REPORT

PG&E Targeted HER

December 5, 2018

Prepared for
Pacific Gas & Electric

Prepared by

Aimee Savage
Consultant II

Dan Thompson
Consultant

Amanda Stansell
Project Analyst

Nexant, Inc.

CALMAC ID PGE0425

Table of Contents

1 Introduction and Key Results

1.1 Introduction

Oracle Utilities (formerly Opower) and PG&E have worked together since 2011 to launch nine experimental waves of home energy reports (HERs), affecting 2.15 million households in total. The reports provide the recipient with neighbor comparisons, energy efficiency tips, and information about energy efficiency programs offered by PG&E. In 2015, the total energy savings of PG&E's HER program was 145 GWh and over 4.6 million therms. While these savings are substantial, program cost-effectiveness is paramount.[1] Therefore, PG&E is investigating approaches to more effective customer targeting and delivery. The research reported herein is part of that effort.

Since residential customers can be quite different from one another[2], it seems likely that the effect of receiving a home energy report also varies from customer to customer. The cost effectiveness of the program could be improved if reports were sent to customers who are likely to provide larger savings and not sent to customers who would provide smaller or even negative savings. Negative savings refers to the case when energy consumption of a participant increases after enrollment in the HER program. A recent PG&E study[3] showed that approximately 30% of current HER participants have greater energy consumption after participating in the program. Unfortunately, at present, one cannot predict the effect of sending a report to any particular customer because energy consumption practices vary over time as a result of a large number of drivers.

Efforts to understand the factors that influence responses to HERs have estimated the average energy savings for subgroups of customers – for example by decile of annual consumption in the prior year. In general, these analyses show that most of the energy savings obtained from HERs is concentrated in the top half of the distribution of energy consumption. An argument can be made for restricting the delivery of HERs to the upper half of the consumption distribution based on program cost-effectiveness; and some utilities do so. However, other utilities are reluctant to restrict delivery based on consumption because it deprives a substantial fraction of the residential population from the benefits that arise from exposure to the HERs. Understandably, PG&E and other utilities could benefit from developing a more refined approach to program targeting.

Like most utilities, PG&E collects a large amount of data on its customers related to load features (e.g., baseload, seasonal, peak, off-peak consumption) or customer attributes (e.g., data from third-party vendors such as Axciom or Experian). One or a combination of these variables may provide a basis to substantially improve the accuracy of HER impact predictions. The purpose of this research is to discover whether this is the case and focuses on electric savings prediction.

---

[1] As illustrated by the CPUC's Proposed Decision on Energy Efficiency Business Plans in April 2018.

[2] This category indeed includes small studio apartments and expansive estates from San Francisco and San Jose to Fresno and Sonoma.

[3] Opinion Dynamics, "PG&E Home Energy Report (HER) Energy Savings Distribution Analysis and Trends Study", August 2018

In this effort, two new machine learning algorithms developed over the last three years were used to identify which customers are more likely to respond strongly to the HER stimulus based on observed customers responses to HERs obtained in a very large scale randomized controlled trial:

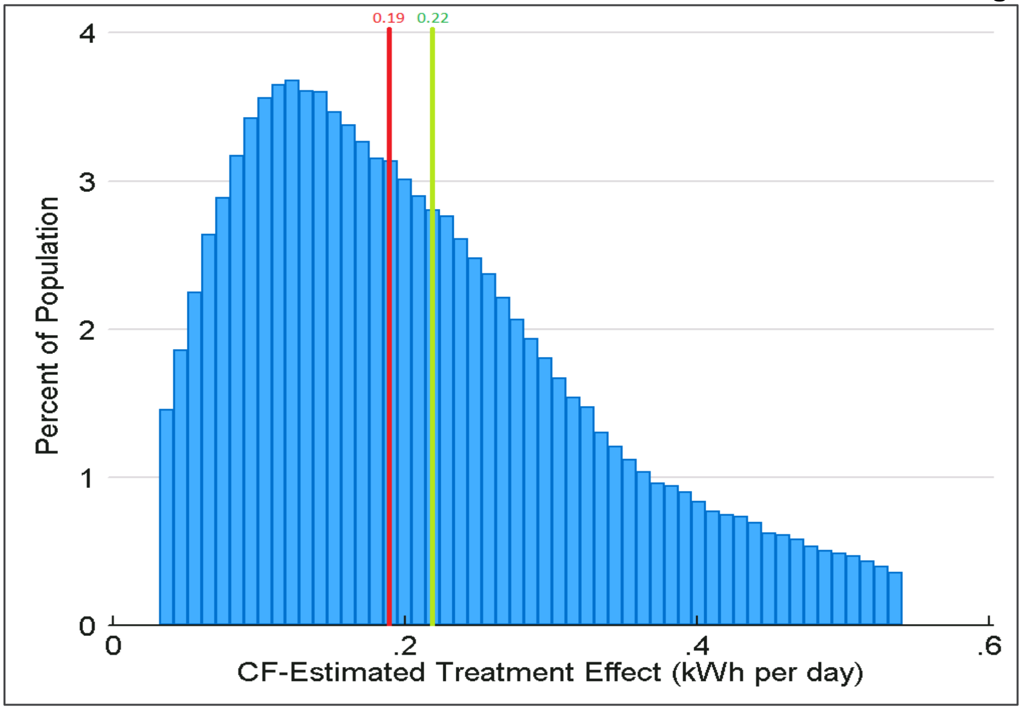
- Causal tree algorithm: this approach iteratively splits the exposed customer population (i.e. treatment group) into smaller and smaller subgroups, searching for partitions that separate customers who are predicted to produce larger energy savings from those who do not.

- Causal forest (a generalization of the causal tree algorithm): this method produces many causal trees using randomly selected variables and randomly selected customers. It then averages those trees together to provide a prediction for the expected energy savings for every individual customer.

1.2 Key Findings

Using the causal forest algorithm and all of the available information about customer characteristics to predict individual customer energy savings results in an interesting statistical distribution of likely HER program impacts. Figure 1-1 presents the distribution of predicted savings ('Estimated Treatment Effects') for the 5th to 95th percentile of customers. The median estimated savings (0.19 kWh per day) is highlighted in red and the mean estimated savings (0.22 kWh per day) is highlighted in green. Using the causal forest algorithm, household daily electric savings ranges from 0.04 kWh per day at the 5th percentile to 0.54 kWh per day at the 95th percentile. About 58% of the customers are predicted to produce less than the average of 0.22 kWh per day.

Figure 1-1: Distribution of Predicted Treatment Effects from Causal Forest Algorithm



Only a small percentage of customers (approximately 2%) are predicted to have negative savings. This indicates that identifying negative savers a priori is unlikely and excluding this group of customers from future participation would not have significant impact on program performance.

Another key insight of this distribution is its pronounced right skew, which means that most customers are predicted to have lower treatment effects than the average. This illustrates both the challenge of targeting for HER programs and the potential opportunities for improving program performance by restricting delivery to customers who are predicted to achieve significantly greater energy savings.

A substantial percentage of HER participants may not provide sufficient savings to justify the cost of delivering the reports to them. Using high-level assumptions about the avoided cost of energy and the cost of delivering the HERs, customers must save over 0.27 kWh per day to justify the cost of delivering the reports to them. Our research shows that 70% of customers are below that threshold. This analysis is yet not precise enough to inform PG&E to stop sending reports to these customers to increase cost-effectiveness.

The potential economic benefits from targeting HERs based on the models described herein are large. Based on this consideration, Nexant recommends PG&E develop a formal experiment designed to test the impact of discontinuing the delivery of HERs to customers who are predicted not to have energy savings sufficient to cost justify their continued delivery. This experiment would be similar to PG&E's experiments testing the persistence of energy savings. Treatment discontinuation would yet be based on customers' model scores, instead of random selection.

2 Causal Tree

2.1 Methodology

The ideal way to target customers for HERs would be to make an accurate prediction of the expected energy consumption reduction for each customer in response to the reports and only enroll customers with the highest predicted reduction. If we believe that a new customer for whom we want to predict energy savings is similar to a set of customers on whom we already ran an HER experiment, we might think of the prediction problem as a number of subgroup analyses. We could gather all of the information we have about these customers (where they live, how much energy they were consuming before the experiment started, how large their home is, etc.), group them with other customers with the exact same characteristics, and subtract the energy consumption of those who received HERs from the consumption of those who did not.

However a problem arises when several variables are used. When the impact of ‘n’ continuous pre-treatment variables is tested for each decile of the population, 10^n different estimates are produced, splitting up the randomized controlled trial into far too many pieces for the estimates to be accurate in any given cell (i.e. subgroup). This problem is thus difficult to solve when several variables are considered. A simpler option is to use algorithms that can search over possible subgroups and find the ones that are important for making good predictions. The logic of regression tree algorithms is very relevant for such analyses.

Regression trees are a class of algorithms that make predictions for unobserved or future characteristics based on observed or historical data. These predictions are made by splitting all of the units (customers in our case) into two groups that look very different with respect to the predicted variable (e.g., energy consumption). This creates two nodes, one for each group (for example, a CARE node and a non-CARE node). This process is then repeated on each node separately until encountering a stopping rule. For example, if the tree reaches a partition that would result in a node containing less than 1% of the total data, we can stop splitting the data into smaller groups. At that point, each final node is called a leaf. Each leaf represents a group of customers that meet a series of binary criteria. The average of the predicted variable within each leaf is estimated, and these average values[4] are treated as our predictions for some new set of units (i.e. customers).

The most common regression tree algorithm, known as CART (Classification and Regression Tree), is a standard machine learning technique for prediction. CART has two basic steps: partitioning and pruning. Partitioning proceeds as follows:

> *Until a partition reaches a certain minimum size:*
>
> > *1. Search every binary split of the customers along every predictor. Calculate and store the average difference between the true value of the outcome variable and the predicted value, known as mean squared error (MSE).*
> >
> > *2. Split the data into two new datasets using the splitting rule that minimizes the MSE based on step 1.*

[4] Some versions of regression trees use something other than an average with leaves, but averages are the most common.

Start over at step 1 with each of the two new datasets, separately.

The partitioning algorithm sometimes makes poor predictions by putting observations with spurious extreme values together in a leaf as the number of customers in a leaf gets smaller. The next step, pruning, addresses this by using cross-validation to determine whether trees with a larger number of leaves (“deep trees”) or a smaller number of leaves (“shallow trees”) have a smaller MSE out of sample. Once the best penalty term on depth is selected via cross-validation, some leaves are removed because they improve the predictions by less than they receive in penalty due to their size.

As an example, imagine we are trying to predict average annual energy consumption for 100 new residential customers but all we know about them is whether they live in a house or apartment and how many people will be living in the home. If we have the same data on all of our existing customers, we could build a regression tree using this data and use the final leaves to make predictions. To do this, we would first look at splitting the data into a ‘house-only’ node and an ‘apartment-only’ node versus splitting it into a ‘one resident’ node and ‘more than one resident’ node versus a ‘one or two resident’ node and a ‘more than two resident’ node, etc. We would choose the splitting rule that minimized the MSE. Imagine that the algorithm chose the apartment versus house split as the biggest difference. In this case, the exercise of finding the best split would then be repeated among customers living in houses and apartments separately. The algorithm might find that the big increase in consumption for apartment dwellers is between apartments with one or two residents and those with more than two residents while houses with fewer than four residents and four or more is the best way to split house-dwellers. If we were to stop the algorithm, we would have four leaves: apartments with one or two residents, apartments with more than two residents, houses with one, two, or three residents, and houses with more than three residents. We would then use the average annual energy consumption within each leaf as the prediction for each of the 100 new customers who had the characteristics defined by that leaf.

The causal tree (CT) method adapts the basic logic of CART to pursue the goal of estimating a causal treatment effect, like the energy savings from a home energy report, within each leaf rather than predicting the value of an outcome variable. To do this, CT diverges from CART in two ways. First, CT changes the partitioning rule to include two components. The rule selects the partition that most effectively separates customers with larger energy savings from those with smaller energy savings. But, the rule also takes into account the fact that smaller partitions have noisier estimates of energy savings than larger ones. For example, imagine there are two possible ways to partition the data. The first splits the data so that 30% of customers are in one leaf and 70% of the customers are in another. The estimate of energy savings in the first leaf is 1 kWh and in the second leaf the estimate is 10 kWh. The second way to partition the data has 50% of the data in one leaf and 50% in another. The estimate of energy savings is now 2 kWh and 9 kWh. Because the second partition found a similar difference in energy savings between its two leaves, but did so with a closer to equal split, the algorithm would tend to select the second partition. This means that the algorithm avoids selection splits with spurious extreme energy savings.

Additionally, CT prunes the tree differently from the CART algorithm. Because CT already accounts for the depth of the tree in its splitting rule (by incorporating the size of the leaves produced by each candidate split), it does not require as much pruning. CT, nevertheless, uses cross-validation to remove some remaining bias that favors deep trees over shallow trees.

We implement CT as described above using the 59 features detailed in Appendix A. We used a 20% random sample[5] of the Waves 1, 2, and 3 for building the tree both to reduce computation time and because we view the causal tree more as an exploratory tool than a final output.

2.2 Results and Output

The most compelling feature of CT is that its output can be visually represented as a tree that details decision rules that, if followed, provide an estimate of energy savings for the customers that meet the criteria up to that node. Instead of presenting every partition, we focused on the first few sections of the CT. Figure 2-1 shows the first few partitions and nodes of the tree displayed in a simplified format.[6] A negative output number represents a load reduction. The first node contains 100% of the customers and the average treatment effect for all of the customers is -0.25 kWh per day. The first partition rule is daily winter usage less or greater than 576 kWh. This value is selected by the algorithm because it minimizes the mean squared error (as described in section 2.1). The first partition is effectively the rule that separates customers with larger energy savings from those with smaller energy savings while limiting the noise of an estimate for a small group. 99% of the population has a daily winter usage value less than 576 kWh and would go to the left node. Only 1% of the population has a daily winter usage value that is greater than 576 kWh.

For the 99% of the population that had a daily winter usage value less than 576 kWh, the average treatment effect is -0.26 kWh per day. The next partition for customers who had a daily winter usage value less than 576 kWh is off-peak daily summer weekend usage being greater than 85 kWh. The next two nodes stemming from this partition split the 99% into two groups. About 33% of the population has an off-peak daily summer weekend usage greater than or equal to 85 kWh while 66% of the population has an off-peak daily summer weekend usage less than 85 kWh.

The customers who had a daily winter usage value greater than 576 kWh had an average treatment effect of -0.14 kWh per day. All nodes and partitions stemming from this node contain less than one percent of the overall population used to create the CT.

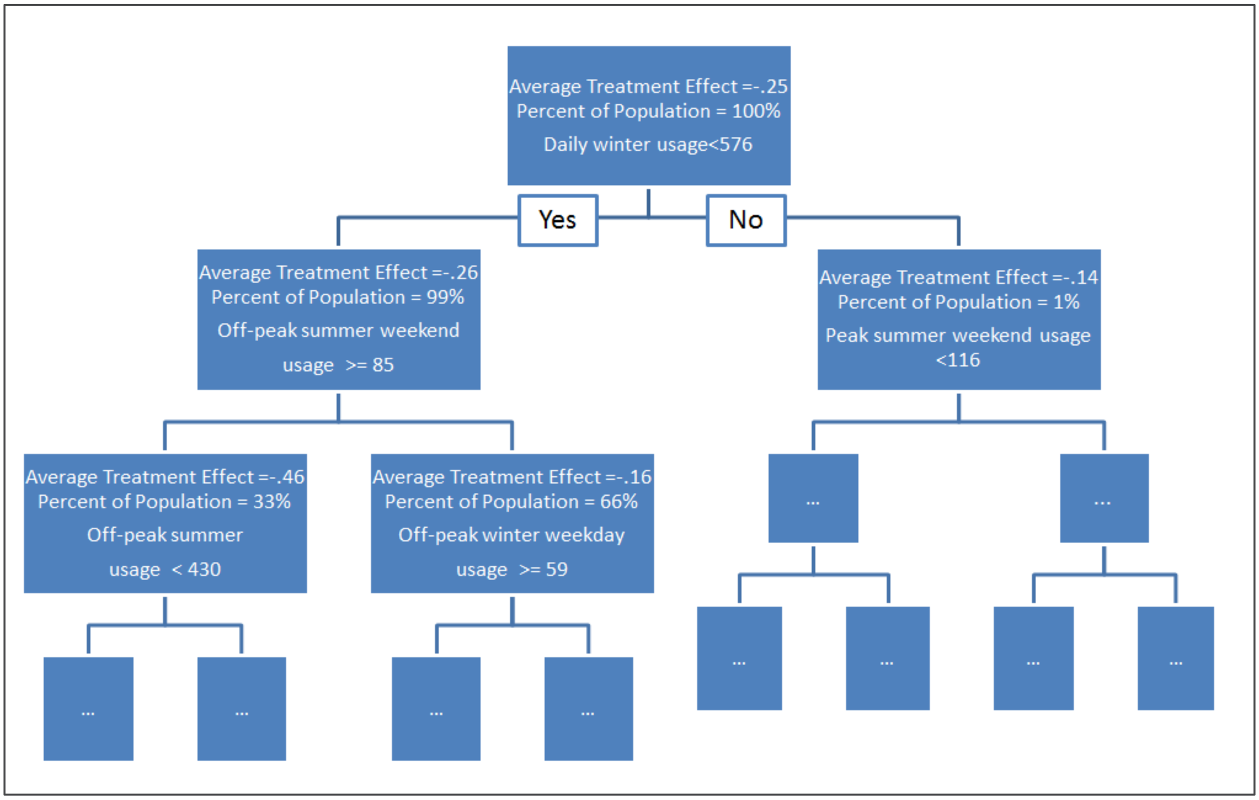
This explains how causal trees are built, seeking out large differences in the estimated treatment effect by partitioning the customers into groups. However, Causal Trees are not the most efficient and reliable approach to predicting future energy savings from customer characteristics. Despite the intuitive and visual appeal of the causal tree algorithm, individual regression trees are known to be highly sample-dependent. In other words, two trees built with the exact same algorithm on two random samples from the exact same population can produce

[5] The random sample was comprised of approximately 262,473 customers.

[6] A continuation of the causal tree is included in Appendix A. The full tree has over 300 nodes and is too large to include in this report.

quite different predictions for any given individual. The most common way to overcome this problem is with a random forest algorithm.

Figure 2-1: CT First Nodes



3 Causal Forest

3.1 Methodology

As the name implies, a random forest is a collection of regression trees. Each tree is built using a bootstrap random sample[7] of the training data, and each split is built using a random sample of the features. The resulting trees are then averaged together, and each observation is provided a predicted treatment effect. The causal forest (CF) method is a simple extension of random forests where each tree is a causal tree rather than a standard regression tree.

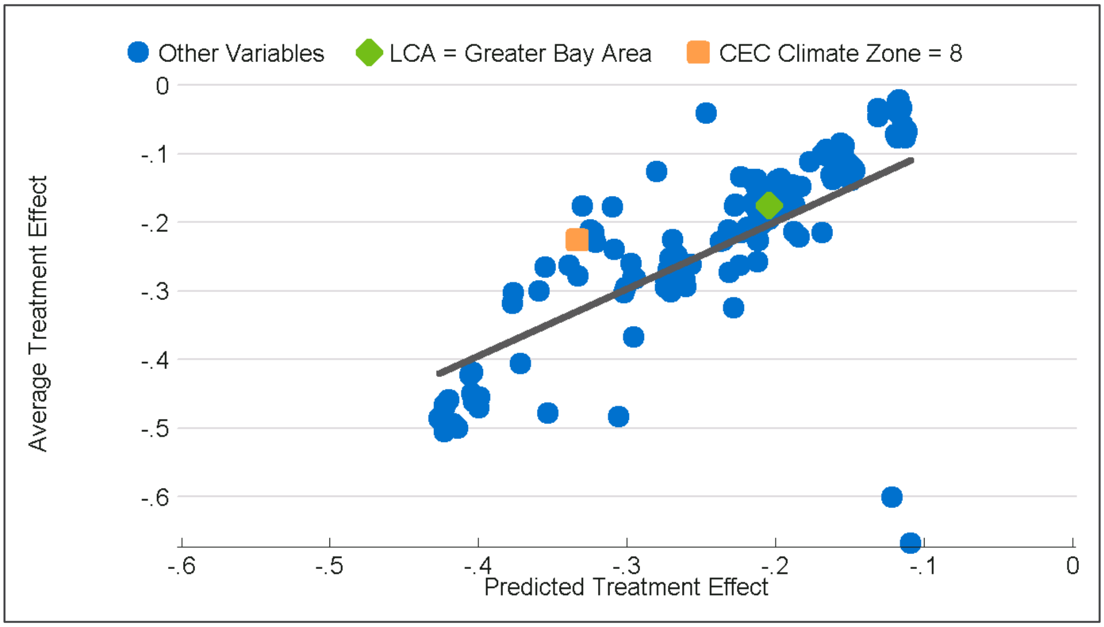
3.2 Calibration

In order to test how well the CF is calibrated, we compare the predicted treatment effects from the CF to the average treatment effects within each of the 5 quintiles of the usage variables used in the model. A similar verification is executed within other categorical variables including CARE, climate zone, local capacity area (LCA), and weather station. Such comparisons test how well predicted and average treatment effects align using a different random sample. Figure 3-1 presents the relationship between the average and the predicted treatment effects for the

[7] A bootstrap random sample is a random sample from a dataset with replacement to match the size of the original dataset. This means that a bootstrap random sample may have multiple copies of the same observation, while other observations are absent.

variables described. A couple examples are highlighted in green and orange. The relationship between the two values is linear, with the average and the predicted treatment effects being very similar in value for most variables. A few outliers (climate zone and the weather station representing the Humboldt and Eureka areas) have large treatment effects while their predicted treatment effects are small.

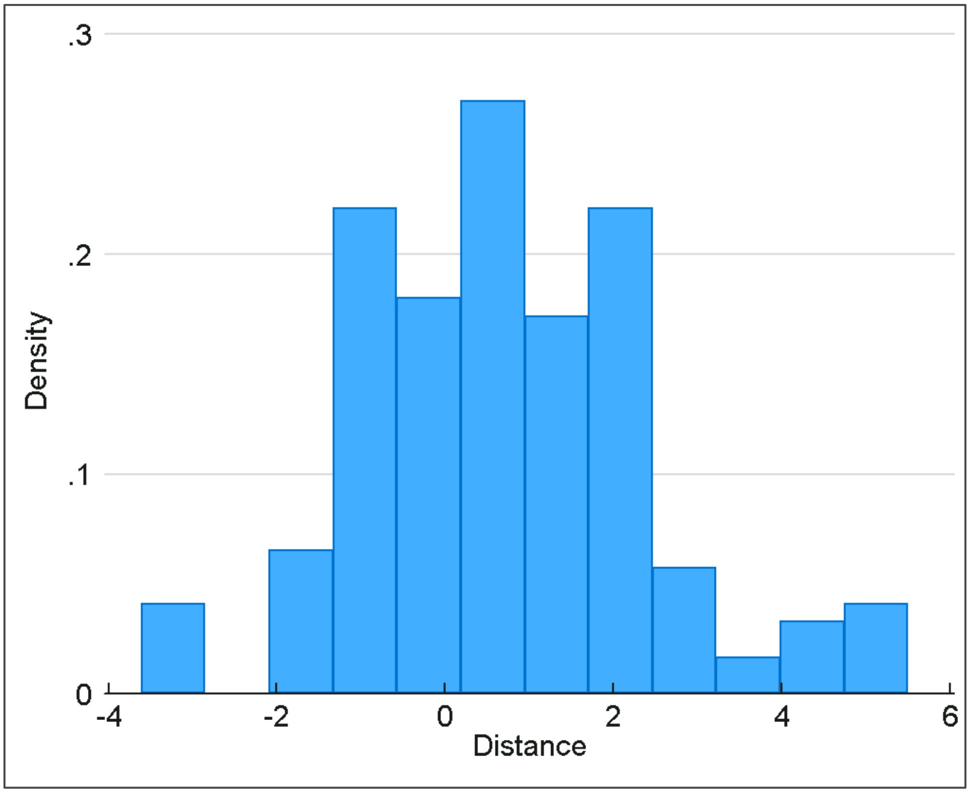
Figure 3-1: Comparison of Average Treatment Effects and Predicted Treatment Effects[8]



The distribution of the difference between the average treatment effect and predicted treatment effect divided by the standard error across all usage and demographic variables is another way to check how well the CF is calibrated. This value is called the “distance.” This tells us whether the predicted treatment effect is a plausible prediction of the true treatment effect in these subgroups (LCA = Greater Bay Area, for example). Figure 3-2 shows that the distribution of the difference across the subgroups is centered around zero and approximately normal. This demonstrates that average predictions from the CF are unbiased when we test them out of sample.

[8] Each point is associated with one specific model variable

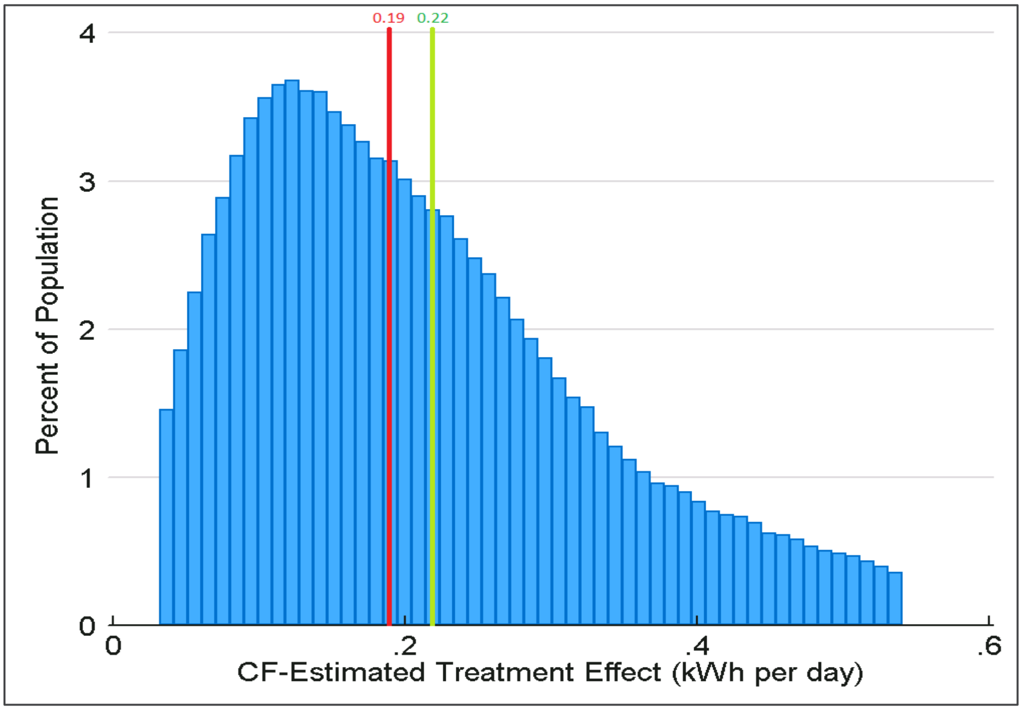
Figure 3-2: Distance Between Treatment Effect and Predicted Treatment Effect



3.3 Results and Output

The causal forest algorithm finds meaningful heterogeneity in average daily energy savings across customers. Figure 3-3 presents the distribution of treatment effects for the 5th to 95th percentile of customers. The individual treatment effects estimated using the causal forest algorithm range from 0.04 kWh per day at the 5th percentile to 0.54 kWh per day at the 95th percentile. Only approximately 2% of customers are predicted to have negative savings. This suggests that, despite the large dataset, there is very little evidence that a particular mix of characteristics is associated with negative savings.

Figure 3-3: Distribution of Treatment Effects from Causal Forest Algorithm



Excluding participants with low future savings would increase program cost-effectiveness. It is thus important to identify participants who are providing savings that justify the per-customer cost. The distribution of expected treatment effects has a pronounced right skew, with a median of 0.19 kWh per day (the red line) and a mean of 0.22 kWh per day (the green line). 58% of customers are predicted to produce savings less than the mean of 0.22 kWh per day. This illustrates both the challenge of targeting for HER programs and the potential opportunities for improving program performance. There may be HER participants that do not provide enough savings to justify the per-participant program cost. The estimates from the causal forest algorithm would allow PG&E to stop sending reports to customers based on their estimated return.

If we assume that the avoided cost of energy is $0.08 per kWh[9] and the marginal cost of HERs is $8 per customer per year[10], a participant must save over 0.27 kWh per day, on average, to offset the cost of their reports. This idea is illustrated in Figure 3-4, where the red line represents an annual cost of $8 per treated customer per year, and the blue line represents the marginal benefit of a treated customer. Customers whose benefits do not outweigh their costs fall below the red line (approximately 70% of customers, in this example).

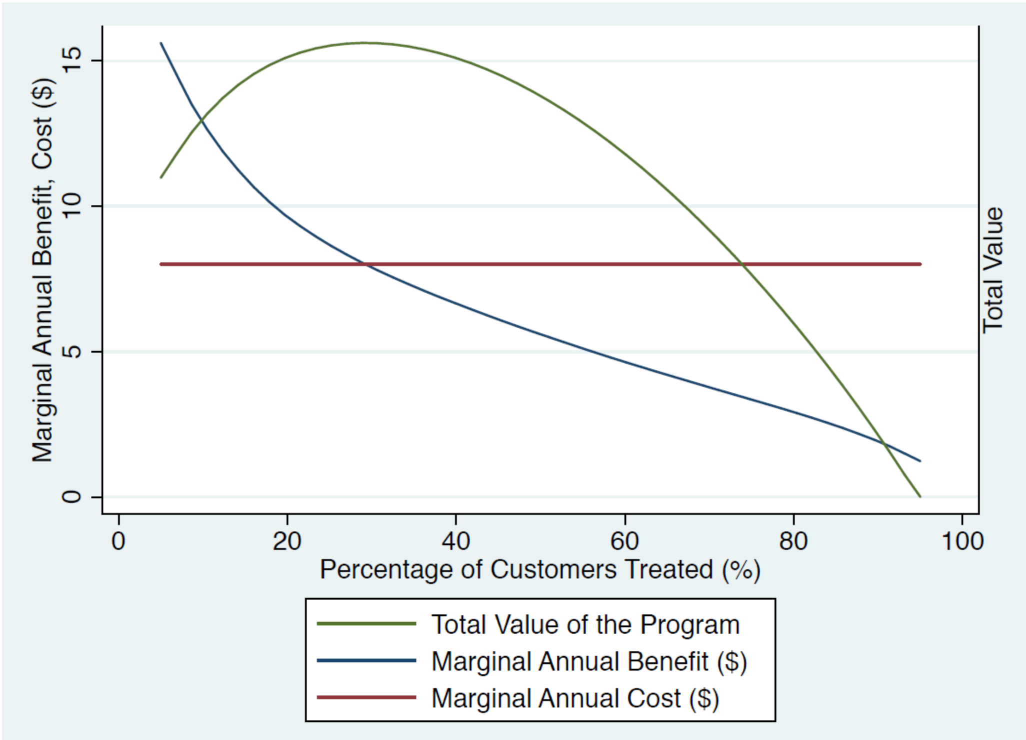
The green line represents the value of the HER program, which peaks at the intersection of the marginal annual benefits and costs per customer, (i.e., when customers in the top 30% of CF-

[9] This value is found in the Avoided Cost Calculator distributed by the CPUC. (http://www.cpuc.ca.gov/General.aspx?id=5267). This is not a cost provided to Nexant by PG&E.

[10] Based on our experience with similar delivery programs, it is reasonable to assume the marginal cost of a HER program is approximately $8 per customer per year. This is not a cost provided to Nexant by PG&E.

predicted energy savings are treated). The value of the program decreases as customers with smaller expected energy savings are included in the program.

Figure 3-4: Example of Marginal Annual Benefits and Costs of HER Program



Based on these simple calculations, approximately 70% of customers are predicted to provide a treatment effect less than 0.27 kWh per day and are therefore not cost-effective to include in the program. A more sophisticated analysis may take into consideration other factors such as PG&E's compensation tied to claimed energy savings (Efficiency Savings and Performance Incentive [ESPI] awards) and other financial inputs. Although this is a simplified cost-effectiveness analysis, it highlights the benefits for PG&E to target customers who are expected to provide more energy savings.

There are three ways to understand what is driving the causal forest output: (1) variable importance, (2) best predicting variables, and (3) simple decision rules.

Variable Importance

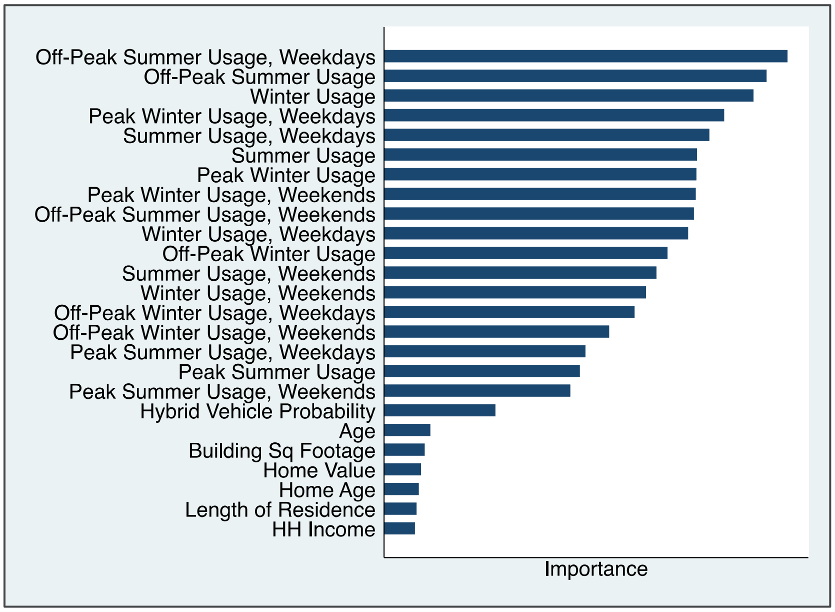
The first is to assess which variables are consistently used to split the data across the forest of trees and those that are used earlier in the splitting. This is called variable importance.[11] Figure 3-5 plots variable importance for the top 25 variables used by a causal forest fit on this data.[12] Usage variables are the most commonly used variables in part because they can be split many

[11] The reported number for variable importance is not important here. We are commenting on the relative values.

[12] Variable importance was calculated using a separate causal forest since this was added after the initial causal forest was trained. Since random forests may vary slightly from one to the next due to different starting values for random number generators, this figure may be slightly different if constructed using the original causal forest that we trained.

times—a binary variable such as a flag for whether a customer is in a particular local capacity area can indeed only be split once. Nevertheless, this suggests that pre-treatment demand measures are more important for predicting energy savings than other demographic characteristics despite the richness of the set of characteristics available.

Figure 3-5: Variable Importance for the Top 25 Variables Used in the Causal Forest



Best Predicting Variables

A second way to understand the results is to determine which variables are best at predicting the treatment effect. This is another way of telling us whether those variables are responsible for most of the predicted treatment effect from the causal forest algorithm.[13] Table 3-1 reports the top ten variables. The single best variables at predicting the causal forest output are all consumption variables, for example "offpeak_summer" which measures average demand during off-peak summer hours.

Two things are highly notable about the variables that appear to be predictive of energy savings from HERs. The first 18 of the top 25 predictive variables are variations on consumption measurements. The conventional demographic factors such as age, home value, length of residence and household income appear to be less important predictors. This finding is consistent with historical findings that the only real predictor of future energy savings from HERs is historical energy consumption.

However a more subtle and interesting finding from the analysis is that a number of the best predictors measure latent consumption (e.g., off-peak summer or off-peak winter consumption) rather than peak consumption. This suggests that consumption patterns at times when energy in the household is not being intensively used (i.e., off peak periods) are better predictors of how

[13] The R^2 values come from 3rd-order polynomial regression for the continuous variables and a fully saturated model for the discrete variables.

much a household will provide in savings, rather than measures that incorporate peak summer consumption.

Table 3-1: Best Variables for Approximating the Causal Forest Output

Variable	R^2
Off-Peak Summer Usage	0.565
Off-Peak Summer Usage, Weekdays	0.557
Off-Peak Winter Usage	0.555
Winter Usage, Weekdays	0.553
Off-Peak Summer Usage, Weekends	0.552
Off-Peak Winter Usage, Weekdays	0.546
Off-Peak Winter Usage, Weekends	0.541
Winter Usage	0.540
Summer Usage	0.537
Summer Usage, Weekdays	0.532

Nevertheless, each of these variables leaves a large percentage of the causal forest output unexplained.[14] Since the causal forest output is perfectly determined by the pre-treatment variables, this analysis suggests that the causal forest algorithm is likely outperforming simple polynomial regression (for continuous variables) or category-by-category prediction (for categorical variables) based on a single variable.

Simple Decision Rules

Lastly, we can evaluate the performance of a simple decision rule for HER participation based on one or two variables versus the causal forest predictions. If the performance is high enough, the simple, more interpretable rule could be used by PG&E. To do this, we divide categorical variables into their given categories and continuous variables into deciles. We then construct every possible decision rule for these variables that uses one or two categories. Table 3-2 reports the results for the top ten, best-performing decision rules. We find that no decision rule explains even half of the variance in the causal forest output. In other words, a simple rule does not perform as well as the causal forest.

[14] The R^2 values reported in Table 3-1 represent the output from linear regressions regression of the causal forest output on a third-order polynomial of the variable named in the table. Categorical predictors, such as CARE status or weather station, were included discretely so that the R^2 represents how well taking the mean causal forest prediction within each category captures the causal forest prediction.

Table 3-2: Best Decision Rules for Approximating the Causal Forest Output

Category 1	Category 2	R^2
Winter > 80th percentile	Off-Peak Summer > 50th percentile	0.400
Winter > 80th percentile	Off-Peak Summer, Weekend > 50th percentile	0.397
Off-Peak Summer > 50th percentile	Off-Peak Winter > 80th percentile	0.397
Off-Peak Summer > 80th percentile	Off-Peak Winter > 50th percentile	0.397
Off-Peak Summer > 80th percentile	Off-Peak Winter > 40th percentile	0.396
Winter, Weekday > 80th percentile	Off-Peak Summer > 50th percentile	0.395
Off-Peak Summer, Weekend > 50th percentile	Off-Peak Winter > 80th percentile	0.394
Winter, Weekend > 80th percentile	Off-Peak Summer > 50th percentile	0.394
Summer > 50th percentile	Winter > 80th percentile	0.394
Winter > 80th percentile	Off-Peak Summer, Weekday > 50th percentile	0.394

4 Recommendations

Given the observed performance of the causal forest algorithm relative to simple rules and its expected performance given prior research, we recommend that PG&E use the predicted treatment effects from the causal forest directly for making decisions about which customers to treat. This output can be pared with a simple decision rule, such as “customers with negative expected energy savings will no longer receive home energy reports” or “customers with expected energy savings worth less than the marginal participant cost will no longer receive home energy reports” to improve upon energy savings forecasts.

Depending on the accurate benefit per kWh and cost per report, a substantial percentage of current HER participants may not be providing sufficient savings to justify the cost of delivering the reports to them. Based on the results from the causal forest algorithm and our assumptions, PG&E could stop sending reports to these customers based on their predicted savings (i.e., likely return on investment). In Section 3.3, an illustrative analysis using assumptions about costs per report and benefits per kWh indicates that 70% of customers provide a treatment effect that is too small to be cost effective. In other words, this analysis reveals that HERs cost more than the benefits they provide for more than 70% of the sampled participants. In this example, delivering HERs only to customers with savings predicted to be in excess of the cost of delivery (over 0.27 kWh per day) could reduce PG&E’s program costs significantly while achieving cost effective energy savings. Excluding customers with trivial or negative energy savings would not have a significant impact on aggregate program impacts (in other words, removing individual customers with zero savings would have zero impact on aggregate savings).

Nexant recommends PG&E develop a formal experiment designed to test the impact of discontinuing the delivery of HERs to customers who are predicted not to have energy savings sufficient to cost justify their continued delivery. This experiment would be similar to PG&E’s experiments testing the persistence of energy savings. Treatment discontinuation would be based on customers’ predicted future savings, instead of random selection. If customers with small expected savings are removed from the program, it is possible that we may not see a difference in energy savings after terminating their reports.

Appendix A

This appendix contains a continuation of the causal tree output (Figure 2-1) in addition to a list of all the variables included in the causal forest.

A.1 Causal Tree Output (continued)

Adobe Acrobat Document

A.2 Causal Forest Variable List

Table A- 1 provides the list of variables used in the causal forest. For each variable, a brief description and the source of the variable are provided. The variables used came from PG&E, Experian, and Axciom. These variables include electricity usage variables created from hourly advanced meter data, demographic variables from PG&E, demographic variables from Axciom, and demographic variables from Experian.

'Summer' lasts from May to October and 'Winter' from November to April. 'Peak hours' is the period from 5 p.m. to 10 p.m.

Table A- 1: Causal Forest Variable List

Variable Name	Description	Source
CARE	California Alternate Rates for Energy (CARE) Program Participation Indicator	PG&E
Res_Class	Residential Dwelling Type	PG&E
Baseline Territories	Breakdown of PG&E's territory	PG&E
C_Sched	Compressed Rate Schedule Indicator	PG&E
BPP	Balance Pay Plan Indicator	PG&E
LCA	Local capacity area	PG&E
Sublap	Sub-load aggregation point	PG&E
Wthrstn	Weather Station	PG&E
Cecclmzn	CEC Climate Zone	PG&E
Swimming Pool Indicator	Flag indicating if there is a swimming pool at the premise	Experian
Presence of Children 0-18	Flag indicating if there are children present	Experian
Dwelling Type	Residential Dwelling Type	Experian
Language	Language Preference Indicator	Experian
Age	Age of the customer	Axciom
Community Involvement Flag	Flag indicating if the customer is involved with the community by financially supporting causes	Axciom

Education Input	Code indicating the education level of the customer	Axciom
Environment or Wildlife	Flag indicating if the customer financially supports environment or wildlife causes	Axciom
Environmental Issues	Flag indicating if someone in the household has an interest in environmental issues	Axciom
Estimated Household Income Ranges	Code indicating the estimated household income range of the customer	Axciom
Green_Living	Flag indicating if the customer participates in green living	Axciom
High_Tech_Living	Flag indicating if the customer participates in high tech living	Axciom
Home_Heating_Cooling	Code indicating the presence of heating and/or cooling in the residence	Axciom
Home_Improvement_Diyers	Flag indicating if the customer participates in do it yourself home improvement	Axciom
Home_Market_Value	Code indicating the home market value range	Axciom
Home_Owner_Renter	Code indicating if the customer is a home owner or renter	Axciom
Home_Pool_Present	Flag indicating if there is a swimming pool at the premise	Axciom
Home_Property Type	Code indicating the home property type	Axciom
Home_Square_Footage_Actual	The actual square footage of the home	Axciom
Home_Year_Built_Actual	Year the home was built	Axciom
Household_Size	The number of people in the household	Axciom
Hybrid_Score	Indicates the likelihood that a customer will purchase a hybrid vehicle	Axciom
Intend_To_Purchase_Home_Improve	Flag indicating if the customer intends to purchase home improvement	Axciom
Language_Preference_Code	Code indicating the customer's language of preference	Axciom

Length_Of_Residence	Code indicating the years of residence at the premise	Axciom
Number_Of_Adults	Number of adults in the household	Axciom
Political	Flag indicating if the customer financially supports political causes	Axciom
Political Conservative	Flag indicating if the customer financially supports conservative political causes	Axciom
Political Liberal	Flag indicating if the customer financially supports liberal political causes	Axciom
Presence Of Children	Flag indicating if there are children present	Axciom
Primary Address	Code indicating if customer premise is the primary address	Axciom
Suppression_Mail_Dma	Flag indicating do not mail to this individual	Axciom
Daily0_Summer	Total daily summer weekday usage	PG&E Collapsed Interval Data
Daily1_Summer	Total daily summer weekend usage	PG&E Collapsed Interval Data
Daily_Summer	Total daily summer usage	PG&E Collapsed Interval Data
Daily0_Winter	Total daily winter weekday usage	PG&E Collapsed Interval Data
Daily1_Winter	Total daily winter weekend usage	PG&E Collapsed Interval Data
Daily_Winter	Total daily winter usage	PG&E Collapsed Interval Data
Peak0_Summer	Peak daily summer weekday usage	PG&E Collapsed Interval Data
Peak1_Summer	Peak daily summer weekend usage	PG&E Collapsed Interval Data
Peak_Summer	Peak daily summer usage	PG&E Collapsed Interval Data
Peak0_Winter	Peak daily winter weekday usage	PG&E Collapsed Interval Data
Peak1_Winter	Peak daily winter weekend usage	PG&E Collapsed Interval Data
Peak_Winter	Peak daily winter usage	PG&E Collapsed Interval Data
Offpeak0_Summer	Offpeak daily summer weekday usage	PG&E Collapsed Interval Data

Offpeak1_Summer	Offpeak daily summer weekend usage	PG&E Collapsed Interval Data
Offpeak_Summer	Offpeak daily summer usage	PG&E Collapsed Interval Data
Offpeak0_Winter	Offpeak daily winter weekday usage	PG&E Collapsed Interval Data
Offpeak1_Winter	Offpeak daily winter weekend usage	PG&E Collapsed Interval Data
Offpeak_Winter	Offpeak daily winter usage	PG&E Collapsed Interval Data