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

Christopher R. Knittel Samuel Stolper*

October 26, 2024

Abstract

We test the ability of causal forests to improve, through selective targeting, the effectiveness of a randomized 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 predicted reductions ranges from roughly 1 to 33 kWh. Pre-treatment electricity consumption and home value are the strongest predictors of differential treatment effects. In a pair of targeting exercises, use of the causal forest increases social net benefits of the nudge program by a factor of 3-5 relative to the status quo. Using earlier program waves to target in later ones, we estimate that the forest produces more benefits than five other alternative predictive models. Bootstrapping to generate confidence intervals, we find the forest's advantage to be statistically significant relative to some but not all of these alternatives.

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, Deputy Director for Policy, MIT Energy Initiative, and NBER, knittel@mit.edu; Stolper: University of Michigan School for Environment and Sustainability, sstolper@umich.edu. Leila Safavi, Paula Meloni, and Shereein Saraf 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 many social and private-sector programs.[1] Concurrent with the rise of RCTs to evaluate programs, following the seminal work of Thaler and Sunstein (2008), there has a been a large increase in the use of "nudges" to move behavior in welfare-enhancing ways;[2] DellaVigna and Linos (2022) report that worldwide, more than 200 government teams are devoted to using behavioral science to improve government programs and outcomes. Yet the interventions used by these agencies are not costless and often have 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, 2019). They offer disciplined ways to search non-parametrically for heterogeneity, as well as strategies for minimizing overfitting and thus maximizing out-of-sample predictive power. There is an active body of research on the design of ML algorithms for causal inference (e.g., Imai and Ratkovic, 2013; Wager and Athey, 2018), program evaluation (e.g., Chernozhukov et al., 2022; Knaus, 2022), and optimal targeting (e.g., Kitagawa and Tetenov, 2018; Athey and Wager, 2021), and empiricists have quickly found use for these tools across a wide variety of settings (e.g., Kleinberg et al., 2018; Deryugina et al., 2019; Allcott and Kessler, 2019). Tree-based methods (Breiman et al., 1984; Breiman, 2001)—in which a sample is repeatedly split into subsets, or "leaves" on a tree—are one class of ML algorithms in which significant progress has been made: researchers have adapted existing methods for causal estimation of conditional average treatment effects (CATEs) via *causal trees* (Athey and Imbens, 2016) and *causal forests* (Wager and Athey, 2018); empirical studies increasingly leverage the causal forest in program evaluation (Davis and Heller, 2020; Gulyas and Pytka, 2020; Ellickson et al., 2021; Knittel and Stolper, 2021).

In this paper, we apply the causal forest algorithm to evaluate a series of large-scale randomized experiments in household energy use. We use the causal forest to predict treatment effects among 700,000 households and investigate the relationship between treatment effects and household attributes. To illustrate the practical value of forest-derived CATEs, we estimate the welfare gains from selective targeting of treatment to maximize a social objective function. We compare

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

[2] Thaler and Sunstein (2008) define a nudge as a "choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives."

the targeting performance of the causal forest to that of the status quo (i.e., actual) treatment assignment, as well as a lasso model and four non machine learning, regression-based predictive models of differing complexities.

Our results contribute to the growing empirical literature on the use of machine learning to target treatment, which features applications to (for example) government tax rebates (Andini et al., 2018), marketing (Ascarza, 2018) and fundraising (Cagala et al., 2021) communications, and programs related to household energy use (Gerarden and Yang, 2022; Christensen et al., 2022). To this literature, we add a case study of causal forests, applied to a large-N experiment in nudges to conserve energy. Understanding the effectiveness of behavioral interventions to reduce energy consumption is also of independent interest. Electricity and heating account for roughly a third of global carbon emissions (World Resources Institute, 2022), and many expect electricity use to increase considerably through the electrification of transportation, heating, and cooking. Policymakers are, therefore, looking for potentially cost-effective ways to reduce electricity and heating demand.

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 spans 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 estimating heterogeneous treatment effects for three reasons: first, the large overall sample size available provides greater statistical power than is normal in randomized control trials (RCTs). Second, intuition and empirical evidence suggest that HERs induce various behavioral responses (Allcott, 2011; Costa and Kahn, 2013). And third, the roll-out of the experiments across both time and geography provides 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 any-

thing, 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 yields an estimated range of household treatment effects of roughly -1 to -33 kWh per month, and the whole treatment effect distribution shifts leftwards (that is, downwards) each successive year. Many households respond by increasing their energy use reductions over time; at the same time, many households have persistently low-magnitude treatment effects. The most commonly-used household attributes in the forest are measures of absolute baseline (that is, pre-treatment) consumption, *relative* baseline consumption (a proxy for the social comparison that each treated household receives), and home value. This suggests that the distribution of households' predictions is caused by both effect heterogeneity and treatment heterogeneity (through the social comparison). We estimate non-parametric group average treatment effects (GATEs; Knaus, 2022) and conduct a classification analysis (CLAN; Chernozhukov et al., 2022) to better understand how attributes differ across the treatment effect distribution. In particular, we find that the size of the treatment effect (i.e., reduction) is monotonically increasing in both absolute *and* relative electricity consumption.

In our targeting exercise, we train a predictive model on one subsample of households and use it to assign treatment in another, held-out subsample, with the rule of only sending reports to those households for which predicted social benefits exceed the marginal cost of sending reports. We then estimate actual social net benefits in this group targeted for treatment, bootstrapping to generate confidence intervals. We "horserace" six predictive models—the causal forest, a lasso, and four regression-based alternatives—so that we are able to not only gauge how the forest performs relative to the status quo treatment assignment but also how the forest performs relative to computationally simpler options. And we run the whole exercise twice—once assessing targeting in a randomly drawn hold-out subsample and a second time training the model on chronologically earlier experimental waves and assessing targeting in later waves with disparate household attributes.

In our primary specifications of the targeting exercise, the forest produces three to five times the social net benefits of the status quo treatment assignment. It also outperforms each of the five alternative predictive models regardless of whether we train on a random sample or split by wave start date. However, not all of these differences are statistically significant. When we split randomly, several of the differences are economically significant, and the forest's gains relative to the lasso and the most complicated regression model are statistically significant (in the latter case, marginally so); when we split by wave start date, the forest's gains are statistically significant relative to the three more complex regression models. The simplest regression-based predictive model of treatment effects, which depends only on absolute baseline consumption, is

competitive with the forest across all our targeting exercises – consistent with the conventional wisdom that baseline consumption is a major determinant of HER treatment effects. In aggregate, the lasso is the second-best performing alternative model. Put together, our results suggest that taking active steps to address overfitting risk is important: the machine learning (forest and lasso) models and the baseline-consumption model – which mitigate overfitting algorithmically and by manually choosing a (very) sparse model, respectively – separate themselves from the more complex regression models in the chronological targeting exercise, which is the more difficult out-of-sample prediction test.

1 Empirical Context

1.1 Home Energy Reports

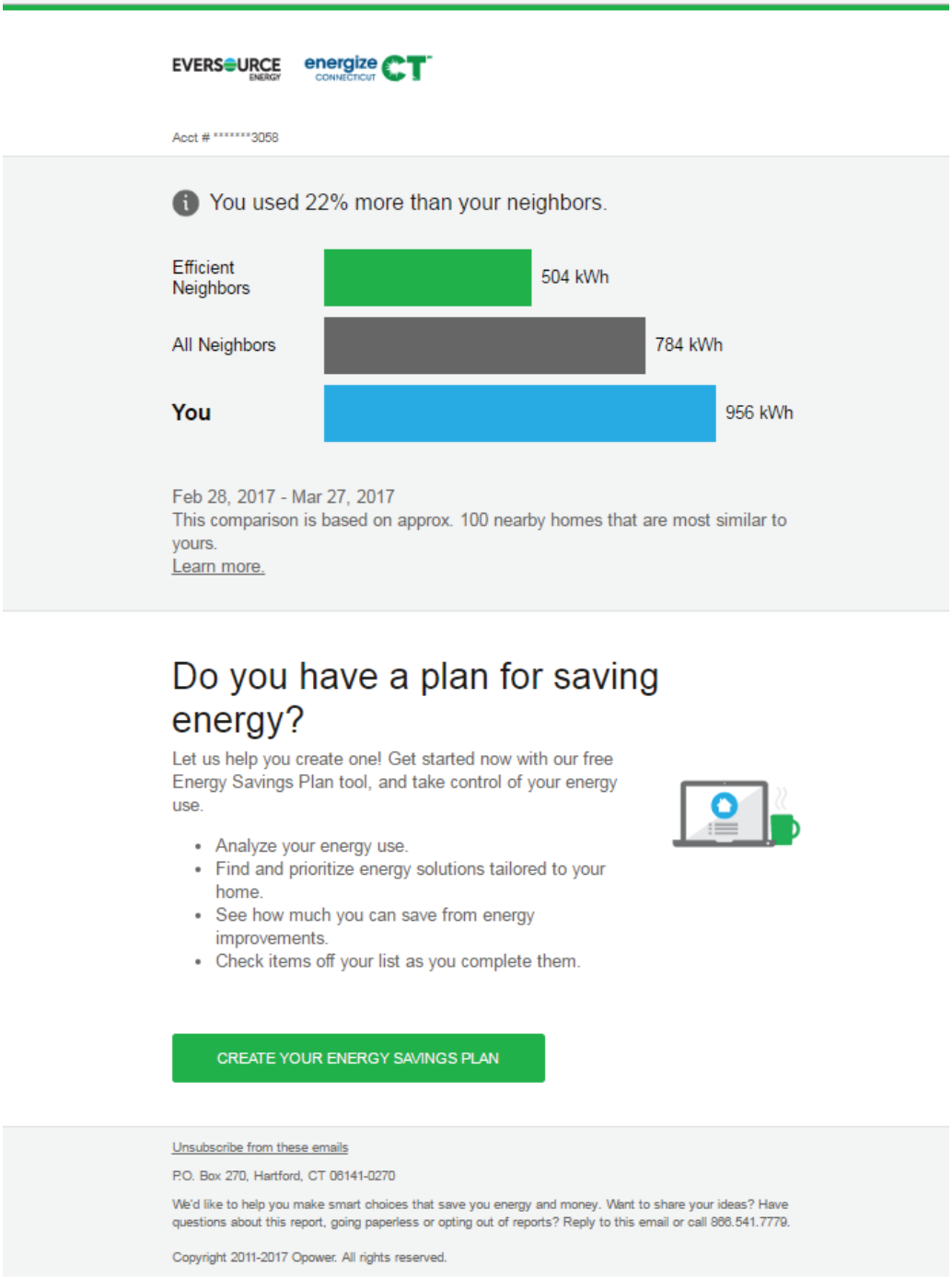
The Home Energy Report (HER) was developed by Opower and rolled out via randomized control trials in participating electric utility service territories beginning in 2008. The initial motivation for the reports came from a field experiment in San Marcos, CA carried out by Schultz et al. (2007), who found 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).[3] 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

[3]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



Source: Eversource.

customers. Perhaps the most frequently discussed reason is compliance with energy efficiency resource standards, which, in 33 states (National Conference of State Legislatures, 2021), 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% (that is, a 2% monthly reduction in electricity consumption).[4] 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.

Several studies document heterogeneous effects of HERs on savings and well-being. 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. Allcott and Kessler (2019) elicit willingness-to-pay for HERs and identify significant heterogeneity across households. Lastly, Gerarden and Yang (2022) estimate significant social benefits of targeting HERs using empirical welfare maximization (Kitagawa and Tetenov, 2018). Our work is most similar to Gerarden and Yang (2022), but is differentiated by its focus on the

[4]In Allcott (2011)'s context, 2.0% is equivalent to 0.62 kilowatt-hours (kWh) per day. A reduction of this magnitude could have been 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.

causal forest, its exploration of the predictors of differential treatment effects, and the setup of its targeting exercise.

1.2 Eversource experiments

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 coverage of these services.

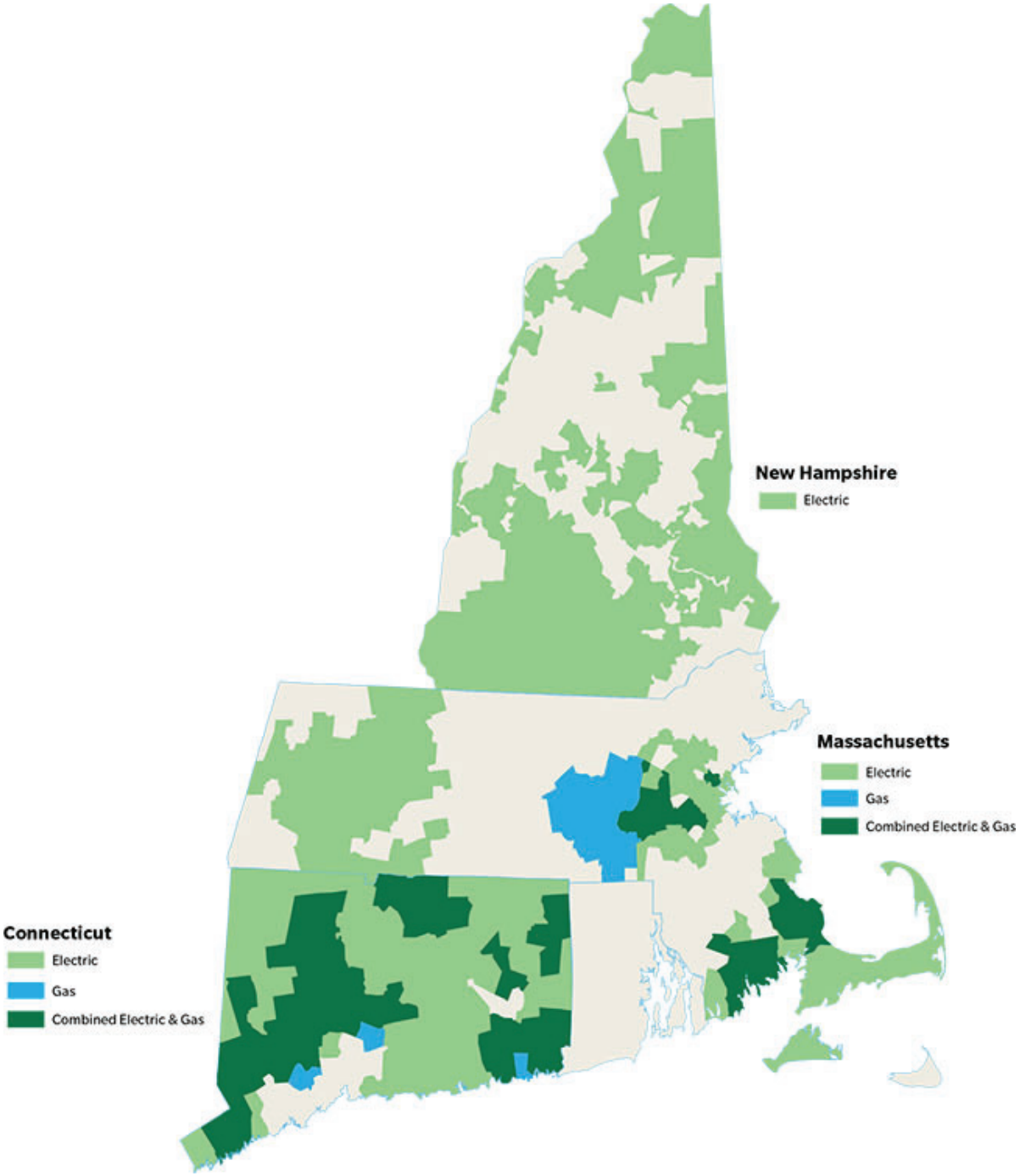
Opower has run 26 waves of home energy report experiments in the Eversource electric service area, with the earliest beginning in February 2011 and the latest beginning in January 2017. We consider 15 of these, dropping 11 waves that either (a) begin outside our five-year period of observation for household 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). 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 (as shown in Figure 1): 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.

According to correspondence with Eversource, Opower strategically targets households with high baseline consumption for HER experimental participation. Indeed, our data confirm that average electricity consumption is higher among households involved in an experiment than among non-participating Eversource electricity customers. Table 1 also 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.

1.3 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. First, we obtain monthly electricity consumption totals (in kilowatt-hours, or kWh)

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	50.0
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	86.5
February 2016	Connecticut	Base	137,896	88.1
February 2016	Connecticut	Low-Income	16,981	53.0
February 2016	Eastern Massachusetts	Base	59,892	76.5
March 2016	Connecticut	Base	17,395	80.1
January 2017	Connecticut	Base	69,517	76.0
January 2017	Eastern Massachusetts	Base	69,517	62.8
	Total		902,581	81.8

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.

for the universe of Eversource customer accounts with residential electricity service in the period from January 2013 to May 2018. After removing accounts with more than one observation (either because the account is associated with multiple properties, or because there are irreconcilable duplicates), we are left with 2,788,369 Eversource accounts ("households"). To these, we merge in treatment assignment data for the fifteen waves that we study. We drop the 2.6 percent of participating households that are enrolled in multiple Opower experiments, as these households contaminate the pure treatment-versus-control comparison.

We combine our consumption and treatment assignment data with cross-sectional home and household predicted 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. After dropping households with outlier values of home square footage or number of rooms, we are left with a main sample of 902,581 households. We feed this sample through a multiple imputation algorithm in order to fill in missing values of the home and household characteristics (see Appendix C.1 for details on this procedure).

Table 2 summarizes the fourteen characteristics and tests for balance across treatment and control observations in our pooled analysis sample. Column 1 presents the fulll-sample mean of each characteristic (with standard deviation in parentheses). Column 2 displays the treatment-control difference in means (and standard error in parentheses) for each characteristic, as the coefficient from a regression of the particular characteristic on the treatment binary variable and a set of wave fixed effects, with robust standard errors. One of the treatment-control differences is significant, at the five percent level; we view this as a typical result of conducting fourteen hypothesis tests. We present wave-specific balance tables in Appendix A.

2 Empirical Strategy

We follow much of the existing literature on Home Energy Reports and begin by using use difference-in-differences regressions, leveraging random assignment of households into treatment and control groups, to estimate average HER 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

Table 2: Treatment-Control Balance

	Sample-wide mean (1)	**T-C difference in means** (2)
Baseline consumption (kWh)	812.161	0.010
	(400.539)	(0.836)
Home value ($)	378,215.758	-1,049.013
	(393,255.443)	(1,002.546)
Home square footage	19.689	-0.019
	(11.857)	(0.035)
Number of rooms in home	7.138	-0.007
	(2.226)	(0.006)
Year home built (1-5)	1,968.060	0.038
	(23.536)	(0.065)
Single-family occupancy (=1)	0.808	-0.001
	(0.394)	(0.001)
Renter (=1)	0.164	0.002**
	(0.370)	(0.001)
Annual income ($)	97,781.877	-199.526
	(68,372.743)	(186.745)
Education (1-5)	3.199	-0.005
	(1.245)	(0.003)
GreenAware score (1-4)	2.163	-0.004
	(1.134)	(0.003)
Number of adults	2.375	-0.002
	(1.347)	(0.004)
Child in home (=1)	0.487	-0.001
	(0.500)	(0.001)
Participated in energy audit (=1)	0.341	0.002
	(0.474)	(0.001)
Age	55.762	0.030
	(14.927)	(0.042)

Notes: Column 1 displays the full-sample mean and (in parentheses) standard deviation of each listed household characteristic. Column 2 displays differences in means and (in parentheses) corresponding standard errors. Column 2 estimates come from linear regression of each characteristic on treatment status, with wave fixed-effects and robust standard errors. * $p < 0.01$, ** $p < 0.05$, *** $p < 0.01$.

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 wave average treatment effects via the following regression:

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

where kWh_{it} is electricity consumption for household i in year-month t. T_{it} is the binary treatment variable taking a value of one for treated households from the program start date onward, X_i is a vector of household characteristics, and θ_i and ω_t are household and year-month fixed effects, respectively. Our primary specification for the ATE is thus a difference-in-differences setup, which is standard in the literature evaluating randomized home energy reports (Allcott, 2011; Ayres et al., 2013; Costa and Kahn, 2013; Gerarden and Yang, 2022); we also provide difference-in-means ATE estimates in Appendix B. We cluster standard errors by zip code. α_2 is the coefficient of interest—the average treatment effect in kWh per month. We calculate a "pooled" ATE as the average of all wave ATEs, weighted by wave sample size.

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:

$$kWh_{iwt} = \beta_1 + \Sigma_{j=-12}^{36}\tau^j D_{iwt}^j + X_i\eta + \theta_i + \omega_{wt} + e_{iwt}. \tag{2}$$

Here, we use all waves simultaneously to estimate a pooled ATE for each period j relative to the event of interest – the beginning of treatment in the relevant wave. D_{wt}^j is a binary variable equaling one if an observation is in wave w, j months after (or before) HER mailings begin in that wave, where $j \in [-12, 36]$.[5] We omit D_{wt}^{-1}, corresponding to the month iimmediately preceding the start of HER mailings, from the estimating equation, so that all coefficients are interpretable as the monthly ATE relative to this month. We employ household and wave-year-month fixed effects and again cluster standard errors at the zip code level.

[5] We also estimate wave-specific event studies and calculate pooled time-specific coefficients in the same fashion as described for the full-period ATE (that is, as averages weighted by wave sample size). The resulting event study plot is qualitatively the same as what we show from our main analysis in Section 3.

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 an ensemble of trees, where each tree is grown from a randomly drawn 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 construct a causal forest using their generalized random forests (*grf*) R package (Tibshirani et al., 2021).

We observe outcomes Y_i, treatment assignment W_i, and household attributes X_i. For a single tree, we start by drawing a random subsample, without replacement, from the full cross-section of Opower households. The algorithm takes this subsample as its "root node" and splits it into two child nodes; the split is defined by some threshold value of one of the household attributes (in X_i). The splitting rule used in the grf package favors splits that increase the heterogeneity of its average treatment effects as fast as possible (Athey et al., 2019). More formally, the objective is to find the single value of a single variable at which splitting the sample minimizes (an indicator of) in-sample prediction error in the child nodes (Athey et al., 2019). Child nodes are then split recursively to form a tree, stopping when there are fewer than a threshold number of households in a given node. The terminal nodes are called "leaves".

The causal forest is a collection of these trees, where each tree has been grown using a new randomly drawn subsample, as well as a new random subset of the splitting variables (X_i). From these trees, we can construct weights $\alpha_i(x)$ that measure how often the ith household falls in the same leaf as x—what Athey et al. (2019) call "the forest-based adaptive neighborhood of x". The weights are then applied to the estimation of treatment effects τ:

$$\hat{\tau} = \frac{\frac{1}{n}\Sigma_{i=1}^{n}\alpha_i(x)(Y_i - \hat{m}(X_i))(W_i - \hat{e}(X_i))}{\frac{1}{n}\Sigma_{i=1}^{n}(W_i - \hat{e}(X_i))^2}, \tag{3}$$

where $\hat{m}(X_i)$ is an estimated expectation of Y_i conditional on X_i, and $\hat{e}(X_i)$ is the propensity score (the causal forest function labels these 'Y.hat' and 'W.hat', respectively). Athey and Wager (2019) describe the estimation procedure for both the weights $\alpha_i(x)$ and treatment effects τ in

further detail.

We grow a forest consisting of 10,000 trees. In our causal forest, each tree is grown with a different random 50-percent subsample of households and a different subset of available characteristics.[6] We employ 'honesty' in training causal forests: after the initial 50-percent subsample is drawn for a given tree, that subsample is split once more in half, and one half is used to grow the tree structure while the other is held out to then repopulate the leaves after tree growth (and define the weights $\alpha_i(x)$). Athey and Imbens (2016) introduce this practice in tree growth as a way to reduce bias and counteract the overstatement of goodness of fit at deeper levels of a tree.

The whole tree-specific procedure can thus be represented as follows:

1. Randomly draw (1) a 50-percent 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.

After all trees are grown, adaptive weights can be calculated, and treatment effects estimated according to Equation 3.

Several additional nuisance parameter values must be chosen. By default, grf first estimates $\hat{m}(X_i)$ and $\hat{e}(X_i)$ via regression forest and makes out-of-bag predictions, which are then used to grow the main causal forest. We retain this default for $\hat{m}(X_i)$, but for $\hat{e}(X_i)$ we use household i's wave-specific treatment fraction, which is the true ('oracle') propensity score (Athey and Wager, 2021). We also choose the default value—0.05—for the 'maximum split imbalance', which stipulates a minimum relative size of each child node in a potential split of some parent node (the option accepts values in [0,0.25]).

The parameter 'minimum node size' sets the minimum number of both treatment and control observations in a leaf required to continue splitting it. The default value is 5, but we decline to use that value because treatment assignment in our context is skewed (in aggregate, 83 percent of households are treated), and we do not want forest leaves with as little as 1 or 0 control observations. Instead, we tune this parameter by growing a set of forests with variable minimum node size between 500 and 10,000 and choosing the minimum node size that minimizes R-loss (Nie and Wager, 2021). We conduct this tuning exercise three times: once for the forest that uses all households (with non-missing input data), and once each for the two forests grown as part of our targeting exercises, which hold out some households entirely from forest growth. We present the R-loss results from our tuning process in Appendix Figure B1; they indicate a minimum node size

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

of 1,000 for the all-household forest, and 3,000 for the first targeting forest, and 500 for the second targeting forest.[7]

The grf package is set up to use cross-sectional data on $\{Y_i, W_i, X_i\}$. To take advantage of our panel data structure, we define Y_i as the difference between average monthly electricity usage in year t of the relevant HER program wave (where $t \in 1, 2, 3$) and average usage in the year prior to wave start date. W_i continues to be the binary treatment variable. To the X_i vector we add one more variable, which is meant to capture variation in the treatment itself (that is, not just in the treatment *effect*). It is a proxy for the social comparison one receives in a Home Energy Report. We do not observe the exact social comparison that each household receives in each month or quarter, but we can develop a proxy inspired by the fact that Opower's algorithm tries to construct comparison groups from similar households in similar locations.

To be precise: we observe 495,136 households with non-missing Y_i in the year prior to as well as the two years following their wave's start date. We draw a random 50 percent of this sample, stratified by zip code, and use that subsample to estimate zip code specific regressions of baseline consumption on home value, home square footage, number of rooms, and year in which the home was built. Then we set this "training" sample aside and predict, in the remaining 50-percent sample, each household's residual baseline consumption according to the calibrated regression model for its zip code. This residual reveals how much larger or smaller a given household's baseline consumption is relative to the average household with those attributes in its zip code—a proxy for the actual social comparison. We include the residual as an element of X in forest growth, and we use as forest input data only those 247,394 households held out of the zip code model prediction.

3 Treatment effect estimates

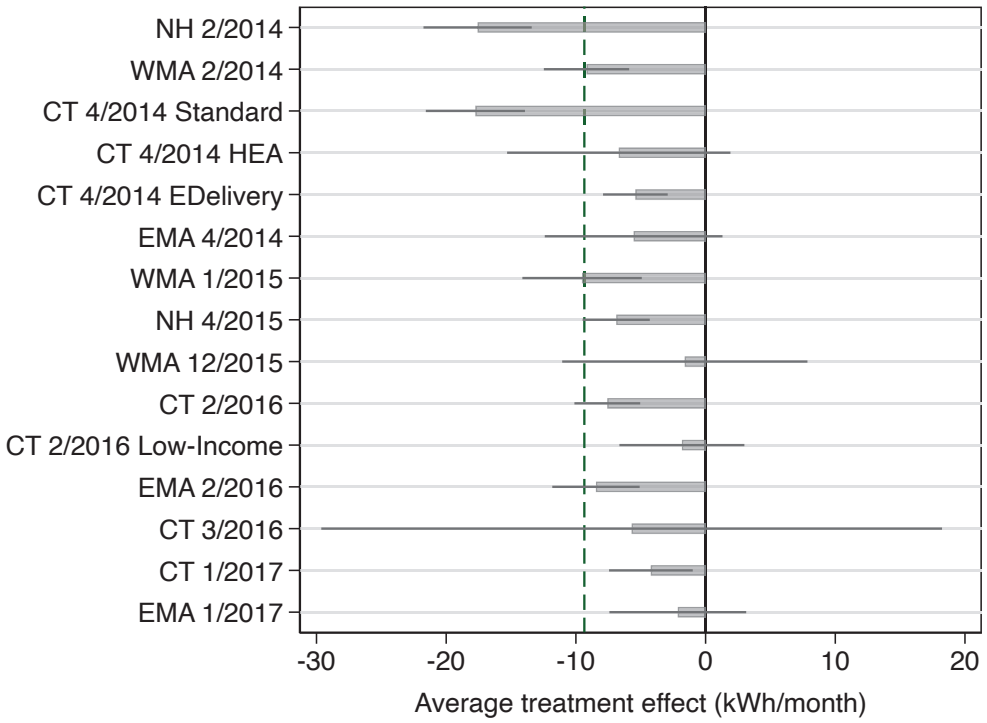
3.1 Average treatment effects

Figure 3 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 -9.35 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. Nine of the fifteen individual program-wave ATEs are

[7] We do not use the 'cluster' option in our preferred forests, but we do grow a forest with clustering and present results for comparison in Appendix Figure B2; see the next section for further reference.

statistically significant at the five-percent level or lower.[8]

Figure 3: Average treatment effects, by wave: consumption



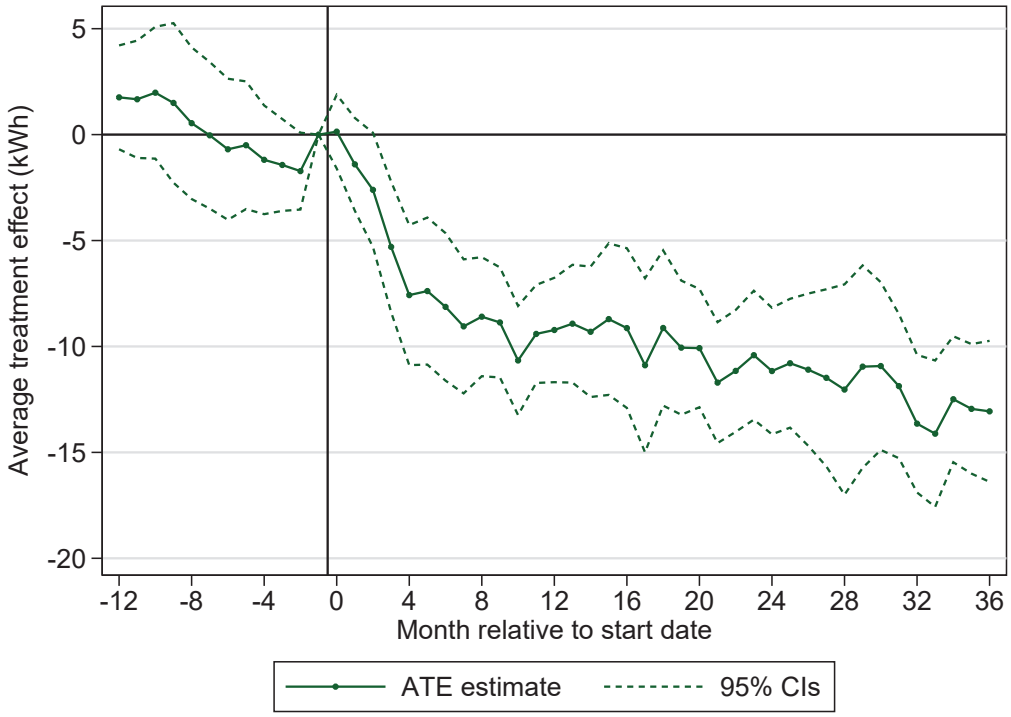
Notes: The y-axis denotes a specific wave (waves are ordered by start date), and the x-axis measures the average treatment effect (ATE) in kilowatt-hours (kWh). The vertical dashed line denotes the weighted pooled average of all wave-specific ATEs, with weights equal to wave sample size. Error bars denote 95% confidence intervals. CT = Connecticut; EMA = Eastern Massachusetts; NH = New Hampshire; WMA = Western Massachusetts. All effects are estimated via panel difference-in-differences regression, using Equation 1 and as described in Section 2.1.

There is an apparent negative trend in ATE estimates over time in Figure 3, which is ordered by wave start date, consistent with Allcott (2015). Differences in the length of the post-period may be a part of the explanation. Figure 4—generated through estimation of Equation 2—sheds light on how the consumption impact of HERs evolves over time, on average. In the twelve months prior to program start date, the point estimate is never statistically different from zero. In the first two months of the treatment (months zero and one of the treatment), there is no drop in consumption. But for the next four months, there is a steady, steep downward trend in average consumption. Month-specific point estimates are statistically significant beginning in month four. The ATE continues to steadily rise in years two and three. In sum, households take time to ramp up their response to reports but continue changing behavior into at least the third year of

[8]In Appendix Table B1, we tabulate difference-in-means estimates of pooled average treatment effect by year. The estimate rises each year, from -6.43 kWh, to -11.46, and to -14.39 by year 3. Each year-specific ATE estimate is statistically significant at the one percent level.

treatment.

Figure 4: 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. The event study binary variable corresponding to the month immediately prior to the start of HER mailings is omitted from the regression and thus set to zero in the figure; all other points are interpretable as predictive effects relative to this omitted month.

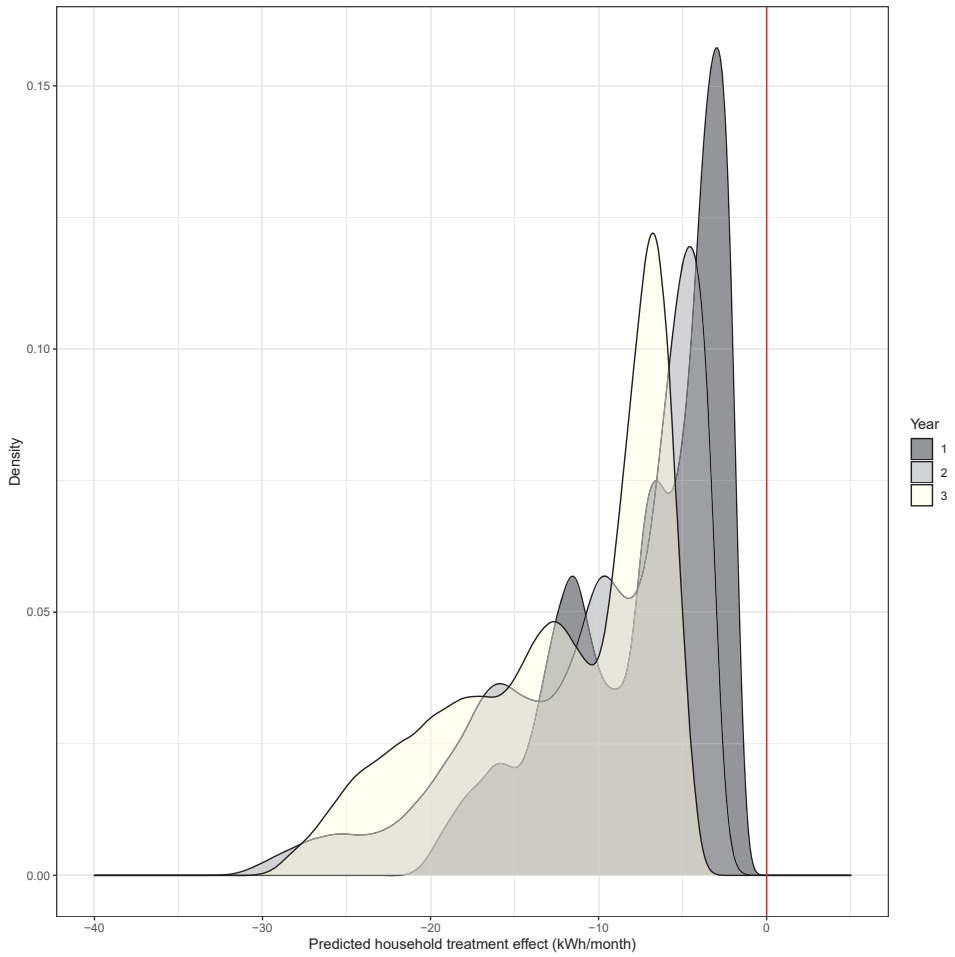
3.2 Conditional average treatment effects

Figure 5 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. Each distribution has the same general shape, with a single prominent peak and a left skew. Each year, the distribution shifts to the left, with the peak getting smaller and the left tail getting thicker. For example, in year one only 0.2 percent of households are predicted to reduce consumption by 20 kWh or more, but by year three that number rises to 15.4 percent. ATEs calculated by the grf function are -7.26 kWh in post-year 1, -10.29 in post-year 2, and -12.5 in post-year 3. The full range of predicted treatment effects extends from roughly -1 to -33 kWh.[9]

What predicts heterogeneity in treatment effects? Figure 6 plots the frequency of selected characteristics' use as a splitting variable in the forest, conditional on being (randomly drawn to

[9]Appendix Figure B2 shows year-specific treatment effect distributions from a forest with zip code level clustering. The results are qualitatively indistinguishable from those of Figure 5.

Figure 5: Distribution of Predicted Treatment Effects

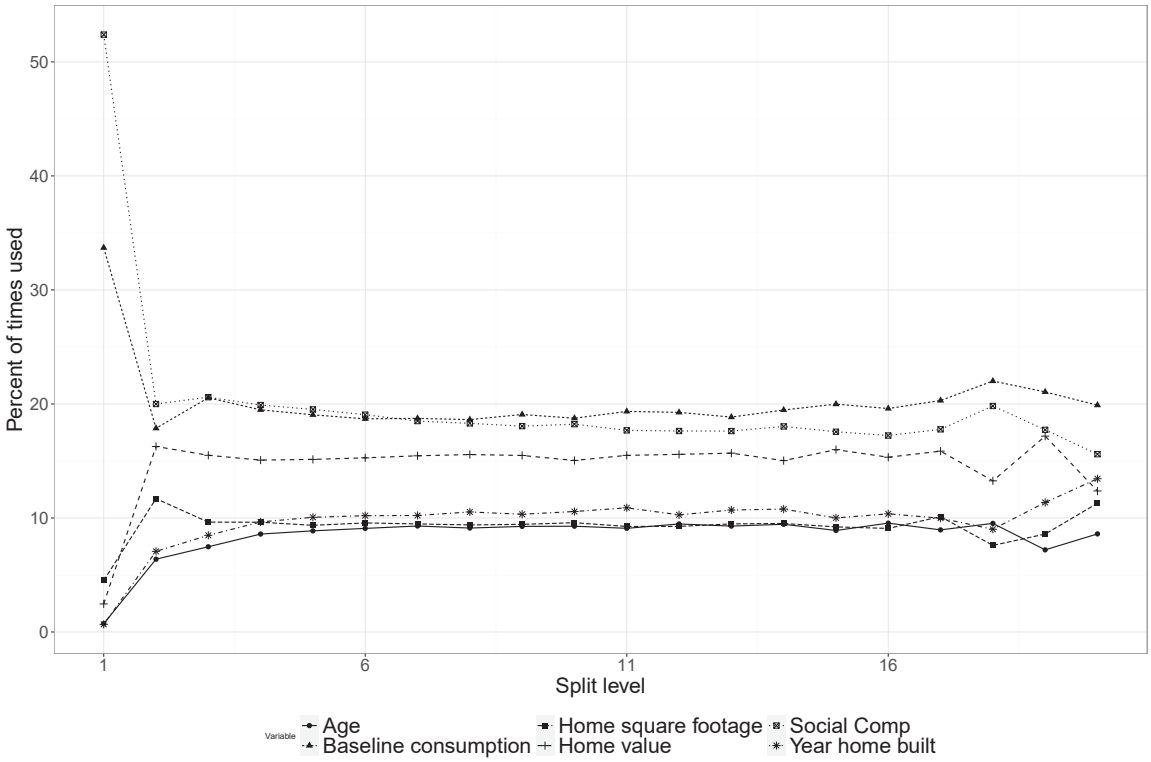


Notes: Each plotted distribution is a kernel density of household treatment effects in a specific year (1, 2, or 3) of HER programming. The sample is fixed across years: only households with non-missing consumption in all three post-years are included. Treatment effect predictions come from our causal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

be) available in splitting. The six characteristics included in this figure—our social comparison variable, baseline consumption, home value, home square footage, the year in which a home was built, and respondent's age—are the most frequently used. The social comparison variable is chosen as the initial splitting variable in 52 percent of trees in which it is eligible. Baseline consumption is chosen about 34 percent of the time. By the third level of the tree, these two attributes have the same frequency of use—just above 20 percent—and they remain in the 15-22 percent range for the duration of tree growth. A third attribute, home value, is also used about 16 percent of the time throughout tree growth. We note, however, that these results may be sensitive to the addition of a new predictor that is correlated with one of the most commonly-used existing ones.

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

Figure 6: 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. We show percentages for the six most frequently-used characteristics: the social comparison proxy, baseline consumption, home value, home square footage, home year built, and age of household respondent. The underlying sample includes all households with non-missing consumption in the year prior to program start and each of the first three years following program start. The dependent variable in the forest is average consumption in post year two minus average consumption in the first pre year. See Section 2.2 for further implementation details.

shed some light on these relationships, we provide two further analyses. The first is a classification analysis (CLAN; Chernozhukov et al., 2022), which is a comparison of average characteristics among those with the largest treatment effects versus those with the smallest, as estimated by the causal forest. We follow the implementation described by Deryugina et al. (2019) to generate treatment effect predictions, which we order and group into quintiles, and then measure differences in mean characteristics between the top quintile and bottom quintile (we describe the procedure in further detail in Appendix C.2).

We present the results in Table 3. The "top 20%" reducers in response to treatment have, on average, higher absolute baseline consumption as well as *relative* consumption (as captured by our social comparison proxy), in comparison to the bottom 20%. Households in the top 20% also have smaller average home values and incomes, and they tend to have more adults but are less likely to have children. Account holders in the top 20% tend to be older but with fewer years of education. Lastly, top-20% homes tend to be newer but less likely to be single-family occupancy.

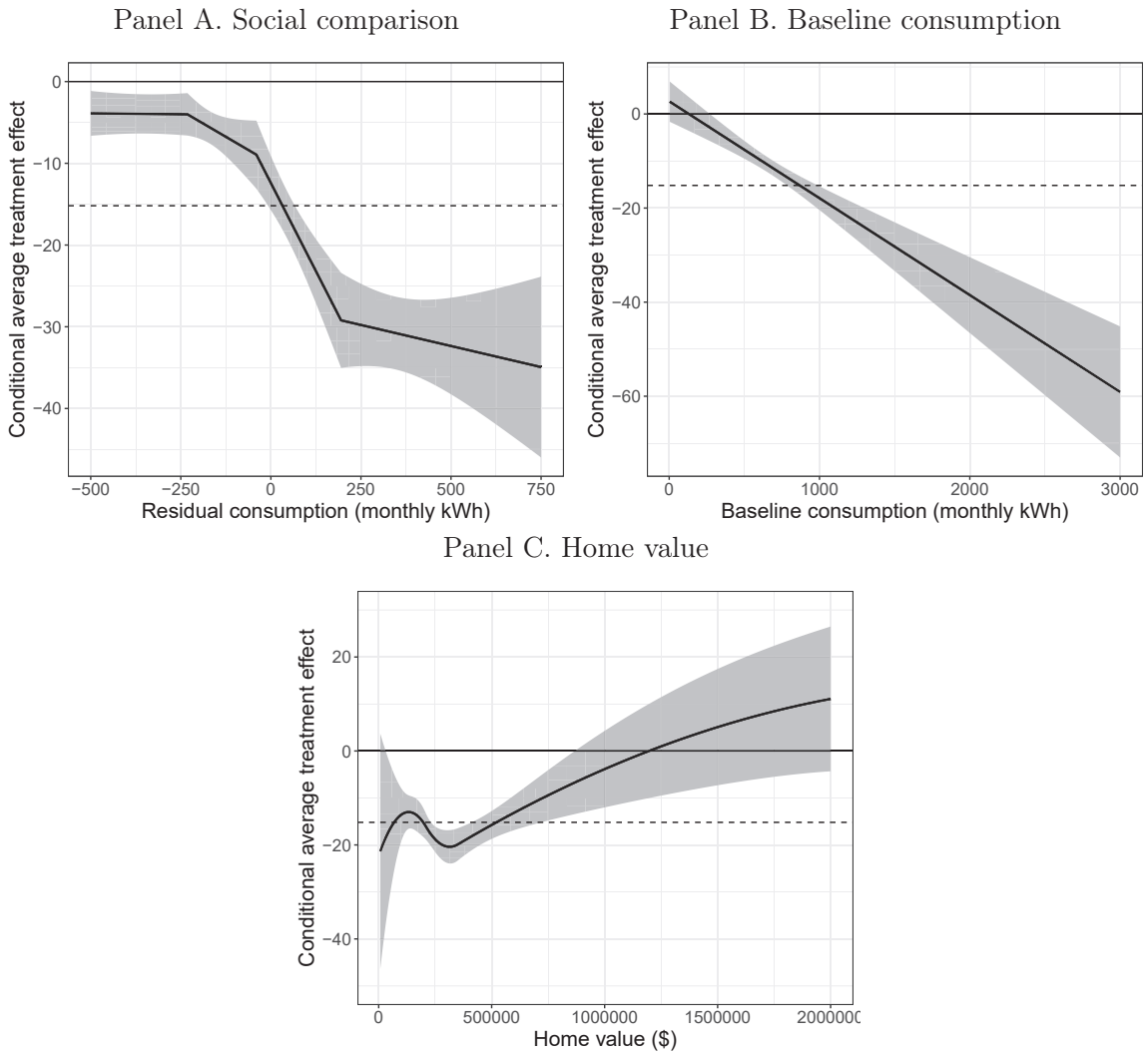
Table 3: Characteristics of high versus low treatment effect households

	Top 20% (1)	Bottom 20% (2)	Difference (3)
Social comparison (kWh)	353.02	-150.51	503.52***
			5.58
Baseline consumption (kWh)	1,260.83	782.84	477.99***
			13.02
Home value ($)	355,758.96	460,184.34	-104,425.38***
			23399.58
Home square footage (100s)	20.62	21.04	-0.41
			0.39
Number of rooms in home	7.26	7.28	-0.02
			0.06
Year home built	1969.81	1967.78	2.03***
			0.43
Single-family occupancy (=1)	0.86	0.91	-0.05***
			0.01
Renter (=1)	0.1	0.1	-0.01
			0.01
Annual income	102,908.33	107,962.35	-5,054.02*
			1900.25
Education (1-5)	3.22	3.35	-0.13***
			0.03
GreenAware score (1-4)	2.21	2.17	0.03
			0.02
Number of adults	2.75	2.43	0.32***
			0.02
Child in home (=1)	0.45	0.55	-0.10***
			0.01
Participated in home energy audit	0.37	0.36	0.00
			0.01
Age	58.07	54.79	3.27***
			0.29

Notes: Columns 1 and 2 display mean characteristics among households in the top 20% and bottom 20%, respectively, of the predicted treatment effect distribution. 'Top 20%' indicates the largest reducers (of electricity consumption); 'Bottom 20%' indicates the smallest reducers (as well as any *in*creasers). Predictions are generated through a procedure developed by Chernozhukov et al. (2022) and implemented by Deryugina et al. (2019); see Appendix C.2 for details. Column 3 displays differences between columns 1 and 2, along with standard errors clustered by zip code in parentheses, estimated via regression of each characteristic separately on a binary variable equaling one if a household is in the top quintile. * $p < 0.01$, $^{**}p < 0.05$, *** $p < 0.01$.

In our second analysis of relationships between household attributes and treatment effect prediction, we construct "heterogeneity curves" non-parametrically. To do so, we follow the procedure of Knaus (2022). We use their "causal double machine learning" strategy (and associated 'cDML' R-package), applied to random forests, to generate treatment effect predictions. Then we estimate spline regressions of treatment effect on a single household attribute based on the R-package 'crs' (Racine and Nie, 2022, and again see Appendix C.2 for further detail on this procedure). In Figure 7, we present one heterogeneity curve for each of the three most commonly split-upon variables: the social comparison variable, baseline consumption, and home value.

Figure 7: Predicted treatment effect vs. household type



Notes: The x-axis measures a different household attribute in each panel: (A) the social comparison variable; (B) baseline consumption; and (C) home value. The y-axis measures predicted treatment effect as a function of the relevant attribute. The solid line and confidence intervals are estimated via spline regression, using the causalDML R-package (Knaus, 2022). The dashed line is a point estimate of the average treatment effect. The sample includes all households with non-missing consumption in the year prior to program start and each of the first three years following program start. See Appendix C.3 for further implementation details.

Figure 7, Panel A shows a trend break around a value of zero for the social comparison residual, with larger positive value predicting a larger treatment-induced reduction in electricity consumption. Panel B shows a near-linear relationship between treatment effect and baseline consumption: the larger the baseline, the greater the reduction in response to treatment. Panel C shows treatment-induced reductions fluctuating modestly below a home value of $350,000 and, above that threshold, dropping steadily in home value. Put together, Figure 7 suggests that targeting for treatment (a) higher baseline consumption, (b) lower home values, and/or (c) less favorable would-be social comparisons may lead to increased social benefits.

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 the 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—for example, 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 the social net benefits produced by treating each household with home energy reports. We use the following equation for predicted annual social net benefits:

$$SNB_i = -TE_i * 12 * SMC_e - MC_{HER} + WTP_i, \tag{4}$$

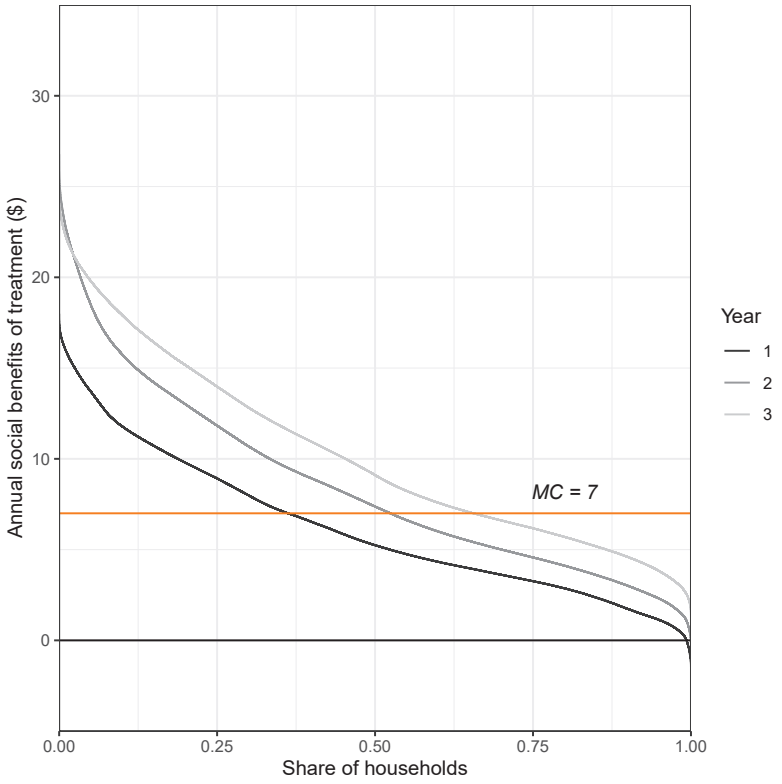
TE_i is the predicted monthly treatment effect for household i, which we multiply by 12 to convert to an annual number; SMC_e is the social marginal cost of electricity (which includes both generation costs and environmental externalities); MC_{HER} is the marginal cost of sending a household HERs for one year; and WTP_i is a household's annual willingness to pay for HERs. TE_i can be taken from any prediction model, such as our causal forest. We set $SMC_e = \$0.065$ per kWh, which is the short-run estimate of Borenstein and Bushnell (2022) for the trio of states in our sample in 2016, shared by the authors. We set $MC = \$7.00$ per household-year, based on consultation with Eversource.[10] 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-

[10]The actual social marginal cost could be below the price charged to Eversource by Opower; in Appendix Figure B4, we replicate our main targeting exercise using a marginal cost of $3.50 per household-year.

specific 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 (we describe this in further detail in Appendix C.4).

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 8 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 the marginal cost line; that is, there are always households whose predicted responses to HERs translate to net negative benefits.

Figure 8: 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.

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 the left edge of the figure to the two curves'

crossing point. However, there is no true hold-out sample in Figure 8; we use all households in the growth of each forest used therein.[11] To mimic the real-life planner's targeting challenge, we build a predictive model with one subset of the entire household sample and target using that model in another entirely different subset. Our general algorithm is as follows:

1. Split the full sample of available households into two: a training set for estimating the model, and a test set for targeting and its evaluation.
2. Estimate a predictive model with the training sample.
3. In the test sample, predict household-level treatment effects (using the model estimated in Step 2) and willingness to pay (using the model with parameters taken from Allcott and Kessler (2019) described above).
4. Calculate predicted social benefits for each household according to Equation 6.
5. Identify all test-sample households whose induced social benefits exceed marginal cost; this is the group "targeted" for treatment.
6. Estimate an average treatment effect (ATE) in the targeted group.
7. Calculate "actual" aggregate social net benefits in the targeted group using Equation 6, but replacing each targeted household's predicted TE with the estimated ATE from Step 6.

The causal forest is our predictive model of primary interest.[12] However, we compare it to several alternative model types that are simpler yet still data-driven, as well as Opower's actual treatment assignment, to assess the relative improvement possible through machine learning. We choose one machine-learning and four regression-based alternatives to test, motivated by convention and prior knowledge of HERs' impact. The machine learning alternative is a lasso, which uses regularized regression to select variables and coefficient values from a set of potential variables consisting of treatment status, the X vector, and all interactions between elements of the two. We think of the lasso as a sort-of bridge between the causal forest and the four regression models, it being a regression-based machine learning method.

Of the four non-ML regression-based alternatives, the first and simplest (which we denote "Baseline") is a linear model in which treatment effect varies 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

[11] Household predictions, however, are still "out-of-bag"—that is, based only on trees grown without the use of the household in question.

[12] In this targeting exercise, we grow forests consisting of 1,000 trees instead of 10,000 for computational tractability: one of our bootstrapping procedures requires regrowing these forests 1,000 times.

in our work as well as that of others before us (Allcott, 2011; Ayres et al., 2013). Given Opower's pre-existing preference for high baseline electricity users as experimental participants, this first model can be thought of as a formal version of what has historically been done in the field.

Each of the next three regression-based models build successively on each other. The second model (denoted "Linear") 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. Throughout this exercise, we predict treatment effects in post-year 2. We describe the five alternative prediction models and other elements of the targeting exercise in further detail in Appendix C.4.

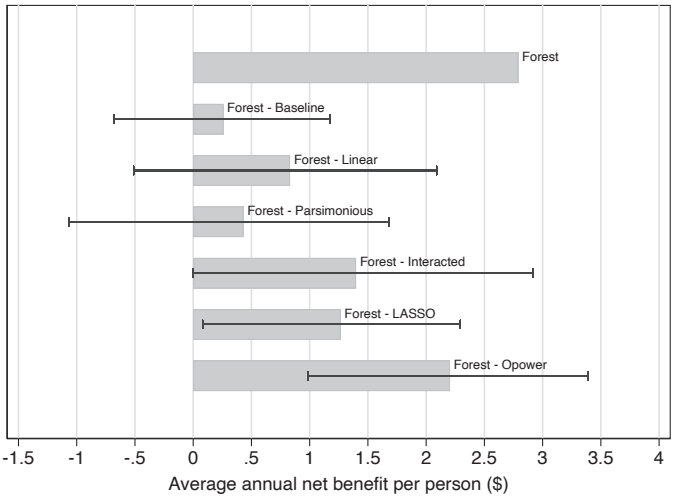
In Step 1 above, we try two different ways of splitting the full sample into training and test sets. In the first, we split the full sample ($N = 247,394$) in half randomly. This splitting rule facilitates a test of whether the forest would be accurate in a held-out group with the same average characteristics; 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 in 2014 (203,248 households), and the hold-out sample is composed of those with wave start date in 2015 or 2016 (44,146 households). 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 groups are very different in this second version of the test—Appendix Table B2 documents this difference. The hold-out group is, on average, a lower absolute consumer of electricity but a higher relative user (as given by our social comparison proxy). It also has larger homes and home values but lower income. We thus view this version of the targeting exercise as shedding light on the *external* validity of each prediction method.

Figure 9 depicts, for each of the two versions of the targeting exercise described above, the performance of the forest relative to each alternative model. Forest-based targeting in a random hold-out sample produces an estimated additional $2.21 in annual social net benefits (SNB) per (sample) household relative to the Opower treatment assignment (the bottom bar in Panel A). This is a product of the forest yielding $2.79 in per-household benefits while the actual Opower program yielded only $0.58, which represents a nearly five-fold (4.8x) increase in net benefits of targeting relative to the status quo. In particular, the Opower distribution treated 102,853 households out of a total 123,698 in this exercise, while the forest finds only 66,256 worth treating according to the targeting rule. This difference is the aggregate effect of the forest switching 46,537 households from treatment to control and 9,940 households from control to treatment.

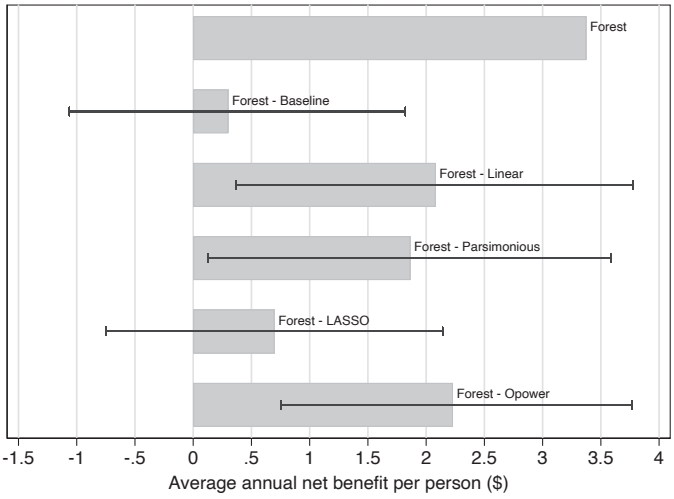
Panel A of Figure 9 also shows that forest performance, as a point estimate, outperforms that

Figure 9: Social net benefits of targeting, by predictive method

Panel A. Training on a random sample



Panel B. Training on 2014 waves



Notes: The top bar is the estimated annual social net benefits (SNB) produced from the treatment assignment chosen by the forest, relative to a no-action counterfactual. Each other bar depicts the estimated annual gain in social net benefits (SNB) produced from targeting using the forest instead of the listed alternative method. The targeting rule is to treat if predicted $SNB > 0$. Net benefits are expressed as an average per household in the full test sample, so that Panels A and B are more comparable. Panel A depicts results from building all predictive models with a 50% random sample of households and targeting in the other 50% "test" 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 "Interacted" model is not included in Panel B because it does not identify households that satisfy the targeting criterion. Confidence intervals are generated via bootstrapping, which we describe in greater detail in Appendix C.5.

of the other predictive models. The "baseline" and "parsimonious" models produce the second- and third-most benefits per person, respectively; the forest outperforms the baseline model by 9.8%. "Linear", "interacted", and the lasso do worse, but all models produce well over twice the social net benefits of the status quo Opower assignment. We follow Gerarden and Yang (2022) and bootstrap to generate confidence intervals on these differences;[13] these are wide for all comparisons in Panel A but significantly different from zero in the case of Opower's actual treatment assignment, the lasso, and the "interacted" model (in this last case, marginally so).

In the version of our test where we use earlier waves to target in later ones (Panel B), the forest continues to outperform both the Opower benchmark and all regression and lasso models in the initial targeting sample. The forest produces 2.95 times the net benefits of Opower's actual distribution in post-2014 waves and outpaces the best alternative (the baseline model) by 9.35%. This time, the differences between forest performance and linear, parsimonious, and interacted model performance are statistically significant (the interacted model fails to find *any* households worth targeting), but forest performance is still indistinguishable from that of the baseline and lasso models. A more conservative bootstrapping procedure yields slightly wider confidence intervals, depicted in Appendix Figure B3. In addition, we find that raising the number of bootstraps in our main procedure to from 1,000 to 10,000 does not appreciably reduce confidence intervals. Our inability to reject the null of equal welfare impacts across certain models is consistent with Gerarden and Yang (2022): they find, in an analogous HER-delivery setting, that the advantage of more complicated treatment rules (using the technique of empirical welfare maximization (Kitagawa and Tetenov, 2018), as a function of attributes) relative to a treatment rule solely determined by baseline consumption is not statistically significant.

We also present results of several variations on the targeting rule in Appendix B. Figure B4 employs a marginal cost of $3.50 instead of $7.00, in acknowledgement of the possibility that the true social marginal cost of HER delivery is lower than what Eversource pays. Figures B5-B7 target households in the top quantile (half, quartile, or decile) of predicted social net benefits for treatment, instead of households with positive predicted net benefits. In these iterations, the forest usually, but not always, performs better than comparison models, with the baseline model continuing to provide the stiffest competition. The lasso, meanwhile, is the clear next-best model when we target in later waves using earlier ones (the stiffer test of external validity). The comparable performance across the board of the one-dimensional baseline model with the forest may reflect the specific treatment in question, as the literature on HERs overwhelmingly finds

[13] Figure 9 uses a "fixed-rule" bootstrapping procedure, in which the treatment assignment chosen by each model in the initial run of the targeting exercise is held fixed across all bootstraps; only the sample changes. Appendix Figure B3 shows confidence intervals from a non fixed-rule procedure in which each bootstrapped sample is used to run a new forest producing a new treatment assignment. See Appendix C.5 for a fuller description of our bootstrapping procedures.

absolute baseline consumption to be the strongest predictor of treatment effects. It may also (or alternatively) reflect our lack of data on the exact social comparison that households receive: our proxy for this comparison is the most frequently used splitting variable in the forest, but using it nonetheless introduces measurement error.

In any case, when using earlier waves to target later, rather different waves, the two machine learning (forest and lasso) models and the simplest regression (baseline consumption) model tend to do better than the three more complicated regression models. This pattern reinforces the notion that avoiding overfitting is valuable, because the better-performing models involve active steps to address overfitting, while the worse-performing models do not. The forest and lasso models counteract overfitting algorithmically through subsampling and regularization, respectively. The baseline regression model does so by restricting the treatment effect to vary only (and linearly) with baseline consumption.

5 Conclusion

This paper brings together two recent innovations in economics: first, machine learning (ML), which holds great promise as a tool for high-resolution evaluation and prediction; and second, the increased use of nudges within government and business programs. We empirically assess the latter using the former, in the context of a 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 ML algorithm, an ensemble method based on classification and regression trees, adapted for causal inference.

The causal forest we estimate reveals several facts about treatment effects in this context. First, there is wide variation in responses to the nudge. The overall average treatment effect is a nine kilowatt-hour monthly reduction in electricity consumption, but individual effects range from -1 to -33 kWh. Second, higher absolute and relative consumption (in comparison to neighbors) are both (simultaneously) predictive of larger treatment-induced reductions in electricity use. These and home value are more frequently used in the forest than all other household attributes. Third, while treatment effects are almost entirely negative (that is, reductions), for many households they are not large enough to be predictive of positive social net benefits from treatment. Put together, the forest-identified variation in treatment effects suggests the potential to improve program effectiveness through targeting and tailoring of treatment.

To test this potential, we develop an out-of-sample prediction exercise that mimics how 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 treatment assignment as well as five alternative targeting methods. When training and hold-out samples are randomly drawn, the forest produces nearly five times the aggregate social net benefits as the Opower program and ten percent more benefits than the best non-forest predictive method. When we train predictive models on 2014 waves and target in post-2014 waves (which are very different), the forest produces nearly three times the benefits of the Opower program and nine percent more benefits than the best alternative. The forest's advantage is statistically significant relative to some but not all alternatives; our pattern of results suggests that the models best set up to minimize overfitting (in our case, the two ML models and the sparsest regression model) tend to perform best.

At a high level, our context, methods, and results, comprise a case study that can be a useful resource for those considering the selective targeting of some treatment, intervention, or program using machine learning methods, especially causal forests. Our results show how forests could potentially be used by firms and policymakers to improve program effectiveness and social welfare through targeting. In addition, they point to the possibility of 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 "disciplined" high-resolution predictive methods like causal forests have the potential to be a helpful tool for improving cost effectiveness, social welfare, and/or distributional equity in a wide 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.

Andini, M., E. Ciani, G. de Blasio, A. D'Ignazio, and V. Salvestrini (2018): "Targeting with machine learning: An application to a tax rebate program in Italy," *Journal of Economic Behavior and Organization*, 156, 86–102.

Ascarza, E. (2018): "Retention Futility: Targeting High-Risk Customers Might Be Ineffective," *Journal of Marketing Research*, LV, 80–98.

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

——— (2017): "The Econometrics of Randomized Field Experiments," *Handbook of Economic Field Experiments*, 1, 73–140.

——— (2019): "Machine Learning Methods that Economists Should Know About," *Annual Review of Economics*, 11, 685–725.

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," *Observational Studies*, 5, 37–51.

——— (2021): "Policy Learning with Observational Data," *Econometrica*, 89, 133–161.

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.

Borenstein, S. and J. B. Bushnell (2022): "Do Two Electricity Pricing Wrongs Make a Right? Cost Recovery, Externalities, and Efficiency," *American Economic Journal: Economic Policy*, 14, 80–110.

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.

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.

Cagala, T., U. Glogowski, J. Rincke, and A. Strittmatter (2021): "Optimal Targeting in Fundraising: A Causal Machine-Learning Approach," *arXiv preprint arXiv:2103.10251*.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2022): "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments," *arXiv preprint arXiv:1712.04802v6*.

Christensen, P., P. Francisco, E. Myers, H. Shao, and M. Souza (2022): "Energy Efficiency Can Deliver for Climate Policy: Evidence from Machine Learning-Based Targeting," Working paper, 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 (2020): "Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs," *Review of Economics and Statistics*, 102, 664–677.

DellaVigna, S. and E. Linos (2022): "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," *Econometrica*, 90, 81–116.

Deryugina, T., G. Heutel, N. H. Miller, D. Molitor, and J. Reif (2019): "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction," *American Economic Review*, 109, 4178–4219.

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

Ellickson, P. B., W. Kar, and J. C. Reeder (2021): "Estimating Marketing Component Effects: Double Machine Learning from Targeted Email Promotions," Working paper, National Bureau of Economic Research.

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

Gerarden, T. and M. Yang (2022): "Using Targeting to Optimize Program Design: Evidence from an Energy Conservation Experiment," *Journal of the Association of Environmental and Resource Economists (forthcoming)*.

Gulyas, A. and K. Pytka (2020): "Understanding the Sources of Earnings Losses After Job Displacement: A Machine-Learning Approach," *Mimeo*.

Hastie, T., J. Qian, and K. Tay (2021): "An Introduction to glmnet," https://glmnet.stanford.edu/articles/glmnet.html.

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

Kitagawa, T. and A. Tetenov (2018): "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616.

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.

Knaus, M. C. (2022): "Double Machine Learning Based Program Evaluation under Confoundedness," *Econometrics Journal*, 25, 602–627.

Knittel, C. and S. Stolper (2021): "Machine Learning about Treatment Effect Heterogeneity: The Case of Household Energy Use," *American Economic Association Papers and Proceedings*, 111, 440–444.

National Conference of State Legislatures (2021): "Energy Efficiency Resource Standards," https://www.ncsl.org/research/energy/energy-efficiency-resource-standards-eers.aspx.

Nie, X. and S. Wager (2021): "Quasi-Oracle Estimation of Heterogeneous Treatment Effects," *Biometrika*, 108, 299–319.

Racine, J. S. and Z. Nie (2022): "crs: Categorical Series Regression," http://cran.r-project.org/web/packages/crs/index.html.

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.

Robins, J., A. Rotnitzky, and L. Zhao (1994): "Estimation of Regression Coefficients when Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

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.

Thaler, R. and C. Sunstein (2008): *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Yale University Press.

Tibshirani, J., S. Athey, R. Friedberg, V. Hadad, D. Hirshberg, L. Miner, E. Sverdrup, S. Wager, and M. Wright (2021): *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.

World Resources Institute (2022): "Climate Watch," www.climatewatchdata.org.