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

Christopher R. Knittel Samuel Stolper*

January 4, 2021

Abstract

We test the ability of causal forests to improve, through selective targeting, the effectiveness of a program providing repeated behavioral nudges towards household energy conservation. The average treatment effect of the program is a monthly electricity reduction of 9 kilowatt-hours (kWh), but the full distribution of responses ranges from -30 to +10 kWh. In a random hold-out sample, selective targeting of treatment using the forest nearly quadruples the social net benefits of the nudge program. In a parallel test where we train the forest on earlier waves and use it for targeting in later waves with substantially different sample characteristics, the forest doubles social net benefits. Targeting using non machine-learning predictive models also raises program net benefits significantly, but the forest outperforms the best of these by 8-9 percent.

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, Meredith Fowlie, Robert Metcalfe, John van Reenen, and seminar participants at Carnegie Mellon, Chicago, UC Berkeley, University of Connecticut, Yale, MIT, Harvard Kennedy School, and University of Chicago 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 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 (that is, 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 700,000 households and investigate the role of observed and unobserved household characteristics in determining outcomes. To illustrate the value of forest-derived CATEs, we estimate the potential welfare gains from selective targeting of treatment to maximize a social objective function. We first train the causal forest on a random 50% sample and test performance in the other held-out half. To investigate external validity, we then train a new forest on earlier experimental waves in our sample and test performance in the held-out latter waves. Finally, we compare the gains from using a causal forest to those from using a series of non machine-learning regression-based methods.

Our results borrow from, build on, and add to an emerging literature on empirical machine learning (e.g., Davis and Heller, 2017; Burlig et al., 2017; Kleinberg et al., 2018; Hussam et al., 2020). Davis and Heller (2017) contribute an early application of the causal forest algorithm to impact evaluation of a randomized experiment—in their case, a youth summer employment

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

program. In comparison, we investigate the heterogeneous impacts of behavioral "nudges" towards energy efficiency and using a much larger (about 100x) sample. Our findings additionally relate to a large literature on the treatment effects of behavioral nudges, which have wide application ranging from 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 three reasons: first, the large overall sample size available to us provides greater statistical power than is normal in randomized control trials (RCTs). Second, intuition and empirical evidence alike suggest that HERs likely induce a wide variety of behavioral responses (Allcott, 2011; Costa and Kahn, 2013). And, third, the roll out of the experiments across both time and geography provide an opportunity to test the external validity of the methods.

Our central estimate of the pooled average treatment effect (ATE) across all HER 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 steady increase in treatment-driven energy conservation throughout program year 1. There is no evidence of attenuation of program impacts in years 2 and 3; if anything, 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 reveals significant heterogeneity and potential for efficiency improvements: multiple modes are apparent in the distribution of predicted individual treatment effects, and predictions range from -30 to +10 kWh per month (the latter being consistent with a "boomerang effect" (Schultz et al., 2007; Bhanot, 2017; Byrne et al., 2018)). The most commonly-used household characteristics in the forest are baseline (that is, pre-treatment) consumption and home value, which indicates that these variables in particular have significant predictive power. However, the bivariate relationships between individual treatment effect and each of these variables are not linear; the forest captures predictive effects that may not be apparent in the results of conventional regression models.

In our targeting exercise, we compare the monetized social net benefits of the actual HER distribution to the net benefits of sending reports only to those households for which predicted benefits exceed the marginal cost of sending reports. This exercise has three parts. First, we grow a causal forest in a training sample of households and use the forest to predict household-specific treatment effects in a hold-out sample. Second, we multiply these latter predictions by an estimate of the social cost of electricity in Eversource's territory and add predictions of households' willingness to pay for HERs based on Allcott and Kessler (2019) to yield household-level predictions of the monetized social benefits of treatment. Third, we identify households whose predicted social benefits exceed the marginal cost of treatment, estimate an actual average treatment effect in that "targeted" subsample (using difference-in-differences), and use this ATE to calculate actual social net benefits of targeting.

Our results suggest that with perfect foresight—that is, observing one year of pre-treatment usage and two years of post-treatment usage—targeting with the forest nearly quadruples the social net benefits of treatment relative to the actual HER rollout. This is because the forest finds far fewer households (27 percent, or nearly 55,000 households, fewer) that produce net benefits than the actual number of treated households in the HER program. We compare these gains from targeting using the forest to the corresponding gains from using four regression-based predictive models. The simplest of these is a model that allows treatment effect to vary only with baseline consumption, while the most complex includes treatment interactions with all characteristics, their squares, and the product of each combination of characteristics. The forest outperforms all of these, including the best competitor—the simple baseline consumption model—by eight percent.

The above results are based on the random splitting of our full sample into training and hold-out subsamples with the same average characteristics, regardless of the empirical timing of treatment. To simulate a more realistic targeting exercise, we next split our sample by timing of treatment and use results from earlier waves (pre-2015) to predict and target in later waves. This version of the exercise has the additional benefit of shedding light on the external validity of forest predictions, since, in our context, later waves have very different average characteristics from earlier ones. We find that targeting using the forest and data from earlier waves doubles social net benefits relative to the actual rollout—less than when training and hold-out samples were randomly drawn, but still a sizeable improvement. Once again, the forest outperforms the non machine-learning prediction methods, by no less than nine percent in every case. All in all, our results seem to suggest that causal forests can yield meaningful benefits through selective targeting.

In identifying the households that do *not* produce social net benefits, forests also point to further potential welfare gains available through tailoring of treatment to better promote privately and socially beneficial responses.

1 Empirical Context

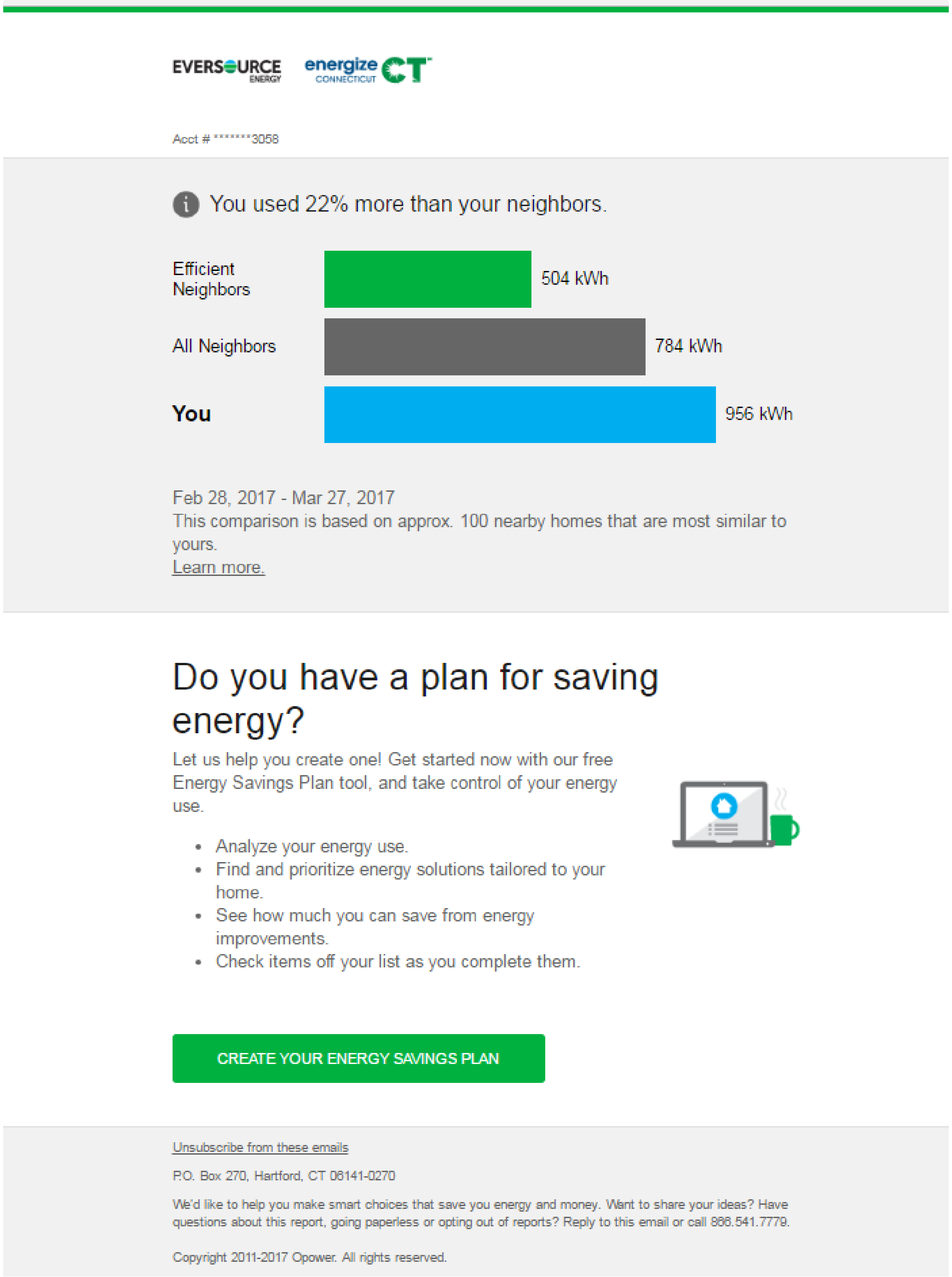
The Home Energy Report (HER) was developed by Opower and rolled out via randomized control trials in participating electric utility service territories beginning in 2008. The initial motivation for the reports came from a field experiment in San Marcos, CA carried out by Schultz et al. (2007), who found 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 1 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 1), or they can displayed directly in the report. Reports are generally sent out either monthly or quarterly. Historically, the great majority of HERs have been delivered by mail in hard-copy form, but Eversource has 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 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.

[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 1: Eversource Home Energy Report



Acct # *******3058

You used 22% more than your neighbors.

Efficient Neighbors 504 kWh

All Neighbors 784 kWh

You 956 kWh

Feb 28, 2017 - Mar 27, 2017
This comparison is based on approx. 100 nearby homes that are most similar to yours.
Learn more.

Do you have a plan for saving energy?

Let us help you create one! Get started now with our free Energy Savings Plan tool, and take control of your energy use.

- Analyze your energy use.
- Find and prioritize energy solutions tailored to your home.
- See how much you can save from energy improvements.
- Check items off your list as you complete them.

Unsubscribe from these emails

P.O. Box 270, Hartford, CT 06141-0270

We'd like to help you make smart choices that save you energy and money. Want to share your ideas? Have questions about this report, going paperless or opting out of reports? Reply to this email or call 866.541.7779.

Copyright 2011-2017 Opower. All rights reserved.

Source: Eversource.

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% (that is, 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 baseline 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 baseline 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. Byrne et al. (2018) identify boomerang effects—that is, unintended positive treatment effects—among low baseline energy users as well as households that overestimate their baseline energy use relative to others. Finally, 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 baseline 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 Eversource'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 2 maps the cov-

[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.

erage 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]

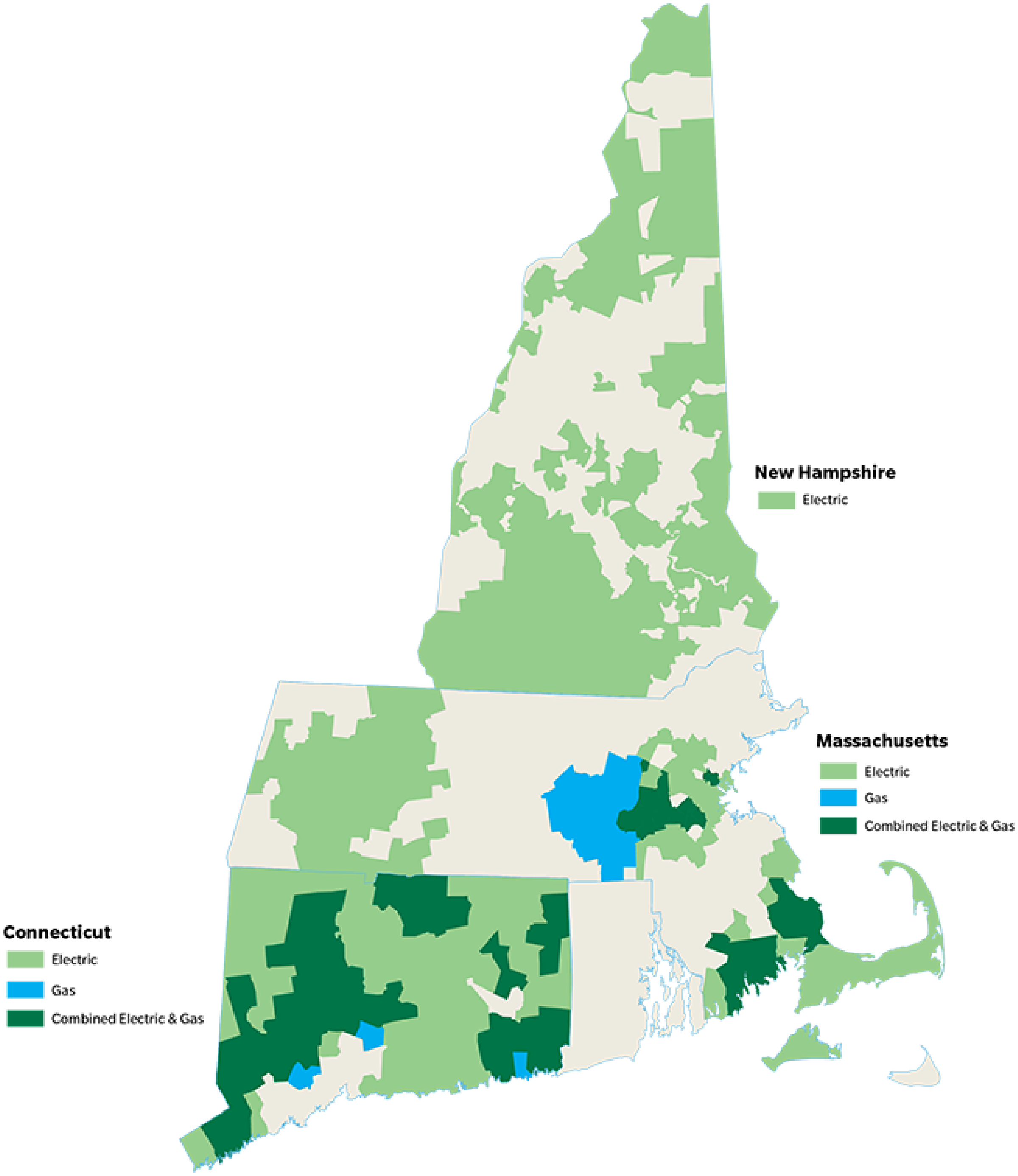
We drop households with outlier values of home square footage and number of rooms, households enrolled in multiple Opower waves, and households that own multiple properties. We further limit our sample to those households for which at least 12 months of pre-experiment data and 12 months of post-experiment 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 fourteen characteristics in our analysis. To capture home attributes, we use home age, value, and square footage, as well as number of rooms. To describe families, we use age of household respondent, 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 baseline consumption, income, educational 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

[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 2: 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.

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 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:

$$kWh_{iwt} = \alpha_1 + \alpha_2 T_{iwt} + X_i \eta + \theta_i + \omega_{wt} + e_{iwt}, \tag{1}$$

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_i and ω_{wt} are household and wave-year-month fixed effects, respectively. We cluster standard errors by wave. α_2 is the coefficient of interest—the average treatment effect in kWh per month.

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:

$$kWh_{iwt} = \beta_1 + \Sigma_{j=-12}^{37} \tau^j D_{iwt}^j + X_i \eta + \theta_i + \omega_{wt} + e_{iwt}. \tag{2}$$

Here, the index j denotes a time period *relative* to the event of interest –the beginning of treatment

[5] Appendix Tables B1-B4 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. * $p < 0.01$, $^{**}p < 0.05$, *** $p < 0.01$.

in the relevant wave. D^j_{wt} 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^0_{wt}—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 fixed effects and clustering 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).

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.

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

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, and following Athey and Wager (2019), we "orthogonalize" 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 3 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:

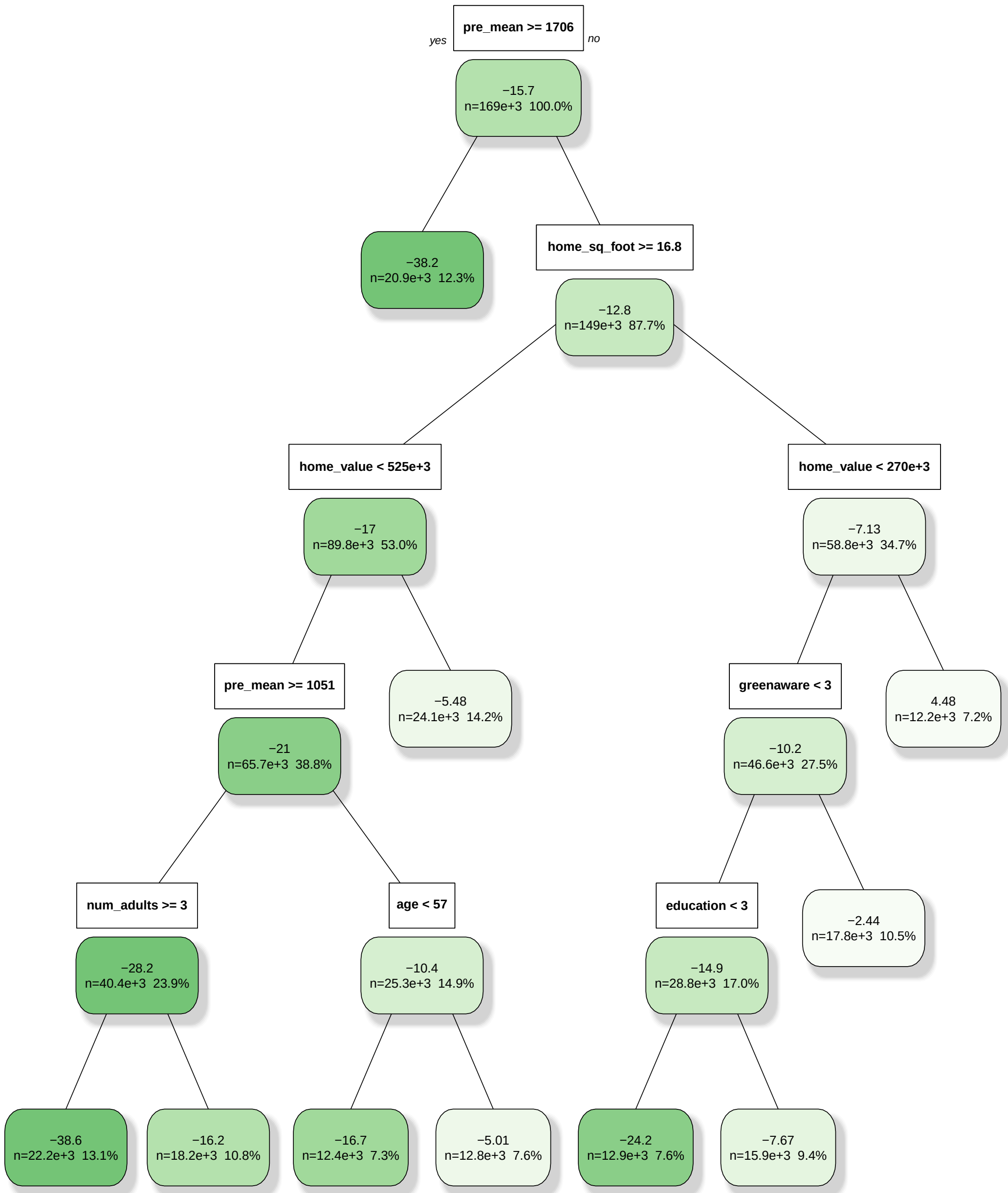
1. Randomly draw (1) a sample of households and (2) a subset of available characteristics.
2. Randomly split the sample in half, creating a "training set" S_{tr} and "estimation" set S_{est}.
3. Using S_{tr}, grow a tree.
4. Match households in S_{est} to leaves of the tree, according to observed characteristics.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

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

[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 3: 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.

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 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 (in our case it is 1,500), as defined in Nie and Wager (2020).

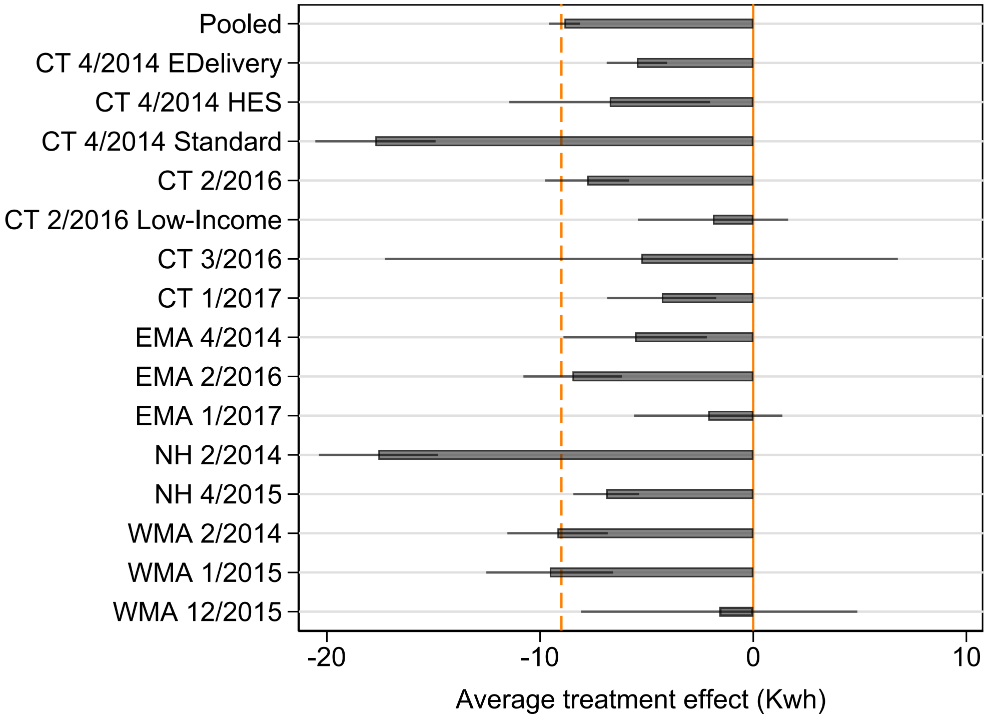
3 Treatment effect estimates

3.1 Average treatment effects

Figure 4 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 part of the explanation. Figure 5—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

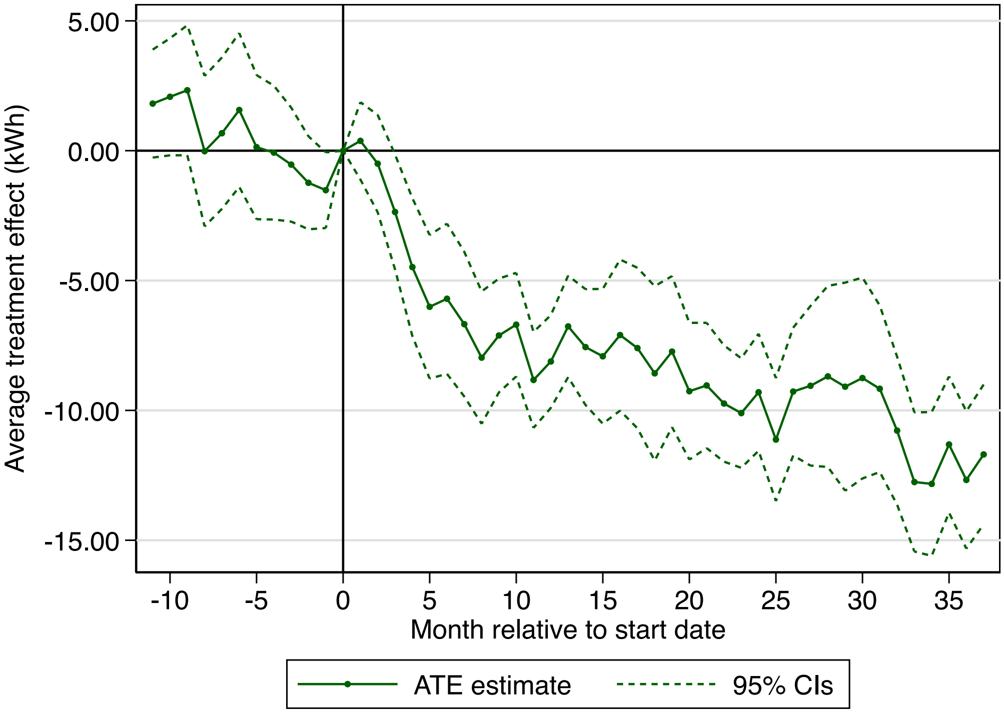
Figure 4: Average treatment effects, by wave: consumption



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.

continue changing behavior into at least the third year of treatment.

Figure 5: Event study of pooled experimental waves: consumption



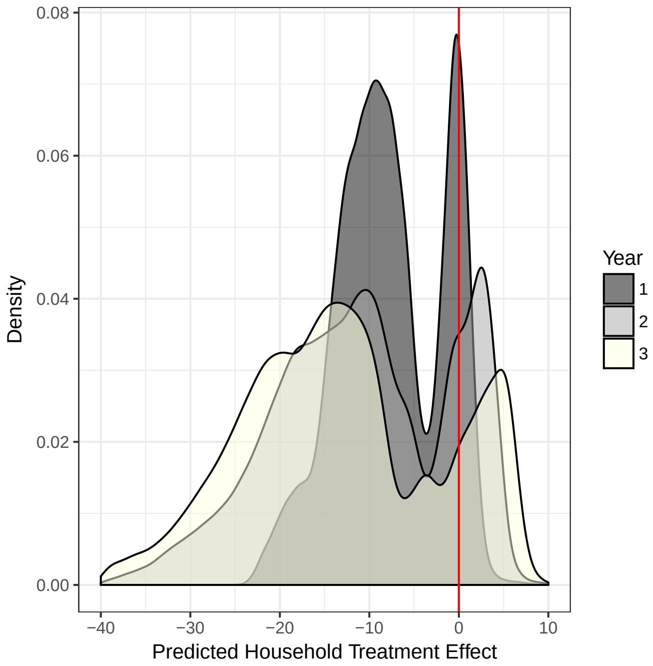
Notes: The solid-line data points are event-study coefficients from estimation of Equation 2. Dashed lines indicate 95% confidence intervals. D^0_{iwt}—which corresponds to the month immediately preceding program start—is omitted.

3.2 Conditional average treatment effects

Figure 6 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.[9] 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] Appendix Figure A1 depicts a single, aggregate distribution of predicted household treatment effects, calculated across program years 1 to 3.

Figure 6: Distribution of Predicted Treatment Effects

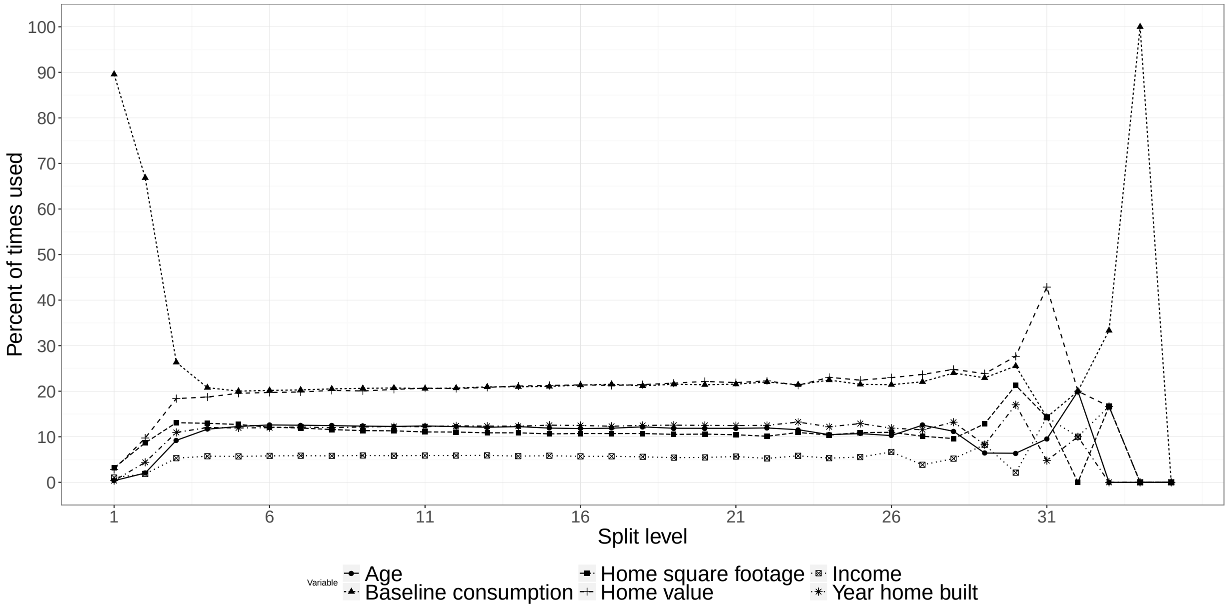


Notes: Each plotted distribution is a kernel density of household treatment effects in a specific year (1, 2, or 3) of HER programming. The sample is fixed across years: only households with non-missing consumption in all three post-years are included. Treatment effect predictions come from our causal forest (Section 2.2).

What drives all this heterogeneity? Figure 7 plots the frequency of selected characteristics' use as a splitting variable in the forest, conditional on being available (recall that only a random subset of characteristics is considered in each tree). The six characteristics included in this figure—baseline consumption, home value, home square footage, the year in which a home was built, income, and respondent's age—are the most frequently used. Among these, the first two are easily the most common splitting variables. Baseline consumption is chosen as the initial splitting variable in 90 percent of trees in which it is eligible. Home value catches up to baseline consumption in frequency of use at the sixth split level. Beyond that point, these two attributes are used about twice as frequently as the other four shown (20 percent of the time vs. 10 percent).

While frequency of use in tree growth provides some insight into the relative predictive power of characteristics, it does not clarify *how* these characteristics are related to treatment effects. To shed some light on these relationships, we zoom in on the two most frequently-used characteristics: baseline consumption and home value. Figure 8 provides evidence on the relationship between the empirical distribution of predicted treatment effects and each of these two 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.

Figure 7: Usage of characteristics in the causal forest

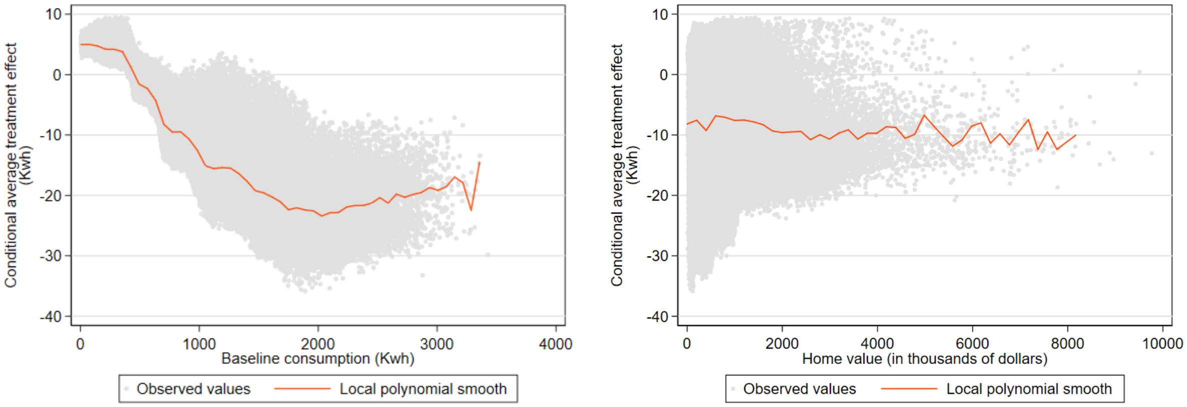


Notes: Each line plots, on the y-axis, the empirical likelihood of a specific characteristic being chosen to define a forest split at split level x, conditional on being available as a splitting variable (only a random subset of characteristics are available for each tree). We show percentages for the six most frequently-used characteristics: baseline consumption, home value, home square footage, home year built, income, and age of household respondent. The underlying sample includes all households with non-missing consumption in the year prior to program start and at least one of the first three years following program start. The dependent variable in the forest is average consumption across the three post years minus average consumption in the first pre year. See Section 2.2 for further implementation details.

Figure 8: Predicted treatment effect vs. household type

Panel A. Baseline consumption

Panel B. Home value



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. Lines display the local smoothed polynomial relationship between ATE and the characteristic. The sample includes all households with non-missing consumption in the year prior to program start and at least one of the first three years following program start. The dependent variable in the forest is average consumption across the three post years minus average consumption in the first pre year. See Section 2.2 for further implementation details.

Both panels of Figure 8 hint at the potential value of non-parametric prediction methods like the causal forest. Relatively simpler predictive models may miss the non-linearity of the relationship between treatment effect and baseline consumption, or they may miss the importance of the home value variable altogether. Panel A exemplifies 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 be predicted to avoid nearly all boomerang effects. Meanwhile, the interpretation of the results in Panel B is more nuanced. Home value, the second most frequently-used attribute in the forest, has almost zero raw correlation with treatment effect. However, if one wanted to better understand the characteristics of the very largest "reducers", Panel B is helpful; such households are confined to the very bottom of the home value distribution. Nobody with home value above 100,000 dollars is predicted to reduce monthly consumption by more than 23 kWh, while the households below that dollar threshold in some cases are predicted to reduce by 30-35 kWh.

4 Selective targeting of treatment

To further investigate the potential gains in program effectiveness from selective targeting, we develop an exercise that simulates a planner's decisions about whom to treat going forward, based on previously observed treatment effects. There are, of course, many possible objective functions that a planner may seek to maximize; we focus here on maximizing aggregate social net benefits produced by treatment. We note, however, that equity is an important component of the true social welfare function. Thus, it is important to keep in mind what might be done for those who do not receive this particular HER treatment (e.g., tailoring the treatment to make the reports more valuable for this group, or increasing spending on other programs for this group (Reames et al., 2018)).

For this exercise, we require estimates of three key values in addition to our treatment effect predictions: the marginal cost of sending HERs; the social marginal cost of electricity (which includes both generation costs and environmental externalities); and customer willingness-to-pay (WTP) for HERs. Based on consultation with Eversource, we assume that the marginal cost of HERs is \$7 per household per year.[10] For the social marginal cost of electricity, we use the short-run estimate of Borenstein and Bushnell (2018) for the New England electricity region in 2016—\$0.065/kWh. To estimate WTP for HERs, we borrow from Allcott and Kessler (2019), who elicit WTP for HERs experimentally and report results from a regression of household-specific

[10] We acknowledge, however, that the true social marginal cost could be below the price charged to Eversource by Opower.

WTP on the logarithm of income, indicators for retirement, marriage, homeownership, and single-family occupancy, and homebuyer's credit worthiness score. We use their regression coefficients 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.

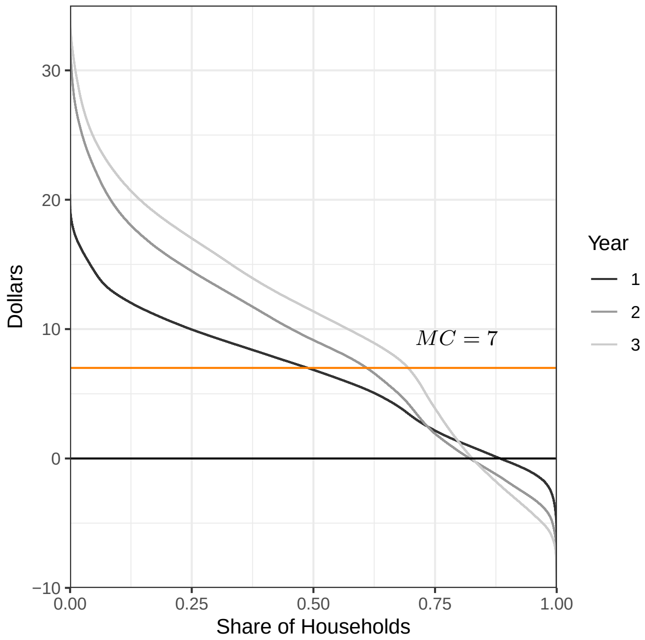
With these estimates, we can then compare the predicted social benefits produced by a given household—here, the sum of its predicted WTP and the estimated social value of its predicted electricity savings—to the marginal cost of sending the reports. Figure 9 graphically depicts this comparison, by plotting the (reverse) cumulative distribution function (CDF) of household-specific, predicted social benefits in each of the first three years of HER programming alongside the (constant) marginal cost curve. In every year, the CDF crosses both the zero line and the marginal cost line; that is, there are always households whose predicted responses to HERs translate to both gross and net negative benefits, respectively.

Taken at face value, the graph suggests that sending only to households whose induced annual social benefits exceed seven dollars would yield aggregate net benefits equivalent to the area between the CDF and marginal cost curve, from their crossing point onwards. However, the CDF here is composed entirely of "in-sample" predictions; that is, all of the households whose predicted treatment effects are plotted in Figure 9 were used to grow the causal forest. To the extent that the forest suffers from overfitting, estimates of the net benefits of targeting calculated directly from Figure 9 will be biased upwards.

We design our targeting exercise to address this issue, by building a predictive model using one sample and testing the effectiveness of targeting in another. The full exercise consists of the following steps:

1. Split the full set of available households into two: a training sample for estimating the model, and a hold-out sample for evaluation of targeting.
2. Choose a predictive modeling method: causal forest, or non machine-learning regression.
3. Estimate the predictive model with the training sample.
4. Predict household-level treatment effects in the hold-out sample using the estimated model and translate into social benefits.

Figure 9: The predicted cumulative distribution of HER-induced social benefits



Notes: Each downward-sloping line is the reverse cumulative distribution function of annual social benefits produced by households in a given HER program year, estimated via our causal forest. The sample is fixed across years: only households with non-missing consumption in all three post-years are included. The line labeled "MC = 7" denotes the estimated marginal cost of sending one year's worth of HERs to a household.

5. Identify all households whose induced social benefits exceed marginal cost; this is the group "targeted" for treatment.
6. Estimate an average treatment effect in the targeted group using difference-in-differences regression.
7. Calculate aggregate social net benefits created in the targeted group and compare to those of alternatives.

Choosing a modeling method in step 1 above allows us to test not only the performance of the causal forest relative to Opower's observed treatment assignment, but also the performance of the forest relative to simpler data-driven alternatives. We choose four non machine-learning, regression-based alternatives to test, motivated by convention and prior knowledge of HERs' impact. The first, and simplest, of these (which we denote "Baseline") is a linear model in which treatment effect is allowed to vary only with baseline electricity consumption. The sparseness of this model means an electric utility could estimate it without the need for demographic, socioeconomic, or home attributes, and it also may lessen the risk of overfitting. Furthermore, baseline consumption stands out as an important predictor of treatment effect in our work as well as that of others before us (Allcott, 2011; Ayres et al., 2013). Anecdotally, Opower too targets high baseline electricity users, so that this first model can be thought of as a formal version of what is already done in the field.

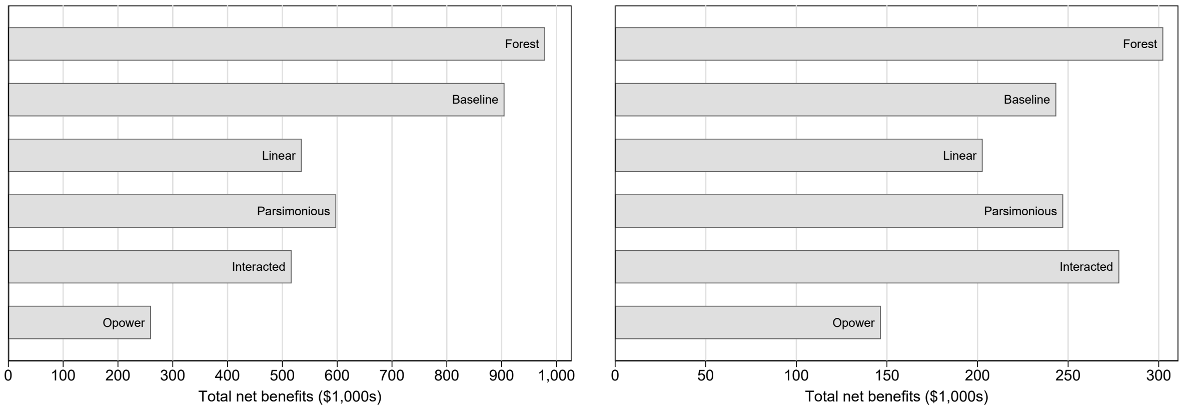
Each of our other three regression-based models builds successively on the previous one. Our second (denoted "Linear") model parameterizes treatment effect to vary linearly with all fourteen of our household characteristics. Our third ("Parsimonious") builds on the second by adding interactions between treatment and the square of each characteristic. And our fourth ("Interacted") adds, on top of that, treatment interactions with the product of each *pair* of characteristics. We specify all of these models so that estimation is as similar as possible to that of the forest. Thus, we run cross-sectional regressions where the dependent variable captures a household's pre-post difference in electricity consumption; in this exercise, we focus on post-year 2. We use the same "orthogonalized" (or "residualized") dependent variable and treatment assignment as are described for the causal forest in Section 2.2.

In step 2 above, we try two different ways of splitting the full sample into training and hold-out groups. In the first, we split the full sample into two randomly, using one to grow the forest and the other to simulate selective targeting. This splitting rule facilitates a test of whether the forest would be accurate in a hold-out sample with the same average characteristics; we therefore think of this first version of the exercise as a good evaluation of each model's *internal* validity. In the second, we split the full sample by the timing of wave start date; the training sample is composed of all households whose program wave started before 2015 (406,637 households), and the hold-out sample is composed of those with wave start date in 2015 or 2016 (83,424 households).[11] By using earlier waves to predict outcomes in later waves, we better approximate the situation in which a utility (or any other service provider) might find itself. In particular, the average characteristics of the two samples are very different in this second version of the test (see Appendix Table B5). We thus view this version as shedding light on the *external* validity of each method.

Figure 10 depicts, for each of the two versions of the exercise described above, the absolute performance of the five different methods and the as-delivered Opower treatment assignment. As displayed in Panel A, forest-based targeting in a random hold-out sample yields aggregate net benefits equal to $979,517 in post-year 2, whereas the actual Opower program in that sample produced only $260,271 in net benefits. This nearly four-fold (3.76x) increase in net benefits is the result of the forest identifying approximately 55,000 fewer households to whom to send treatment, relative to the actual Opower program. In particular, this number represents the aggregate effect of switching 77,452 households from treatment to control, and 22,348 households from control to treatment.

[11] We omit one wave in Western Massachusetts that begins in December 2015, because all other waves begin in the first four months of a year, and omitting this wave thus reduces treatment effect variation coming from the season of in which a wave starts.

Figure 10: Social benefits of targeting, by predictive method


Panel A. Training on a random sample

Panel B. Training on pre-2015 waves

Notes: Bars are estimated net benefits of sending a year's worth of HERs to only those households predicted to yield social benefits exceeding the marginal cost of sending HERs. Bar labels denote the model used to predict household-level benefits. Panel A depicts results from building all predictive models with a 50% random sample of households and targeting in the other 50% "hold-out" sample. Panel B depicts results from building all predictive models exclusively with households in HER waves beginning in 2014 and targeting among waves beginning in 2015 or later. The sample includes all households with non-missing consumption in the year prior to program start and in post-year 2. The dependent variable in the forest is average consumption in post-year 2 minus average consumption in the first pre year. See Section 4 for implementation details.

The forest also produces more net benefits than any of the alternative, non-ML predictive models in this first version of the targeting exercise. Only the "Baseline" model comes close to matching the forest's performance; the latter outperforms the former by 8.2 percent. Both of these models do far better than the other three non-ML models. Including successively more predictors in the regression-based models apparently leads to significant overfitting in this case.

The aggregate net benefits of targeting are smaller in the second version of the exercise, depicted in Panel B, as a mechanical result of their being far fewer households in the hold-out sample. However, the forest still outperforms both the Opower benchmark and all four regression-based models. The forest produces twice the net benefits of Opower's actual distribution in post-2014 waves. Again, the forest outperforms the best alternative by a little more than 8 percent (8.7), but this time the best alternative is the "Parsimonious" model. In both versions, then, all five targeting methods do significantly better than Opower's actual program, and the forest produces the largest aggregate net benefits of all targeting methods.[12]

[12] We present the results of three other targeting exercises in Appendix A. In each exercise, we estimate the aggregate net benefits produced from targeting the top quantile (half, quartile, or decile) using each predictive model.

5 Conclusion

Machine learning holds great promise as a tool for high-resolution evaluation and prediction. In this paper, we test that promise in the context of large-scale experiment promoting household energy conservation. We leverage fifteen experimental waves covering more than 900,000 households, in which the treatment is a periodic social comparison message designed to nudge households to reduce electricity consumption. We use the causal forest machine-learning algorithm, an ensemble method based on classification and regression trees that has been adapted for causal inference.

The causal forest that we estimate reveals several facts about treatment effects in this context. First, there is wide variation in responses to Home Energy Reports. The overall average treatment effect is a nine kilowatt-hour monthly reduction in electricity consumption, but individual effects range from -40 to +10 kWh. Second, there are multiple statistical modes, and these evolve differentially over time. Those who reduce consumption in program year 1 tend to reduce even more in subsequent years, while those who respond minimally or not at all to treatment in year 1 tend to *raise* consumption later on. Third, baseline consumption and home value are the household characteristics most frequently used to grow the forest. Altogether, these facts illustrate the potential for improved program effectiveness through targeting and tailoring of treatment.

To test this potential, we develop an out-of-sample prediction exercise that mimics the way targeting may be done in the real world. The exercise allows us to compare the benefits of forest-based targeting to those of the actual Opower distribution as well as non-ML based targeting. In the *internal validity* version of this test—in which the training and hold-out samples are randomly drawn—the forest produces nearly four times the aggregate social net benefits as the Opower program and eight percent more benefits than the best non-ML predictive method. In the *external validity* version of this test—where we train the predictive model on 2014 waves and target in post-2014 waves (which are very different)—the forest produces twice the benefits of the Opower program and, again, eight percent more benefits than the best non-ML predictive method.

Our results show how causal forests can be used to improve program effectiveness and social welfare through targeting. In addition, they point to the potential for even further welfare gains through selective *tailoring* of treatment. Those whom targeting identifies as undesirable to treat as is are strong candidates to receive adjusted treatments or programming that meets their specific needs. All told, we believe that high-resolution predictive methods like the forest are a useful tool for improving cost effectiveness, social welfare, and distributional equity in a broad variety of settings.

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, and S. Wager (2019): "Generalized random forests," *The Annals of Statistics*, 47, 1148–1178.

Athey, S. and S. Wager (2019): "Estimating Treatment Effects with Causal Forests: An Application," *Working paper.*

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. Fernández-Val (2018): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," Working Paper 24678, 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. M. and S. B. Heller (2017): "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. Roth (2020): "Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field," Working Paper 20-082, Harvard Business School.

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 paper, World Bank.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018): "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 133, 237–293.

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

Nie, X. and S. Wager (2020): "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.

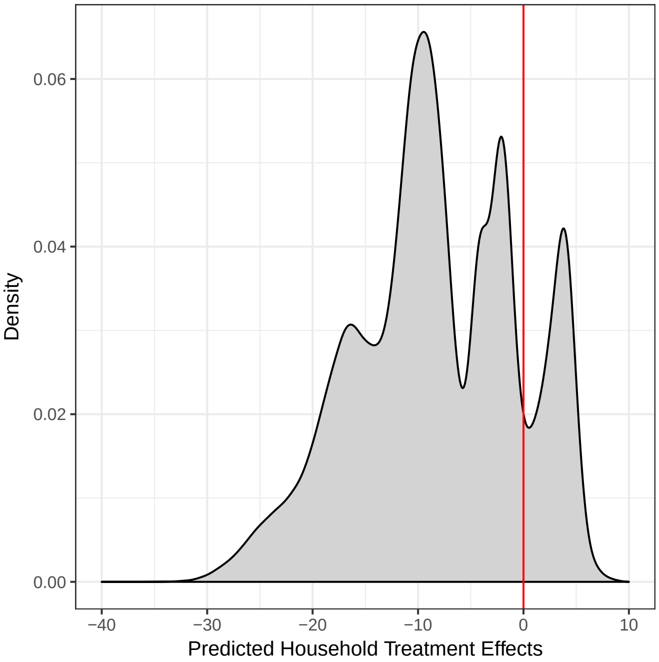
Tibshirani, J., S. Athey, S. Wager, R. Friedberg, L. Miner, and M. Wright (2018): *Package 'grf'*.

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.

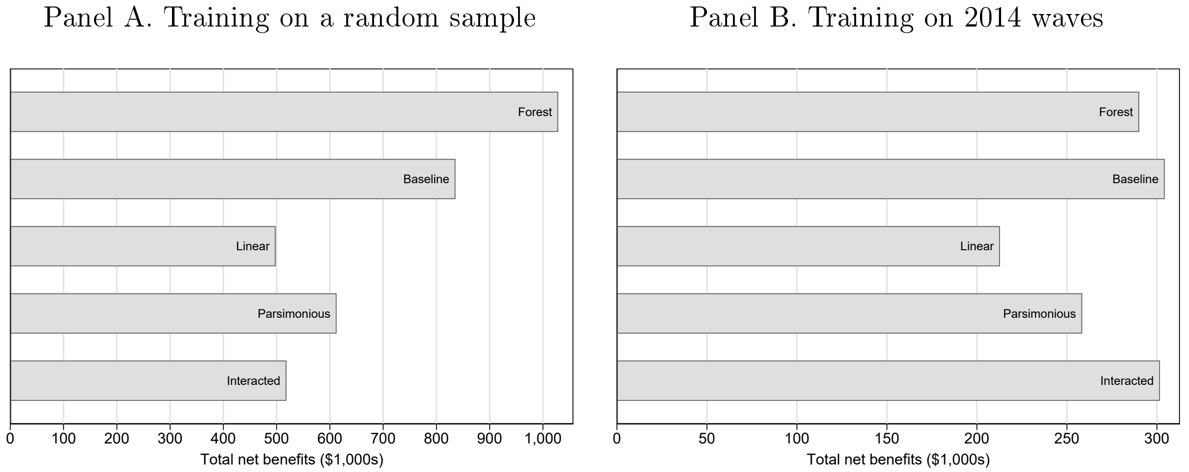
Appendix A: Additional forest-based results

Figure A1: Distribution of Predicted Treatment Effects: 3-Year Average



Notes: The plotted distribution is a kernel density of predicted household treatment effects. The sample includes all households with non-missing consumption in the year prior to program start and at least one of the first three years following program start. The dependent variable in the forest is average consumption across the three post years minus average consumption in the first pre year. See Section 2.2 for details.

Figure A2: Estimated total net benefits produced by the top half of households, by method

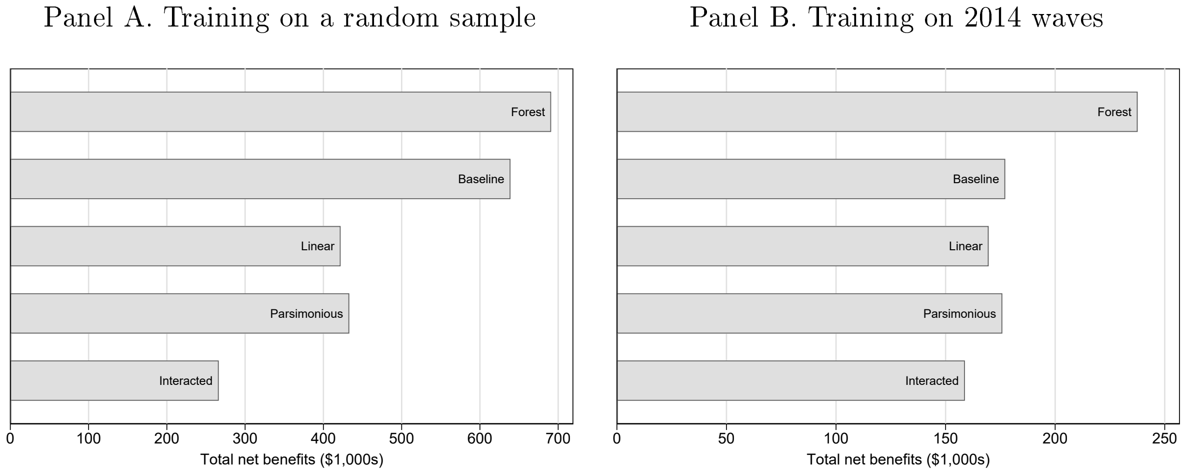


Panel A. Training on a random sample

Panel B. Training on 2014 waves

Notes: Bars are estimated net benefits of sending a year's worth of HERs to only those households predicted to yield social benefits in the top half of the distribution. Bar labels denote the model used to predict household-level benefits. Panel A depicts results from building all predictive models with a 50% random sample of households and targeting in the other 50% "hold-out" sample. Panel B depicts results from building all predictive models exclusively with households in HER waves beginning in 2014 and targeting among waves beginning in 2015 or later. The sample includes all households with non-missing consumption in the year prior to program start and in post-year 2. The dependent variable in the forest is average consumption in post-year 2 minus average consumption in the first pre year. See Section 4 for implementation details.

Figure A3: Estimated total net benefits produced by the top quartile of households, by method

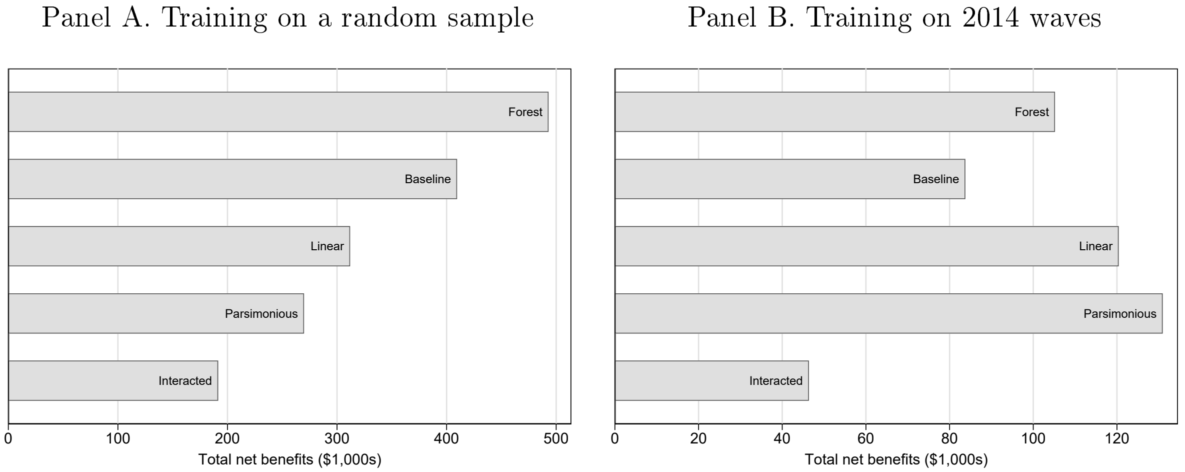


Panel A. Training on a random sample

Panel B. Training on 2014 waves

Notes: Bars are estimated net benefits of sending a year's worth of HERs to only those households predicted to yield social benefits in the top quartile of the distribution. Bar labels denote the model used to predict household-level benefits. Panel A depicts results from building all predictive models with a 50% random sample of households and targeting in the other 50% "hold-out" sample. Panel B depicts results from building all predictive models exclusively with households in HER waves beginning in 2014 and targeting among waves beginning in 2015 or later. The sample includes all households with non-missing consumption in the year prior to program start and in post-year 2. The dependent variable in the forest is average consumption in post-year 2 minus average consumption in the first pre year. See Section 4 for implementation details.

Figure A4: Estimated total net benefits produced among by top decile of households, by method



Panel A. Training on a random sample

Panel B. Training on 2014 waves

Notes: Bars are estimated net benefits of sending a year's worth of HERs to only those households predicted to yield social benefits in the top decile of the distribution. Bar labels denote the model used to predict household-level benefits. Panel A depicts results from building all predictive models with a 50% random sample of households and targeting in the other 50% "hold-out" sample. Panel B depicts results from building all predictive models exclusively with households in HER waves beginning in 2014 and targeting among waves beginning in 2015 or later. The sample includes all households with non-missing consumption in the year prior to program start and in post-year 2. The dependent variable in the forest is average consumption in post-year 2 minus average consumption in the first pre year. See Section 4 for implementation details.

Appendix B: Additional summary statistics

Table B1: 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 B2: Summary Statistics for Eastern Massachusetts

	Total	Unenrolled	Enrolled	Balance
	(1)	(2)	(3)	(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 B3: 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 B4: 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. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B5: Summary Statistics for Training and Prediction Samples - Group 3

	Training	**Hold-out**	
	Mean	Mean	Difference
Home value ($)	373,917.884	392,921.314	-19,003.430***
	(402,062.109)	(351,003.166)	(1,473.418)
Home square footage	19.610	20.056	-0.446***
	(11.091)	(14.690)	(0.061)
Annual income	103,470.842	91,397.997	12,072.845***
	(68,038.072)	(65,449.361)	(250.816)
Education (1-5)	3.266	3.172	0.094***
	(1.231)	(1.276)	(0.005)
Num Adults	2.602	2.356	0.246***
	(1.360)	(1.358)	(0.005)
Number of Rooms in Home	7.064	7.089	-0.026*
	(2.167)	(2.699)	(0.012)
Year home built	1,969.490	1,969.215	0.275*
	(24.675)	(29.327)	(0.126)
GreenAware score (1-4)	2.139	2.169	-0.030***
	(1.142)	(1.110)	(0.004)
Renter (=1)	0.086	0.187	-0.101***
	(0.281)	(0.390)	(0.001)
Single-family occupancy (=1)	0.882	0.737	0.145***
	(0.322)	(0.440)	(0.002)
Child in home (=1)	0.450	0.467	-0.016***
	(0.498)	(0.499)	(0.002)
Participated in EA (=1)	0.349	0.450	-0.100***
	(0.477)	(0.497)	(0.002)
Age	57.964	55.493	2.471***
	(14.534)	(15.646)	(0.063)
Baseline Consumption (kwh)	901.648	722.313	179.335***
	(452.727)	(368.487)	(1.460)
F-test			935.309
			(0.000)
N	406,637	83,424	

Notes: Columns 1 and 2 display the mean of the listed household characteristic for the training sample and hold-out sample, respectively. Standard deviations are listed beneath in parentheses. Column 3 reports the difference between the two sample means, which we estimate via bivariate regressions, and robust standard error in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. We additionally report an F-statistic and corresponding p-value from a test of the joint significance of the these differences.

Appendix C: 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.