Machine-Learning the Impacts of Behavioral Interventions

Evidence from Household Energy Use

Samuel Stolper (joint with Chris Knittel and Leila Safavi)

University of Michigan

March 30th, 2019

Machine learning to improve program evaluation

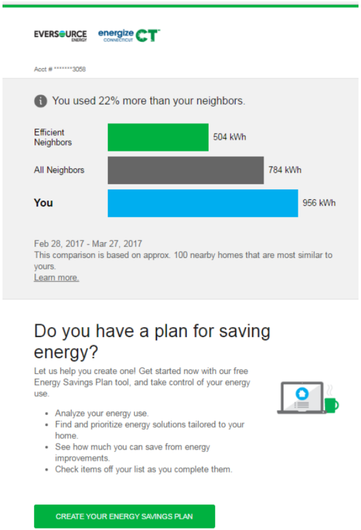
Understanding treatment effect heterogeneity facilitates improvements to program effectiveness

- Can selectively target those who respond “best”
- Can tailor treatment where it is not having the desired effect

Our aim: use random forests to estimate the distribution of responses to a widely used behavioral nudge

- The “Home Energy Report” (HER), which aims to encourage energy efficiency

Report objective and components



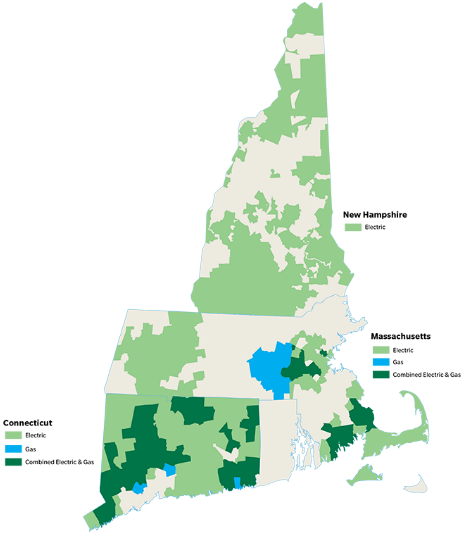
- The aim:
  - Nudge consumers to reduce usage
  - Increase customer satisfaction
- The format
  - Social comparison of usage
  - Ways to save

Background

Some facts about HERs:

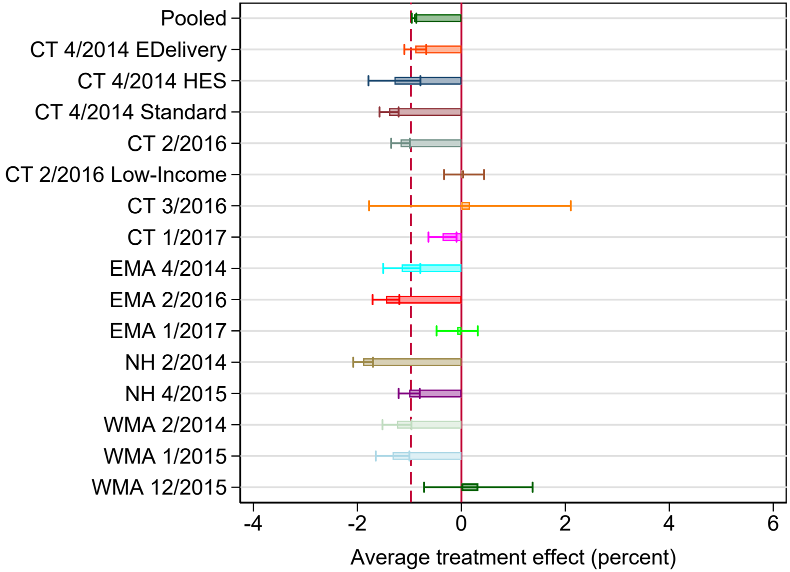
- Used by over a hundred utilities in at least nine countries
- Repeatedly been proven effective at reducing consumption on average
  - ATEs: 1-2% of monthly household consumption (Allcott 2011; Ayres et al. 2013)
- Some evidence of heterogeneity in impacts
  - Allcott (2011); Costa and Kahn (2013); Allcott and Kessler (2017)

Context

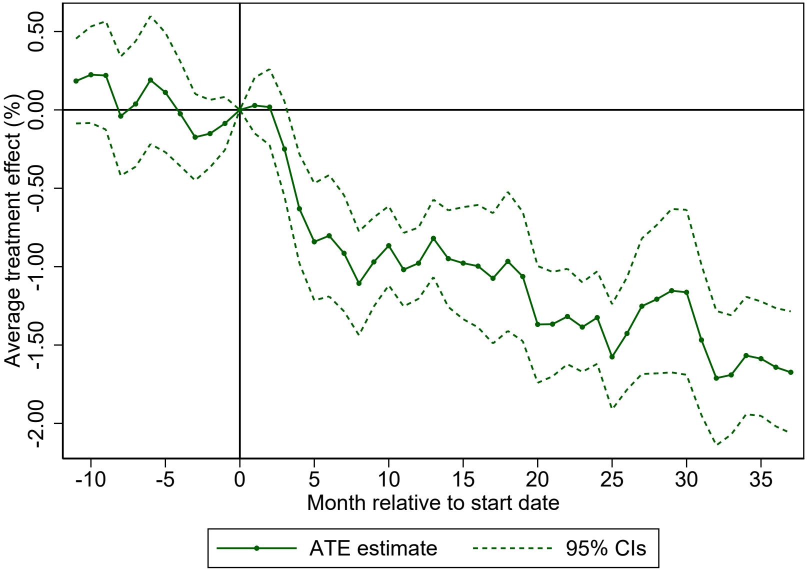


- Eversource residential electricity customers
  - Monthly usage (kWh) from 2013-2018
  - 900k households enrolled in an experiment
    - 50m household-months
  - Household characteristics from Experian

ATEs by wave: consumption



Event study of pooled experimental waves: consumption



Random forest algorithm

We use the generalized random forest algorithm, developed by Wager, Tibshirani, and Athey (2018)

- Grow a collection of (10,000) trees using recursive partitioning
  - Each tree splits the sample into “leaves” defined by ranges of characteristic values
  - Splits are made to maximize cross-split differences in ATE

- Predict household i’s treatment effect using a weighted average of nearest neighbors
  - Weights equal to the likelihood of being in the same leaf as household i

Tree-growing procedure

1. Draw random 50% sample of households for use in tree-growing
   - Split the sample into a “training set” and an “estimation set” of equal size
2. Draw a random subset of household characteristics to use in splitting
   - This and (1) de-correlate the trees
3. Split the training set recursively to create a tree, whose terminal leaves identify unique, disjoint sets of characteristics
4. Match estimation-set households to leaves based on their characteristics
   - So one set is used to grow tree structure; the other is used to estimate ATEs
5. Estimate within-leaf ATEs

Implementation details

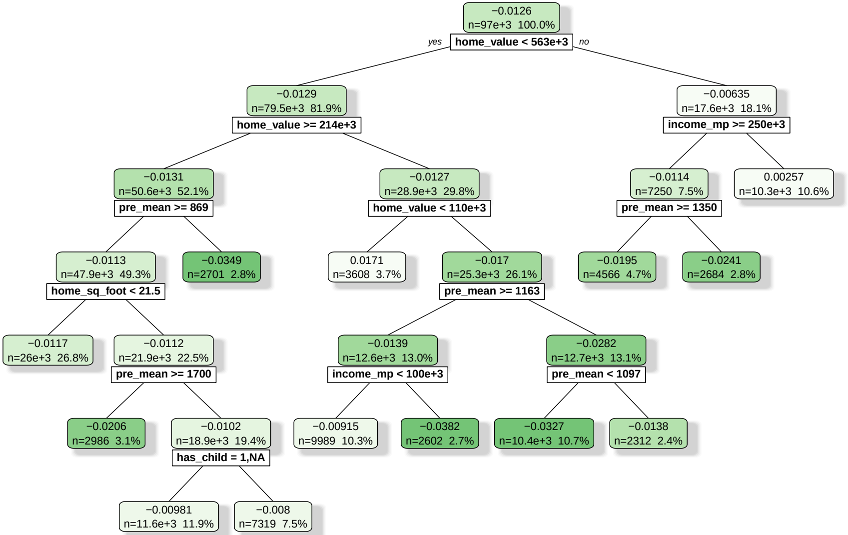
Parameter choices

- Size of sample and characteristic vector drawn
- Minimum node size
- Imbalance limit and penalty

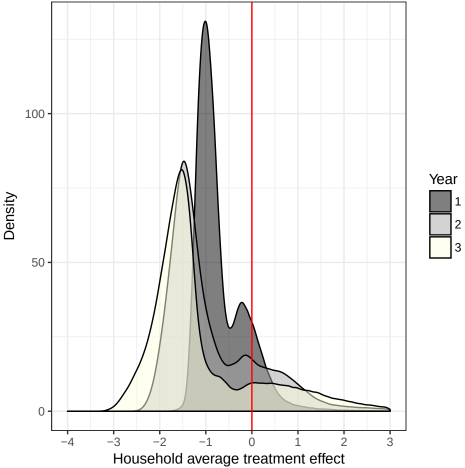
Pre-processing:

- Dependent variable: post - pre consumption
  - Regress Y and W on characteristics and wave FE and use residuals
- Weights: inverse p(treatment) by wave

A sample tree

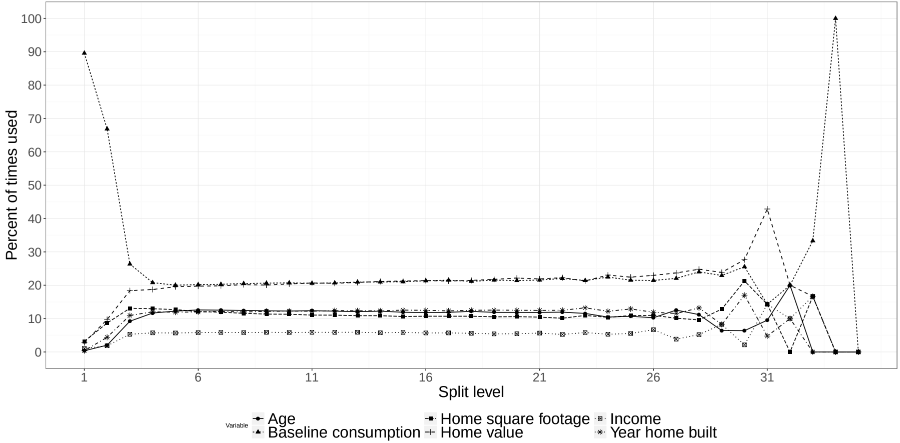


Kernel density of household predictions

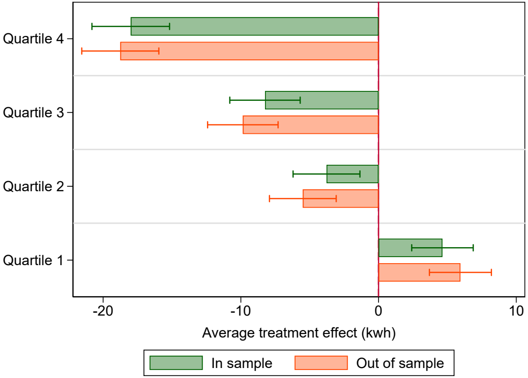


- Multiple distinct peaks
- Long right tail of positive treatment effects
- Peak-shifting over time

Usage of characteristics in random forest



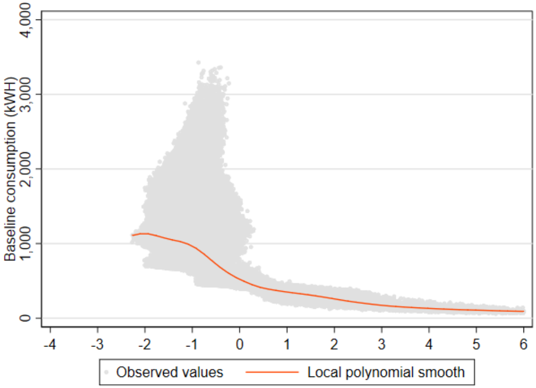
Test of out-of-sample performance



- Grow forest using only half the sample
  - Predict treatment effects in hold-out sample using forest
- Regress usage on treatment X (predicted ATE quartile)
- Forest appears to do well out-of-sample

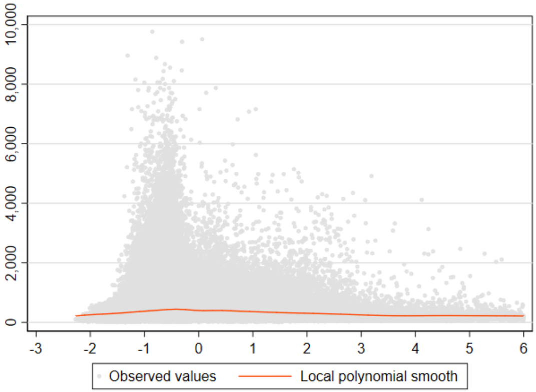
$$Y_{it} = \alpha_0 + \alpha_1 T_{it} + \Sigma_{j=2}^{4}\left(\alpha_j T_{iwt} * 1[Q_i = j]\right) + \theta_i + \omega_t + e_{it}$$

Predicted treatment effect vs. baseline usage



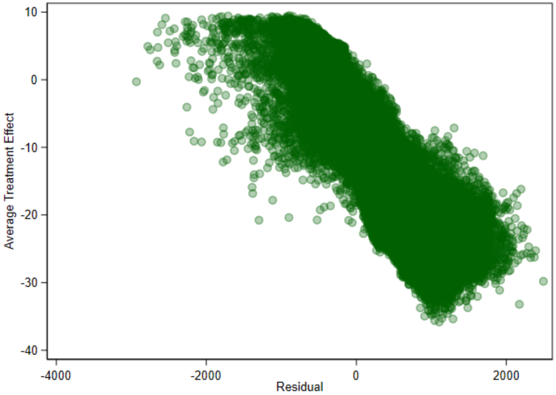
- Positive treatment effects are exclusively found among households with the lowest baseline consumption
- Above bottom quartile, tight relationship breaks down

Predicted treatment effect vs. home value



- No slope to the treatment effect - home value relationship
- But the largest drops in consumption are confined to the low end of the home value distribution

Predicted treatment effect vs. pre-consumption residual



- “Residual” indicates consumption *relative* to an average household with similar characteristics
  - This may be correlated with social comparison messaging
- Graph suggests a “boomerang effect”
- Graph suggest a continuous relationship b/w treatment effect and residual

Learning about machine learning

- Opower experiments span multiple states and time periods
  - Can we leverage this fact to test the “pace” of machine learning?
  - Examine how performance evolves as we add waves to the training set
  - Compare different prediction methods:
    - Random forests
    - LASSO
    - Traditional regression

The “horse race” procedure

1. Divide waves into 3 groups, by time period
2. Build predictive model for the first group chronologically using each method
3. Predict treatment effects among HHs in next group
4. Estimate “actual” effects among these same HHs with diff-in-diff

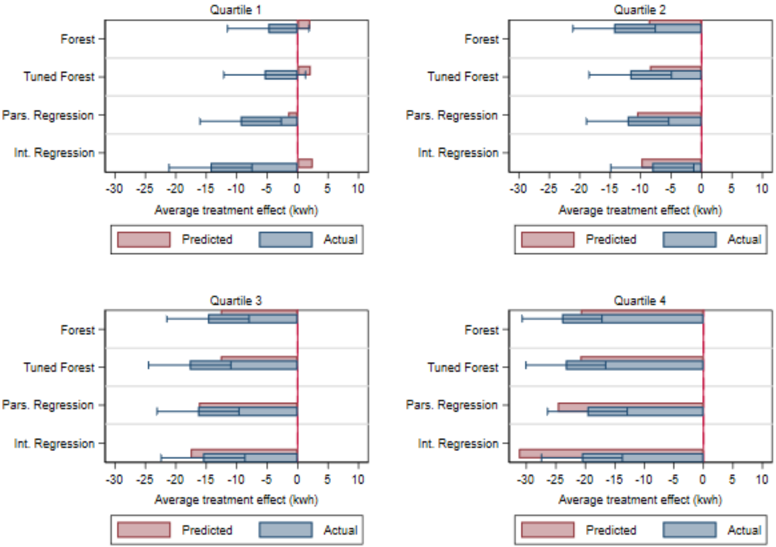
Can replicate this predicted vs. actual comparison in group 3, using models built from groups 1 and 2

Comparing predicted to actual by quartile

Multiple metrics for performance:

1. Relationship of actual ATE magnitude to predicted quartile
2. Magnitude of actual ATE in top quartile
3. Accuracy of predicted ATE in each quartile

Horserace performance for Group 2



Horserace performance for Group 3