Machine Learning & Targeted Marketing

Stephan Seiler (Imperial College London & CEPR)

Based partly on work with ...
Adam Smith (UCL School of Management)
Ishant Aggarwal (Lloyds Bank)

Katia Campo Retailing symposium 2023

How to best design personalized marketing?

- Targeting and personalization is increasing
  - Better data
  - Technological advances

- **This talk:** Combining machine learning with causal inference & managerial objectives
  - Targeting and incrementality
  - Deriving policy from an economic objective function
  - Choosing an ML model that generates the best policy

Incrementality

Example: Churn Management (Ascarza, 2019)

Marketers can also use big data to identify which customers are at highest risk of churn—and re-engage them before they defect.

—AIMIA Institute (Rogers 2013)

The challenge, of course, is to identify customers who are at the highest risk of churn before they switch to another carrier.

—*Analytics Magazine* (2016)

More sophisticated predictive analytics software use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

—"Customer Attrition," Wikipedia

Predicting Churn Risk

- Current industry approach
  - Use historic data customers to understand risk of churn
  - Then target current customers with highest risk
  - Could be done via regularized regression (or any other ML approach)

$$Retention_i = Z_i'\beta + \varepsilon_i$$

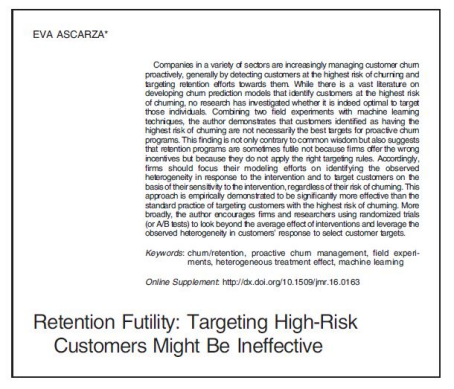
- What does this regression tell us?
  - What customer characteristics predict whether customer is likely to churn
  - We can target high risk customer with marketing (e.g. reminder, discount)
  - But: not clear that high risk customers are sensitive to marketing

Causal inference + Prediction

- Combine prediction model with A/B test
  - Denote $T_i = 1$ if customer receives marketing action ($T_i = 0$ otherwise)
  - Run regression that estimates effect heterogeneity
  - Then compute conditional average treatment effect (CATE) for a given set of characteristics $\tau(Z_i)$

$$
\begin{aligned}
Retention_i &= Z_i'\beta + (T_i \times Z_i)'\gamma + \varepsilon_i \\
\tau(Z_i) &= \mathbb{E}[Retention_i | Z_i, T_i = 1] - \mathbb{E}[Retention_i | Z_i, T_i = 0]
\end{aligned}
$$

Incrementality really matters



EVA ASCARZA*

Companies in a variety of sectors are increasingly managing customer churn proactively, generally by detecting customers at the highest risk of churning and targeting retention efforts towards them. While there is a vast literature on developing churn prediction models that identify customers at the highest risk of churning, no research has investigated whether it is indeed optimal to target those individuals. Combining two field experiments with machine learning techniques, the author demonstrates that customers identified as having the highest risk of churning are not necessarily the best targets for proactive churn programs. This finding is not only contrary to common wisdom but also suggests that retention programs are sometimes futile not because firms offer the wrong incentives but because they do not apply the right targeting rules. Accordingly, firms should focus their modeling efforts on identifying the observed heterogeneity in response to the intervention and to target customers on the basis of their sensitivity to the intervention, regardless of their risk of churning. This approach is empirically demonstrated to be significantly more effective than the standard practice of targeting customers with the highest risk of churning. More broadly, the author encourages firms and researchers using randomized trials (or A/B tests) to look beyond the average effect of interventions and leverage the observed heterogeneity in customers' response to select customer targets.

Keywords: churn/retention, proactive churn management, field experiments, heterogeneous treatment effect, machine learning

Online Supplement: http://dx.doi.org/10.1509/jmr.16.0163

Retention Futility: Targeting High-Risk Customers Might Be Ineffective

- Ascarza (2019) uses data from two companies and compares baseline risk targeting to the incremental approach
  - Baseline Risk: target top 10% of customers with highest risk
  - Incremental: target top 10% with highest treatment effect (i.e. most responsive to marketing)
- Incremental leads to 5-times larger reduction in churn

From Prediction to Policy

How to translate prediction into marketing policy?

- How can we derive a policy?
  - Ad hoc: target top 10% of customers with highest responsiveness (why 10%??)
  - Explicitly write down economic objective function

- Optimal policy for churn management setting:

$$\begin{aligned} \pi_i(Z_i, T_i = 1) &= Pr(Retention_i = 1 | Z_i, T_i = 1) \times Fee - AdCost \\ \pi_i(Z_i, T_i = 0) &= Pr(Retention_i = 1 | Z_i, T_i = 0) \times Fee \end{aligned}$$

$$\Delta\pi_i(Z_i) = \tau(Z_i) \times Fee - AdCost$$

Optimal Policy Derivation

$$\Delta\pi_i(Z_i) \quad = \quad \tau(Z_i) \times Fee - AdCost$$

- So what is the optimal policy?
  - Target all consumers with $\Delta\pi_i(Z_i) > 0$, i.e. $\tau(Z_i) > AdCost/Fee$
  - Justifies targeting consumers with highest treatment effect
  - Derives explicit threshold for how many consumer should be targeted!

Model (and Policy) Choice

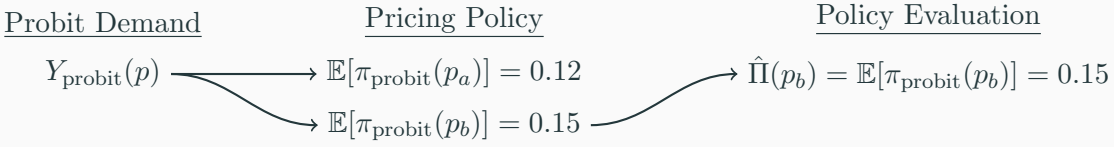
How to choose the best model?

$$\Delta\pi_i(Z_i) \quad = \quad \hat{\tau}(Z_i) \times Fee - AdCost$$

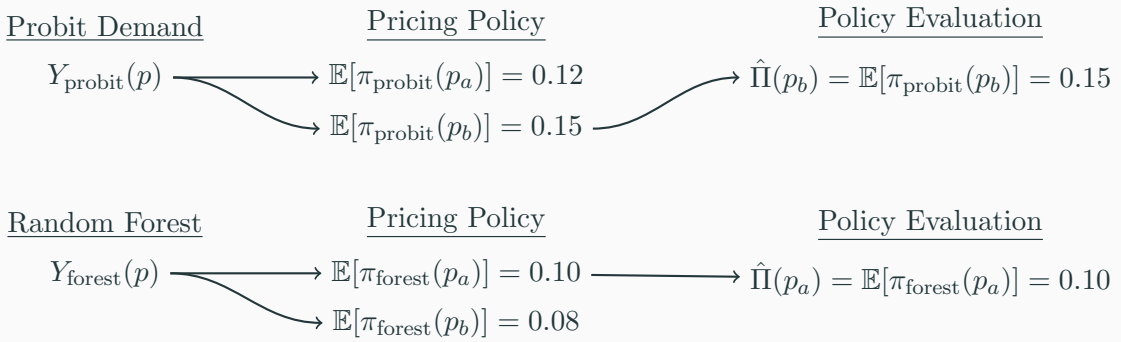
- We know how to derive the optimal policy conditional on knowing the conditional average treatment effect
- But we need to estimate $\hat{\tau}(Z_i)$ first
  - Which data inputs Z_i to use (demographics, past behavior, ...)
  - Choose an estimation method (lasso, random forest, neural net ...)
- How do we decide which model "works best"?
  - Typical ML approach: compare model out-of-sample fit
  - But: fit comparison has no direct relationship to policy and profits

How to evaluate different policies

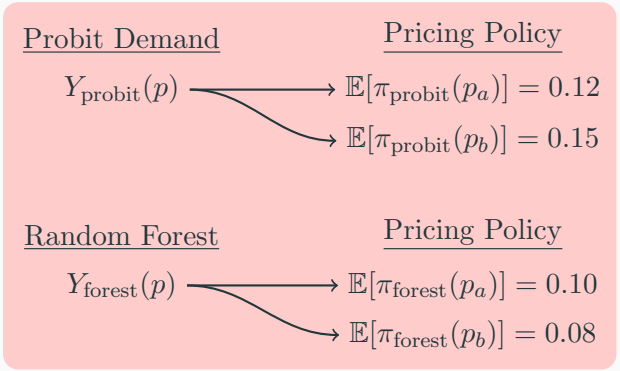
- Setting
  - Optimal pricing policy (either discount or regular price)
- Typical approach
  - Estimate model → compute pricing policy → use model to calculate profits
  - Simple example: choose between two pricing policies p_a or p_b
  - Problem: model is used twice



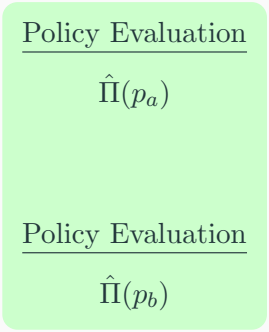
Challenge: Cross-model comparison



Our Approach: de-couple evaluation



Demand Estimation & Policy Generation
(training sample)



Evaluation
(test sample)

Solution: Do Evaluation "In-Sample"

- High-level idea
  - We use only observations where (price observed in the data) = (price prescribed by the policy)
  - Then we re-weight observations to account for rate at which prices don't match
  - $\rightarrow$ Inverse probability weighted profit estimator

Simple Example

- Simple example (based on time series for one consumer)
  - Two price levels: regular and discount
  - Consumer i is observed for $T_i = 30$ trips, 20 at regular price, 10 at discounted price
- How to use our estimator
  - If regular price is prescribed: only 20 observations are usable
  - Dividing by 2/3 re-scales them to 30 observations
  - We then divide by $T_i = 30$ to obtain consumer-specific average profits
  - If prescribed price is discount: use 10 observations, divide by 1/3

IPW Estimator

A targeting policy maps customer characteristics $\mathbf{z}_i \in \mathcal{Z}$ into prices $p \in \mathcal{P}$:

$$d : \mathcal{Z} \to \mathcal{P}$$

Inverse probability weighted estimator (for customer i):

$$\hat{\Pi}_i(d) = \frac{1}{T_i} \sum_t \sum_{p \in \mathcal{P}} \mathbf{1}(d(\mathbf{z}_i) = p) \times \frac{\mathbf{1}(p_{it} = p)}{e_p} \times \pi_{it}(p)$$

$$e_p = \mathbb{P}(p_{it} = p)$$

- Only use rows where a person sees their targeted price $d(\mathbf{z}_i) = p$
- Need to re-weight by propensity to see a given price

→ Estimator is independent of any demand-side modeling assumptions!

Smith, Seiler, Aggarwal (2022)

What do we do in the paper?

- Price targeting w/ consumer panel data
  - Specifically: panel data of mayo purchases
- Derive pricing policies from different models of demand
- Compare profitability to measure value of data inputs and model flexibility

Set of models we estimate

- "Traditional" approach in marketing
  - Bayesian logit choice model
- Machine learning models
  - Lasso regression
  - Elastic net
  - Neural Network
  - Deep Neural Network
  - Random forest
  - XGBoost
  - k-nearest-neighbor
- Use different data inputs in each model:
  - Basic demographics, extended demos, purchase histories

Model comparison: Average profits per 100 customers

	Customer Characteristics		
Policy	demos	+ more demos	+ purchase vars
Targeted			
Logit (normal heter.)	6.47	6.16	
Logit (mixtures of normals)	6.50	5.99	
Lasso	5.52	5.71	6.33
Elastic Net	5.63	5.68	6.33
Neural Net	5.61	4.35	6.33
k-nn	4.38	4.78	5.68
Random Forest	5.15	4.54	6.08
Blanket	5.53		
None	4.52		

- We use blanket coupon profits as benchmark
- Expect targeting policies to do better (but not guaranteed)

Model comparison: Average profits per 100 customers

	Customer Characteristics		
Policy	demos	+ more demos	+ purchase vars
Targeted			
Logit (normal heter.)	6.47	6.16	
Logit (mixtures of normals)	**6.50**	5.99	
Lasso	5.52	5.71	6.33
Elastic Net	5.63	5.68	6.33
Neural Net	5.61	4.35	6.33
k-nn	4.38	4.78	5.68
Random Forest	5.15	4.54	6.08
Blanket	5.53		
None	4.52		

- Logit (with basic demos) performs well compared to ML models
- 17.5% improvement for logit, only 3% for k-nn

Model comparison: Average profits per 100 customers

	Customer Characteristics		
Policy	demos	+ more demos	+ purchase vars
Targeted			
Logit (normal heter.)	6.47	6.16	
Logit (mixtures of normals)	6.50	5.99	
Lasso	5.52	5.71	6.33
Elastic Net	5.63	5.68	6.33
Neural Net	5.61	4.35	6.33
k-nn	4.38	4.78	5.68
Random Forest	5.15	4.54	6.08
Blanket	5.53		
None	4.52		

- Adding demographics has small (sometimes detrimental) impact
- Adding purchase variables increases profits across all models

Model comparison: Average profits per 100 customers

	Customer Characteristics		
Policy	demos	+ more demos	+ purchase vars
Targeted			
Logit (normal heter.)	6.47	6.16	
Logit (mixtures of normals)	**6.50**	5.99	
Lasso	5.52	5.71	6.33
Elastic Net	5.63	5.68	6.33
Neural Net	5.61	**4.35**	6.33
k-nn	4.38	4.78	5.68
Random Forest	5.15	4.54	6.08
Blanket	5.53		
None	4.52		

- Large heterogeneity across models
- From 21% profit loss to 17.5% gain

Profits and model fit

	Customer Characteristics		
Policy	demos	+ more demos	+ purchase vars
Targeted			
Logit (normal heter.)	6.47	6.16	
Logit (mixtures of normals)	6.50	5.99	
Lasso	5.52	5.71	6.33
Elastic Net	5.63	5.68	6.33
Neural Net	5.61	4.35	6.33
k-nn	4.38	4.78	5.68
Random Forest	5.15	4.54	6.08
Blanket	5.53		
None	4.52		

- We generate a similar table for fit statistics (hit probs/log-likelihood)
- Out-of-sample profits uncorrelated with both fit stats

Summary of findings

- Demographics less valuable than purchase histories
- Large heterogeneity across models: From 21% profit loss to 17.5% gain relative to blanket coupon
  - Bayesian hierarchical choice model very attractive although not used much in practice
- Fit statistics provide poor guidance to model performance

Conclusion

Re-cap: machine learning and targeted marketing

- Need to combine ML techniques with insights from causal inference and marketing / economic theory
  - Focus on differences in causal effects across customers ($\rightarrow$ incrementality)
  - Evaluate models based on decision-theoretic objective function
  - Both steps lead to large changes in performance

- But there are some additional costs
  - We need A/B test or other causal approach for incremental targeting
  - Need model-free eval framework to compare profits

Thank You !!!