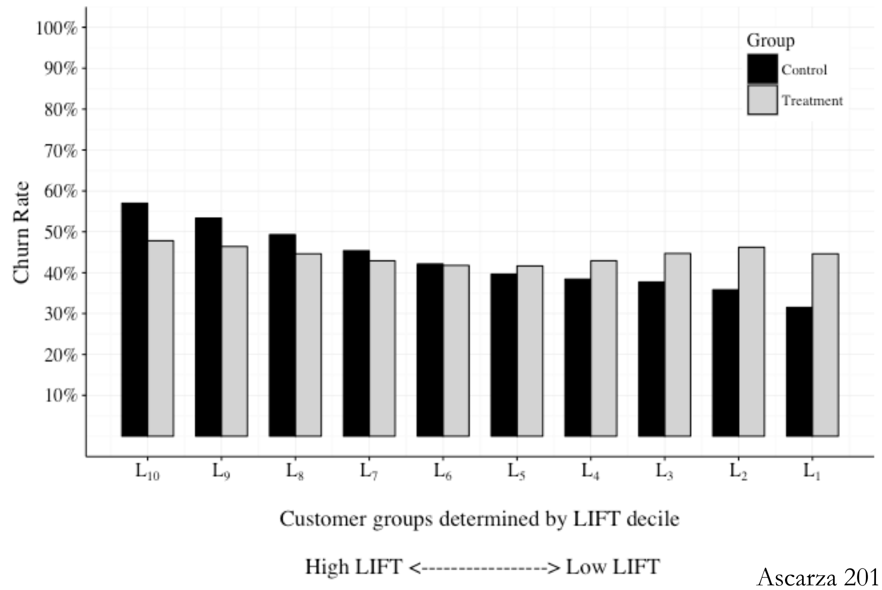
Churn rates, by levels of RISK



High RISK <----------------> Low RISK

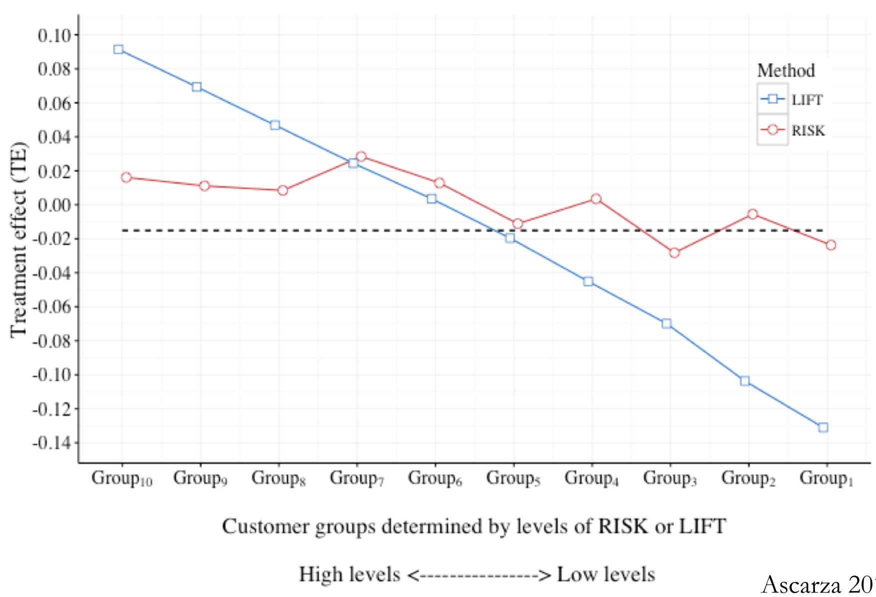
Ascarza 2018

Churn rates, by levels of LIFT



Ascarza 2018

Treatment effect



Ascarza 2018