Using Machine Learning to Target Treatment: The Case of Household Energy Use

Christopher R. Knittel Samuel Stolper*

November 26, 2019

Abstract

We use causal forests to evaluate the heterogeneous treatment effects (TEs) of repeated behavioral nudges towards household energy conservation. The average response is a monthly electricity reduction of 9 kilowatt-hours (kWh), but the full distribution of responses ranges from -30 to +10 kWh. Selective targeting of treatment using the forest raises social net benefits by 12-120 percent, depending on the year and welfare function. Pre-treatment consumption and home value are the strongest predictors of treatment effect. We find suggestive evidence of a "boomerang effect": households with lower consumption than similar neighbors are the ones with positive TE estimates.

Keywords: machine learning, program evaluation, targeting, energy efficiency.

JEL Codes: C53; Q40; D90

*Knittel: George P. Shultz Professor Sloan School of Management, Director Center for Energy and Environmental Policy Research, Co-Director Electric Power Systems Low Carbon Energy Center, MIT and NBER, knittel@mit.edu; Stolper: University of Michigan School for Environment and Sustainability, sstolper@umich.edu. Leila Safavi and Paula Meloni provided outstanding research assistance. We thank Hunt Allcott and seminar participants at Carnegie Mellon, UC Berkeley, University of Connecticut, Yale, and MIT for valuable feedback. Alberto Abadie, Jonathan Davis, Peter Christensen, Stefan Wager, and Susan Athey gave valuable advice on implementing the causal forest algorithm. This research would not be possible without the work of Amy Findlay and colleagues at Eversource, who supplied the necessary data and background on the Home Energy Report Program.

Introduction

The rise of randomized controlled trials (RCTs) in economics has produced a wealth of evidence on the average causal effect of a great number of social and private-sector programs.[1] Yet such programs quite often have widely divergent impacts across the treated population. Understanding how different subgroups respond to a given treatment has the potential to unlock large increases in program effectiveness, by allowing for improved targeting of the existing treatment (i.e., identifying *whom* to treat) as well as improved design of the treatment itself (e.g., tailoring treatment for specific subgroups).

Machine-learning (ML) methods are an attractive option for estimating heterogeneous treatment effects (Athey and Imbens, 2017). They offer disciplined ways to search non-parametrically for heterogeneity, and are especially useful when the researcher observes a large number of baseline characteristics. They also offer tools for minimizing overfitting and thus maximizing out-of-sample predictive power. However, ML algorithms have traditionally been built for *prediction* of y from x, rather than *parameter estimation* of treatment effects β (Mullainathan and Spiess, 2017). Consequently, there is an active body of research on the use of ML algorithms for causal inference (e.g., Imai and Ratkovic, 2013; Chernozhukov et al., 2018). Tree-based methods (Breiman et al., 1984; Breiman, 2001) are one class of ML algorithms in which significant progress has been made. Athey and Imbens (2016) propose methods for causal estimation of conditional average treatment effects (CATEs) from regression trees, which they denote "causal tree" estimators. Wager and Athey (2018) extend these methods to the estimation of "causal forests."

In this paper, we apply the causal forest algorithm to the evaluation of a series of large-scale randomized experiments in household energy use. We predict treatment effects among more than 900,000 households and investigate the role of observed and unobserved household characteristics in determining outcomes. To illustrate the value of forest-derived CATEs, we measure the potential welfare gains from selective targeting of treatment to maximize, alternatively, social and private (i.e., electric utility) objective functions. Finally, we construct tests of internal and external validity to assess absolute and relative performance of the causal forest method in this context.

Our results borrow from, build on, and add to an emerging literature on empirical machine learning (e.g., Davis and Heller, 2017b; Burlig et al., 2017; Kleinberg et al., 2017; Hussam et al., 2018). Davis and Heller (2017b) are the first to apply the causal forest algorithm to impact evaluation of a randomized experiment—in their case, a youth summer employment program. In comparison, we investigate the heterogeneous impacts of behavioral "nudges" towards energy efficiency and using a much larger (10x) sample. Our findings additionally relate to a large literature on the treatment effects of behavioral nudges, which have wide application ranging from

[1] The list of RCTs in economics is far too long to detail here, but see, for example, Duflo et al. (2007).

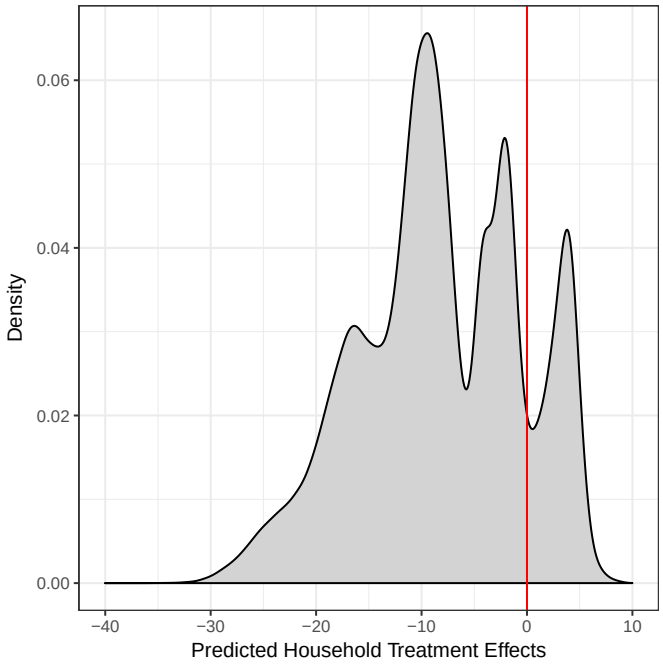
water use (Ferraro and Price, 2013), to tax compliance (Kettle et al., 2016), to charitable giving (Andreoni et al., 2017).

Our empirical setting is the retail electricity service territory of Eversource, the largest electric utility in New England. Eversource's flagship behavioral energy efficiency product is the Home Energy Report (HER), a short, regular mailing that compares a customer's electricity (and natural gas) consumption to that of similar, nearby households and provides information on ways to save energy. Since 2011, the company has been experimentally rolling out HER programming in waves. Our program evaluation leverages data from 15 experimental waves covering 902,581 Eversource residential customers. We observe monthly household electricity consumption from 2013-2018 and cross-sectional characteristics pertaining to homes and their occupants. This context is especially ripe for estimation of heterogeneous treatment effects for two reasons: first, the large overall sample size available to us provides greater statistical power than is normal in randomized control trials (RCTs); and second, intuition and empirical evidence alike suggest that HERs likely induce a wide variety of behavioral responses (Allcott, 2011; Costa and Kahn, 2013).

Our central estimate of the pooled average treatment effect (ATE) across all Opower program waves—which we estimate via panel regression—is a reduction in monthly electricity usage of 9 kilowatt-hours (kWh), or 1 percent. This ATE is consistent with the lower end of the range of existing estimates (Allcott, 2011; Ayres et al., 2013; Allcott, 2015). However, the pooled average masks heterogeneity across waves and over time, because sample makeup varies across waves and the household response to HERs evolves with repetition, respectively. Our event study of Eversource's HER program shows a monotonic rise in the absolute value of month-specific ATEs from months 1 through 5 of the intervention and a further net rise in the latter half of program year 1. There is no evidence of attenuation of program impacts in years 2 and 3; if anything, rather, the reductions in electricity consumption continue to increase. The year-three pooled ATE in our sample is -14 kWh, or -1.5 percent.

Our causal forest methods reveal significant heterogeneity and potential for efficiency improvements. In Figure 1 below, we show the estimated distribution of household-level, three-year average treatment effects. At least three distinct modes are apparent, and the estimates range from roughly -30 to +10 kWh per month. What accounts for this sizable heterogeneity? We find that the most commonly-used household characteristics in the forest are baseline (i.e., pre-treatment) consumption and home value, which indicates that these variables have significant predictive power. In addition, we find suggestive evidence that the social comparison embedded in HERs induces a "boomerang effect" (Bhanot, 2017; Schultz et al., 2007; Byrne et al., 2018): the households that are predicted to raise their consumption appear to be the ones that receive "positive" messaging about their own consumption relative to others (i.e., are told that they are consuming less than

Figure 1: Distribution of Predicted Treatment Effects: 3-Year Average



Notes: Estimates are from a causal forest grown from all households with three years of post-treatment data, using three-year average consumption as the dependent variable. See Section 2.2 for details.

other, similar households).

In our targeting exercise, we compare the monetized net benefits of the actual HER distribution to the net benefits of sending reports only to those households for which benefits exceed the marginal cost of sending reports. We replicate this comparison with three different objective functions: one inspired by the utility's desire to help customers save money, and two that value energy conservation according to its social value. In all cases, program net benefits can be increased significantly through selective targeting; in every program year and with every objective function, treatment leads to negative net benefits among at least 15 percent of households. Between \$500K and \$1.2M of deadweight loss can hypothetically be avoided each year. These avoided losses are particularly large as a percentage of the program's social net benefits: for instance, according to our preferred social objective function, the welfare gains from targeting are 66 percent in year 1, 36 percent in year 2, and 25 percent in year 3.

To check for internal validity of our forest results, we split the full sample into two random subsamples, grow a forest with one of them, and compare actual CATEs (estimated via Ordinary Least Squares regression) in the other with predictions derived from the forest. We observe small differences and conclude that the forest produces internally valid estimates. To check for external validity, we use a similar procedure but split the full sample non-randomly into three chronological groupings and additionally compare the forest to lasso and traditional regression methods (without

"learning"). All methods perform well in this test when in- and out-of-sample households are relatively similar, but prediction errors rise significantly when these two sets of households exhibit relatively more differences. These results suggest that selective targeting may be difficult at the outset of an intervention, unless a previously-treated sample with similar characteristics is available. We find, however, that one can learn a lot from the first year of treatment; household-specific responses are persistent over time. From the stylized perspectives of society and the utility, 85 and 99 percent, respectively, of the welfare gains achieved in program years 2 and 3 through targeting with perfect information can be realized by sending HERs only to those households that show positive net benefits in the first year.

1 Empirical Context

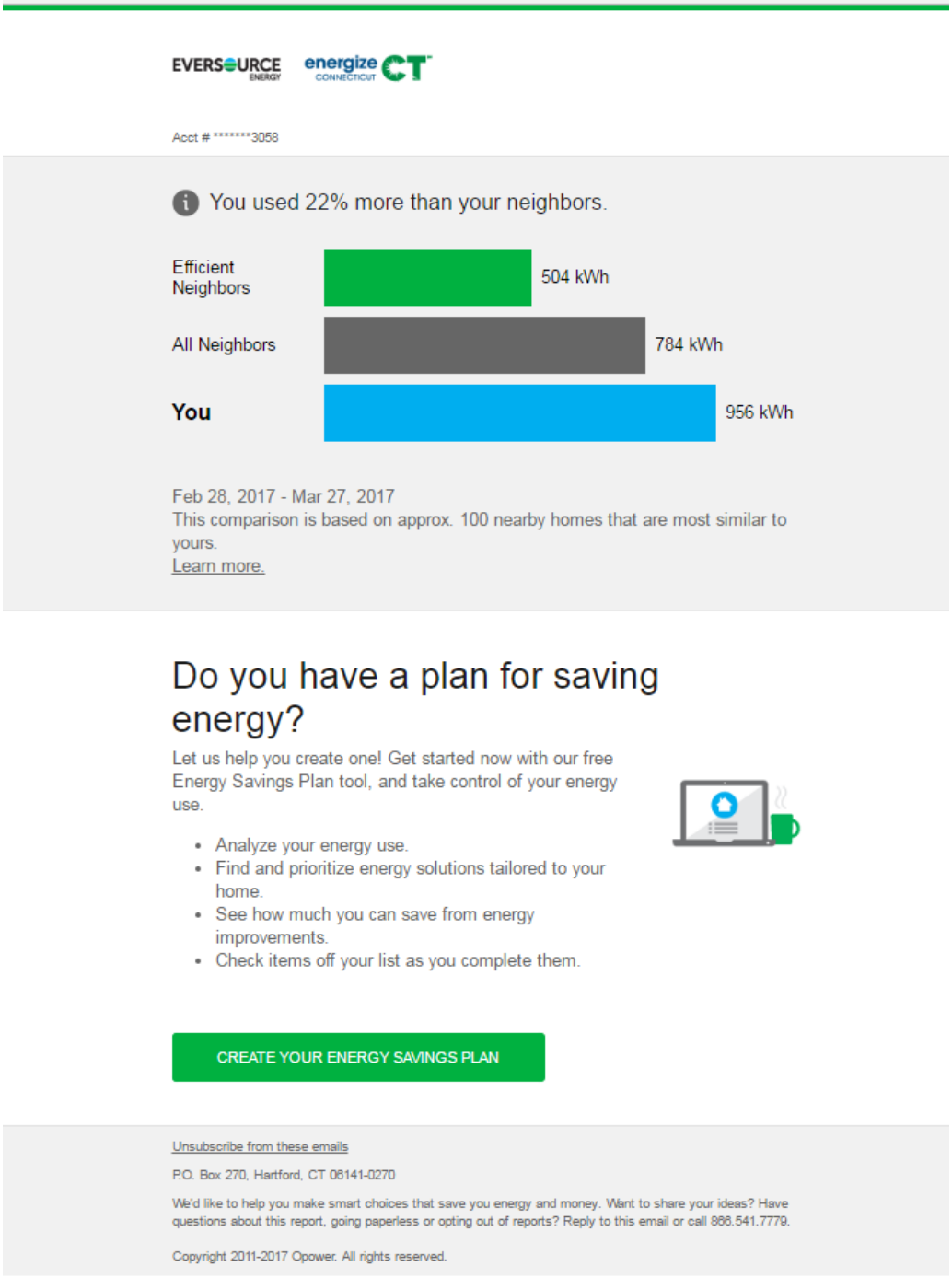
The Home Energy Report (HER) was developed by Opower and rolled out via randomized control trials in participating electric utility service territories beginning in 2008. The initial motivation for the reports came from a field experiment in San Marcos, CA carried out by Schultz et al. (2007), who found social norms messaging to be effective in reducing home energy consumption. The Opower HER is characterized by two components. The first is information about absolute and relative energy consumption. Usually, the HER lists a household's consumption in the last month and compares it (numerically and graphically) to a sample of similar, nearby households. In the context of social norm theory, peer-rank information can serve as a non-financial incentive to "nudge" individuals towards socially desirable behavior. By providing a relevant reference point, households are able to compare their behavior to that of others when no other social standard is available, inducing convergence towards the displayed social norm (Festinger, 1954).[2] See Figure 2 for an example Eversource HER.

The second component of the HER is a set of action steps—suggestions for how to conserve energy, both through changes to a household's stock of energy-using durables and changes in the use of that capital stock. Action steps can be made accessible through a customer portal (as in Figure 2), or they can displayed directly in the report. Reports are generally sent out either monthly or quarterly. The great majority of HERs have been delivered by mail in hard-copy form, but Opower has recently experimented with email HERs. Customers can and (infrequently) do opt out of the HER program, but it is unclear how many households are aware of the opportunity to do so.

There are several potential reasons why an electric utility may choose to send HERs to its customers. Perhaps the most frequently discussed reason is compliance with energy efficiency

[2]The algorithm that identifies "similar" households is an Opower trade secret, but we believe it is a function of, at least, home location and home size.

Figure 2: Eversource Home Energy Report



Source: Eversource.

standards, which, in 26 states, requires utilities achieve a certain amount of new cost savings through energy efficiency measures every year. HERs may provide a cost-effective way to comply with such standards. Another reason to send HERs is to improve customer satisfaction by keeping households informed about their bill and ways to potentially reduce it. Research on HER impacts has, to date, focused almost exclusively on energy consumption rather than customer satisfaction, perhaps due to limitations on the latter's data availability.

Allcott (2011) studies the electricity usage impacts of the first wave of Opower experiments and estimates a short-run average treatment effect (ATE) of -2.0% (i.e., a 2% monthly reduction in electricity consumption).[3] Ayres et al. (2013) concurrently study the effects of two other Opower interventions and find ATEs of -2.1% and -1.2%, respectively (the latter is an aggregate estimate for home electricity and natural gas usage). Allcott (2015) identifies "site selection bias" in HER experiments: using results from the first ten Opower experiments to predict results in the next 100 experiments significantly overstates program effectiveness. Allcott and Rogers (2014) study the long-run impacts of HERs and shed light on the time-pattern of a household response. Initially, treated households reduce energy use right after receiving a report but slide back upwards over time until receiving the next report. This "action and backsliding" pattern dissipates over time, but the monthly conservation effect continues rising even after two years of repeated treatment. Finally, the conservation effect is relatively persistent after reports are stopped: the decay rate of the effect is 10-20% per year.

While it is intuitive that HERs' impact on actions, savings, and well-being will vary across households, there is limited evidence of such heterogeneity. Allcott (2011) finds that the treatment effect varies with pre-treatment electricity consumption: the top decile has an ATE of 6.3%, while the bottom decile's ATE is statistically indistinguishable from zero. Ayres et al. (2013) similarly find a positive correlation between pre-treatment usage and HER-induced reductions in usage. Costa and Kahn (2013) show that politically liberal households reduce energy usage in response to HERs two to four times more than politically conservative ones. Allcott and Kessler (2019) elicit willingness-to-pay for HERs and identify significant heterogeneity across households. According to correspondence with Eversource, Opower's only strategy for targeting customers for HER experimental participation is high pre-treatment consumption.

1.1 Data

We combine three types of data in order to estimate the impacts of home energy reports: household monthly electricity consumption from Eversource; treatment assignment and timing of Ev-

[3] In Allcott (2011)'s context, 2.0% is equivalent to 0.62 kilowatt-hours (kWh) per day. A reduction of this magnitude could be achieved, for example, by turning off a typical air conditioner for 37 minutes per day, or by switching off a 60-watt incandescent lightbulb for 10.4 hours per day.

ersource's HER experiments; and cross-sectional demographic and socioeconomic characteristics of participants. Eversource's service territory is divided into four regions: Eastern Massachusetts, Western Massachusetts, Connecticut, and New Hampshire. Some of its customers receive both electric and natural gas service, while others receive only one or the other; Figure 3 maps the coverage of these services. We obtained monthly electricity consumption totals (in kilowatt-hours, or kWh) for the universe of Eversource customer accounts ("households") with residential electricity service in the period from January 2013 to December 2017. The raw total number of accounts is 3,055,682.

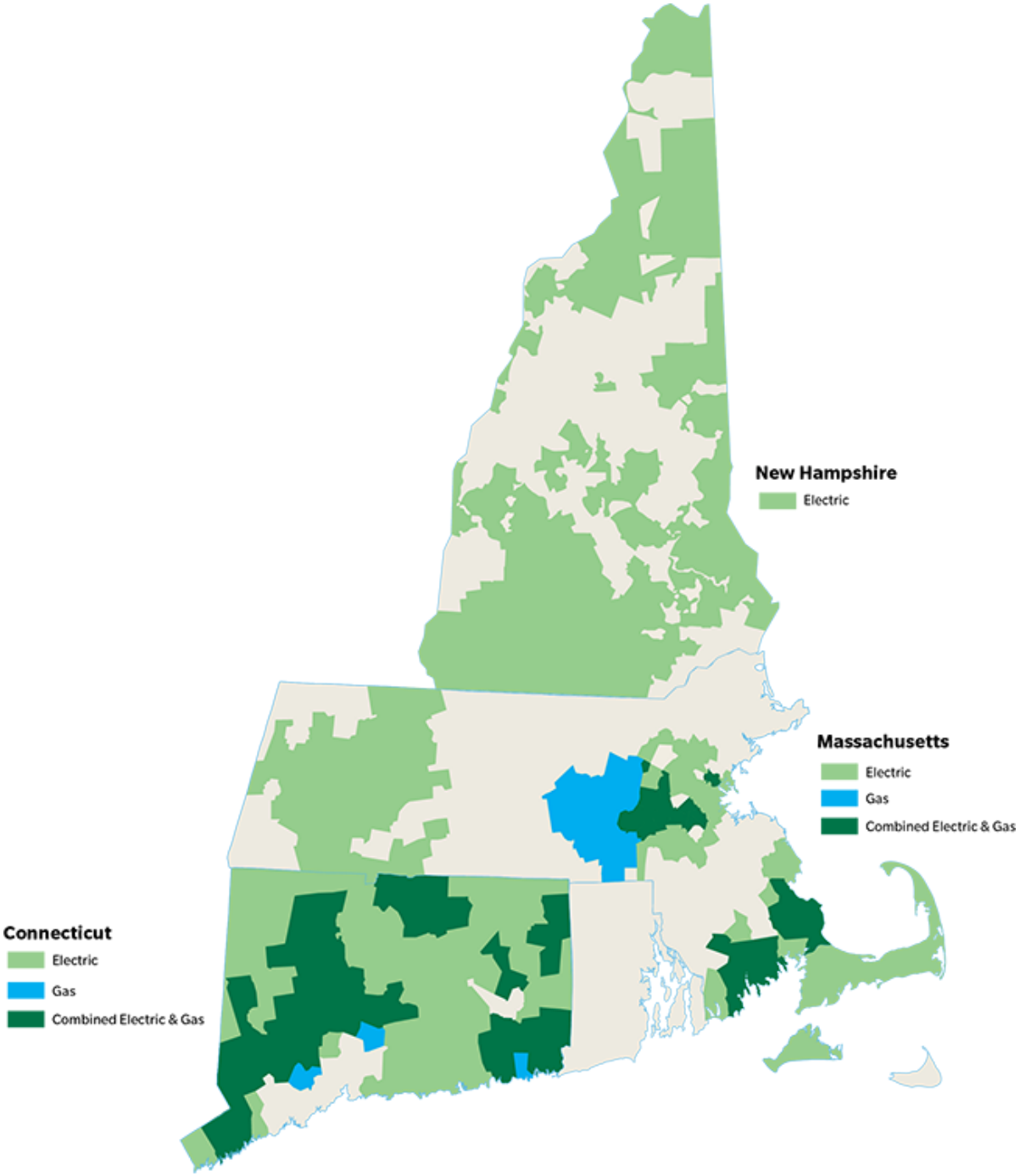
Opower has run 26 waves of home energy report experiments in the Eversource electric service area, with the earliest beginning in February 2011 and the latest beginning in January 2017. We drop 11 waves that either (a) begin outside our five-year period of observation for household energy consumption, (b) target natural gas customers, or (c) target households that have just moved into new homes (who, in these waves, receive different HERs that additionally vary over time). This leaves us with fifteen waves with which to conduct our analysis. Table 1 details the timing, location, and size of each wave that we use in our analysis. Twelve of these waves use the standard, or "base," Eversource treatment: a periodic, hard-copy mailed report showing the customer's electricity consumption last month, average consumption among "similar" nearby households, and a textual comparison of the two. Three program waves deviate from this standard treatment: one of these replaces hard-copy reports with emailed ones; another exclusively covers households that have previously received "home energy assessments" aimed at providing recommendations on how to save energy; and the third targets households with, on average, significantly lower incomes than the norm for Opower. All waves use either monthly or quarterly report frequency.[4]

We drop households with outlier values of home square footage and number of rooms, households enrolled in multiple Opower waves, and households that own multiple properties. We further limit our sample to those households for which at least 12 months of pre-experiment data and 12 months of post-experiment data are available. This leaves us with 902,581 households and a total of 49,491,297 household-monthly observations.

We combine these consumption and treatment assignment data with cross-sectional home and household characteristics from Experian, via Eversource. We include thirteen characteristics in our analysis. To capture home attributes, we use home age, value, and square footage, as well as number of rooms. To describe family size, we use the number of adult residents and an indicator for the presence of children. We further include indicators for single-family occupancy and owner occupancy. Finally, we include average pre-enrollment consumption, income, educational

[4]Table 1 shows that treatment-control ratio varies significantly across wave and is always at or above 50:50. Opower chose such high treatment probabilities in order to meet its electricity savings goals while keeping the number of waves low.

Figure 3: Eversource service territory map



Source: Eversource.com.

Table 1: Summary of experimental Home Energy Report program waves

Date	Location	Type	N	% Treatment
February 2014	New Hampshire	Base	42,709	50
February 2014	Western Massachusetts	Base	95,455	91.9
April 2014	Connecticut	E-Delivery	85,360	83.3
April 2014	Connecticut	HEA	11,883	66.4
April 2014	Connecticut	Base	199,802	91.7
April 2014	Eastern Massachusetts	Base	49,610	88.4
January 2015	Western Massachusetts	Base	24,837	71.1
April 2015	New Hampshire	Base	32,571	71.5
December 2015	Western Massachusetts	Base	11,272	86.6
February 2016	Connecticut	Base	137,896	88.1
February 2016	Connecticut	Low-Income	16,981	53
February 2016	Eastern Massachusetts	Base	59,892	76.5
March 2016	Connecticut	Base	17,395	80.0
January 2017	Connecticut	Base	69,517	75.9
January 2017	Eastern Massachusetts	Base	47,401	62.8

Notes: "Base" indicates the standard Opower treatment. "E-Delivery" indicates an email-only treatment. "HEA" indicates a sample of participants who have previously received a home energy assessment, aimed at providing recommendations on how to save energy. "Low-Income" indicates a lower-income sample of participants.

attainment, an index for "green awareness", and an indicator for take-up of a subsidized home energy assessment. We fill in missing values of these characteristics using multiple imputation (see Appendix D for details on this procedure).

Table 2 tests for covariate balance across treatment and control observations in our pooled analysis sample. Columns 1 and 2 present raw means for the characteristics that we use in our main analysis. In column 3, we calculate the difference in means for each characteristic as the coefficient from a regression of the particular variable on the treatment dummy and a set of wave fixed effects, with weights equal to inverse treatment probability by wave and standard errors clustered at the household level. Only one characteristic (home value) exhibits a statistically significant difference across treatment and control ($p = 0.07$).[5]

2 Empirical Strategy

We use conventional difference-in-differences regression, leveraging random assignment of households into treatment and control groups, to estimate average Home Energy Report program effects on electricity consumption. To test for heterogeneity in these effects and investigate the role of household characteristics in predicting them, we use the causal forest algorithm, implemented with Tibshirani et al.'s (2018) generalized random forest package. This algorithm yields a distribution of predicted, individual household impacts on consumption, as well as information about the use of each characteristic in growing the forest from which those impacts are predicted.

2.1 Estimation of average treatment effects

We use our household-monthly panel data on electricity consumption to estimate the average treatment effect via the following regression:

$$kWh_{iwt} = \alpha_1 + \alpha_2 T_{iwt} + X_i\eta + \theta_w + \omega_t + e_{iwt}, \tag{1}$$

where kWh_{iwt} is electricity consumption for household i from program wave w in year-month t. T_{iwt} is the binary treatment variable, X_i is a vector of household characteristics, and θ_w and ω_t are wave and year-month fixed effects, respectively. We cluster standard errors by wave, and we account for different treatment probabilities across waves by using inverse probability weights. α_2 is the coefficient of interest—the average treatment effect in kWh per month.

[5] Appendix Tables C1-C4 report summary statistics separately for each of Eversource's four service regions, to provide a glimpse of Opower's selection strategy. As a general rule, Opower appears to target households with higher baseline usage, more wealth, and more education.

Table 2: Average Characteristics and Treatment-Control Balance

	Treatment Mean/SD	**Control** Mean/SD	**Balance** Difference/SD
Baseline consumption (kWh)	849.685	745.597	0.110
	(412.996)	(376.888)	(0.979)
Home value ($)	363,281.560	343,071.887	-2,062.288*
	(370,144.602)	(339,779.476)	(1,071.788)
Home square footage	19.370	19.225	-0.014
	(10.983)	(11.226)	(0.036)
Annual income	99,592.697	93,693.277	-226.122
	(67,443.015)	(65,175.334)	(208.421)
Education (1-5)	3.211	3.138	-0.005
	(1.238)	(1.238)	(0.004)
Number of rooms in home	7.060	7.046	-0.008
	(2.142)	(2.214)	(0.007)
Year home built	1,968.271	1,969.043	0.037
	(23.463)	(23.613)	(0.074)
GreenAware score (1-4)	2.144	2.158	-0.000
	(1.135)	(1.119)	(0.004)
Renter (=1)	0.122	0.162	0.001
	(0.328)	(0.368)	(0.001)
Single-family occupancy (=1)	0.850	0.811	-0.002
	(0.357)	(0.392)	(0.001)
Child in home (=1)	0.475	0.461	-0.001
	(0.499)	(0.499)	(0.002)
Participated in EA (=1)	0.335	0.380	0.000
	(0.472)	(0.485)	(0.002)
Age	57.366	57.224	-0.028
	(14.650)	(14.911)	(0.048)

Notes: Columns (1) and (2) display the mean of each listed household characteristic for the treatment and control groups, respectively. Standard errors are listed beneath in parentheses. Column (3) checks for balance between the control and treatment groups with respect to the given characteristic. Results are from a linear regression of the characteristic on treatment status with wave fixed-effects and robust standard errors. * $p < 0.01$, $^{**}p < 0.05$, *** $p < 0.01$.

With variation in the timing of wave start dates, we use an event study model to investigate the evolution of HER impacts over time. The estimating equation is:

$$kWh_{iwt} = \beta_1 + \Sigma_{j=-12}^{37} \tau^j D_{iwt}^j + X_i\eta + \theta_w + \omega_t + e_{iwt}. \quad (2)$$

Here, the index j denotes a time period *relative* to the event of interest –the beginning of treatment in the relevant wave. D_{wt}^j is thus a binary variable equaling one if an observation is in wave w, j months after (or before) HER mailings begin in that wave, where $j \in [-12, 37]$.[6] We omit D_{wt}^0—corresponding to the month immediately preceding the start of mailings—from the estimating equation, so that all coefficients are interpretable as the monthly ATE relative to this month. We employ the same clustering and weighting as in Equation 1.

2.2 Causal Forests

The causal forest algorithm (Athey et al., 2019) is an adaptation of random forests (Breiman, 2001) for the measurement of causal effects. Random forests are themselves an ensemble method applied to classification and regression trees (CART) (Breiman et al., 1984), which employ recursive partitioning to split a sample into subgroups that maximize heterogeneity across splits. A tree is a single run of recursive partitioning; a forest is a collection of trees, where each tree is grown from a randomly drawn (bootstrapped) subsample of the data.

CART was originally developed for prediction of outcomes $\hat{y}$ as a non-parametric function of covariates. Athey and Imbens (2016) adapt CART for prediction of treatment effects $\hat{\beta}$, enabling the construction of valid confidence intervals for these effects. Wager and Athey (2018) do the same for random forests, establishing the consistency and asymptotic normality of their "causal" forest estimators. Athey et al. (2019) nest causal forests in a "generalized random forest" framework; we implement the causal forest algorithm using the generalized random forests (*grf*) R package (Tibshirani et al., 2018).

The basic building block of the causal forest is a regression tree. For a single tree, we start by drawing a random subsample, without replacement, from the full cross-section of Opower households. A single root node is created containing this random subsample. The root node is split into child nodes, and child nodes are split recursively to form a tree. Splits are chosen to maximize heterogeneity in subgroup ATEs, subject to penalties for within-node variance in ATEs and treatment-control imbalance. If splitting a given node would not result in an improved fit, that node is not split further and forms a "leaf" of the final tree (Tibshirani et al., 2018).

[6] We include 37 post-period months because some households begin being treated towards the end of the month in which the program starts.

Conventional regression tree algorithms use the same dataset to both grow tree structure and estimate ATEs at each node. Athey and Imbens (2016), however, show that this practice tends to overstate goodness of fit with deeper and deeper trees; they introduce the practice of "honest estimation", in which the full random subsample is split in half, one subset is used to grow the tree structure, and the other subset is used to estimate leaf ATEs. We employ this honest estimation in our trees.

Within-leaf ATE estimation in the generalized random forest package is implemented as a cross-sectional, difference-in-means comparison between treatment and control group. To take advantage of our panel data structure, we define our dependent variable as the difference between average monthly electricity usage in year X of the relevant HER program wave (where $X \in 1, 2, 3$) and average usage in the year prior to wave start date. Additionally, we residualize our dependent variable and treatment assignment, by regressing each of these on observable characteristics and wave fixed effects and recovering the residuals (again using weights by inverse probability of treatment).

Figure 4 shows a sample causal tree constructed using data from the April 2014 Connecticut "base" wave. The top node is the root: it contains 169,000 randomly chosen households, whose ATE is -15.7 kWh. The first split is made at a baseline consumption ("pre_mean") value of 1,706 kWh, and it creates two child nodes with different size and CATE. The algorithm can (a) split on the same variable in two successive branches, (b) split on different covariates across branches at the same level, and (c) stop branches at different depths.

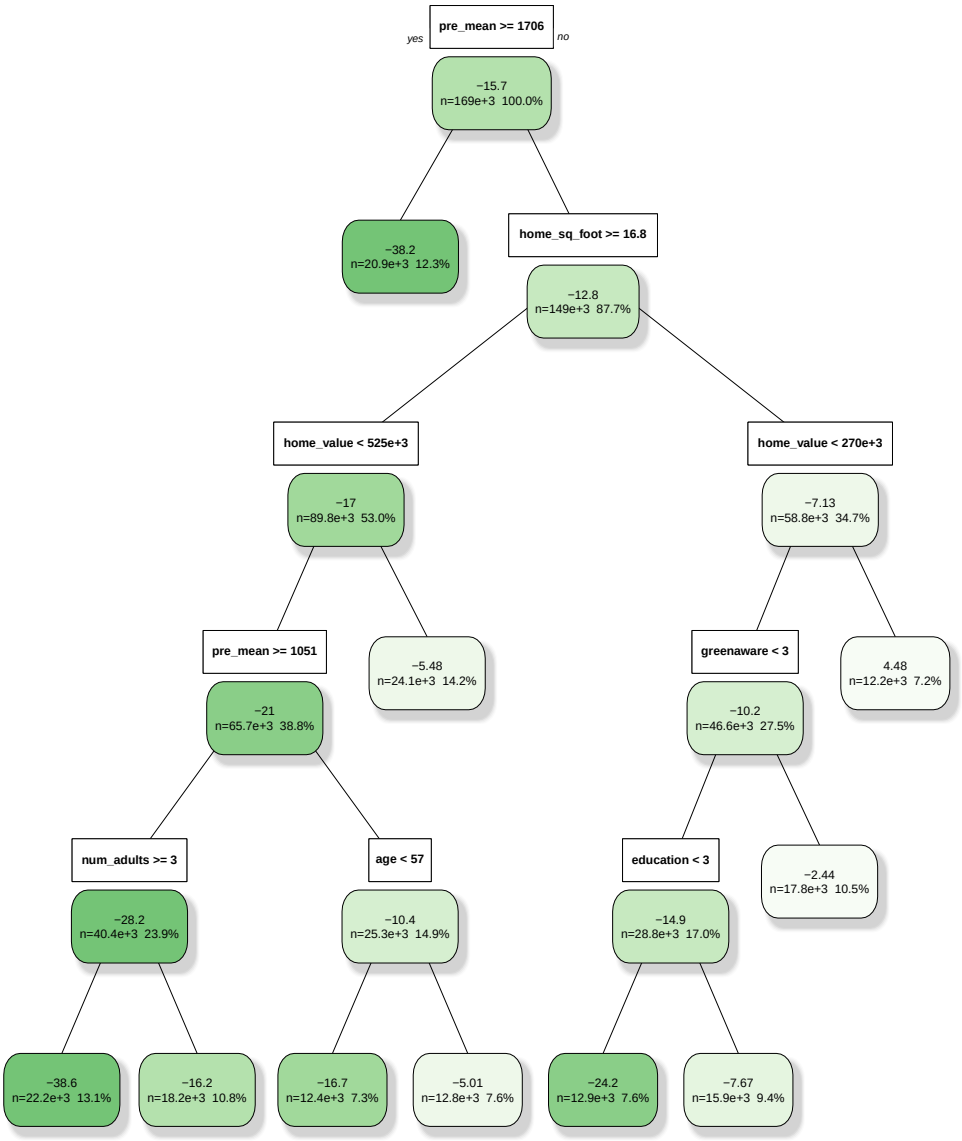
The terminal nodes, or leaves, report the estimated average treatment effect for households of the corresponding type. For example, if we follow the right-most set of branches, households that have baseline consumption less than 1,706, home square footage less than 1,680, and home value greater than 270,000 have an ATE of +4.48 kWh. In this particular tree, the right-most leaf is the only one with a positive ATE. The remainder of terminal-leaf ATEs range from -2.44 to -38.6 kWh.

We grow a forest consisting of 10,000 trees. In our causal forest, each tree is grown with a different random 50% subsample of households and a different subset of available characteristics.[7] The whole, tree-specific procedure can thus be represented as follows:

1. Randomly draw (1) a sample of households and (2) a subset of available characteristics.
2. Randomly split the sample in half, creating a "training set" S_{tr} and an "estimation" set S_{est}.
3. Using S_{tr}, grow a tree.
4. Match households in S_{est} to leaves of the tree, according to observed characteristics.

[7] The number of characteristics chosen varies by tree according to a draw from a Poisson distribution.

Figure 4: A sample causal tree



Notes: The tree is constructed from the Connecticut "base" wave beginning in April 2014. The dependent variable is the difference between average monthly electricity usage in program year 2 and the year prior to program start. Reported numbers in each box are leaf-specific ATE (in kWh), the number (n) of households falling into this leaf, and the corresponding proportion (in %) of total households used.

5. Estimate ATEs in each leaf using the matched observations from S_{est} in that leaf.[8]

For each of the 10,000 trees, we predict treatment effects for all households not used at all in the tree-growing procedure (i.e., not selected in Step 1 above). We thus obtain a large number of predictions for each household (in expectation, 5,000). We aggregate these predictions into a single, central estimate of a household's treatment effect using adaptive neighborhood estimation (Tibshirani et al., 2018). For each household i, we assign every other household a weight corresponding to the frequency with which it falls into the same leaf as i. These weights define the forest-based adaptive neighborhood. We then estimate household i's treatment effect as the weighted average of all other households' average predictions.

In addition to the relative size of the bootstrapped sample and the number of characteristics used, a few other parameters influence the forest algorithm and thus the estimates that emerge from it: minimum node size (a threshold number of observations in a node, below which no further splits can be made); maximum split imbalance (between child-node treatment and control N); and the penalty for split imbalance. For all of these parameters except minimum node size, we use the default values provided by the generalized random forest algorithm. The distribution of household treatment effect predictions is sensitive to minimum node size; we tune this parameter by training forests with different minimum node size values and choosing the value that minimizes R-loss, as defined in Nie and Wager (2017).

3 Results

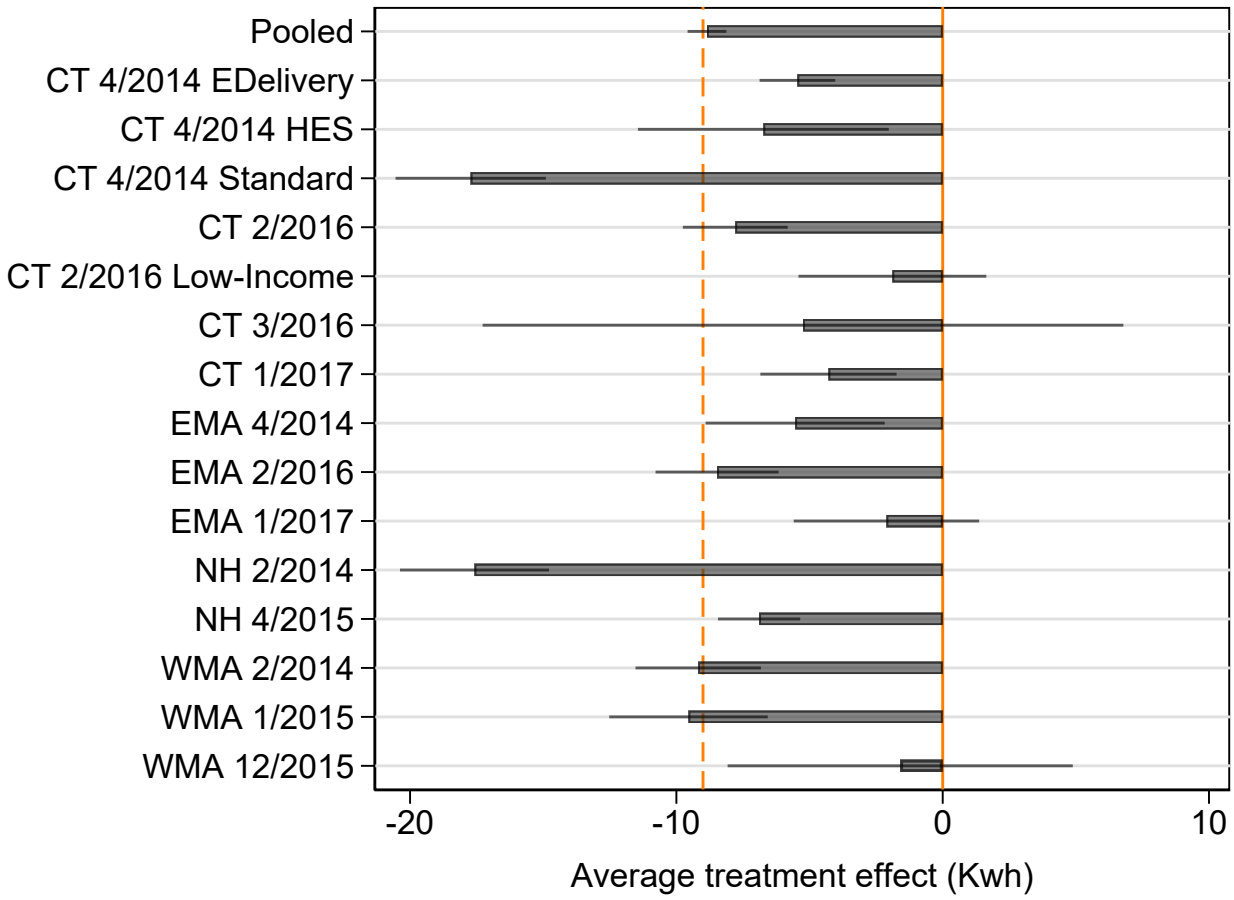
3.1 Average treatment effects

Figure 5 displays ATE estimates in each individual Opower wave as well as for the full, pooled sample. These results correspond to Equation 1. The pooled ATE is -8.85 kWh (per month), or -1 percent. While this is somewhat lower than the ATEs found in earlier Opower experiments (Allcott, 2011; Ayres et al., 2013; Costa and Kahn, 2013), the difference may be explained at least in part by "site selection bias" (Allcott, 2015): earlier Opower experiments systematically targeted areas and households with larger potential to reduce consumption. Wave-specific ATEs range in magnitude from -1.6 to -17.7 kWh. The pooled ATE and 12 of the 15 individual program-wave ATEs are statistically significant at the five-percent level or lower.

While the timing and household makeup of each program wave likely explain some of the heterogeneity in wave-specific ATEs, differences in the length of the post-period may also be a

[8]Due to computational considerations, an approximate criterion is computed using gradient-based approximations of the in-sample conditional average treatment effect estimators of the child nodes.

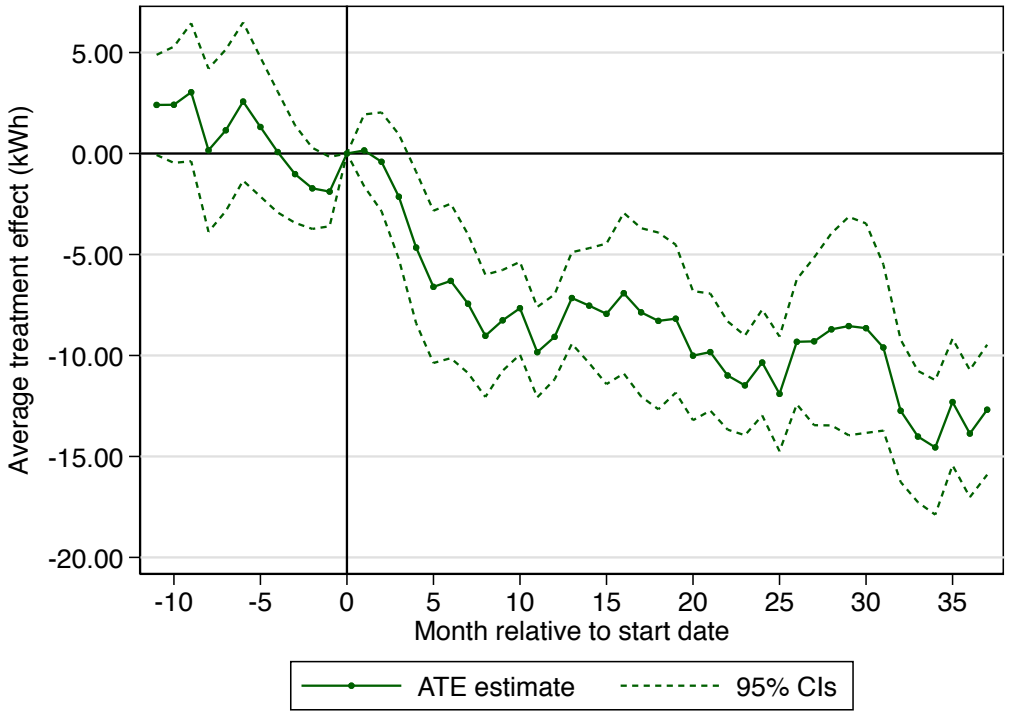
Figure 5: Average treatment effects, by wave: consumption



Notes: The y-axis denotes a specific wave ("Pooled" indicates all waves put together). The x-axis measures the treatment effect. Error bars denote 95% confidence intervals. CT = Connecticut; EMA = Eastern Massachusetts; NH = New Hampshire; WMA = Western Massachusetts. All effects are estimated using Equation 1 as described in Section 2.

part of the explanation. Figure 6—generated through estimation of Equation 2—sheds light on how the consumption impact of HERs evolves over time, on average. In the 12 months prior to program start date, none of the point estimates are statistically different from zero. In months 1 and 2, too, there is no discernible impact on consumption. But from months 2 through 8, there is a consistent, steep downward trend in average consumption. Month-specific point estimates are statistically significant beginning in month 4. The ATE in each successive year is larger than that of the previous one. In sum, households take time to ramp up their response to reports but continue changing behavior into at least the third year of treatment.

Figure 6: Event study of pooled experimental waves: consumption



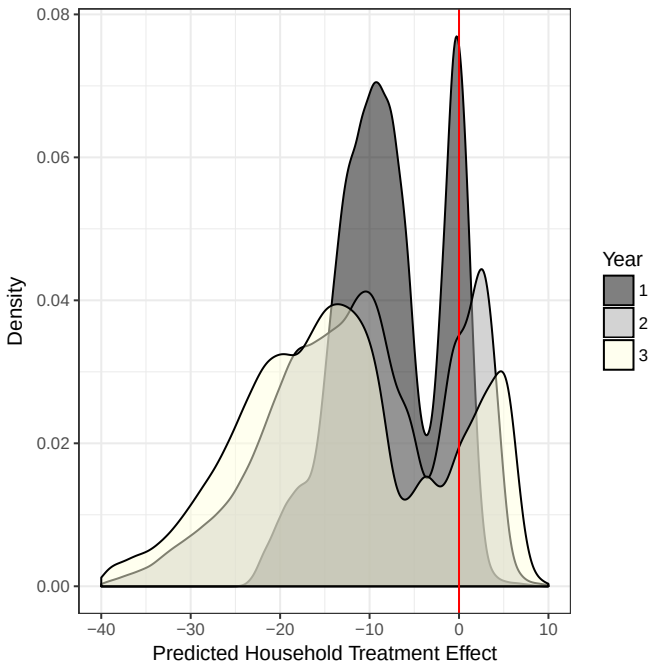
Notes: The solid-line data points are event-study coefficients from estimation of Equation 2. Dashed lines indicate 95% confidence intervals. D^0_{iwt}—which corresponds to the month immediately preceding program start—is omitted.

3.2 Conditional average treatment effects, via causal forest

Figure 7 depicts the distribution of household treatment effect predictions produced by the causal forest. We plot separate distributions for each of the first three years of treatment. It is immediately clear from this graph that the distribution of treatment effects is multi-modal. In year 1 of treatment, there is a large peak centered on -10 kWh, as well as an even larger, albeit narrower, peak centered on zero. This zero peak implies that a significant number of households don't initially respond to, or perhaps even read, their home energy reports. In years 2 and 3 of treatment,

both peaks progressively widen and shift away from zero. Households that respond by reducing consumption appear to learn to do more of that over time, but a sizeable subset of the sample (18 percent) is predicted to *raise* its consumption. The full range of predicted treatment effects in Year 3 extends from roughly -40 to +10 kWh.[9]

Figure 7: Distribution of Predicted Treatment Effects

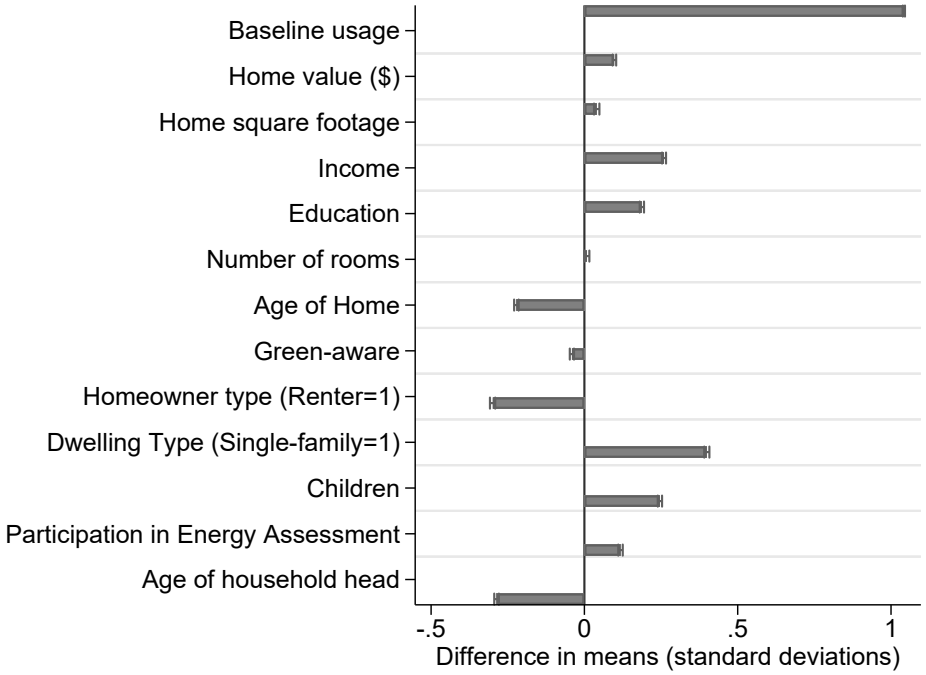


Notes: Each plotted distribution is a kernel density of household treatment effects in a specific year (1, 2, or 3) of HER programming. Treatment effect predictions come from our causal forest (Section 2.2).

Figure 8 provides a picture of the average difference in households predicted to reduce their consumption versus those predicted to raise it. There are significant differences in every characteristic used in the forest: "reducers" are more likely to own their residence; their homes tend to be larger, newer, and more valuable; and they tend to have more children, higher income, and younger and more educated heads of household. Finally, those who reduce their consumption have significantly higher average baseline electricity consumption, and the difference dominates all others in magnitude. This finding validates Opower's documented strategy of targeting high-consumption users for HER delivery, and it is consistent with the idea that there is more "room" for energy savings when baseline consumption is larger. In Appendix Figure C1, we measure the predictive power of our household characteristics in a different way, by plotting the frequency of each characteristic's use as a splitting variable in the forest. The results are highly consistent with the story told by Figure 8: baseline consumption is used far more frequently than any other

[9]In Appendix A, we describe a test of internal validity, based on Davis and Heller (2017a), that compares forest predictions from a training set of households to actual estimates in a test set. The differences are minimal, which suggests that the forest's household-specific treatment effects are internally valid.

Figure 8: Characteristics of "reducers" vs. "increasers"



Notes: Bars denote differences in mean between households with negative predicted treatment effects and those with positive ones, for each listed characteristic. Units are standard deviations of the relevant characteristic. Error bars denote 95% confidence intervals.

characteristic and is chosen for the first split in 90 percent of trees. Home value is the second most often-used, and it catches up to baseline consumption in frequency of use at the sixth split level.

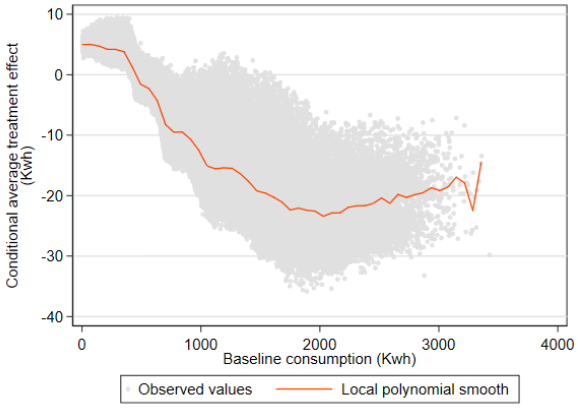
Figure 9 provides evidence on the relationship between the empirical distribution of predicted treatment effects and three key household attributes. Each panel presents a scatterplot of individual values: the y-axis measures predicted treatment effect, and the x-axis measures the attribute in question. We fit smooth, local polynomial functions to each scatterplot's data. The first two panels give us a more refined understanding of how treatment is related to the two most important predictors, baseline consumption and home value, while the third panel sheds light on the role of the HER's social comparison.

Panel A illustrates the potential for improved program outcomes through selective targeting on observable characteristics. The overwhelming majority of households with positive treatment effects have baseline consumption less than 800 kWh per month; setting the threshold for program inclusion at this level would thus have avoided nearly all adverse consumption impacts. According to the fitted line, treatment effect increases steadily with baseline consumption up to about 1800 kWh, at which point the trend flattens out and baseline consumption ceases to distinguish treatment effects on its own.

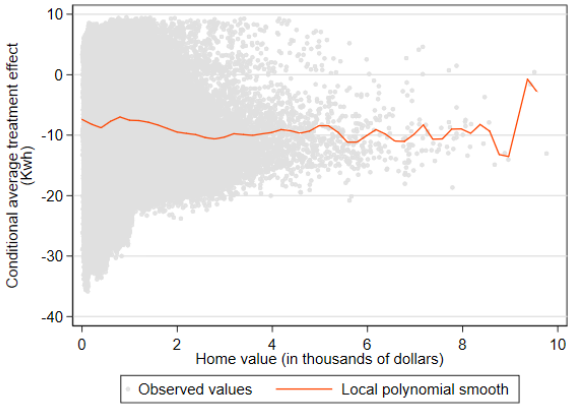
The relationship between treatment effect and home value in Panel B is flat throughout the

Figure 9: Average treatment effect vs. household type

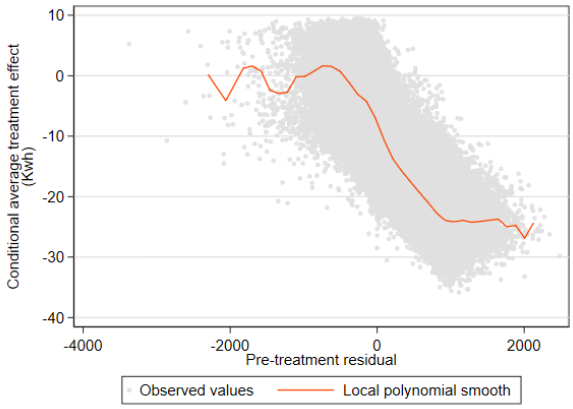
Panel A. Baseline consumption



Panel B. Home value



Panel C. Pre-treatment residual



Notes: Each plotted point represents a household. In each panel, the x-axis measures the value of the indicated household characteristic, while the y-axis measures treatment effect (TE) predicted by our causal forest; see Section 2.2 for implementation details. Lines display the local smoothed polynomial relationship between ATE and the characteristic. In Panel C, the x-axis "characteristic" is a household's residual from a zip-code specific regression of pre-treatment consumption on home characteristics.

range of observed home values. However, the largest predicted reductions in consumption are confined to the very bottom of the home value distribution. Nobody with home value above 100,000 dollars is predicted to reduce consumption by more than 23 kWh, while the households below that dollar threshold in some cases are predicted to reduce by 30-35 kWh. Panels A and B together thus imply that the largest "reducers" have high baseline consumption but low home value.

Panel C provides suggestive evidence of the mechanism by which certain households (who, according to Panel A, have low baseline consumption) are driven to *raise* their consumption: low electricity consumers are more likely to receive positive Opower feedback. Previous studies have documented a so-called "boomerang effect" (Bhanot, 2017; Schultz et al., 2007), in which a social comparison aiming to reduce the use of of some good inadvertently ends up increasing it. In our context, a household that discovers it is relatively more energy-efficient than other households may actually *raise* its electricity consumption. While Schultz et al. (2007) find evidence of a boomerang effect of social comparisons in energy consumption among 290 California households, subsequent, larger-scale evaluations of HERs (for example, Allcott, 2011) have not found evidence of such an effect.[10]

We are not able to *directly* test for a boomerang effect, because we do not have access to the specific social comparison received by each household each month (that is, we do not know whether a given household was told they were consuming more or less than their respective comparison group). To work around this limitation, we construct a proxy for a household's comparison to its similar neighbors: for each zip code, we regress pre-treatment household average consumption on household characteristics. The residual of this regression provides a measure of how a given household's consumption compares to an average household with the same home characteristics and in the same zip code. We expect that households with negative residuals are more likely to have received a home energy report stating that they are consuming less than their comparison group.

It is apparent from Panel C that households predicted to raise their consumption in response to HERs are overwhelmingly likely to have a negative residual. The result is suggestive of a boomerang effect at scale in household energy conservation. From the utility's perspective, it also exemplifies the potential gains of tailoring its treatment, to prevent "adverse" consumption outcomes.[11] More generally, predicted treatment effect is tightly correlated with the calculated residual (the correlation coefficient is -0.9). Over a large range of residual values (approximately

[10](Byrne et al., 2018), however, find a boomerang effect in HER-treated households who specifically are observed to have *overestimated* their own energy use.

[11]Some HERs (especially earlier versions) include "injunctive norms"—that is, smiley faces accompanying message of low relative use—in an effort to reduce boomerang. The HERs delivered by Opower to Eversource customers, however, do not use injunctive norms, in order to avoid adverse customer satisfaction outcomes.

-1,000 to 1,000), households appear to respond in a continuous way to nudges: the larger the disparity with one's comparison group, the larger the predicted response.

4 Economic Benefits of Targeting

High-resolution conditional average treatment effects (CATEs) are potentially very useful because they point to possible improvements to program effectiveness through targeting and tailoring. We illustrate this by evaluating the gains to targeting according to three different objective functions. The first of these is meant to approximate the utility's perspective. In this case, we assume that the utility's objective is to maximize electricity savings, net of the cost of creating and sending Home Energy Reports.[12] This assumption is motivated by the existence of standards requiring Eversource to show evidence of new electricity savings from energy efficiency annually, as well as the notion that customers like to save money. On the other hand, it is clear from conversations with Eversource that this assumption is an oversimplification; the utility's objective function has other components, such as broadly maintaining (and increasing) goodwill among customers.

To optimize this objective function, we require estimates of both the value of electricity savings and the marginal cost of sending HERs. For the value of kWh reductions, we use Eversource's average retail electricity rate of $0.21 per kWh. We multiply this number by a household's forest-predicted treatment effect to find the gross benefit of sending an HER to that household. Based on consultation with Eversource, we assume that the marginal cost of HERs is $7 per household per year.

Alternatively, targeting may be structured to maximize social welfare. To model this perspective, we adjust the valuation of electricity savings to reflect the social marginal cost of electricity, which includes both generation costs and environmental externalities.[13] We value kWh reductions at the short-run social marginal cost estimated by Borenstein and Bushnell (2018) for the New England electricity region in 2016—$0.065/kWh. In this iteration of the analysis, we continue to assume that the social marginal cost of one year of treatment is $7, though we acknowledge that the true social marginal cost could be below the price charged by Opower.

Our third and final variant of the objective function augments the social welfare function described above to account for customer willingness-to-pay (WTP) for HERs. Allcott and Kessler

[12] It is important to note that Eversource is predominantly an electricity retailer; it owns very little generation. Therefore, the moral hazard concerns about asking a vertically-integrated utility to reduce consumption are less of an issue in our context.

[13] We calculate social welfare here as the sum of individual net benefits of treatment, but the true social welfare function embeds distributional preferences. Thus, socially "optimal" targeting decisions may deviate from simple benefit-cost comparisons. Reames et al. (2018) find that, in Michigan, utilities spend four times more money on energy efficiency programming for middle- and high-income customers than for low-income ones.

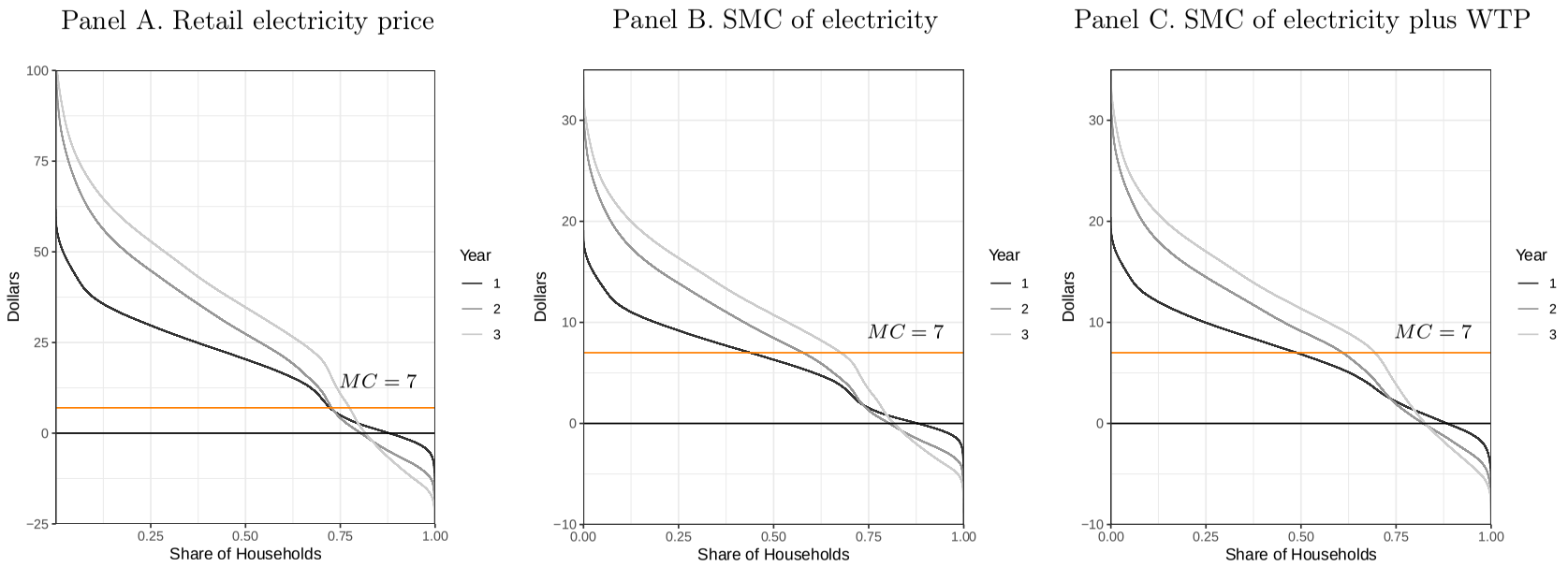
(2019) elicit WTP for HERs experimentally, and they report results from a regression of household-specific WTP on the logarithm of income, indicators for retirement, marriage, homeownership, and single-family occupancy, and homebuyer's credit worthiness score. We use the regression coefficients of Allcott and Kessler (2019) to predict household-specific WTP in our sample, given the characteristics of each household. Our data do not match up perfectly to theirs, but we do have measures of income, age, number of adults in the household, homeownership, and single-family occupancy. We define households with a head-of-household that is older than 65 as "retired." We define households with at least two adults living in the household as "married." Allcott and Kessler (2019) do not report a constant term for the regression but do report an average WTP. We thus use, as our own constant term, the difference between their reported mean WTP and the fitted mean value in our data using their regression coefficients. Social benefits in this last objective function are then equal to the sum of our predicted WTP value and the social value of electricity savings.

We restrict our sample to those households that received at least 36 months of HERs and the corresponding control groups, which yields 449,824 households. Figure 10 plots, for each objective function, the (reverse) cumulative distribution functions (CDF) of predicted household savings (in \$) in each of the first three years of HER programming. In every year and with every objective function, the CDF crosses both the MC line and the zero line; that is, there are always households whose responses to HERs translate to both net and gross negative benefits. This, in turn, implies that in every scenario there are potential gains to selective targeting.

According to the utility's objective function (Panel A), the total welfare gains created by the actual HER program in years 1, 2, and 3 are \$5.5M, \$9.1M, \$11.5M, respectively; this increase is due to larger treatment effects over time (in both directions— but most households decrease, rather than increase, consumption, so the net effect is to increase aggregate program benefits). These numbers include deadweight losses of approximately \$1M each year from sending reports to households whose responses do not produce positive net benefits. With perfect foresight, the utility could elect to send reports to the 72% of participating households that lay to the left of the crossing point between the CDF and the MC in the first year. In that case, welfare gains would increase by 14% to \$6.3M. In the second year, the utility could use the same rule to send HERs to 73% of households, and welfare would increase by 14% to \$10.4M. Finally, year-3 welfare would increase by 12%, to \$13.0M, due to such targeting.

Using our two social objective functions (Panels B and C) yields much smaller estimated welfare gains from the actual program, because the social marginal cost of electricity is significantly lower than the retail price of electricity. This fact also means that far fewer households respond in ways that are net-beneficial to society. According to the first social objective function (Panel B), over

Figure 10: The distribution of HER net benefits



Notes: Each downward-sloping line is the reverse cumulative distribution function of annual savings in a given HER program year, estimated via our causal forest. In Panel A, electricity consumption reductions are valued at the average retail price of electricity, \$0.21/kWh. In Panel B, they are valued at the social marginal cost (SMC) of electricity in New England in 2016, \$0.065/kWh (Borenstein and Bushnell, 2018). In Panel C, they are valued at the same SMC plus private willingness-to-pay for HERs estimated based on the findings of Allcott and Kessler (2019). Note that the y-axis range is larger in Panel A than in Panels B and C. The two horizontal lines are drawn at \$0 and \$7, the latter of which is an estimate of the marginal cost of sending one year's worth of HERs. See Sections 2.2 and 4 for further details.

50% of homes produce negative net benefits in the first year, while 44 and 32% of homes produce negative net benefits in years 2 and 3, respectively. The gains from targeting with perfect foresight are correspondingly large as a percentage of baseline welfare. Absent targeting, total welfare gains in the first year of treatment amount to -$477K; targeting raises total welfare in the first year to $616K. In the second year, targeting increases welfare from $633K to $1.8M, a rise of over 120%. In the third year, targeting increases welfare from $1.4M to over $2.4, a rise of over 70%.

Our second social objective function serves as an intermediate case relative to the other two functions, because the inclusion of WTP raises the social benefits of HERs somewhat. We estimate that households are willing to pay, on average, $2.97 for HERs each year. With perfect foresight, a social planner would treat 65, 68, and 74% of households in years 1, 2, and 3, respectively. Absent targeting, social welfare gains from treatment in the first year are $860K; targeting increases this by 66% to $1.4M. Welfare increases from $2.0M to $2.7M in year two—a rise of 36%. Finally, in year 3 welfare increases from $2.7M to $3.4M, or 25%.

The reported magnitudes of welfare gains from selective targeting rely on the assumption of perfect foresight—that is, that household treatment effects are known before treatment even begins. This assumption may be appropriate in certain specific situations, such as when a causal forest can be grown on a set of households that is very similar to the new set of households about to be treated. In many cases, however, this is unlikely to be the case. In Appendix B, we test the out-of-sample predictive accuracy of our causal forest along with several other methods, and we find suggestive evidence that large differences between training and test datasets significantly impedes forecasting accuracy out of sample.

Consequently, we investigate the efficacy of selective targeting in program years 2 and 3 based on causal forest estimates from program year 1. We rely on the conservative rule that the utility continues to send HERs to households with benefits greater than the marginal cost of HERs in the first year; this ignores the fact that, over time, treatment effect magnitudes tend to increase. Because this strategy does not require perfect foresight, it is a more realistic option for the utility or social planner. In principle, however, it is vulnerable to false positives (sending HERs to households that don't produce net benefits in years 2 and 3) and false negatives (*not* sending HERs to households that do produce net benefits).

Whether a household's net benefits exceed the MC of a HER is quite stable over time. Table 3 displays, for each of the first two objective functions, the frequencies of false positives and false negatives in the second and third year of treatment based on predictions from the household's savings in year 1 (the errors are the off-diagonals). According to Panel A, if the utility were to base its targeting decisions off of the first year of treatment, it would send reports to 325,080 (321,722+3,358) households in year 2 and not send HERs to 124,744 (5,398+119,346) households.

The share of mistakes from relying solely on the first year of the experiment is less than 2%. Because ATEs tend to grow over time, there are more false negatives in year 2 (5,398) than false positives (3,358). There are relatively more targeting "mistakes" in program year 3, but the overwhelming majority of decisions (95 percent) produce net benefits.

Table 3: Targeting based on outcomes in program year 1

Panel A. Using retail electricity price

		Year 2		Year 3	
		Send	Do Not Send	Send	Do Not Send
Year 1	Send	321,722	3,358	324,707	373
	Do Not Send	5,398	119,346	23,652	101,092

Panel B. Using social marginal cost of electricity

		Year 2		Year 3	
		Send	Do Not Send	Send	Do Not Send
Year 1	Send	186,744	10,409	196,811	342
	Do Not Send	72,898	179,773	107,510	145,161

Panel C. Using social marginal cost of electricity plus WTP

		Year 2		Year 3	
		Send	Do Not Send	Send	Do Not Send
Year 1	Send	211,301	8,314	219,289	326
	Do Not Send	62,825	167,384	93,140	137,069

Notes: The table summarizes the consequences of selective targeting in program years 2 and 3 based on outcomes in program year 1. Each panel's results rely on the use of a different objective function; see Section 4 for definitions. Counts corresponding to (Send, Send) include all households whose HER-induced benefits are larger than the marginal cost of HERs in year 1 and also year X, where $X \in \{2, 3\}$. Those corresponding to (Do Not Send, Do Not Send) include all whose HER-induced benefits are lower than marginal cost in year 1 and also in year X. Those with mismatched combinations are false positives and negatives—i.e., those for whom year-1 outcomes would suggest one decision (ex ante) while year-X would suggest the other (ex post). Counts are derived from year-specific, forest-based predictions of household treatment effects; see Section 2.2 for implementation details.

With the social objective functions (Panels B and C), there are far more false negatives—for example, 62,825 in year 2 and 93,140 in year 3 in Panel C, where estimated private WTP for HERs is included in the measurement of program benefits. This is a consequence of there being far more "Do Not Send" households in year 1 when the social objective function is used. In spite of this fact, targeting based on program year 1 estimates achieves most of the available year-2 and year-3 welfare gains available with perfect foresight. This is partly driven by the fact that the benefits lost due to false negatives and positives are small, since such households do not reduce their consumption very much. Targeting based on year-1 estimates achieves 99 percent of theoretically available welfare gains according to the utility objective function and 85 percent of such gains according to the social objective function.

5 Conclusion

Machine learning is fast becoming a powerful tool for high-resolution program evaluation. In this paper, we provide an early example of its capability by applying one of the most recent machine-learning methods—causal forests—to the evaluation of a large-scale behavioral intervention. Home Energy Reports have long been studied as an example of a successful "nudge" towards behavior that is both privately and socially beneficial. But despite consistent findings of modest, significant reductions in electricity consumption, relatively little is known about the mechanisms that govern the household response to HERs. Through estimation of a causal forest, we begin to shed light on these mechanisms.

Across fifteen experimental waves of Opower's HER program, the average household reduces its monthly electricity consumption by approximately 9 kWh. The random forest reveals the rich heterogeneity that underlay the consumption ATE: the distribution of household effects is left-skewed, so that 81 percent of households reduce consumption by more than the monthly mean. The largest reductions are three times the mean, while some households actually *in*crease their consumption. Pre-treatment consumption and home value are the strongest predictors of individual responses, but several other characteristics have predictive power as well, and the relationship between treatment effect and these characteristics is non-linear.

The forest results illustrate how machine learning might be used to improve the effectiveness of interventions. We find large welfare gains from targeting treatment according to the perspective of both the utility and society. While it may be difficult to accurately target treatment in a previously untreated sample of households, outcomes in the first year of the intervention provide valuable information for the targeting task. In our context, at least 85 percent of available welfare gains from treatment in years 2 and 3 are achievable through the use of year-1 estimates. Among households that do respond in privately or socially beneficial ways, it may be possible to raise welfare through tailoring of treatment to include different information or rely on a different framing of the nudge.

References

Allcott, H. (2011): "Social Norms and Energy Conservation," *Journal of Public Economics*, 95, 1082–1095.

——— (2015): "Site Selection Bias in Program Evaluation *," *The Quarterly Journal of Economics*, 130, 1117–1165.

Allcott, H. and J. B. Kessler (2019): "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons," *American Economic Journal: Applied Economics*, 11, 236–76.

Allcott, H. and T. Rogers (2014): "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *American Economic Review*, 104, 3003–37.

Andreoni, J., J. M. Rao, and H. Trachtman (2017): "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving," *Journal of Political Economy*, 125, 625–653.

Athey, S. and G. Imbens (2016): "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

Athey, S. and G. W. Imbens (2017): "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31, 3–32.

Athey, S., J. Tibshirani, S. Wager, et al. (2019): "Generalized random forests," *The Annals of Statistics*, 47, 1148–1178.

Ayres, I., S. Raseman, and A. Shih (2013): "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage," *The Journal of Law, Economics, and Organization*, 29, 992–1022.

Bhanot, S. P. (2017): "Rank and Response: A Field Experiment on Peer Information and Water Use Behavior," *Journal of Economic Psychology*, 62, 155–172.

Borenstein, S. and J. B. Bushnell (2018): "Do Two Electricity Pricing Wrongs Make a Right? Cost Recovery, Externalities, and Efficiency," Working Paper 24756, National Bureau of Economic Research.

Breiman, L. (2001): "Random Forests," *Machine Learning*, 45, 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984): *Classification and Regression Trees*, Routledge.

Burlig, F., C. R. Knittel, D. Rapson, M. Reguant, and C. Wolfram (2017): "Machine Learning from Schools about Energy Efficiency," Working Paper 23908, National Bureau of Economic Research.

Byrne, D. P., A. L. Nauze, and L. A. Martin (2018): "Tell me something i don't already know: Informedness and the impact of information programs," *Review of Economics and Statistics*, 100, 510–527.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," Tech. rep., National Bureau of Economic Research.

Costa, D. L. and M. E. Kahn (2013): "Energy Conservation "Nudges" and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment," *Journal of the European Economic Association*, 11, 680–702.

Davis, J. and S. B. Heller (2017a): "Using causal forests to predict treatment heterogeneity: An application to summer jobs," *American Economic Review*, 107, 546–50.

Davis, J. M. and S. B. Heller (2017b): "Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs," *Review of Economics and Statistics*, 1–47.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using Randomization in Development Economics Research: A Toolkit," *Handbook of development economics*, 4, 3895–3962.

Ferraro, P. J. and M. K. Price (2013): "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment," *The Review of Economics and Statistics*, 95, 64–73.

Festinger, L. (1954): "A Theory of Social Comparison Processes," *Human relations*, 7, 117–140.

Hussam, R., N. Rigol, and B. R. HBS (2018): "Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field," .

Imai, K. and M. Ratkovic (2013): "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics*, 7, 443–470.

Kettle, S., M. Hernandez, S. Ruda, and M. Sanders (2016): "Behavioral Interventions in Tax Compliance," Research papers, World Bank.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017): "Human Decisions and Machine Predictions," Working Paper 23180, National Bureau of Economic Research.

Mullainathan, S. and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

Nie, X. and S. Wager (2017): "Quasi-Oracle Estimation of Heterogeneous Treatment Effects," *arXiv preprint arXiv:1712.04912*.

Reames, T. G., M. A. Reiner, and M. B. Stacey (2018): "An incandescent truth: Disparities in energy-efficient lighting availability and prices in an urban US county," *Applied energy*, 218, 95–103.

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2007): "The Constructive, Destructive, and Reconstructive Power of Social Norms," *Psychological Science*, 18, 429–434, pMID: 17576283.

Tibshirani, J., S. Athey, S. Wager, R. Friedberg, L. Miner, and M. Wright (2018): "Package 'grf'," .

Wager, S. and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

White, I. R., P. Royston, and A. M. Wood (2011): "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in medicine*, 30, 377–399.

Appendix A - Test of Internal Validity

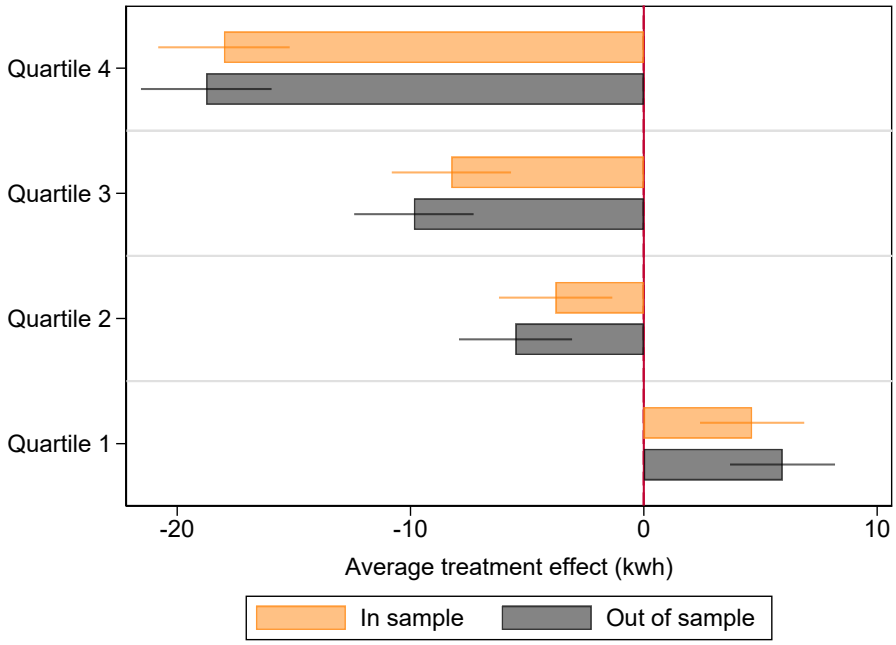
We can test the predictive power of our forest in a random hold-out subsample of our Eversource households. We follow the procedure of Davis and Heller (2017b). First, we split the full set of households randomly in half to create in- and out-of-sample groups S_{in} and S_{out}. Second, we run the causal forest procedure only using S_{in}. Third, we predict treatment effects (TEs) in S_{out} and group them by quartile of predicted TE. Lastly, we regress electricity consumption on the treatment dummy as well as its interaction with TE-quartile dummies, separately for both S_{in} and S_{out}. The estimating equation is:

$$Log(kWh)_{iwt} = \alpha_0 + \alpha_1 T_{iwt} + \Sigma_{j=2}^{4}\Big(\alpha_j T_{iwt} * 1[Q_i = j]\Big) + X_i\eta + \theta_w + \omega_t + e_{iwt}. \tag{3}$$

where Q_i is a household's quartile of predicted treatment effect, j indexes quartile, and the bottom quartile is the omitted group. The coefficients of interest are α_1, α_2, α_3, and α_4. We compare the magnitudes of each of these coefficients in S_{in} versus S_{out}. Significant differences imply that the causal forest procedure suffers from overfitting.

Figure A1 plots the in-sample and out-of-sample estimated ATEs across forest-predicted ATE quartiles. Specifically, the x-axis indicates an ATE estimated using Equation 3, while the y-axis shows the corresponding quartile of forest-predicted impacts. ATEs increase with quartile, using both the in-sample data and the out-of-sample data. Top-quartile ATEs are nearly -18 kWhs, while bottom-quartile ATEs are approximately 4 kWhs. Furthermore, in-sample and out-of-sample quartile-specific ATEs are very similar across the board. These results imply that the forest method produces internally valid TE estimates. We note, however, that the hold-out sample here is, by design, similar to the training sample, so this test does not provide insight into the forest's predictive accuracy in populations with different average characteristics.

Figure A1: Test of out-of-sample performance



Notes: The y-axis denotes quartile of forest-predicted treatment effect on electricity consumption. The x-axis measures the treatment effect. Lines show 95% confidence intervals. "In sample" refers to S_{in}, a 50% subsample of all households used in forest building; "Out of sample" refers to S_{out}, the remaining 50% subsample that is entirely omitted from forest building in this exercise. All effects are estimated using Equation 3. Appendix A explains the exercise in further detail.

Appendix B - A Test of External Validity

To test the external validity of the forest results, we follow the previous section's procedure without forcing training and test sets to have the same average characteristics. Instead, we take advantage of the staggered nature of our fifteen Opower waves. Our conceptual strategy is to calibrate a predictive model with earlier data, use it to predict outcomes in later data, and then compare predicted to actual. We can assess the relative benefits of the causal forest out-of-sample by "horseracing" it against other prediction methods.

We divide our sample of Opower waves into three chronological groups according to program start date: 2/2014-4/2014, 1/2015-4/2015, and 2/2016-3/2016. For each predictive method in consideration, we use the following algorithm:

1. Build a predictive model of treatment effects exclusively using households in group 1.[14]
2. Use that model to predict treatment effects for each household in group 2.
3. Aggregate group-2 households into quartiles by size of "predicted" treatment effect and calculate quartile-specific averages.
4. Estimate "actual" average treatment effects in each of these quartiles via Ordinary Least Squares (OLS) using group-2 data, and compare to "predicted".

We replicate this procedure using groups 1 and 2 together as the input to the causal forest (step 1 above) and group 3 as the sample in which to compare predicted with actual (steps 2-4). There are thus two rounds in which to assess performance.

We complete iterations of the above procedure with each of two causal forests calibrated in slightly different ways. The first forest uses the same minimum node size as before (1,500) and default values of all other parameters as provided by the *grf* package. The second uses the same minimum node size but is otherwise "tuned" using the *grf* package's built-in tuning algorithm. We compare these two forest methods to five other methods: a lasso (least absolute shrinkage and selection) estimator and four variants of a conventional regression-model approach (i.e., methods without any machine learning).

Our lasso estimator considers a large number of candidate predictors of the outcome variable (the pre-post difference in consumption). We include: the treatment dummy, all forest characteristics, and dummies for service territory, zip code, revenue class, and tariff rate; interactions between treatment and all other aforementioned characteristics; the squares of all continuous forest

[14]We define the dependent variable as the average monthly consumption in year 2 of HER receipt minus the average monthly consumption one year prior to program start. We choose year 2 rather than year 3 because the former yields a relatively larger sample size.

characteristics as well as their interactions with treatment; decile dummies for all continuous forest characteristics and their interactions with treatment; double interactions between treatment and every pair of forest characteristics; and splines of each forest characteristic (using deciles) interacted with treatment.

In each of the four regression methods, we estimate treatment effects via OLS with a different set of explanatory variables, all interacted with the treatment dummy. Method 1 ("Interacted") includes treatment interactions with all forest characteristics, their squares, and the product of each combination of characteristics. Since this risks oversaturation and spurious correlation, we define method 2 ("Parsimonious") to be a simpler systematic model: a second-order Taylor series expansion based on all of the variables included in the forest. Method 3 ("Linear") is more parsimonious still: we include all forest variables linearly. Method 4 ("Pre-mean") is meant to mimic Opower's targeting strategy by estimating treatment effect only as a function of baseline consumption.

We judge the models along three dimensions that we believe might be relevant for a utility, government, or other social planner. First, we check for monotonicity in within-quartile "actual" treatment effects. Second, we inspect the magnitude of "actual" treatment effects in the top quartile. Third, we calculate the average prediction error and average square prediction error of each method.[15]

Table B1 reports the key results of our horserace: within-quartile predicted and actual average treatment effects for each method in each round. For conciseness, we display results from only three methods: the default forest, the second-order Taylor regression method, and the linear regression method. These choices are motivated by the desire to compare a "standard" forest algorithm with the best-performing alternative methods. Results from the remainder of methods are presented in Table B2. Lasso is omitted from both tables; in our context, the lasso algorithm fails to uncover any treatment effect heterogeneity because it never chooses to keep interactions between treatment and household characteristics.

According to group 1 results (Panel A), none of the three methods stands out as clearly superior. "Actual" ATEs drop monotonically in predicted quartile (from top to bottom) using all three methods.[16] The forest does identify a top quartile with a higher ATE than that of the two displayed regression methods: -23.997 versus -19.535 and -19.692, respectively. More generally, the forest method appears to do better at identifying households that belong in the outer quartiles (1 and 4), while the regression methods sometimes do better at the inner ones (2 and 3). The final

[15]From our discussions with utilities, we have learned that they frequently submit energy efficiency plans for future reduction in demand based on predictions of the Opower treatment effect.

[16]Note from Appendix Figure B2, however, that the tuned forest maintains monotonicity, while the other two regression-based methods do not.

Table B1: Horserace results for next-wave prediction, by method

	Default Forest		Parsimonious Regression		Linear Regression	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
Panel A: Group 2						
Quartile 4	-20.771	-23.997***	-24.700	-19.535***	-24.597	-19.692***
		(4.427)		(4.399)		(4.393)
Quartile 3	-12.613	-14.703***	-16.203	-16.397***	-16.047	-15.469***
		(3.633)		(3.405)		(3.430)
Quartile 2	-8.668	-14.348***	-10.595	-12.195**	-10.525	-11.859***
		(3.120)		(3.182)		(3.132)
Quartile 1	2.081	-4.852**	-1.537	-9.309***	-1.753	-10.693***
		(2.259)		(2.492)		(2.505)
Panel B: Group 3						
Quartile 4	-11.009	-9.350	-13.587	-11.892	-13.342	-11.858**
		(6.034)		(5.218)		(5.247)
Quartile 3	-0.666	-7.783**	-5.139	-3.686	-5.254	-2.179
		(3.869)		(3.393)		(3.273)
Quartile 2	2.988	-4.565*	-0.502	-3.507***	-0.894	-4.793*
		(2.401)		(2.918)		(2.854)
Quartile 1	5.342	-3.457	8.027	-4.229	7.705	-5.255
		(2.284)		(3.965)		(4.021)

Notes: The "Predicted" column lists ATEs for the corresponding method, wave and percentile. The "Actual" column lists the results of an OLS regression of the difference between year-2 post-treatment and pre-treatment average consumption on treatment status, using wave fixed-effects and robust standard errors. Standard errors are listed in parentheses. See the text of Appendix B for explanations of the three methods. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

metric—prediction error—does not differ substantially across methods. For the default forest, the average absolute-value prediction error is 4.48 kWh, while it is 3.68 kWh and 3.94 kWh for the second-order and linear regression models, respectively. The corresponding comparison of average squared prediction error is 23.78 versus 22.42 and 26.52.

All methods perform significantly more poorly in group 3 (Panel B). Here, the default forest is the only method that generates monotonically falling actual ATEs in prediction quartile. However, its top-quartile ATE is smaller than that of the other two methods (-9.35 versus -11.892 and -11.858), and it does not perform better in terms of prediction accuracy. Average prediction errors for the three methods are 6.28, 4.6, and 5.35, respectively, while average squared prediction errors are 46.96, 41.06, and 48.71, respectively. The relative difference between group-2 performance and group-3 performance is stark: there is far less statistical significance of within-quartile ATE estimates, and prediction error is significantly higher. It is worth noting, however, that each method correctly recognizes the relatively low actual savings of group 3.

Table B2: Horserace results using alternative methods

	Tuned Forest		Interacted Regression		Pre-mean Regression	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
Panel A: Group 2						
Quartile 4	-20.847	-23.399***	-31.676	-20.598***	-21.385	-18.998***
		(4.436)		(4.149)		(4.541)
Quartile 3	-12.629	-17.699***	-17.560	-15.592***	-14.845	-19.262***
		(3.595)		(3.518)		(3.611)
Quartile 2	-8.468	-11.661***	-9.879	-8.095***	-11.696	-15.622***
		(3.122)		(3.273)		(3.005)
Quartile 1	2.179	-5.516**	2.698	-14.231***	-4.963	-4.256*
		(2.288)		(2.902)		(2.208)
Panel B: Group 3						
Quartile 4	-10.824	-11.171*	-23.514	-6.308**	-16.395	-8.985
		(5.915)		(4.678)		(6.232)
Quartile 3	-0.861	-4.394	-9.104	-4.416	-8.197	-8.766**
		(3.790)		(3.962)		(3.635)
Quartile 2	2.250	-4.973**	-1.006	-14.196	-5.930	-2.618
		(2.441)		(3.459)		(2.424)
Quartile 1	4.938	-4.316	13.315	1.013	-4.703	-4.400**
		(2.661)		(3.800)		(2.097)

Notes: The table displays horserace results for the three alternative predictive methods not displayed in Table B1. The "Predicted" column lists ATEs for the corresponding method, wave and percentile. The "Actual" column lists the results of an OLS regression of the difference between year-2 post-treatment and pre-treatment average consumption on treatment status, using wave fixed-effects and robust standard errors. Standard errors are listed in parentheses. See the text of Appendix B for explanations of the three methods. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

To understand why all of the methods perform poorly in round 2, we turn to Tables B3 and B4, which summarize the differences between the training set ("in-sample") and the test set ("out-of-sample") in each round of the horserace. The results reveal how different the households are, in-sample versus out of sample. All thirteen characteristics show statistically significant differences in both round 1 and 2, and both F-statistics for joint significance are very high (576 and 1,016, respectively). However, differences-in-means of several characteristics—including the two most often-used characteristics in our forests—become significantly larger in round 2. For instance, the difference in home value is twice as large in round 2 as in round 1, and the analogous difference in baseline consumption is six times as large. Thus, round 2 of the horserace asks all methods to make predictions on a test set (group 3) characterized by comparatively little overlap with the training data (groups 1 and 2).

Table B3: Summary Statistics for Training and Prediction Samples - Group 2

	Training Mean/SD	**Predicted** Mean/SD	Difference/SD
Home value ($)	373,917.884	259,937.190	113,980.694***
Home square footage	19.610	20.746	-1.136***
Annual income	103,470.842	85,237.971	18,232.871***
Education (1-5)	3.266	2.939	0.327***
Num Adults	2.602	2.521	0.081***
Number of Rooms in Home	7.064	7.094	-0.030*
Year home built	1,969.490	1,973.687	-4.197***
GreenAware score (1-4)	2.139	2.299	-0.161***
Renter (=1)	0.086	0.186	-0.099***
Single-family occupancy (=1)	0.882	0.858	0.024***
Child in home (=1)	0.450	0.489	-0.039***
Participated in EA (=1)	0.349	0.504	-0.154***
Age	57.964	56.716	1.247***
Baseline Consumption (kwh)	901.648	848.410	53.238***
F-test			576.239 (0.000)
Number of HH	406,637	49,192	
Treatment propensity	85.1	71.13	

Notes: Columns (1) and (2) display the mean of the listed household characteristic for the treatment and control groups, respectively. Standard deviations are listed beneath in parentheses. Column (3) checks for the difference between the sample used in training and the one used in prediction with respect to the household characteristic. Results are from a regression with robust standard errors. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. P-values for the F-test are listed beneath the F-statistic in parenthesis.

Table B4: Summary Statistics for Training and Prediction Samples - Group 3

	Training Mean/SD	**Predicted** Mean/SD	Difference/SD
Home value ($)	362,349.622	586,541.115	-224,191.493***
Home square footage	19.722	19.041	0.681***
Annual income	101,502.777	100,247.262	1,255.516**
Education (1-5)	3.231	3.508	-0.277***
Num Adults	2.593	2.120	0.473***
Number of Rooms in Home	7.066	7.083	-0.017
Year home built	1,969.914	1,961.840	8.074***
GreenAware score (1-4)	2.156	1.982	0.174***
Renter (=1)	0.097	0.189	-0.092***
Single-family occupancy (=1)	0.880	0.562	0.318***
Child in home (=1)	0.455	0.432	0.022***
Participated in EA (=1)	0.366	0.372	-0.006*
Age	57.833	53.604	4.229***
Baseline Consumption (kwh)	895.903	541.111	354.792***
F-test			1,016.023 (0.000)
Number of HH	455,829	34,232	
Treatment propensity	83.6	76.35	

Notes: Columns (1) and (2) display the mean of the listed household characteristic for the treatment and control groups, respectively. Standard deviations are listed beneath in parentheses. Column (3) checks for the difference between the sample used in training and the one used in prediction with respect to the household characteristic. Results are from a regression with robust standard errors. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. P-values for the F-test are listed beneath the F-statistic in parenthesis.

Appendix C - Additional Tables and Figures

Table C1: Summary Statistics for Connecticut

	Total (1)	Unenrolled (2)	Enrolled (3)	Balance (4)
Monthly consumption (kWh)	667 (763)	459 (892)	942 (405)	0.31 (1.25)
Home value ($)	328,597 (407,528)	298,403 (408,132)	364,200 (403,926)	-2,910* (1,742)
Home square footage	1,881 (1,292)	1,807 (1,501)	1,947 (1,071)	-1.56 (4.94)
Annual income ($)	89,971 (67,346)	78,625 (63,585)	104,736 (69,215)	-564* (291)
Education (1-5)	3.01 (1.25)	2.85 (1.23)	3.22 (1.24)	-0.007 (0.005)
Number of rooms in home	6.99 (2.49)	6.92 (2.87)	7.05 (2.11)	-0.014 (0.010)
Year home built	1,969 (24)	1,966 (25)	1,971 (23)	0.020 (0.112)
GreenAware score (1-4)	2.18 (1.11)	2.19 (1.07)	2.17 (1.16)	0.001 (0.005)
Renter (=1)	0.171 (0.377)	0.240 (0.427)	0.102 (0.302)	0.003** (0.001)
Single-family occupancy (=1)	0.788 (0.409)	0.704 (0.457)	0.877 (0.329)	-0.003* (0.002)
Child in home (=1)	0.444 (0.497)	0.407 (0.491)	0.489 (0.500)	-0.002 (0.002)
Participated in EA (=1)	0.298 (0.457)	0.301 (0.459)	0.294 (0.456)	-0.002 (0.002)
Age	57.7 (16.6)	58.3 (18.3)	57.2 (14.8)	-0.064 (0.071)
Observations	1,017,854	580,152	437,702	

Notes: This table lists summary statistics for all HH in Connecticut (Column (1)), for HH that are not enrolled in a HER program (Column (2)), for HH that are enrolled in a HER program and participated in a wave with available pre-enrollment data (Column (3)). Column (4) checks for balance between treatment and control. Baseline consumption for the unenrolled HH corresponds to average consumption for the entire analysis period. Results are from a linear regression of the listed HH characteristic on treatment status with wave fixed-effects and robust standard errors. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table C2: Summary Statistics for Eastern Massachusetts

	Total (1)	Unenrolled (2)	Enrolled (3)	Balance (4)
Monthly consumption (kWh)	503	497	558	-1.01
	(557)	(578)	(289)	(2.39)
Home value ($)	592,696	591,035	607,259	-5,744
	(443,623)	(444,486)	(435,717)	(4,138)
Home square footage	2,060	2,071	1,973	-11.64
	(1,926)	(1,965)	(1,611)	(15.84)
Annual income ($)	97,388	96,721	103,486	353
	(70,902)	(70,775)	(71,769)	(597)
Education (1-5)	3.44	3.43	3.51	0.003
	(1.27)	(1.27)	(1.29)	(0.011)
Number of rooms in home	7.35	7.35	7.29	-0.053*
	(3.09)	(3.10)	(3.05)	(0.031)
Year home built	1,963	1,964	1,960	-0.023
	(30)	(30)	(31)	(0.329)
GreenAware score (1-4)	2.05	2.06	1.98	0.003
	(1.09)	(1.09)	(1.08)	(0.009)
Renter (=1)	0.216	0.220	0.188	-0.001
	(0.412)	(0.414)	(0.391)	(0.004)
Single-family occupancy (=1)	0.612	0.612	0.610	0.002
	(0.487)	(0.487)	(0.488)	(0.004)
Child in home (=1)	0.342	0.334	0.411	-0.001
	(0.474)	(0.472)	(0.492)	(0.004)
Participated in EA (=1)	0.290	0.280	0.371	0.000
	(0.454)	(0.449)	(0.483)	(0.004)
Age	56.3	56.4	55.5	-0.188
	(17.2)	(17.2)	(17.1)	(0.157)
Observations	922,802	832,851	89,951	

Notes: This table lists summary statistics for all HH in Eastern Massachusetts (Column (1)), for HH that are not enrolled in a HER program (Column (2)), for HH that are enrolled in a HER program and participated in a wave with available pre-enrollment data (Column (3)). Column (4) checks for balance between treatment and control. Baseline consumption for the unenrolled HH corresponds to average consumption for the entire analysis period. Results are from a linear regression of the listed HH characteristic on treatment status with wave fixed-effects and robust standard errors. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table C3: Summary Statistics for Western Massachusetts

	Total (1)	Unenrolled (2)	Enrolled (3)	Balance (4)
Monthly consumption (kWh)	599 (1,273)	534 (2,040)	637 (347)	1.98 (3.32)
Home value ($)	220,368 (153,057)	215,627 (173,478)	222,984 (140,464)	-1,538 (1,500)
Home square footage	1,803 (1,465)	2,037 (1,978)	1,723 (1,232)	13.58 (15.96)
Annual income ($)	67,663 (52,110)	60,280 (52,024)	71,917 (51,682)	-149 (471)
Education (1-5)	2.82 (1.21)	2.69 (1.20)	2.90 (1.22)	0.010 (0.011)
Number of rooms in home	6.93 (2.58)	7.58 (3.24)	6.70 (2.27)	0.005 (0.028)
Year home built	1,961 (28)	1,959 (30)	1,962 (27)	0.149 (0.300)
GreenAware score (1-4)	2.24 (1.07)	2.41 (1.03)	2.14 (1.08)	-0.013 (0.010)
Renter (=1)	0.206 (0.404)	0.338 (0.473)	0.156 (0.363)	-0.000 (0.004)
Single-family occupancy (=1)	0.819 (0.385)	0.704 (0.457)	0.866 (0.340)	-0.003 (0.004)
Child in home (=1)	0.407 (0.491)	0.465 (0.499)	0.377 (0.485)	-0.007 (0.005)
Participated in EA (=1)	0.417 (0.493)	0.397 (0.489)	0.426 (0.495)	0.003 (0.005)
Age	57.5 (17.0)	49.9 (17.7)	60.1 (15.9)	0.075 (0.166)
Observations	173,311	64,233	109,078	

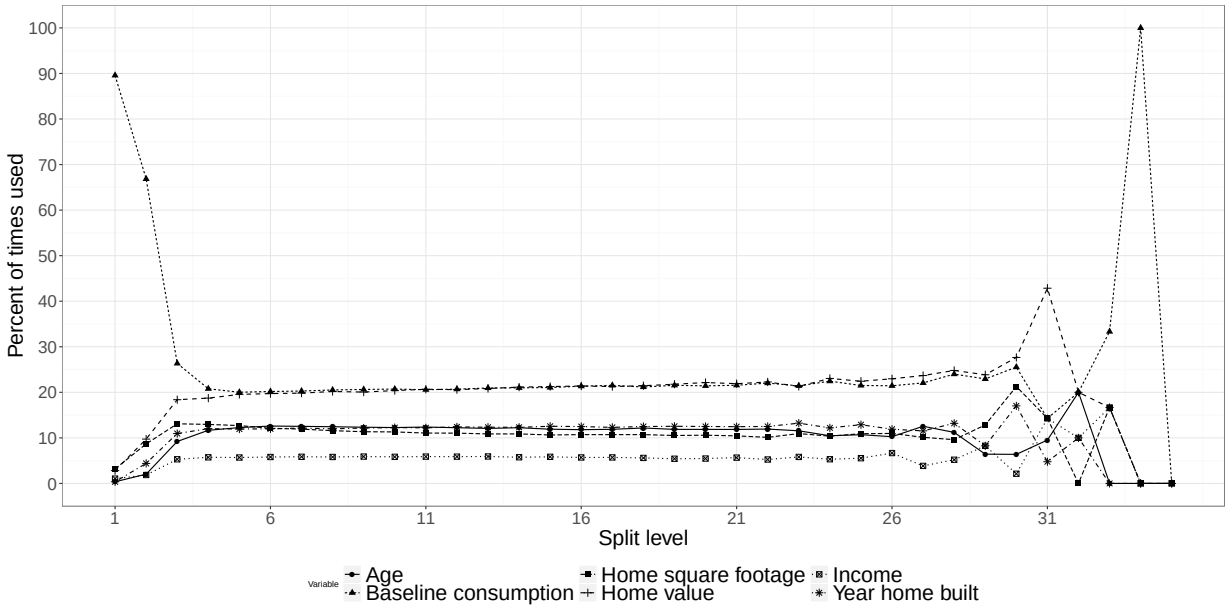
Notes: This table lists summary statistics for all HH in Western Massachusetts (Column (1)), for HH that are not enrolled in a HER program (Column (2)), for HH that are enrolled in a HER program and participated in a wave with available pre-enrollment data (Column (3)). Column (4) checks for balance between treatment and control. Baseline consumption for the unenrolled HH corresponds to average consumption for the entire analysis period. Results are from a linear regression of the listed HH characteristic on treatment status with wave fixed-effects and robust standard errors. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table C4: Summary Statistics for New Hampshire

	Total (1)	Unenrolled (2)	Enrolled (3)	Balance (4)
Monthly consumption (kWh)	558	505	795	-0.77
	(442)	(440)	(364)	(2.48)
Home value ($)	245,744	238,545	275,751	1,042
	(166,378)	(161,232)	(183,283)	(1,459)
Home square footage	1,885	1,844	2,017	1.79
	(1,304)	(1,370)	(1,050)	(8.94)
Annual income ($)	80,855	77,520	95,737	269
	(57,082)	(56,157)	(58,780)	(451)
Education (1-5)	2.95	2.91	3.14	-0.013
	(1.13)	(1.12)	(1.17)	(0.009)
Number of rooms in home	6.59	6.52	6.80	0.022
	(2.29)	(2.39)	(1.95)	(0.019)
Year home built	1,979	1,979	1,980	0.111
	(24)	(24)	(22)	(0.193)
GreenAware score (1-4)	2.29	2.32	2.19	0.001
	(1.12)	(1.11)	(1.13)	(0.009)
Renter (=1)	0.165	0.187	0.084	-0.000
	(0.371)	(0.390)	(0.278)	(0.002)
Single-family occupancy (=1)	0.795	0.768	0.896	-0.003
	(0.404)	(0.422)	(0.305)	(0.003)
Child in home (=1)	0.377	0.372	0.396	0.002
	(0.485)	(0.483)	(0.489)	(0.004)
Participated in EA (=1)	0.414	0.398	0.477	0.004
	(0.493)	(0.490)	(0.499)	(0.004)
Age	57.5	57.1	58.6	0.071
	(15.3)	(15.8)	(13.5)	(0.112)
Observations	393,075	321,699	71,376	

Notes: This table lists summary statistics for all HH in New Hampshire (Column (1)), for HH that are not enrolled in a HER program (Column (2)), for HH that are enrolled in a HER program and participated in a wave with available pre-enrollment data (Column (3)). Column (4) checks for balance between treatment and control. Baseline consumption for the unenrolled HH corresponds to average consumption for the entire analysis period. Results are from a linear regression of the listed HH characteristic on treatment status with wave fixed-effects and robust standard errors. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure C1: Usage of characteristics in random forest for all post-treatment years



Notes: The x-axis denotes the split level of a tree. The y-axis measures the percentage of trees that use a household characteristic at the indicated split level. We plot percentages for the six most frequently-used characteristics: baseline consumption, home value, home square footage, home year built, income, and age of household respondent. See Section 3 for an explanation of the depicted results.

Appendix D - Multiple Imputation

We use multiple imputation (MI) to fill in missing values of household characteristics. We implement MI through the multivariate imputation by chained equations (MICE) approach. The process can be broken down into the following steps:

1. We define a set of variables $X_1, \ldots, X_n$ to be used in the imputation model. Every missing value is filed in at random to act as a placeholder.

2. The placeholder values for the first variable with at least one missing value, X_1, are returned to missing and the observed vales of X_1 are regressed on $X_2, \ldots, X_n$ using a regression model (e.g., linear, logistic) based on the data type of X_1. Predictive mean matching (e.g., known-nearest neighbor) can also be performed.

3. The missing values of X_1 are replaced by simulated draws from the posterior predictive distribution of X_1. In the remaining steps, X_1 consists of the observed and imputed values.

4. Repeat steps 2-3 for the remaining $n-1$ variables where the value of each variable is updated. For example, the next step would be to regress X_2 is regressed on the newly imputed values of X_1 and $X_3, \ldots, X_n$ and estimate missing values of X_2 with draws from its posterior predictive distribution. A "cycle" is said to have passed when all variables have been imputed.

5. Repeat steps 2-4 for 20 cycles to stabilize the results. The placeholder values at the start of each cycle are the imputed values from the previous cycle. A single imputed dataset is produced at the end of all 10 cycles.

6. Repeat steps 1-5 M number of times. (White et al., 2011) suggests that a rule of thumb for deciding M is that M should be a least equal to the percentage of incomplete cases in the dataset.