RESOURCES *for the* FUTURE

RCTs Against the Machine: Can Machine Learning Prediction Methods Recover Experimental Treatment Effects?

Brian C. Prest, Casey J. Wichman, and Karen Palmer

Working Paper 21-30
September 2021

About the Authors

Brian C. Prest is a fellow at Resources for the Future specializing in climate change, electricity markets, and oil and gas economics. Prest uses economic theory and econometric models to improve energy and environmental policies by assessing their impacts on markets and pollution outcomes. His recent work includes evaluating the impacts of federal tax credits for coal use. He is also working to establish an empirical basis for determining discount rates used in the social cost of carbon. His past work includes econometric analysis of the US oil and gas industry, understanding the economic effects of rising temperatures, modeling the market dynamics of climate change policy under policy uncertainty, and assessing household responses to time-varying electricity pricing. His work has appeared in the *Journal of the Association of Environmental and Resource Economists*, *Energy Economics*, and *The Energy Journal*.

Casey J. Wichman is an assistant professor in the School of Economics at the Georgia Institute of Technology and a university fellow at Resources for the Future. He is an applied microeconomist working on issues at the intersection of environmental and public economics. His research focuses on how people interact with the natural and built environment and what that behavior reveals about the value of environmental amenities. Wichman's research spans water and energy demand management, valuation of environmental resources and infrastructure, urban transportation, public goods provision, demand for outdoor recreation, climate policy, and assessment of informational constraints in decision-making.

Karen Palmer is a senior fellow at Resources for the Future and an expert on the economics of environmental, climate, and public utility regulation of the electric power sector. Her work seeks to improve the design of environmental and technology regulations in the sector and the development of new institutions to help guide the ongoing transition of the electricity sector.

Acknowledgements

The authors are grateful to Vincent Gonzalez and Derek Wietelman, who provided exceptional research assistance on this project. The authors thank Matt Harding, Grant Jacobsen, Joshua Blonz, Ken Gillingham, Jesse Burkhardt, and seminar and conference participants at the 2020 Duke Energy Data Analytics Symposium, Resources for the Future, and the 2021 Virtual AERE conference for helpful comments and suggestions. We also appreciate data access and advice provided by Suzanne Russo, Fisayo Fadelu, and Rachel Jenkins at Pecan Street. This research was supported by the Alfred P. Sloan Foundation.

About RFF

Resources for the Future (RFF) is an independent, nonprofit research institution in Washington, DC. Its mission is to improve environmental, energy, and natural resource decisions through impartial economic research and policy engagement. RFF is committed to being the most widely trusted source of research insights and policy solutions leading to a healthy environment and a thriving economy.

Working papers are research materials circulated by their authors for purposes of information and discussion. They have not necessarily undergone formal peer review. The views expressed here are those of the individual authors and may differ from those of other RFF experts, its officers, or its directors.

Sharing Our Work

Abstract

We investigate how well machine learning counterfactual prediction tools can estimate causal treatment effects. We use three prediction algorithms—XGBoost, random forests, and LASSO—to estimate treatment effects using observational data. We compare those results to causal effects from a randomized experiment for electricity customers who faced critical-peak pricing and information treatments. Our results show that each algorithm replicates the true treatment effects, even when using data from treated households only. Additionally, when using both treatment households and nonexperimental comparison households, simpler difference-in-differences methods replicate the experimental benchmark, suggesting little benefit from ML approaches over standard program evaluation methods.

Keywords: machine learning, causal inference, electricity demand, time-varying pricing

1. Introduction

In the last three decades, a central focus of applied economics research has been estimating causal effects of policies, market rules, or other factors on market outcomes, including consumer behavior. Randomized controlled trials (RCTs) are the gold standard for understanding the causal effects of such interventions and have been applied in many settings, including development economics (e.g., Banerjee, 2009), education (e.g., Krueger, 1999), labor economics (e.g., Jones et al., 2019; List, 2011), and environmental and energy economics (e.g., Allcott and Rogers, 2014; Jessoe and Rapson, 2014; Ferraro and Price, 2013). However, RCTs are costly, can take substantial amounts of time, and may not be feasible in certain settings due to regulations prohibiting the requisite randomization of subjects into treatment and control groups or firm or government unwillingness to implement experimental designs. More generally, researchers seeking to estimate causal treatment effects may only have access to nonexperimental data.

In such settings lacking randomized treatment, machine learning (ML) algorithms paired with program evaluation tools offer a potential alternative approach to estimate treatment effects. In data-rich settings, ML algorithms can be used to explain variation in a variable of interest and predict that variable out of sample through flexible functions that can capture complex interactions and nonlinearities in covariates. These prediction methods can, in turn, be used as a credible counterfactual that, when compared to observed data on particular outcomes, can reveal the effect of a policy change. These methods are particularly valuable in settings where outcomes are cyclical or seasonal and detailed, high-frequency data is readily accessible by researchers (Burlig et al., 2020; Christensen et al., 2021). Although some nonexperimental designs have been shown to recover experimental results in certain settings (e.g., Dehejia and Wahba, 2002; Ferraro and Miranda, 2017), the degree to which counterfactual prediction methods can recover causal treatment effects is not clear. We explore the performance of these methods alongside more standard program evaluation methods in replicating treatment effects from RCTs.

This paper compares treatment effect estimates from a randomized electricity pricing and information experiment to nonexperimental effects estimated using both standard difference-in-differences methods and three commonly used ML counterfactual prediction methods (XGBoost, random forests, and Least Absolute Shrinkage and Selection Operator (LASSO)) in a "prediction-error" framework, where predicted outcomes serve as counterfactuals. We partnered with the nonprofit Pecan Street Inc. to access high-frequency energy-use data matched with household characteristics for a set of households in Austin, Texas, that were part of this experiment (also analyzed by Burkhardt, Gillingham, and Kopalle 2019, henceforth "BGK") and similar households that were not subject to the experiment. We train our ML models using pretreatment hourly consumption data, household characteristics, weather, and interactions among these variables to make out-of-sample predictions for electricity consumption during the treatment period that we use as household-by-hour-specific counterfactual consumption. We then compare these counterfactuals to the observed consumption in a fixed effects regression framework to estimate nonexperimental treatment effects. We implement this approach

for four treatments, including ones with both true significant effects and null effects according to the experimental benchmark, to assess the ability to replicate both kinds of results.

To assess the value of the ML counterfactual prediction methods in various contexts, we implement our procedure using three different samples in which we pair the treatment groups' data with (a) the original experimental control group data, (b) a nonexperimental comparison group comprising similar households that were not part of the randomized experiment, and (c) no additional comparison group—that is, with only data on treated households. The first sample forms our experimental benchmark using both observed and predicted data. The second sample forms a nonexperimental replication sample, which we use to estimate treatment effects in a difference-in-differences framework with observed data and a prediction-error framework with ML-predicted data as counterfactual consumption. The third sample follows the same format but uses data from the treatment group only, forgoing the nonexperimental comparison group entirely, with predicted consumption serving as a counterfactual.

We uncover several important results. First, each ML counterfactual prediction approach replicates the RCT's true treatment effects (both true significant and null effects), even with only data on treated households. This result suggests that ML counterfactual prediction methods can construct a valid counterfactual when researchers do not have a credible set of comparison units. Second, when using treatment households along with nonexperimental comparison households, we find that simpler difference-in-differences methods can replicate the experimental benchmark, while ML-predicted counterfactuals offer little benefit relative to traditional program evaluation approaches with the nonexperimental comparison group. Third, we find little difference in the ability of each ML algorithm to replicate the experimental benchmark, despite nontrivial differences in predictive accuracy across algorithms. This last result suggests that predictive accuracy is an imperfect proxy for how well ML prediction methods will perform in estimating causal treatment effects. Additionally, we explore what combination of variables (e.g., weather, household characteristics, and covariates representing diurnal, weekly, and seasonal patterns) are necessary for replication, which provides actionable insight for applied researchers. We find that ML counterfactual prediction methods can replicate experimental treatment effects remarkably well with relatively little data on predictors.

These methods and results may be particularly useful in analyzing energy demand and how pricing and other policy interventions intended to change consumer demand choices affect actual consumption behavior. The growing penetration of smart electricity meters has increased the temporal granularity of consumption data, which both presents researchers with potentially valuable opportunities for better policy evaluation (Ghanem and Smith 2021) and makes it increasingly possible to implement time-varying prices that could increase the efficiency of electricity market outcomes. In practice, most policy evaluation is performed with nonexperimental data, so new ML-based methods applied to newly available rich datasets offer an implementable opportunity for improving the estimation of causal effects.

2. Related Literature

Over the last 35 years, a series of researchers have evaluated how well common program evaluation methods using observational data perform in recovering experimental treatment effects. This work has provided general insight into the usefulness of different empirical approaches in different settings (Imbens and Wooldridge, 2009; Cook et al., 2008). These "design-replication studies"—sometimes referred to as "within-study designs"—focused initially on replicating the results from an experimental job-training program, starting with LaLonde (1986). In these settings, researchers attempt to replicate the results of an experimental benchmark, typically generated from a randomized field experiment, without using the experimental control group. To do so, authors gather a *nonexperimental* comparison group to serve as a credible replacement for the experimental control group and then apply common research designs, including difference-in-differences, propensity-score matching, and covariate matching, as well as regression discontinuity and instrumental variables in certain settings.

The initial conclusions from this line of inquiry suggested that early propensity-score matching designs failed to replicate randomized results, although this conclusion has been subject to lively debate, finding cases in which the method does succeed, particularly among subsamples where treatment and comparison groups are demonstrably similar (e.g., Heckman et al., 1998; Dehejia and Wahba, 1999, 2002; Smith and Todd, 2005; Dehejia, 2005).[1] More recently, design-replication studies that pair covariate matching with panel designs have fared comparatively better than early results that focused on propensity-score matching, likely due to the ability to control for both observable and unobservable heterogeneity by design (e.g., Ferraro and Miranda, 2017; St. Clair et al., 2014). However, some have found that nonexperimental results from matching and panel data fail to replicate even when exhibiting strong balance and common pretrends (Wichman and Ferraro, 2017).

In recent years, the advent of high-frequency and feature-rich data sets has spurred a rapid development and adoption of ML tools to predict counterfactuals that can be used for estimation of causal effects. While the idea of estimating causal effects in reference to an unobserved counterfactual is not new (Rubin, 1974), recent advances in ML algorithms and increased availability of rich datasets have greatly improved economic researchers' prediction accuracy (Varian, 2014, 2016). Although counterfactual prediction using ML has the flavor of synthetic control approaches (e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010), ML approaches permit much more flexible nonlinear, high-dimensional models that can deliver superior predictions.

Recent ML-based methods used to predict counterfactuals include LASSO and random forests (e.g., Burlig et al. 2020; Liberman et al. 2018; Abrell et al. 2019; Dueñas et al. 2021), XGBoost (e.g., Souza 2019; Christensen et al. 2021; Dueñas et al. 2021), matrix completion (Athey et al., 2021), and deep neural networks (Hartford et al., 2016). The application in

[1] Replications via regression discontinuity have performed reasonably well near the cutoff (e.g., Wing and Cook, 2013), although they are not informative for settings in which no plausible cutoff in a forcing variable exists.

this paper is energy, but these methods are used in many areas in the literature just referenced, ranging from studying consumer credit in developing countries to the effects of COVID-19 on international trade. In contrast to other recent applications of ML tools in the context of causal inference, such as to estimate heterogeneous treatment effects (e.g., Athey and Imbens, 2016; Prest, 2020; Knittel and Stolper, 2019; Cunningham et al., 2021) or firm unobservables (e.g., Davis et al., 2020), the focus of this paper is counterfactual prediction. These approaches are promising, particularly in settings in which an observational, nonexperimental comparison group cannot be found. However, as far as we are aware, no research has formally compared these ML counterfactual prediction tools to a randomized benchmark.

We apply ML counterfactual prediction methods to study policies intended to reduce electricity consumption at times of peak demand. Specifically, we consider when such ML methods can replicate the results of an electricity pricing experiment that was evaluated by BGK. Economists have long been advocates for electricity prices that reflect temporal variations in the costs of electricity production to encourage consumers to shift their use away from high-cost hours (Borenstein and Holland, 2005; Borenstein and Bushnell, 2019; Harding and Sexton, 2017). Real-time pricing would produce the most substantial efficiency gains relative to time-invariant pricing (Borenstein, 2005), but it can face significant political hurdles (Joskow and Wolfram, 2012; Borenstein, 2007) and is rarely implemented. Other approaches, such as critical-peak pricing, can approximate real-time pricing and yield much of its welfare gains (Blonz, 2020). Consumers will respond to various forms of time-varying prices (Wolak, 2011); however, in many cases, consumer responses have been quite small (Fabra et al., 2021; Jessoe and Rapson, 2015), which could be rational if attention costs are large (Sallee, 2014). Consumer responsiveness is also often heterogeneous in a policy's effects (Fowlie et al., 2020; Prest, 2020), and information can help increase price responsiveness (Jessoe and Rapson, 2014), as can automation (Gillan, 2018; Harding and Lamarche, 2016; Blonz et al., 2021). We contribute to this literature by demonstrating how and when ML counterfactual prediction methods can be useful in evaluating energy demand policies.

3. Empirical Setting and Data

We use data from a randomized experiment conducted by Pecan Street, Inc., a nonprofit research institute based in Austin, Texas. Pecan Street collects detailed use and generation (for rooftop solar) data to enable study of various aspects of electricity, ranging from the effects of pricing programs to smart home technology and electric vehicles. Pecan Street monitors more than 1,000 homes at high temporal resolution (down to the minute level). Throughout 2013 and 2014, Pecan Street conducted a randomized controlled trial (RCT) on about 250 households in Austin's Mueller neighborhood to assess the effect of four alternative pricing and information treatments on electricity use during peak demand hours on particularly hot summer days.

3.1. Benchmark experimental estimates

The Pecan Street RCT that we use as a benchmark was evaluated by "BGK." Their approach offers "ground-truth" estimates of a set of treatment effects to which we can compare and thereby evaluate our ML counterfactual prediction approaches.

Pecan Street recruited households into the experiment by offering a $200 participation incentive; 280 households ultimately participated, with the majority in the Mueller neighborhood in Austin. The goal was to test whether (and what type of) incentives and information about critical-peak demand events would generate short-run conservation behavior during the peak period (4–7PM). The four treatment arms were as follows:

- **Passive information ("Portal"):** Households were given access to an online portal with information about their energy use.[2]
- **Active information ("Text"):** Households were sent text messages the day before each critical-peak event, which stated: "A Pecan Street Project critical peak event is taking place tomorrow from 4 PM to 7 PM."
- **Active information + recommendation ("Text+Rec."):** Households received the same text message as the "Text" treatment arm, with the addition of one of the following recommended actions: "Pre-cool your home," "Reduce your air conditioning usage," or "Do not use your clothes dryer."
- **Pricing ("Pricing"):** Households received a text message the day before each pricing event, which stated: "Tomorrow is a Critical Peak Pricing event. Your experimental electric rate will be $0.64 per kilowatt hour from 4 PM–7 PM. Pecan Street Inc. Pricing."

Customers experienced 27 critical-peak pricing events between June 2013 and September 2014.

BGK evaluated the experiment using minute-level electricity consumption data at the household level. To estimate treatment effects, BGK regressed electricity consumption on a dummy variable, T_{ijt}, that indicates that household i is in treatment group j and received the treatment during "event" time t (which captures whether a critical-peak pricing event took place on a given day in the posttreatment period). In addition to the time-varying treatment indicator, BGK included household and quarter-hour-of-sample fixed effects as well as residual time and treatment interactions that are standard in difference-in-difference-style regressions. Standard errors are clustered at the household level. As a result, BGK's analysis produces a "triple-difference" estimate of an average treatment effect that can be interpreted as the change in electricity use across (i) treatment-control households, (ii) event and nonevent days, and (iii) critical-peak and non-critical-peak hours. BGK find that the Pricing treatment reduced hourly electricity use by 0.39 kWh

[2] All treatment arms had access to this passive information; households in the control group did not.

(∼14%) per hour, although the Portal, Text, or Text+Rec. treatments were all estimated to have no economically or statistically significant impact on consumption.

We acquired the experimental data through Pecan Street independently in 2020. Between the time that BGK acquired the data and when we did, Pecan Street implemented small changes to the historical data to improve data quality; as a result, our data (and, hence, our benchmark results) differ slightly from those presented in BGK's analysis. The major difference is that our data set features 249 households in the RCT (versus 256 in BGK) because Pecan Street found some data problems for a small number of homes. Our sample differs slightly from BGK's, but the resulting treatment effects are nearly identical.

Additionally, we extend their analysis by drawing upon more historical data to train our ML models and also considering an additional, nonexperimental sample of comparison households that were monitored by Pecan Street during the same time frame but were not part of the experiment, thereby creating a larger sample. The final difference is that for computational reasons, in training our ML models, we aggregate minute-level data to an hourly time step. Testing shows that the minute-level data produces the same treatment effect estimates as the hourly-level aggregation, which is intuitive because the treatment was imposed over three-hour windows. With data aggregated to the hourly level, we cannot use the quarter-hour-of-sample fixed effects in BGK's difference-in-differences specification, so we use hour-of-sample fixed effects instead.

With slight changes in our sample relative to the BGK benchmark, our empirical approach is adjusted accordingly. We aggregate electricity use data to the household-by-hour level but retain the central triple-difference approach. As a result, the finest level of time fixed effects we use is hour-of-sample (which is discussed in detail in Section 5). Our results, however, exploit the same variation across households, event days, and event hours.

3.2. Data

In addition to the high-resolution household electricity consumption data described earlier, the dataset includes household characteristics: square footage, house age, whether solar PV installed and how much, and whether the house is a single-family residence. Finally, it includes local weather variables at the hourly resolution: temperature, humidity, dew point, visibility, apparent temperature, pressure, and precipitation intensity. We use all of these characteristics in our ML predictive models.

Because we identify nonexperimental treatment effects using a difference-in-differences strategy, we assume parallel trends in energy consumption across the treatment and control groups. We visually assess the validity of this assumption in Figure 1 by graphing average weekly consumption for each group both before and after the treatment period begins. Because the treatments are primarily focused on changing behavior during peak hours, we show not only average consumption across all hours of the day (top panel) but also average consumption only during peak hours (4–7pm) (bottom panel).

The seasonal pattern appears similar across control groups, suggesting that the parallel trends assumption is reasonable. However, the treatment groups appear to have somewhat higher baseline consumption (and, indeed, we find these differences to be statistically significant). To the extent these differences are time invariant, they will be absorbed by household fixed effects in the difference-in-differences treatment effect estimator. Further, even time-varying differences will be absorbed to some degree by the triple differences because of the triple difference setup (which compares event days to nonevent days in the treatment period).

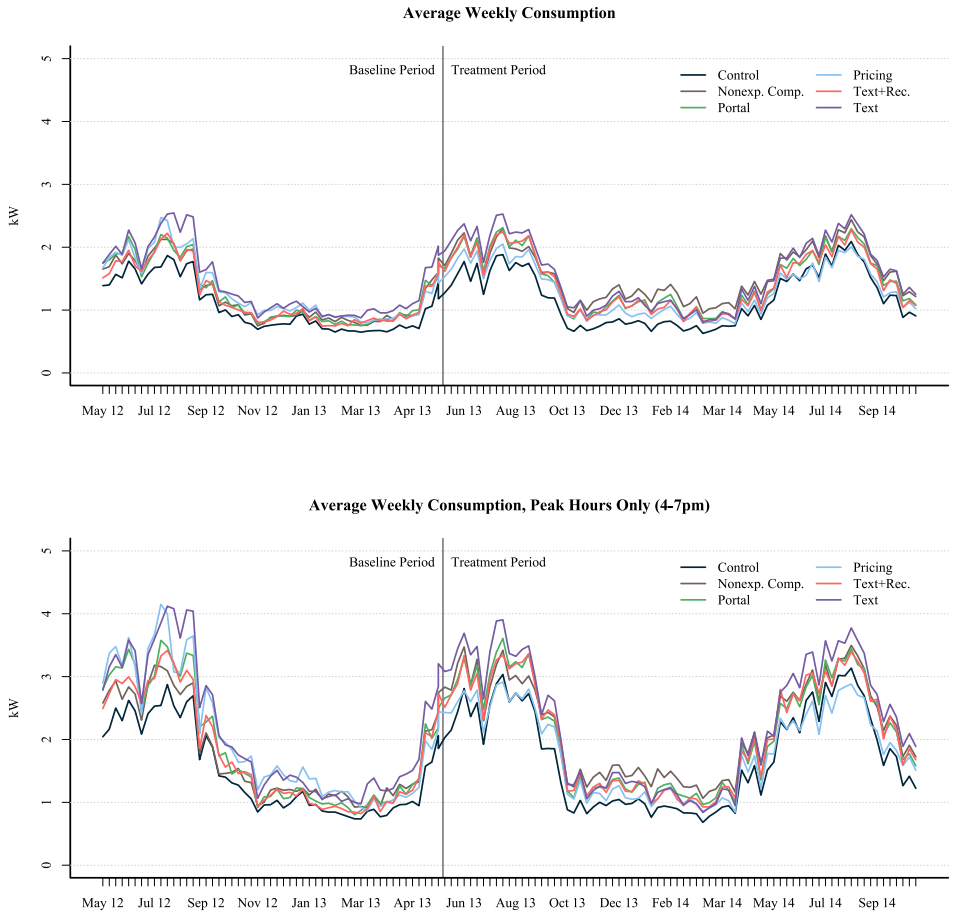

Figure 1. Weekly Average Energy Consumption for All Hours (top) and Peak Hours (bottom), by Group

4. Methods

In this section, we describe our empirical approach in detail. We first describe how we implement our counterfactual predictions via three ML algorithms that are used to generate a "prediction error" that reflects the difference between the ML counterfactual and our observed data. We then show how we use this prediction error in a difference-in-

differences-style regression framework to estimate nonexperimental treatment effects using different samples of comparison households. Last, we discuss the criteria used to compare nonexperimental treatment effects with our randomized benchmark.

We train three common ML models (XGBoost, random forest, and LASSO) on a year of historical hourly, household-level electricity consumption data. The training period is June 2012 through May 2013, which are the 12 months immediately preceding the rollout of the RCT. Each model is used to generate out-of-sample household-by-hour predictions during the experimental period, June 2013 through October 2014. These predictions serve as a counterfactual prediction for what a household would have consumed.

Each ML approach uses unit-level household characteristics and hourly weather data to inform our prediction models. The household characteristics that we include are square footage, house age, whether solar PV is installed and how much, and whether the house is a single-family residence. Weather data includes temperature, humidity, apparent temperature (i.e., the "feels like" temperature), dew point, visibility, pressure, and precipitation intensity. We include squared terms of all weather variables except visibility. We also capture predictable diurnal and seasonal consumption patterns by including flexible time covariates: hour of day, day of week, week of year, month of year, weekend, and peak hour. We include interactions of all covariates, either implicitly (in random forests and XGBoost, which by their nature construct flexible interactions) or explicitly (in LASSO). This allows for complex, highly nonlinear relationships between the variables.

4.1. ML Training

ML models simultaneously allow for flexible, highly nonlinear relationships and avoid overfitting through cross validation (CV). Conceptually, CV splits a dataset into "training" and "test" sets; each model is estimated on the training set and then used to generate predictions on the test set. One can compare the out-of-sample predictive accuracy of alternative models by evaluating which perform best on the test set, such as in terms of root mean square error (RMSE). Models with few covariates tend to lack sufficient information to generate good predictions, whereas overly complex models tend to overfit training data and produce poor predictions out of sample. CV allows one to find the model (or models) that strike the right balance by evaluating the out-of-sample predictive accuracy for a series of increasingly complex models. This process is called "tuning."

We tune each model using K-fold CV, where each observation is assigned to one of K "test" folds that are iteratively held out for evaluation. We use $K = 10$ for LASSO and random forest and $K = 5$ for XGBoost due to computational constraints. As in Burlig et al. (2020), we assign observations to folds in weekly blocks to ensure that the training data incorporates the observed temporal autocorrelation in energy consumption.

To avoid potential confusion, we pause to be explicit about our CV and counterfactual prediction periods. Our goal is to generate counterfactual predictions for the RCT's treatment period, June 2013 through October 2014. We call this the "prediction period," to distinguish it from the test set. The test sets are based entirely on pretreatment data during

the preceding year, which we refer to as the "training period": June 2012 to May 2013. In K-fold CV, the training period is divided into training and test sets, but those test sets consist entirely of data during the "training period." This approach correctly ensures that the counterfactual predictions during the prediction period are not improperly derived using prediction period data.

Our predictive models are "pooled," meaning that each method uses the same model across households. We also considered estimating household-specific predictive models, but those models tended to achieve worse out-of-sample predictive accuracy than the pooled models. This finding suggests that household-specific models can overfit by failing to incorporate cross-household information. While our "pooled" models are the same across households (for a given ML approach and sample), the predictions vary by household because of the use of household characteristics and their interactions with other variables.[3]

We now describe the details of each ML approach: XGBoost, random forest, and LASSO. In each case, we tended to favor default, "off-the-shelf" model options in the relevant statistical packages in R because these are the methods that many practitioners would likely use. We tune hyperparameters for each model as described in the subsequent sections; for hyperparameters or options not explicitly discussed, we use the default options in the relevant R package.

4.1.1. XGBoost

XGBoost, also referred to as a "Gradient Boosted Machine" (GBM) or boosted trees, uses regression trees to generate predictions. Regression trees use recursive partitioning of a dataset along each covariate (for example, is the temperature above or below 80 degrees?). For a given tree, the predicted value of the dependent variable in each partition (e.g., energy use when temperature is above 80 degrees) equals the sample average of observations in a node. The dataset is split recursively many times to automatically form flexible interactions between covariates (e.g., generating a distinct prediction for energy use when the temperature is above 80 degrees on a Sunday at 2 pm in a house larger than 1,500 square feet). This recursive splitting process allows regression trees to automatically form flexible, nonlinear interactions of explanatory variables.

XGBoost grows many regression trees sequentially that each aim to predict the residuals not yet explained by the previous trees. Specifically, boosting starts by estimating an initial regression tree, $\hat{f}^1(x_i)$, for the outcome, y_i, then shrinks it by the learning rate denoted $\eta \in (0, 1)$. The resulting residual is calculated as $\epsilon_i^1 = y_i - \eta\hat{f}^1(x_i)$. Next, a new regression tree $\hat{f}^2(x_i)$ is estimated on the residual, ϵ_i^1, instead of the outcome y_i, to explain some of the remaining variation, generating a new residual, and so on. This process

[3] In a sensitivity analysis discussed in Section 6., we use household fixed effects instead of household characteristics when training the ML model, which allows for some household specificity in the prediction models because the fixed effects can interact with other covariates.

Table 1. XGBoost Hyperparameters

Parameter	Values
Number of Trees (T)	1, 2, ..., 999, 1,000
L1 Regularization Coefficient on Sample Weights (α)	0, 0.5, 1
L2 Regularization Coefficient on Sample Weights (λ)	0, 0.5, 1

is repeated T times, where on iteration b the residual is updated as $\epsilon_i^b = \epsilon^{b-1} - \lambda \hat{f}^b(x_i)$. The final predictive model is the aggregation of the T trees $\hat{f}(x_i) = \sum_{b=1}^{T} \lambda \hat{f}^b(x_i)$.

We implement XGBoost using the `xgboost` package in R. We tune the hyperparameters shown in Table 1, leaving the remaining ones at their default values. The first hyperparameter is the number of trees, T. The other two hyperparameters penalize overly variable trees by applying some combination of the L1 penalty (α, analogous to LASSO) and the L2 penalty (λ, analogous to ridge regression). These can each be applied at full strength ($\alpha = 1$ or $\lambda = 1$), at half strength ($\alpha = 0.5$ or $\lambda = 0.5$), not at all ($\alpha = 0$ or $\lambda = 0$), or some combination thereof.[4]

4.1.2. Random Forest

Random forests also use regression trees, but whereas XGBoost grows many trees *sequentially* and sums them, the random forest algorithm grow many trees *independently* and averages them. We estimate 500 individual trees each using bootstrapped draws of training data. Random forests also decorrelates trees from each other by randomly sampling subsets of available covariates in each tree, ensuring that not all trees end up focusing on the same group of covariates. The predicted value for each observation is calculated separately for each of the 500 trees and averaged across trees to generate a single prediction for each test observation. Random forests have four key hyperparameters that we consider, shown in Table 2. We implement random forests using the `ranger` package in R.

Table 2. Random Forest Hyperparameters

Parameter	Values
Fraction of Covariates Available for Splitting	$\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}$
Num. Observations in Terminal Node Required for Split	2, 3, ..., 10
Perform Sampling with or without Replacement	With, Without
Fraction of Observations to Sample	0.60, 0.65, 0.70, 0.75, 0.80

[4] For more information on `xgboost` parameters, see `https://xgboost.readthedocs.io/en/latest/parameter.html`

4.1.3. LASSO

LASSO is a regularized version of linear regression. LASSO estimates regression coefficients by minimizing the sum of squared residuals (like ordinary least squares) plus a penalty equal to a penalty parameter λ times the L1 norm of the standardized coefficient vector. The value of λ, known as the "LASSO hyperparameter," is chosen to minimize out-of-sample prediction error as measured by the RMSE in the test sets.

As is well known, by penalizing the L1 norm of the coefficient vector, some (often many) coefficients can be driven exactly to zero, meaning the LASSO effectively performs some model selection by completely dropping variables to avoid overfitting. While XGBoost and random forests automatically capture interactions between explanatory variables, LASSO only considers linear functions of the input variables provided. Therefore, we form interactions manually by considering all possible interactions between the variables discussed at the beginning of this section. This results in a covariate space of 1,631 terms for LASSO to consider, although many are dropped in CV. We implemented LASSO using the `glmnet` package in `R`.

4.1.4. Other Counterfactual Prediction Approaches

In summary, we use three common but powerful ML approaches: XGBoost, random forests, and LASSO. These three are commonly used by economists in the literature for counterfactual prediction in similar settings, such as in Burlig et al. (2020) (LASSO and random forests) and Souza (2019) and Christensen et al. (2021) (XGBoost). We also considered the following approaches to predicting counterfactuals but ultimately did not adopt them for the reasons explained here.

- **Matrix completion**: Matrix completion, as proposed by Athey et al. (2021), "fills in" missing cells of a matrix of outcomes (such as energy use) to generate counterfactual predictions. It typically uses no covariates other than the outcome of interest and bases its prediction solely on the cross-unit and temporal patterns. This limits the use of covariates, such as temperature, that are valuable for predicting counterfactuals in energy applications. Further, matrix completion only works when one has *both* "horizontal" data, such as observing the same unit at different times, *and* "vertical" data, such as observing both treated and untreated units at the same time. The latter is not possible without a control group, which is a key consideration of this study.
- **Neural networks**: Neural network packages are not designed to handle panel data "out of the box." Further, they are not commonly used by economists in the causal inference literature, whereas the other approaches we use are.
- **Synthetic control**: Synthetic control is typically used to estimate the effect of a policy that is enacted for a single treated unit at a single time and remains in place continuously. This structure is not applicable in our setting, which has hundreds

of treated households and some treatments only occurring on certain hours of certain "event" days during the experimental period. For example, the critical-peak pricing treatment was only applied from 4–7PM on those event days.

4.1.5. RMSE Estimates

The RMSE estimates for each optimally tuned model are shown in Table 3. The RMSE values pertain to the pre-treatment period (June 2012-May 2013) as estimated using CV, meaning they are indicative of each algorithm's out-of-sample performance. XGBoost yields the lowest RMSE under all three samples, followed closely by random forests, and LASSO performs substantially worse. For this reason, we focus primarily on the results for XGBoost.

Table 3. Root Mean Square Error, by ML Approach and Sample (in kW)

	Randomized Control	Nonexperimental Comparison	Treatment Only
XGBoost	0.6689	0.7421	0.6693
Random Forests	0.6825	0.7592	0.6972
LASSO	0.8724	0.9049	0.8878

4.2. Generating Counterfactuals and Prediction Errors

After training and tuning each ML model on a year of pre-RCT data (i.e, June 2012-May 2013), we use each of the models to generate household-by-hour predictions for counterfactual energy consumption, denoted $\hat{Y}_{it}$ for each hour t in the experimental period (June 2013-October 2014). We then generate a "prediction error" for each such hour reflecting the difference between observed consumption and this counterfactual, or $Y_{it} - \hat{Y}_{it}$. We estimate the effect of treatment on this prediction error. Intuitively, if a treatment reduces observed consumption relative to the counterfactual, the prediction error will be negative, and the coefficient representing the treatment effect will reflect this negative effect. Alternatively, one could interpret the counterfactual prediction $\hat{Y}_{it}$ as a new pseudo-control observation in a difference-in-differences framework.

To illustrate this design, Figure 2 depicts the average consumption profile on event days for the RCT control group and the critical-peak pricing groups, alongside the counterfactual prediction for the pricing group under the preferred ML approach, XGBoost. The counterfactual prediction shown corresponds to the "treatment only" case where neither the control group nor the nonexperimental comparison group is used to generate the ML prediction, meaning the counterfactual is driven by only data from treated households. The divergence between energy consumption between the pricing and control group during the treated hours (4–7PM, shown as the blue shaded area) is clear and presents visual evidence for a treatment effect. Absent the comparison to the control group, however,

no strong, compelling evidence would exist for the existence or magnitude of the effect. This situation is where the counterfactual predictions can substitute as a pseudo-control group. Comparing the observed consumption to its predicted counterfactual serves as evidence for the treatment effect, without requiring an actual control group.

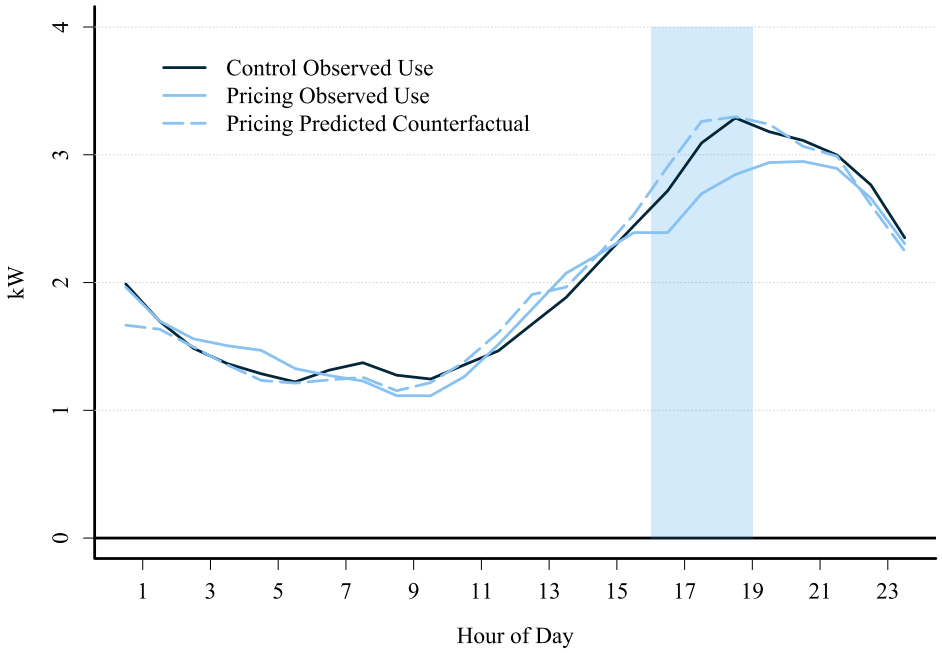

Figure 2. Event Day Average Consumption Profile for Randomized Control Group and Pricing Group, Actual and XGBoost Predicted Counterfactual (Using Only Treatment Data)

Note: The blue region reflects peak hours (4–7pm). As in BGK, the profiles are plotted after conditioning on household fixed effects and recentering on the grand mean to highlight the source of identification under a difference-in-differences regression.

4.3. Treatment Effect Estimation

4.3.1. Difference-in-Differences Specification

Once the prediction error is generated, we use standard difference-in-differences approaches to estimate treatment effects, analogous to BGK but with the prediction error as our dependent variable. Namely, we regress the prediction error (observed consumption minus predicted) during the experimental period on indicators for treatment group, event day, and event hour indicators, and their interactions serve as our regressors. We also include household fixed effects and flexible time fixed effects. Thus, we estimate

the following equation:

$$Y_{it} - \hat{Y}_{it} = \sum_j \beta_j T_{ijt} + \mathbf{X}_{ijt}\gamma + \rho_i + \phi_t + \epsilon_{it}, \quad (1)$$

where Y_{it} and $\hat{Y}_{it}$ are actual and predicted consumption for household i at time t. The treatment arm is denoted by j for the four arms (pricing, text message, text message + recommendation, and portal). T_{ijt} is an indicator variable equal to one for treated households in treatment arm j during peak hours (4–7PM) on treatment ("event") days and zero otherwise. $\mathbf{X}_{ijt}$ are peak hour and treatment day indicators, also interacted with treatment arm. ρ_i is a household fixed effect. ϕ_t represents time fixed effects. As discussed in the next section, we consider cases with and without control or comparison data. In the former case, we use hour-of-sample fixed effects. Without a comparison group, we cannot use these fixed effects because they would be perfectly collinear with the treatment variable (i.e., 4–7PM on event days for treated households, which, without a comparison group, is all households). In this case, we use hour-of-day and week-of-sample fixed effects to capture diurnal and seasonal patterns. The results are not sensitive to alternative approaches to time fixed effects. We cluster standard errors at the household level.

With a comparison group, this specification reflects a triple difference on the prediction error, analogous to BGK: treatment versus comparison group, treated day versus nontreated day, and treated hour versus nontreated hour. Without a comparison group, the first difference (treated versus comparison group) vanishes, meaning the specification is a double difference. In both cases, the prediction error can be interpreted as an additional difference (observed consumption versus predicted counterfactual).

4.3.2. Use of Control or Comparison Groups

To explore when the prediction error approach is valuable for estimating treatment effects, we consider three different cases:

1. **Experimental Control Group**: We use data from households with pricing or information interventions and households that were in the experimental (randomized) control group (249 households total: 193 treated, 56 control).
2. **Nonexperimental Comparison Group**: We continue to use households that were treated, but we replace the experimental (randomized) control group with households not part of the experiment but in the same neighborhood, creating a nonexperimental comparison group (334 households total: 193 treated, 141 comparison).
3. **No Control Data**: We only use treated households. In this case, the predicted counterfactual consumption for each household-hour effectively serves as its own "control" (193 households total, all treated).

In each case, we retrain our three ML models using the pre-RCT data and generate prediction errors using only the relevant households. For example, this methodology ensures

that the “no control” case corresponds to what would be practically implementable by a researcher who had access only to data on treated households and no control group, such as when a policy is simultaneously rolled out to an entire region or class of customers.

4.4. Criteria for a Successful Replication

Several criteria have been proposed to evaluate the quality of design-replication studies. Heckman et al. (1998) lay out four criteria of design that minimize bias from nonexperimental studies: (H1) treatment and comparison units have the same distribution of *unobserved* attributes, (H2) treatment and comparison units have the same distribution of *observed* attributes, (H3) outcomes and characteristics are measured in the same way for both groups, and (H4) treatment and comparison units are drawn from a common economic environment. Cook et al. (2008) lay out several additional criteria for successful design-replication studies, which we adopt and interpret within the context of our nonexperimental framework. The primary Cook et al. (2008) criteria are (C1) a randomly chosen control group and a nonexperimental comparison group are included, (C2) the experiment and observational study should estimate the same causal quantity (e.g., an average treatment effect), (C3) the contrast between control and comparison groups should not be correlated with other variables related to the study outcome, (C4) analysts of the experimental and nonexperimental data should be blind to each other’s results, and (C5) and (C6), the experiment and nonexperimental analysis, respectively, should each meet standard criteria of technical adequacy.

In our context, the nonexperimental analysis meets all but one criteria of Heckman et al. (1998) because the first criterion is fundamentally untestable, although we control for time-invariant characteristics of the households with fixed effects (H1). Our treatment and comparison households possess similar levels and trends in pretreatment energy use (see, e.g., Figure 1) and our empirical design includes a variety of fixed effects as controls (H2); the data for all groups is generated from the same data-collection method (H3); and each of our groups is drawn from households in a similar neighborhood in Austin, TX, that all share a willingness to participate as Pecan Street households (H4). Our design also satisfies all but the fourth criterion of Cook et al. (2008). We construct a nonexperimental sample using Pecan Street households that were not subject to the experiment (C1) and our nonexperimental analysis estimates an average treatment effect, similar to the experiment (C2). The choice of experimental control and nonexperimental comparison sample is not correlated with the measured outcomes (C3). We do not satisfy C4, as we have knowledge of the experimental treatment effect estimates. The experiment was conducted soundly, and we found no reason to doubt the validity of its treatment effects (C5). Last, we conduct our nonexperimental analysis following best practices in the ML and program evaluation literature, documenting any design choices (e.g., choice of tuning parameters) and evaluating alternative specifications (C6).

The criteria described thus far focus primarily on study *design*, but they are not instructive for how to compare the resulting estimates from the nonexperimental analysis to the randomized benchmark. Cook et al. (2008) note that some judgment is required on how comparable the results of the experimental and observational analysis are. We adopt the

following criterion explicated by Ferraro and Miranda (2017): the nonexperimental point estimate must fall within the 95% confidence interval of the experimental treatment effect.[5] We also require that our nonexperimental estimates achieve the correct sign and significance as the randomized benchmark (providing correct inference for both nonzero and null effects), thus ensuring that these estimates additionally replicate the qualitative policy relevance (or lack thereof) of estimated treatment effects.[6]

5. Results

We present our results in two formats. First, we present a table summarizing the detailed regression results for treatment effects from our preferred ML approach, XGBoost, which achieves the best predictive performance as measured by RMSE in CV.[7] After discussing the detailed results from this approach, we graphically summarize the results from the other ML approaches, which generally perform similarly. Detailed regression coefficients for each method are provided in the appendix.

In Table 4, we compare the results from the XGBoost prediction error regressions (in even-numbered columns) to those from the standard econometric difference-in-differences approach (in odd-numbered columns) applied to the electricity use data from the experiment. Column (1) shows the "ground-truth" estimates of the treatment effects of each intervention, reflecting the results from our standard difference-in-differences regression using the original experimental sample (i.e., the original randomized treatment and control groups). Consistent with the results in BGK, the pricing treatment produced a statistically significant reduction in peak consumption (about -0.36 kWh per hour, or -0.36 kW), but the other three treatments yielded no significant reduction.[8]

[5]We do not formally test the similarity of treatment effects between the experimental and nonexperimental estimates because we might fail to reject equivalence between the two if the nonexperimental treatment effect were estimated with a large standard error. It would, in principle, be quite easy to find statistical equality between experimental and nonexperimental treatment effects with a very large standard error on the nonexperimental effect.

[6]Ferraro and Miranda (2017) specify an additional criterion concerning the sensitivity of sample choice by incorporating a bootstrapping procedure to estimate nonexperimental treatment effects. We do not adopt this last criterion because resampling in our setting would require re-estimating our prediction models several hundred times, which, given current computational runtimes, would be measured in years.

[7]XGBoost's predictive accuracy is only slightly better than that of random forest. Both are superior to LASSO's RMSE, but we include LASSO because it is common in the literature.

[8]As noted, our sample is slightly different from the one used by BGK due to data updates by Pecan Street. In addition, to ease computational burden in training our ML models, we aggregate energy-use data to the hourly time step, whereas BGK used minute-level data. As a result of both of these differences, our treatment effect estimates are slightly different in magnitude from theirs (e.g., a pricing effect of -0.36 here versus -0.39 in their analysis). Nevertheless, each treatment's statistical significance (or lack thereof) is the same as in their study.

Column (3) shows the results of an alternative regression that replaces the randomized experimental control with a nonexperimental comparison group. The results are similar to the experimental results in column (1), indicating that the nonexperimental control group serves as a good substitute in this setting. Column (5) presents estimated coefficients based on a regression that includes only data from treated households, where treatment effects are identified solely based on differences between treatment periods and nontreatment periods for treated households, conditional on household and time fixed effects.[9] This column shows that when limiting the sample to just treated households, we do not recover the experimental results (i.e., column (1)); in fact, the results erroneously suggest that all the interventions increase energy use, because the interventions (such as critical-peak pricing) are triggered by periods of high demand. This structure means that without a comparison group, one would expect consumption for members of the treatment group to indeed be higher during periods of treatment than in other periods. The existence of a control or comparison group reveals that untreated households also have higher demand during these periods, allowing us to net out this effect. Without a comparison of some sort—whether an explicit control or comparison group or a predicted counterfactual—a comparison of treated and untreated periods would erroneously include this effect.

The results from the XGBoost prediction error regressions are displayed in columns (2), (4), and (6) of Table 4 for the different samples. They indicate that this ML approach provides a good replication of the ground-truth results in all cases—a significant negative effect of the pricing intervention (at $p < 0.001$) and no effect of any of the information interventions ($p > 0.10$)—using both the experimental control and nonexperimental comparison groups. In all cases, the estimated 95% confidence intervals overlap the "ground-truth" results. Finally, with no comparison group, the XGBoost approach is able to replicate the experimental results using only data on treated households (column (6)), whereas the simple difference approach fails to do so (column (5)). Tables of treatment effect results for the other ML approaches are included in the appendix (Tables A.1 and A.2).

We show the results for XGBoost only in Table 4, but we present results for other ML approaches in figures for each treatment to make comparisons across algorithms visually. Figure 3 shows the results for the estimated coefficient on the pricing treatment using our three ML approaches and for each of the three samples (with the experimental control group, the nonexperimental comparison group, and treatment group data only). The first dark blue dot shows the "ground-truth" treatment effect estimate, based on the a standard difference-in-differences regression on energy use (i.e., Table 4, column (1)), to which all the other estimates should be compared. The figure reveals that all the ML approaches we use replicate the experimental findings for the pricing treatment and that the ML approaches do not perform substantially better in replication than using a nonexperimental comparison group in a standard program evaluation framework.

[9] As previously discussed, all of these regressions include household fixed effects and time fixed effects (hour-of-sample in cases with a control group, or hour-of-day and week-of-sample for regressions without a control group).

While the ML approaches replicate the ground truth when the researcher has access to a reasonable comparison group, they do not appear to add much value to the standard difference-in-differences approach. By contrast, the rightmost panel of Figure 3 shows that ML approaches do add value in replicating experimental results when no comparison group is available. The simple differences regression erroneously suggests a significant positive effect of the pricing treatment on consumption (see Table 4, column (5)), but all the ML approaches replicate the ground truth even without any control or comparison group. This finding suggests that ML approaches can be highly useful in situations where data are only available for treated observations. Intuitively, a counterfactual prediction can serve as a pseudo-control to replace control households.

Figure 4 provides analogous illustrations of the coefficient on the Portal treatment with and without comparison groups and shows similarly that (a) the difference-in-differences approach with a nonexperimental control replicates the ground truth finding of no effect, (b) all the ML approaches with a control or nonexperimental comparison group replicate the ground truth, (c) the simple differences approach with no comparison group erroneously estimates a positive and significant treatment effect, and (d) all the ML approaches replicate the findings of no effect on consumption from access to the portal, as revealed by the original experiment. The corresponding figures for the remaining two treatment groups are similar and are shown in the appendix (Figures A.1 and A.2). Across the 36 treatment effect estimates (four treatments times three ML methods times three approaches to the control or comparison group), only two yield incorrect significance levels. These two both arise under the LASSO case (the worst-performing ML model) without a control or comparison group, where Text+Rec. and Portal treatments are both incorrectly found to be significant, at the 5 and 10% levels, respectively. At a 5% significance threshold, two false positives out of 36 is to be expected (2/36=0.055).

As seen in column (5) of Table 4, a naive regression using only treated households yields estimates of the wrong sign (positive) and significance. One reason for this is that the treatments, such as peak pricing or notifications, occur on particularly hot days when electricity demand is high due to air conditioning needs. The ML models account for this implicitly because they use data on temperature as an input and accordingly predict high counterfactual consumption. An alternative, non-ML approach to account for this would be to simply include flexible controls for temperature in the treatment effect estimation (equation (1)). The results of this approach, which includes a quadratic temperature polynomial or flexible 5°F bins of temperature, are shown in Table A.3. While this approach improves the estimates somewhat, it still generally fails to replicate experimental treatment effects. Even with flexible temperature controls, nearly all treatments, including those with true null effects, are incorrectly found to be significant at the 5% level or better. ML counterfactual prediction performs better across the board.

Table 4. XGBoost Prediction Error Difference-in-Differences and Regressions

	Exp. Control		Nonexp. Comparison		Treatment Only	
	Use	Pred. Error	Use	Pred. Error	Use	Pred. Error
	(1)	(2)	(3)	(4)	(5)	(6)
Pricing	-0.358***	-0.404***	-0.305***	-0.281***	0.203**	-0.361***
	(0.091)	(0.092)	(0.080)	(0.082)	(0.076)	(0.078)
Text Message + Rec	-0.026	-0.092	0.027	0.040	0.540***	-0.048
	(0.072)	(0.073)	(0.058)	(0.060)	(0.049)	(0.055)
Text Message	0.057	-0.083	0.111	0.016	0.618***	-0.048
	(0.080)	(0.078)	(0.068)	(0.066)	(0.060)	(0.061)
Portal	0.023	-0.100	0.076	0.026	0.580***	-0.030
	(0.075)	(0.072)	(0.062)	(0.061)	(0.054)	(0.051)
Observations	2,627,869	2,627,869	2,996,735	2,996,735	2,056,604	2,056,604
Households	249	249	334	334	193	193

Note: *** = p<0.001, ** = p<0.01

Notes: Columns (1) and (2) are estimated using the original experimental sample. Columns (3) and (4) are estimated using the experimental treatment group and a nonexperimental comparison group. Columns (5) and (6) are estimated using the treatment group only. In the "Use" columns, the dependent variable is electricity use at the household-hour level. In the "Pred. Error" regressions, the dependent variable is the prediction error, also at the household-hour level. Standard errors are clustered at the household level.

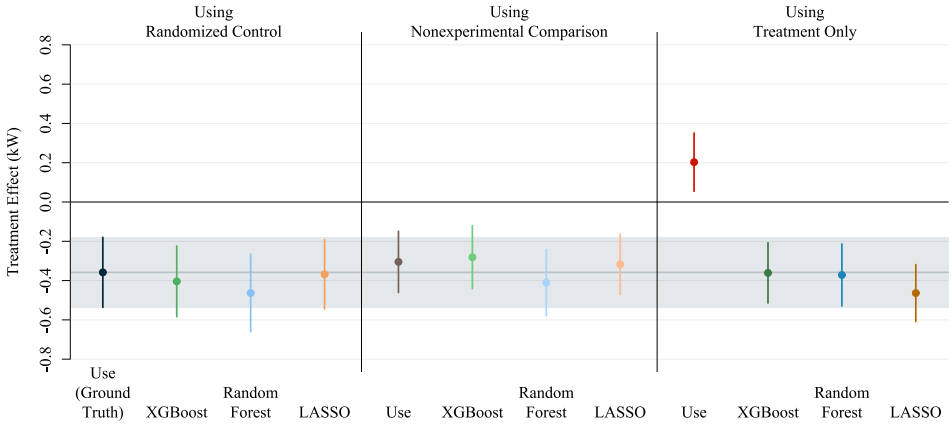

Figure 3. Critical-Peak Pricing Treatment Effect, by ML algorithm and sample

Notes: Labels along the bottom indicate the dependent variable used in the specification (equation 1). "Use" indicates the dependent variable is electricity use. For the other cases, the dependent variable is the prediction error from the specified ML model. Point estimates and 95% confidence intervals are presented for each treatment effect estimate. The shaded area represents 95% confidence interval on ground-truth estimate.

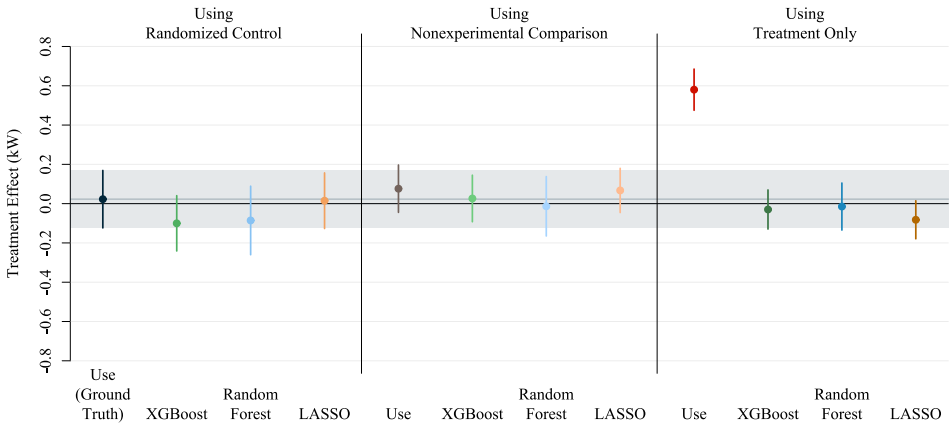

Figure 4. Portal Treatment Effect, by ML algorithm and sample

Notes: Labels along the bottom indicate the dependent variable used in the specification (equation 1). "Use" indicates the dependent variable is electricity use. For the other cases, the dependent variable is the prediction error from the specified ML model. Point estimates and 95% confidence intervals are presented for each treatment effect estimate. The shaded area represents 95% confidence interval on ground-truth estimate.

6. Discussion

Our results provide strong evidence that ML prediction algorithms can replicate experimental benchmarks in several settings, for both null and nonzero treatment effects. In the most surprising case, we find that when we only use information on treated households, we can generate a predicted counterfactual that, when used in a prediction error regression, works as well as the randomized control group. This finding is powerful because all realms of economics have many instances in which policy changes occur for all units at the same time. Our finding suggests that ML counterfactual prediction algorithms provide some optimism for recovering causal effects of those policies even without an observational comparison group. Synthetic control methods are often used in scenarios like this, although our findings suggest that treatment effects can be recovered even without "donor" observations, and so it may be easier to meet the data requirements for implementing ML prediction algorithms.

In the case in which we have access to a group of nonexperimental units (i.e., units that were not included in the randomized experiment), we similarly find that we can replicate the experimental benchmarks using our ML-assisted prediction algorithms. Although this appears to be valuable result, we find that a simpler, more standard difference-in-differences design also recovers the experimental treatment effects. Therefore, if one has access to a reasonable comparison group, the ML-prediction-based methods do not improve upon simpler, more common, and easier-to-implement program evaluation tools. Fortunately, both ML and standard difference-in-differences methods produce similar

results, which would allow a researcher to be confident in their nonexperimental results if both frameworks were used. Our nonexperimental comparison group satisfies standard assumptions for difference-in-differences (i.e., common pretrends), and our experimental treatment turns on at the same time for all units. If a researcher has a nonexperimental comparison group with different pretrends and poor balance among observables, then ML prediction methods (or other appproaches to preprocessing a control group, such as covariate matching) may improve their ability to estimate causal effects.

Interestingly, because each of our ML algorithms generally replicate treatment effects in each scenario, we are unable to rank the algorithms' ultimate effectiveness in estimating causal effects. We rely on standard CV and goodness-of-fit (i.e., RMSE) to provide guidance on choosing the algorithm with the best predictive qualities. Although we find that XGBoost performs better than random forest, which performs better than LASSO, all three reproduce the experimental benchmark, suggesting that marginal improvements in model fit are unlikely to be decisive for estimating causal effects.

Our findings lead to a fundamental question: in what settings are these results most likely to be applicable? Or, put another way, what features of a research design are most likely to benefit from ML counterfactual prediction algorithms? We explore these questions in our setting and speculate on the general features that will be most useful for researchers aiming to import our findings to other settings.

One aspect of this question revolves around what type of threats to identification the researcher is concerned about. In our setting, treatment is plausibly exogenous, conditional on household characteristics (or household fixed effects) and weather (or granular time fixed effects). The primary threat to exogeneity of treatment is that hot, summer days tend to be days when treatment kicks on to curtail consumption. Conditioning on weather or time fixed effects permits a plausible interpretation of exogeneity, and so our results can speak primarily to similar settings.

A common, alternative threat to identification is selection bias. It is important to highlight what kinds of selection bias our approach can and cannot mitigate. One type of selection concern is that households who sign up to participate in an experiment (whether they end up randomized into a treatment or control group) may be more likely to respond to the treatment, relative to the general population. This threatens the external validity of the findings, even if an internally valid treatment effect for recruited households can be reliably estimated. Our approach cannot solve this problem, which fundamentally stems from the design of a policy's recruitment process. Another type of selection bias is where households can select into treatment relative to nontreatment. This concern would bias comparisons between treated and untreated groups, threatening internal validity. The ML counterfactual approach can mitigate this problem because it allows for dropping the untreated group entirely, substituting ML counterfactual predictions.

Therefore, we anticipate that ML counterfactual approaches have the most promise to improve upon standard approaches when treatment is conditionally exogenous or selection into treatment is a concern, and in the latter case, the two methods are likely to produce different estimates. This distinction could explain why our panel and ML

counterfactual results are comparable whereas, for example, the results from Burlig et al. (2020) for energy-efficiency upgrades in schools diverge between the two approaches. As a result, additional research in different settings, testing different threats to internal validity, is important to conduct before extrapolating our findings to settings in which selection into treatment is the primary concern.

The study of electricity demand has seen an increase in the application of ML algorithms primarily because (a) researchers can easily access high-frequency consumption data and (b) the data-generating process is predictable (e.g., with daily and seasonal patterns). Within this context, one question is how important granular data is (Ghanem and Smith, 2021). Our experiment does not allow for answering this question, because the treatment period is a three-hour window during event days. Thus, we cannot aggregate our data, say, to the weekly or monthly level to evaluate how well the algorithms perform. The data can only be as aggregated as the treatment itself, so we must limit our focus to subdaily changes in energy consumption.

A second question is what *type* of data researchers need access to in order to generate causal estimates. How far would they get if they were limited in the number of explanatory variables to include in a study? We explore this question by evaluating variable importance plots from our XGBoost algorithm, presented in Figure A.4. These plots iteratively exclude each variable (and hence its associated interactions) from the ML model and compute the resulting deterioration in predictive power. Across each sample, the most important variable is the square footage of the home, followed by the temperature, apparent temperature, capacity of solar PV installed on the home, and home age. The remaining top 10 characteristics include week or evening hour indicators and humidity. Notably, the structural characteristics of the home are time invariant and can be easily replaced by household fixed effects. Some combination of the weather variables is almost always included in electricity demand specifications based on theory and intuition and these data are easily accessible from public databases or controlled for using temporal fixed effects. Recall that we replicated the experimental treatment effects in a standard difference-in-differences design with fixed effects, which requires little information about household characteristics. As a result, it appears that little information about households is necessary for recovering treatment effects with nonexperimental data.

To explore this question a step further, we re-estimate our primary XGBoost counterfactual prediction models several times, each time excluding certain categories of data. We then re-estimate treatment effects using prediction errors from each restricted model. This is meant to emulate potential real-world data availability constraints that could be faced by researchers. These exclusions are summarized as follows:

1. Without household characteristic variables, using household fixed effects instead
2. Without weather variables
3. Without variables on weather or household characteristics, using household fixed effects (and time variables)
4. Using only time variables (time fixed effects and a weekly time trend)

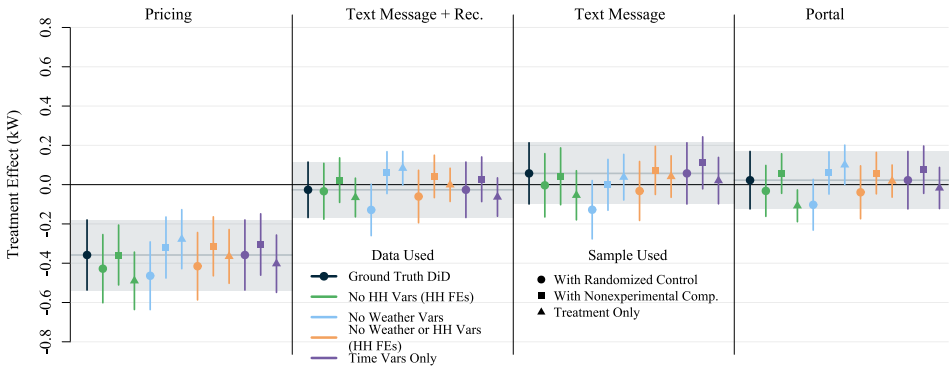

Figure 5. Prediction Error-Based Treatment Effect Estimates with Limited Data in ML Training

Notes: "HH FEs" = Household Fixed Effects. Point estimates and 95% confidence intervals are presented for each treatment effect estimate. Shaded areas represent 95% confidence intervals on ground-truth estimate.

In each case, we rerun the prediction using the three samples in our main analysis: with the randomized control group, the nonexperimental comparison group, and treated households only. The results are shown in Figure 5, alongside the experimental benchmark. Even with limited data, the ML prediction models perform well. Across 48 treatment effect estimates (four covariate constraints by three samples by four treatments), at the 5% significance level, we find no false negatives and only one false positive. At the 10% level, another four of the 48 estimates yield false positives, all of which appear when we drop weather data. The full regression results are shown in appendix Tables A.4-A.7. This set of results indicates that the ML counterfactual predictions can do quite well even with little data in the context of electricity demand. This result is likely because electricity demand shows strong diurnal and seasonal patterns, making it an attractive application for counterfactual prediction approaches. It is difficult to make strong claims about the full set of necessary conditions for ML counterfactual prediction to work well, other than to say that that processes that can be predicted well with available data are good candidates for this approach.

7. Conclusion

The importance of sound program evaluation to policymakers and society at large cannot be overstated. Our work yields valuable policy insights and methodological contributions at the intersection of economics, energy policy, and ML that can inform that policy evaluation. While RCTs have played a growing role in policy evaluation, such approaches are not always feasible. At the same time, ML approaches are becoming more prevalent in such research. No formal comparisons exist of treatment effects estimated via counterfactual

prediction methods that use ML and those estimated from randomized experiments. Our work begins to fill that void. The results of our research help define best practices for evaluations of existing programs that may not have been implemented with randomized trials and can guide future analysis of consumer-focused energy policies.

When RCTs are not possible, identifying causal effects involves picking the right tools for the setting. Our results suggest that the contributions of ML methods to program evaluation and RCT replication depend importantly on the setting and on whether one has access to a suitable comparison group. When comparison group data are available, then familiar difference-in-differences approaches perform comparably to an RCT, and thus may be a preferred approach. On the other hand, when data are only available for treated subjects, as may often be the case, then ML can provide a superior alternative to analysis based on simple differences in outcomes before and after treatment. Sometimes the right tool is a machine.

References

Abadie, A., A. Diamond, and J. Hainmueller. 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A. and J. Gardeazabal. 2003. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132.

Abrell, J., M. Kosch, and S. Rausch. 2019. How Effective Is Carbon Pricing? Emissions and Cost Impacts of the UK Carbon Tax. 19(317).
URL `https://ssrn.com/abstract=3372388`

Allcott, H. and T. Rogers. 2014. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *The American Economic Review*, 104(10):3003–37.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. 2021. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 1–41.

Athey, S. and G. Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Banerjee, E. D., Abhijit V. 2009. The experimental approach to development economics. *Annual Review of Economics*, 1:151–78.

Blonz, J., K. Palmer, C. J. Wichman, and D. C. Wietelman. 2021. Smart thermostats, automation, and time-varying pricing. Working paper, Resources for the Future.

Blonz, J. A. 2020. Making the best of the second-best: Welfare consequences of time-varying electricity prices. Technical report, Working Paper.

Borenstein, S. 2005. The long-run efficiency of real-time electricity pricing. *The Energy Journal*, 26:93–116.
URL http://dx.doi.org/10.2307/41319500

Borenstein, S. 2007. Wealth Transfers Among Large Customers from Implementing Real-Time Retail Electricity Pricing. *The Energy Journal*, 28(2):131–150.
URL https://ideas.repec.org/a/aen/journl/2007v28-02-a06.html

Borenstein, S. and J. Bushnell. 2019. Do two electricity pricing wrongs make a right? cost recovery, externalities and efficiency. Technical report, Energy Institute Working Paper, 294R.

Borenstein, S. and S. Holland. 2005. On the efficiency of competitive electricity markets with time-invariant retail prices. *RAND Journal of Economics*, 36(3):469–494.

Burkhardt, J., K. Gillingham, and P. K. Kopalle. 2019. Experimental evidence on the effect of information and pricing on residential electricity consumption. Technical report, National Bureau of Economic Research.

Burlig, F., C. Knittel, D. Rapson, M. Reguant, and C. Wolfram. 2020. Machine learning from schools about energy efficiency. *Journal of the Association of Environmental and Resource Economists*, 7(6):1181–1217.

Christensen, P., P. Francisco, E. Myers, and M. Souza. 2021. Decomposing the wedge between projected and realized returns in energy efficiency programs. Technical report, E2e Working Paper, 046.

Cook, T. D., W. R. Shadish, and V. C. Wong. 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4):724–750.

Cunningham, B., J. LaRiviere, and C. J. Wichman. 2021. Clustered into control: Heterogeneous causal impacts of water infrastructure failure. *Economic Inquiry*, 59(3):1417–1439.

Davis, R. J., J. S. Holladay, and C. Sims. 2020. Coal-Fired Power Plant Retirements in the US.

Dehejia, R. 2005. Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics*, 125(1–2):355–364.

Dehejia, R. H. and S. Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Dehejia, R. H. and S. Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161.

Dueñas, M., V. Ortiz, M. Riccaboni, and F. Serti. 2021. Assessing the Impact of COVID-19 on Trade: A Machine Learning Counterfactual Analysis.

Fabra, N., D. Rapson, M. Reguant, and J. Wang. 2021. Estimating the elasticity to real-time pricing: Evidence from the Spanish electricity market. *AEA Papers and Proceedings*, 111:425–29.
URL http://dx.doi.org/10.1257/pandp.20211007

Ferraro, P. J. and J. J. Miranda. 2017. Panel data designs and estimators as substitutes for randomized controlled trials in the evaluation of public programs. *Journal of the Association of Environmental and Resource Economists*, 4(1):281–317.

Ferraro, P. J. and M. K. Price. 2013. Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1):64–73.

Fowlie, M., C. Wolfram, C. A. Spurlock, A. Todd, P. Baylis, and P. Cappers. 2020. Default effects and follow-on behavior: evidence from an electricity pricing program. Technical report, Energy Institute Working Paper, 280R.

Ghanem, D. and A. Smith. 2021. What are the benefits of high-frequency data for fixed effects panel models? *Journal of the Association of Environmental and Resource Economists*, 8(2):199–234.

Gillan, J. M. 2018. Essays in energy and environmental economics. Technical report, University of California Berkeley.

Harding, M. and C. Lamarche. 2016. Empowering consumers through data and smart technology: Experimental evidence on the consequences of time-of-use electricity pricing policies. *Journal of Policy Analysis and Management*, 35(4):906–931.

Harding, M. and S. Sexton. 2017. Household response to time-varying electricity prices. *Annual Review of Resource Economics*, 9:337–359.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy. 2016. Counterfactual prediction with deep instrumental variables networks. *arXiv preprint arXiv:1612.09596*.

Heckman, J. J., H. Ichimura, and P. Todd. 1998. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294.

Imbens, G. W. and J. M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Jessoe, K. and D. Rapson. 2014. Knowledge is (less) power: Experimental evidence from residential energy use. *American Economic Review*, 104(4):1417–38.
URL http://dx.doi.org/10.1257/aer.104.4.1417

Jessoe, K. and D. Rapson. 2015. Commercial and industrial demand response under mandatory time-of-use electricity pricing. *The Journal of Industrial Economics*, 63(3):397–421.
URL http://dx.doi.org/https://doi.org/10.1111/joie.12082

Jones, D., D. Molitor, and J. Reif. 2019. What do workplace wellness programs do? evidence from the Illinois workplace wellness study. *The Quarterly Journal of Economics*, 134(4):1747–1791.

Joskow, P. L. and C. D. Wolfram. 2012. Dynamic pricing of electricity. *American Economic Review*, 102(3):381–85.
URL http://dx.doi.org/10.1257/aer.102.3.381

Knittel, C. R. and S. Stolper. 2019. Using machine learning to target treatment: The case of household energy use. Technical report, National Bureau of Economic Research.

Krueger, A. B. 1999. Experimental estimates of education production functions. *The Quarterly Journal of Economics*, 114(2):497–532.

LaLonde, R. J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604–620.

Liberman, A., C. Neilson, L. Opazo, and S. Zimmerman. 2018. The equilibrium effects of information deletion: Evidence from consumer credit markets. Working Paper 25097, National Bureau of Economic Research.
URL http://dx.doi.org/10.3386/w25097

List, I. R., John A. 2011. Field experiments in labor economics. *Handbook of Labor Economics*, 4 Part A:103–228.

Prest, B. C. 2020. Peaking interest: How awareness drives the effectiveness of time-of-use electricity pricing. *Journal of the Association of Environmental and Resource Economists*, 7(1):103–143.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Sallee, J. 2014. Rational inattention and energy efficency. *Journal of Law and Economics*, 57(3):781–820.

Smith, J. A. and P. E. Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1–2):305–353.

Souza, M. 2019. Predictive counterfactuals for event studies with staggered adoption: Recovering heterogeneous effects from a residential energy efficiency program. Technical report, Working paper.

St. Clair, T., T. D. Cook, and K. Hallberg. 2014. Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*, 35(3):311–327.

Varian, H. R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Varian, H. R. 2016. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315.

Wichman, C. J. and P. J. Ferraro. 2017. A cautionary tale on using panel data estimators to measure program impacts. *Economics Letters*, 151:82–90.

Wing, C. and T. D. Cook. 2013. Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management*, 32(4):853–877.

Wolak, F. A. 2011. Do residential customers respond to hourly prices? evidence from a dynamic pricing experiment. *American Economic Review*, 101(3):83–87. URL http://dx.doi.org/10.1257/aer.101.3.83

A Supplemental Results

A.1 Treatment Effect Replication Figures

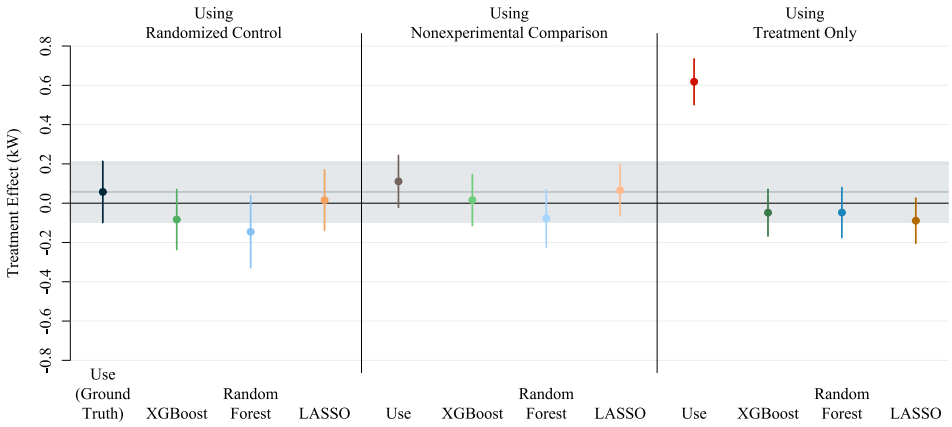

Figure A.1. Text Message Treatment Effect, by ML algorithm and sample

Notes: Labels along the bottom indicate the dependent variable used in the specification (equation 1). "Use" indicates the dependent variable is electricity use. For the other cases, the dependent variable is the prediction error from the specified ML model. Point estimates and 95% confidence intervals are presented for each treatment effect estimate. The shaded area represents the 95% confidence interval on the ground-truth estimate.

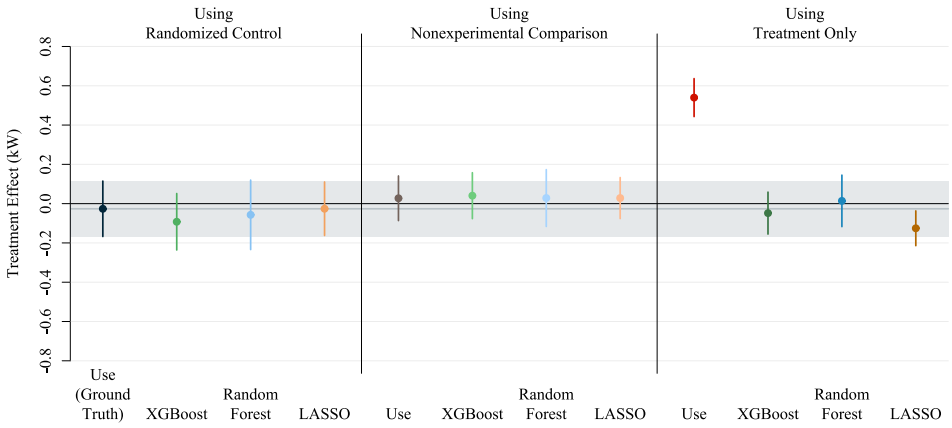

Figure A.2. Text + Recommendation Treatment Effect, by ML algorithm and sample

Notes: Labels along the bottom indicate the dependent variable used in the specification (equation 1). "Use" indicates the dependent variable is electricity use. For the other cases, the dependent variable is the prediction error from the specified ML model. Point estimates and 95% confidence intervals are presented for each treatment effect estimate. The shaded area represents the 95% confidence interval on the ground-truth estimate.

A.2 Treatment Effect Replication Regression Tables

Table A.1. Random Forest Prediction Error Regressions

	Exp. Control		Nonexp. Comparison		Treatment Only	
	Use	Pred. Error	Use	Pred. Error	Use	Pred. Error
	(1)	(2)	(3)	(4)	(5)	(6)
Pricing	-0.358***	-0.454***	-0.305***	-0.411***	0.203**	-0.379***
	(0.091)	(0.101)	(0.080)	(0.086)	(0.076)	(0.081)
Text Message + Rec	-0.026	-0.057	0.027	0.029	0.540***	0.014
	(0.072)	(0.090)	(0.058)	(0.074)	(0.049)	(0.067)
Text Message	0.057	-0.146	0.111	-0.078	0.618***	-0.047
	(0.080)	(0.094)	(0.068)	(0.075)	(0.060)	(0.065)
Portal	0.023	-0.086	0.076	-0.014	0.580***	-0.015
	(0.075)	(0.089)	(0.062)	(0.077)	(0.054)	(0.061)
Observations	2,627,869	2,627,869	2,996,735	2,996,735	2,056,604	2,056,604
Households	249	249	334	334	193	193

Note: *** = p<0.001, ** = p<0.01

Table A.2. LASSO Prediction Error Regressions

	Exp. Control		Nonexp. Comparison		Treatment Only	
	Use	Pred. Error	Use	Pred. Error	Use	Pred. Error
	(1)	(2)	(3)	(4)	(5)	(6)
Pricing	-0.358***	-0.368***	-0.305***	-0.317***	0.203**	-0.464***
	(0.091)	(0.091)	(0.080)	(0.079)	(0.076)	(0.074)
Text Message + Rec	-0.026	-0.026	0.027	0.028	0.540***	-0.125**
	(0.072)	(0.070)	(0.058)	(0.053)	(0.049)	(0.045)
Text Message	0.057	0.015	0.111	0.066	0.618***	-0.089
	(0.080)	(0.079)	(0.068)	(0.066)	(0.060)	(0.059)
Portal	0.023	0.015	0.076	0.067	0.580***	-0.082+
	(0.075)	(0.072)	(0.062)	(0.058)	(0.054)	(0.049)
Observations	2,627,869	2,627,869	2,996,735	2,996,735	2,056,604	2,056,604
Households	249	249	334	334	193	193

Note: *** = p<0.001, ** = p<0.01, * = p<0.05, + = p<0.10

Table A.3. Treatment-Only Regressions with Temperature Controls

	Use - Treatment Only		
Temperature Control:	None	Quadratic	5°F Bins
	(1)	(2)	(3)
Pricing	0.203**	-0.231**	-0.218**
	(0.076)	(0.073)	(0.071)
Text Message + Rec.	0.540***	0.088+	0.106**
	(0.049)	(0.048)	(0.046)
Text Message	0.618***	0.169**	0.188***
	(0.06)	(0.057)	(0.056)
Portal	0.580***	0.138**	0.155**
	(0.054)	(0.053)	(0.052)
Temperature (°C)		-0.1192***	
		(0.005)	
Temperature^2		0.0011***	
		(0.00004)	
Observations	2,056,604	2,056,604	2,056,604
Households	193	193	193

Note: *** = p<0.001, ** = p<0.01, * = p<0.05, + = p<0.10

A.3 Actual and XGBoost Predicted Counterfactual Consumption Profiles

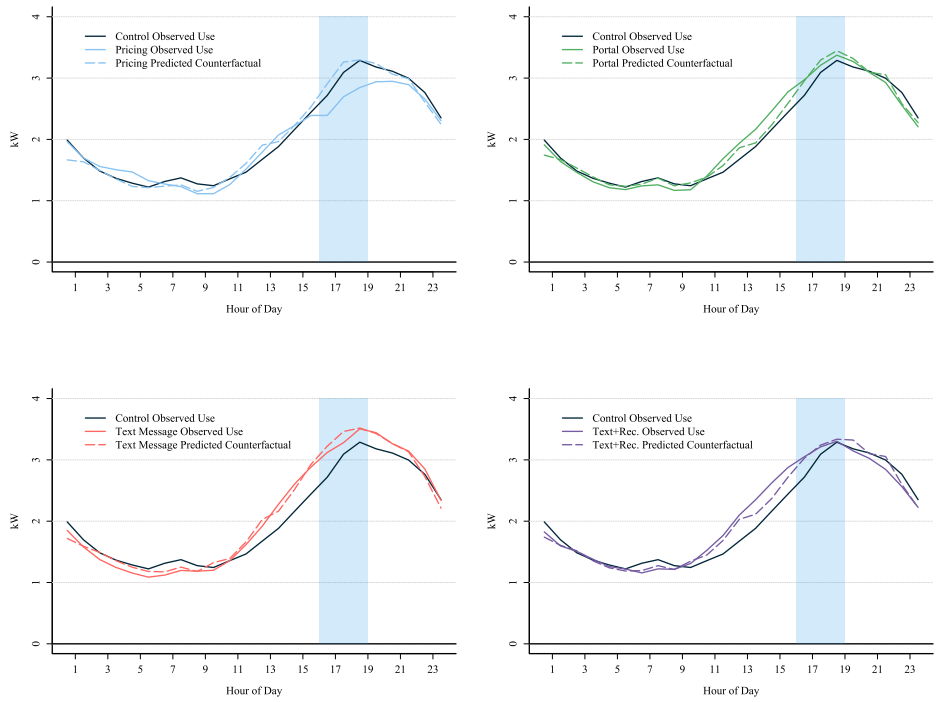

Figure A.3. Event Day Average Consumption Profile by Group, Actual and XGBoost Predicted Counterfactual (Using Only Treatment Data)

Note: The blue region reflects peak hours (4–7pm). As in BGK, the profiles are plotted after conditioning on household fixed effects and recentering on the grand mean to highlight the source of identification under a difference-in-differences regression.

A.4 XGBoost Variable Importance Plot

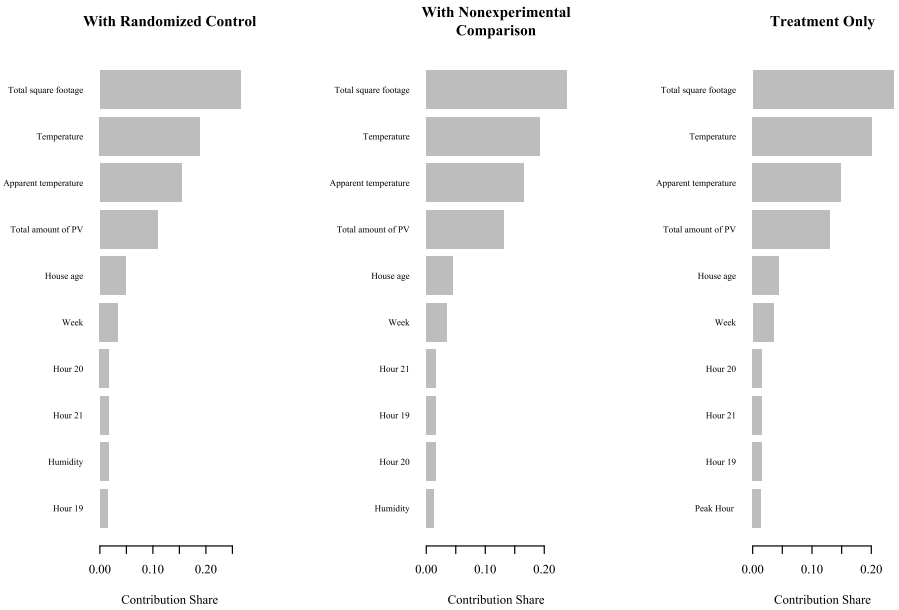

Figure A.4. Relative Contributions of Variables to Predicting Consumption, Top 10 Most Important Variables

Note: Figures show the top 10 most important variables in predicting consumption under each dataset. Values indicate each variable's relative contribution (i.e., the sum across variables is 100%). In each case, the top 10 variables account for 86-88% of predictive power.

Table A.4. XGBoost Prediction Error Difference-in-Differences and Regressions, Use Household Fixed Effects in Training

	Exp. Control		Nonexp. Comparison		Treatment Only	
	Use	Pred. Error	Use	Pred. Error	Use	Pred. Error
	(1)	(2)	(3)	(4)	(5)	(6)
Pricing	-0.358***	-0.428***	-0.305***	-0.358***	0.203**	-0.490***
	(0.091)	(0.089)	(0.080)	(0.078)	(0.076)	(0.075)
Text Message + Rec	-0.026	-0.034	0.027	0.023	0.540***	-0.065
	(0.072)	(0.073)	(0.058)	(0.058)	(0.049)	(0.050)
Text	0.057	-0.003	0.111	0.042	0.618***	-0.054
	(0.080)	(0.082)	(0.068)	(0.074)	(0.060)	(0.064)
Portal	0.023	-0.032	0.076	0.056	0.580***	-0.108**
	(0.075)	(0.066)	(0.062)	(0.051)	(0.054)	(0.041)
Observations	2,627,869	2,627,869	2,996,735	2,996,735	2,056,604	2,056,604
Households	249	249	334	334	193	193

Note: *** = $p<0.001$, ** = $p<0.01$. Household fixed effects are used in place of household characteristics (e.g., square footage).

Table A.5. XGBoost Prediction Error Difference-in-Differences and Regressions, No Weather Data in Training

	Exp. Control		Nonexp. Comparison		Treatment Only	
	Use	Pred. Error	Use	Pred. Error	Use	Pred. Error
	(1)	(2)	(3)	(4)	(5)	(6)
Pricing	-0.358***	-0.464***	-0.305***	-0.320***	0.203**	-0.278***
	(0.091)	(0.088)	(0.080)	(0.079)	(0.076)	(0.077)
Text Message + Rec	-0.026	-0.129+	0.027	0.061	0.540***	0.083+
	(0.072)	(0.067)	(0.058)	(0.054)	(0.049)	(0.044)
Text	0.057	-0.128+	0.111	-0.001	0.618***	0.038
	(0.080)	(0.076)	(0.068)	(0.066)	(0.060)	(0.059)
Portal	0.023	-0.102	0.076	0.059	0.580***	0.101+
	(0.075)	(0.066)	(0.062)	(0.055)	(0.054)	(0.051)
Observations	2,627,869	2,627,869	2,996,735	2,996,735	2,056,604	2,056,604
Households	249	249	334	334	193	193

Note: *** = $p<0.001$, ** = $p<0.01$, + = $p<0.10$

Table A.6. XGBoost Prediction Error Difference-in-Differences and Regressions, Only Household FEs and Time Variables in Training

	Exp. Control		Nonexp. Comparison		Treatment Only	
	Use	Pred. Error	Use	Pred. Error	Use	Pred. Error
	(1)	(2)	(3)	(4)	(5)	(6)
Pricing	-0.358***	-0.415***	-0.305***	-0.314***	0.203**	-0.365***
	(0.091)	(0.088)	(0.08)	(0.077)	(0.076)	(0.070)
Text Message + Rec.	-0.026	-0.060	0.027	0.042	0.540***	-0.001
	(0.072)	(0.068)	(0.058)	(0.055)	(0.049)	(0.043)
Text	0.057	-0.032	0.111	0.073	0.618***	0.041
	(0.080)	(0.077)	(0.068)	(0.063)	(0.06)	(0.054)
Portal	0.023	-0.039	0.076	0.059	0.580***	0.018
	(0.075)	(0.069)	(0.062)	(0.054)	(0.054)	(0.042)
Observations	2,627,869	2,627,869	2,996,735	2,996,735	2,056,604	2,056,604
Households	249	249	334	334	193	193

Note: *** = p<0.001, ** = p<0.01. Household fixed effects are used in place of household characteristics (e.g., square footage).

Table A.7. XGBoost Prediction Error Difference-in-Differences and Regressions, Only Time Variables in Training

	Exp. Control		Nonexp. Comparison		Treatment Only	
	Use	Pred. Error	Use	Pred. Error	Use	Pred. Error
	(1)	(2)	(3)	(4)	(5)	(6)
Pricing	-0.358***	-0.358***	-0.305***	-0.305***	0.203**	-0.403***
	(0.091)	(0.091)	(0.080)	(0.080)	(0.076)	(0.075)
Text Message + Rec.	-0.026	-0.026	0.027	0.027	0.540***	-0.063
	(0.072)	(0.072)	(0.058)	(0.058)	(0.049)	(0.050)
Text	0.057	0.057	0.111	0.111	0.618***	0.020
	(0.080)	(0.080)	(0.068)	(0.068)	(0.060)	(0.060)
Portal	0.023	0.023	0.076	0.076	0.580***	-0.017
	(0.075)	(0.075)	(0.062)	(0.062)	(0.054)	(0.054)
Observations	2,627,869	2,627,869	2,996,735	2,996,735	2,056,604	2,056,604
Households	249	249	334	334	193	193

Note: *** = p<0.001, ** = p<0.01. Only temporal variables used in training: hour of day, day of week, week of year, month of year, weekend, and peak hour.