

Energy Efficiency Can Deliver for Climate Policy: Evidence from Machine Learning-Based Targeting

Peter Christensen,^{*} Paul Francisco,^{*} Erica Myers,[†]
Hansen Shao,^{*} and Mateus Souza[‡]

November 4, 2021

Abstract

Building energy efficiency has been a cornerstone of greenhouse gas mitigation strategies for decades. However, impact evaluations have revealed that energy savings typically fall short of engineering model forecasts that currently guide funding decisions. This creates a resource allocation problem that impedes progress on climate change. Using data from the largest U.S. energy efficiency program, we demonstrate that machine learning combined with energy billing data provides more accurate ex-ante predictions of retrofit outcomes than the engineering approach. Targeting high-return interventions based on these predictions dramatically increases net social benefits, from \$0.93 to \$1.23 per dollar invested.

Key words: energy efficiency, machine learning, cost-effectiveness, targeting

JEL Classification: C45, C53, Q48, Q56

^{*}University of Illinois at Urbana-Champaign; [†]University of Calgary; [‡]Universidad Carlos III de Madrid. Emails: Christensen (pchrist@illinois.edu), Francisco (pwf@illinois.edu), Myers (erica.myers@ucalgary.ca), Shao (hshao4@illinois.edu), and Souza (mateus.nogueira@uc3m.es). We thank Mick Prince, Chad Wolfe, the PRISM Climate Group, and student assistants in the University of Illinois Big Data and Environmental Economics and Policy (BDEEP) Group at the National Center for Supercomputing Applications for assistance with data and computational resources related to this project. We acknowledge excellent feedback and comments from Tatyana Deryugina, Arik Levinson, Lucija Muehlenbachs, Stefan Staubli, and Bruce Tonn. We acknowledge generous support from the Alfred P. Sloan foundation and the Illinois Department of Commerce and Economic Opportunity's Illinois Home Weatherization Assistance Program. Souza gratefully recognizes the support from the European Research Council (ERC, under the European Union's Horizon 2020 research and innovation programme, Grant Agreement No. 772331).

Engineering models often project that residential energy efficiency is one of the most cost-effective strategies to reduce greenhouse gas emissions (International Energy Agency, 2019; McKinsey & Co, 2009). Consequently, it has become a key component of climate and energy policy worldwide, with billions of dollars invested each year to unlock its potential (European Parliament, 2012; EEA, 2018; ARB, 2017; Barbose et al., 2013). Energy efficiency projects are also often considered to be environmentally responsible stimulus for addressing an economy weakened by COVID-19.¹ However, savings have been found to typically fall short of projections.² As a result, some economists have begun to caution against prioritizing energy efficiency when applied to climate and economic stimulus policies (Fowlie, 2020; Auffhammer, 2021), as alternative approaches may be more cost-effective for these objectives (Gillingham and Stock, 2018).

The goal of this paper is to contribute to an ongoing debate on whether energy efficiency programs have a role for climate policy. In a recent study, Christensen et al. (2021) find significant heterogeneity in benefits, demonstrating that many energy efficiency projects are cost-effective and that modeling bias explains much of the wedge between realized and projected savings. An implication is that better modeling may increase cost-effectiveness of these programs by improving assignment of funds across projects. While economists and internal program evaluators have produced consistent evidence of upward bias in projections (e.g. Fowlie, Greenstone, and Wolfram, 2018; Allcott and Greenstone, 2017; Berry and Gettings, 1998; Dalhoff, 1997; Sharp, 1994), researchers have not developed or tested the effects of more accurate prediction strategies.

The present study asks whether better ex-ante predictions can substantially increase the cost-effectiveness of residential energy efficiency retrofit programs through improved allocation of funds. Currently, the vast majority of energy efficiency programs use engi-

¹For example, President Biden’s “American Jobs Plan” proposes to invest “\$213 billion to produce, preserve, and retrofit more than two million affordable and sustainable places to live” (The White House, 2021). See also European Commission (2020) and Hepburn et al. (2020).

²This has been shown across a wide range of energy efficiency initiatives, including home retrofit programs (Fowlie, Greenstone, and Wolfram, 2018; Allcott and Greenstone, 2017; Zivin and Novan, 2016; Berry and Gettings, 1998; Dalhoff, 1997; Sharp, 1994), appliance rebate programs (Houde and Aldy, 2014; Davis, Fuchs, and Gertler, 2014), and in efficient new construction (Levinson, 2016; Bruegge, Deryugina, and Myers, 2019; Davis, Martinez, and Taboada, 2020).

neering models to predict savings and determine which retrofits should be done.³ However, projecting the impacts of multiple retrofits in individual diverse buildings presents major modeling challenges, such as accounting for heat and air exchanges between a building and the surrounding environment, and interactions between different retrofits. The exercise is further complicated by the fact that current modeling techniques often do not incorporate homes' energy consumption data, which could improve predictions.

We aim to improve on this approach by using machine learning (ML) with observed pre- and post-retrofit consumption from billing data. ML is well suited for this type of prediction exercise because energy consumption is a function of many complex, high-order interactions among the various observable aspects of a home and household. Our analysis uses data obtained for the Illinois Home Weatherization Assistance Program (IHWAP), which is the Illinois implementation of the U.S. Department of Energy's Weatherization Assistance Program (WAP). WAP's focus is on reducing energy costs for low-income households while also maintaining health and safety. It was started in the 1970's and carbon abatement is not one of the program's original mandates.⁴ Given the potential to deliver low-cost reductions, however, residential energy efficiency has increasingly become a part of climate policy platforms. Because many retrofit initiatives (including IHWAP) rely on a common set of accepted structural engineering equations to make their predictions (Edwards et al., 2013; Sentech, 2010), our findings also have implications for determining the cost-effectiveness of these retrofits as a carbon abatement strategy. We focus on energy-related benefits, but the method we propose here could be extended to capture the benefits from effects on other measurable outcomes.

The first step of our analysis is to determine whether machine learning (ML) combined with pre-retrofit billing data can improve *ex-ante* prediction accuracy in energy efficiency programs. The *ex-ante* prediction framework is a form of out-of-sample prediction for which machine learning algorithms are optimized. It differs from the use of

³These models are based on equations describing the physical relationships between energy consumption, weather and home characteristics. They also incorporate demographic information, modelling effects of characteristics such as the number of occupants on energy consumption.

⁴The program's enabling statute mentions energy security, health and safety, and a focus on low-income households, without any reference to carbon abatement or climate goals (US Code, 1964).

ML in ex-post program evaluations, as initially introduced by Burlig et al. (2020) to the literature on energy efficiency programs. The researcher’s goal is to mimic the role of a program implementer who is trying to predict the magnitude of treatment effects prior to an intervention. In an energy efficiency program, the implementer would build her model using data on home characteristics, observed weather, consumption, and upgrades performed for homes that have previously been retrofitted. However, she needs to predict outcomes for new homes based on information available only prior to the retrofits. We use a neural net algorithm with nested-cross validation to predict each home’s energy consumption under three conditions: 1) pre-retrofit; 2) post-retrofit; 3) post-retrofit counterfactual, i.e. what consumption would have been in absence of the upgrades.⁵ Within this ex-ante framework and data-rich setting, we find that the ML approach accurately predicts household energy usage. On average, the predictions are statistically indistinguishable from true energy consumption both pre- and post-retrofit.

We then turn to the primary goal of the paper, which is to test whether ML-based predictions of *net present benefits* (NPB) can be used to increase cost-effectiveness by targeting investments to the highest return projects. We begin by predicting the net present benefits for each home with its associated household by summing over the discounted predicted savings, which are determined by the difference between post-retrofit counterfactuals and post-retrofit predictions. We find that only about half of the homes in the IHWAP sample have positive private net present benefits. This despite the fact that most retrofits (excluding health and safety) performed by the program should have a savings-to-investment ratio (SIR) greater or equal to one, according to the engineering models. Ex-post analyses of IHWAP and other residential energy efficiency programs have also found low or negative net social benefits, at least on average (Christensen et al., 2021; Fowle, Greenstone, and Wolfram, 2018; Allcott and Greenstone, 2017).

To maximize the total predicted NPB from the program, the implementer would choose to treat only those homes (or even individual measures within homes) that have positive expected returns. We estimate predicted benefits at the home level, which is the

⁵With nested-cross validation, predictions for a given home omit that home’s data from both the training and the validation sample.

unit of treatment within IHWAP. We compare the ML-based targeting strategy to one that uses projections from the program’s current engineering model. We evaluate the performance of both strategies against a set of ex-post NPB estimates from Christensen et al. (2021). The *ex-post* estimates are informed by both pre-treatment as well as post-treatment data, acting as a benchmark for assessing the accuracy of the two sets of *ex-ante* predictions. We find that the ML strategy significantly outperforms the engineering model and could have a drastic impact on program cost-effectiveness. In this sample, targeting funds to the 43% of projects with positive predicted energy-related benefits dramatically increases social net benefits of a dollar spent from \$0.93 to \$1.23.

These findings are relevant to a broad literature on the development of data-driven approaches to identify heterogeneous treatment effects for optimal policy targeting (Athey and Wager, 2021; Wager and Athey, 2018). These methods have been applied to targeting studies that identify the most responsive subgroups across a range of public programs (Knittel and Stolper, 2019; Davis and Heller, 2020; Johnson, Levine, and Toffel, 2019; Erel et al., 2018). Given that we observe comprehensive project cost data and are able to accurately predict individual-level heterogeneity in benefits, our work belongs to a subset of the literature that is able to study the implications of a targeting approach that captures the full net present benefits of a program (Finkelstein and Notowidigdo, 2019; Deshpande and Li, 2019; Lieber and Lockwood, 2019). We demonstrate that improving the accuracy of ex-ante predictions substantially increases cost-effectiveness in a large intervention. This may generalize to a wide range of settings, where program effects are largely driven by measurable characteristics. In settings that involve considerable unobserved heterogeneity in underlying behavioral mechanisms, such as government loan programs for small businesses, accurately predicting and targeting outcomes can be more challenging (e.g. McKenzie and Sansone, 2019).

Our results also contribute to a literature that quantifies the costs of greenhouse gas abatement technologies (Gillingham and Stock, 2018), which is important given recent calls for consideration of second or third-best mechanisms to address climate change (Stiglitz, 2019). In considering various technologies, most previous work has focused ex-

clusively on the average performance of GHG investments. Recent work has demonstrated that even interventions that are cost-effective on average, such as home energy reports, can achieve gains in net benefits by better targeting program participants (Knittel and Stolper, 2019; Gerarden and Yang, 2021). Our findings suggest that improvements in predictive modeling could dramatically change outcomes for residential energy efficiency retrofit programs: they can shift from net negative social benefits to one of the lowest cost vehicles for achieving greenhouse gas reductions. These improvements could be realized at relatively low cost, given that funding prioritization software could readily be updated to include household billing data as an input and to base its predictions on the results from ML models.⁶

I Background

The WAP is the U.S.'s largest residential weatherization program. It aims to lower energy bills for low-income homes while maintaining health and safety.⁷ Energy savings are achieved through a variety of measures including insulation, air sealing, heating/cooling system repair or replacement, and electric baseload measures such as lighting and refrigerators. In order to qualify for IHWAP, applicants must demonstrate household income below 200% of the poverty guidelines established by the US Department of Health & Human Services (2020). After initial screening, energy audits are conducted for the homes of successful applicants. During those audits, detailed information on housing structure is collected (variables presented in Table 1). All data are then entered into a program management software called WeatherWorks, which streamlines the steps required to complete a weatherization project, including: determining eligibility, assigning contractors, producing work orders, and determining retrofits to be implemented in each

⁶Many energy efficiency programs are sponsored by utilities, almost all of which now have the data infrastructure to store, query, and serve household billing data (EIA, 2017a). Even for programs not sponsored by utilities, it is becoming increasingly easier for consumers to access and share their consumption data (The Green Button Alliance, 2020). The resulting predictions from ML models like those developed here could replace the engineering equations in the current funding prioritization software. These models would only need to be run periodically, perhaps by consulting or academic teams.

⁷Health and safety is addressed through measures such as installation of smoke/carbon monoxide alarms, installation of ventilation, and ground covers over bare dirt in foundation spaces.

home.

The engineering model embedded in WeatherWorks projects the impacts of a given retrofit on energy savings using data collected during a pre-retrofit audit and a widely-used set of structural equations (Edwards et al., 2013; Sentech, 2010).⁸ It estimates savings-to-investment ratios (SIR) for the full set of potential retrofits for each eligible home by dividing the projected life cycle benefits for a candidate retrofit by its installation costs.⁹ For each home, the WeatherWorks system then ranks all possible retrofits from highest to lowest SIR. Retrofits are performed in order of SIR until the per-home funding is exhausted or until there are no retrofits with $\text{SIR} \geq 1.0$.¹⁰ As each measure is chosen, the algorithm recalculates the conditional SIR for all remaining measures to determine the next most cost-effective measure.

II Data

We use comprehensive data on building structure characteristics, household demographics, the labor/materials costs of all retrofits considered and those completed, and monthly energy use from over thirteen thousand homes served by IHWAP between 2006-2016. All homes served by the IHWAP undergo extensive energy audits prior to treatment. The Illinois' Department of Commerce & Economic Opportunity provided the universe of measurements collected during pre-treatment audits, including information on: family size, age, income and sex of householder; home's floor area, number of bedrooms, number of windows, presence of multiple stories, presence of attic, attic insulation, air sealing (blower door test); building vintage and shielding class; type, age, size and operation status of home's main heating equipment; operation status and setting of water heater; presence of air-conditioning; location (county) of home. Table 1 reports descriptive statistics for these measurements as well as information on audit and retrofit

⁸The Weatherworks model does not incorporate information on household energy consumption from utility data. More details on this model are presented in Appendix B.

⁹These SIR estimates are based on private – not social – benefits. Private benefits are quantified using savings from retail electricity rates, whereas social benefits encompass the total benefits of avoided consumption, including avoided generation, transmission, and distribution costs, and pollution damages (Borenstein and Bushnell, 2018).

¹⁰The program occasionally must resolve a serious health or safety issue even if it has a low SIR.

dates and retrofit-specific costs. The sample is comprised of low-income families (average income less than \$16,800), with an average householder age of 53 years. We observe significant variation in housing structure across the program. For example, air-tightness (as measured by blower door tests) varied from 980 CFM50 (cubic feet per minute, at 50 Pascals) to over 13,600 CFM50. Similarly, we observe substantial variation in retrofit-specific expenditures across homes.

In addition to measurements from pre-retrofit audits, this study incorporates monthly energy consumption from IHWAP-treated homes served by a major utility in Illinois. We restrict the sample to homes that use either natural gas or electricity as their main heating fuel (representative of approximately 88% of homes in the state according to the US Census Bureau, 2013) and focus our analyses on the combined energy consumption from both fuels in MMBtu. Figure 1 plots the distribution of monthly energy use separately for non-winter (Panel a) and winter (Panel b) months, and both before and after energy efficiency retrofits. In winter months, the median home consumes 15.1 MMBtu per month preceding retrofits, whereas the median home following retrofits consumes 12 MMBtu – a 20% difference in the raw distributions. During non-winter months, the median home consumes 5.1 MMBtu before retrofits and 4.6 MMBtu following retrofits – a 10% shift. We note that IHWAP primarily targets home heating, but it can also improve the efficiency of home cooling among homes that have air conditioning. Efficiency improvements in both heating and cooling are important in Midwestern climates.

Finally, for each home and month, we collected data on minimum outdoor temperature, maximum outdoor temperature, and precipitation from the PRISM Climate Group (2018).¹¹ We use these measurements to calculate heating degree days (with bases 60F and 65F), and cooling degree days (with base 75F). Summary statistics of the observed weather variables are reported in Appendix Table A.1 Panel A. We project future weather realizations for use in predicting post-treatment energy usage, which retains the conceptual consistency of one of the main purposes of this study: to predict energy sav-

¹¹The PRISM Climate Group (2018) provides interpolated weather data for the US. Those were matched with the location of each home in our sample based on geocoded coordinates. Addresses from the homes in our sample were geocoded using an API from Google (2018).

ings *ex-ante*, before weather can be observed. We first created daily “climate normals” for each home by calculating day-of-the-year average pre-treatment temperatures and precipitation. We then aggregated those measures up to the monthly (bill cycle) level for a given home by taking the average of the daily climate normals across the month.¹² We use these weather projections as the post-treatment weather data for use in our prediction model.¹³

III Empirical Strategy and Results

A Ex-Ante Estimates of Savings

The first step of our analysis evaluates the accuracy of ML-based predictions of the effects of energy efficiency programs in an *ex-ante* setting – before any retrofits have been installed. As it would be for planners making projections that guide funding decisions, our objective is to generate accurate estimates of expected savings for each home, conditional on building characteristics, household characteristics, predicted weather, and the measures to be performed.

The ex-ante model of the effects of retrofits takes the form:

$$b_{it}^{EA} = \hat{Y}_{it}(1) - \hat{Y}_{it}(0) \quad , \quad (1)$$

where b_{it}^{EA} is an ex-ante prediction of the reduction in energy use resulting from energy efficiency retrofits for home i in month t , $\hat{Y}_{it}(1)$ is an ex-ante prediction of energy use in the presence of treatment, and $\hat{Y}_{it}(0)$ is an ex-ante counterfactual prediction of energy use in the absence of treatment. Negative values of b_{it}^{EA} represent energy savings (i.e. the home used less energy after treatment, compared to the counterfactual). We used a machine learning algorithm to obtain $\hat{Y}_{it}(1)$ and $\hat{Y}_{it}(0)$. We then calculated the average annual savings for each home in the sample by summing per-home month-of-year averages

¹²For heating and cooling degree days we sum over all days in a given bill cycle, rather than taking averages.

¹³Summary statistics of the projected weather variables are presented in Table A.1 Panel B, which closely approximate the observed variation presented in Panel A.

of b_{it}^{EA} across twelve months.

We selected the best performing ML algorithm from the following set of candidates: neural network, gradient boosting, random forest and Lasso. We trained each model using: (1) the characteristics of homes and households obtained from an audit conducted prior to determining retrofit funding for each home, (2) the expenditures scheduled for each retrofit measure, (3) the monthly-level utility data for homes other than the ones for which predictions are being made, and (4) the projections of monthly temperature and precipitation prior to retrofit decisions. The neural network exhibited the highest out-of-sample accuracy, assessed through *nested cross-validation*. The properties of two-step approaches, with neural networks as the first step, are also well-understood in the econometric literature (Farrell, Liang, and Misra, 2021a; Farrell, Liang, and Misra, 2021b).

We used nested cross-validation to guarantee that all the model parameters and hyperparameters were estimated for a subsample that was entirely distinct from the subsample for which energy consumption outcomes were predicted. Nested cross-validation is also shown to significantly reduce bias of out-of-sample prediction errors (Varma and Simon, 2006). Figure 2 illustrates our nested cross-validation design. We first split the full set of homes into four equally-sized random subsamples.¹⁴ We then trained the neural net algorithm using three of the subsamples (‘training set’), while holding a fourth out as the ‘test set.’ Within the training set, we performed an inner-layer of cross-validation where, in each iteration, we used two subsamples for training and a third for validation (i.e. to assess prediction accuracy). We selected the best models on the basis of lowest mean squared error in the validation set (the configuration of the best-performing neural network is presented in Appendix C.1). We repeated this process four times, such that each subsample serves once as the test set.¹⁵

Figure 3 plots the results of our predictions. We find that the ML method is able to recover remarkably accurate predictions of household energy usage both pre- and post-

¹⁴We use stratified sampling to assure that all monthly observations from a given home will be in either one of the subsamples.

¹⁵The machine learning models were trained with a sample of over 13 thousand homes (as shown in Table 1), while results were assessed in a restricted sample of 3,913 homes for which a complete year of post-retrofit data was available.

retrofit, likely owing to the richness of the household/home characteristics and usage data. The x-axis depicts months before (negative) and months after (positive) weatherization. To make estimates comparable across homes that were treated at different points in the year, we normalized monthly household consumption data to account for seasonality. First, we calculated the mean of the observed pre-retrofit energy usage for each month in the sample. We then subtracted that mean from the corresponding predicted (blue and green lines) or observed (orange lines) monthly energy usage of a given home. The y-axis thus represents deviations from the mean pre-retrofit monthly usage.

The blue line and blue shaded area represent the normalized mean and 95% confidence interval corresponding to observed data. The orange line and orange vertical bars represent the normalized mean and 95% confidence interval for the out-of-sample ML predictions. The green line depicts counterfactual consumption, or predicted energy use in the absence of treatment. We note a slight drop in predictions of counterfactual monthly energy use when compared to pre-treatment consumption. This can largely be attributed to discrepancies between observed and projected weather.¹⁶

Estimates of the treatment effect for the average home in the sample are given by the difference between the green and blue lines illustrated by diagonal shading in Figure 3. Post-treatment predictions and realized energy usage fall substantially in the months following retrofit installation. The overlap between the orange bars and the blue shading in the months both preceding and following the installation of retrofits indicates that the ML predictions are statistically indistinguishable from the observed values. While the prediction errors are minimal on average, in Appendix C.3, we investigate differences in their distribution. Prediction errors in the pre-retrofit and post-retrofit periods are similar in magnitude and sign, reducing any bias in estimated effects of retrofits b_{it}^{EA} .

¹⁶We normalized the monthly observed and predicted values relative to the observed pre-retrofit consumption in the corresponding month of sample. In Appendix Figure A.1, panel b, we show results for predictions with observed weather, revealing closer alignment between counterfactual and pre-treatment consumption. Importantly for our approach, discrepancies between projected and observed weather difference out, since they affect counterfactual and post-treatment predictions in the same direction. Both approaches yield highly similar estimates of treatment effects.

B Increased Cost-Effectiveness from Targeting

The ultimate goal of this study is to examine whether ML-based predictions of net present benefits can be used to more effectively target investments to increase their cost-effectiveness. To that end, we define a project as a home with its suite of measures that were determined by the standard engineering model. We then examine whether program cost-effectiveness could be improved by choosing to treat (or not treat) homes based on ex-ante predictions of savings.¹⁷

We convert predicted savings into monetary benefits by estimating the social benefits of avoided energy consumption, including avoided generation, transmission and distribution costs, as well as benefits from reduced GHG and local air pollution (Borenstein and Bushnell, 2018; Davis and Muehlegger, 2010).¹⁸ The study focuses on energy-related benefits, although there may be other potential benefits such as improvements related to the health and safety of the occupants of treated homes (Tonn, Rose, and Hawkins, 2018; Pigg, Cautley, and Francisco, 2018).

We estimate the social net present benefit (NPB) of a project as follows:

$$\text{NPB}_i = \sum_{t=1}^{T_i} \left[\frac{\hat{\beta}_i^e \times \text{cost}_{\text{elec},t}}{(1+r)^t} + \frac{\hat{\beta}_i^g \times \text{cost}_{\text{gas},t}}{(1+r)^t} \right]_i - \text{TotalCost}_i$$

where $\hat{\beta}_i^e$ and $\hat{\beta}_i^g$ are our ex-ante annual estimates of electricity and natural gas savings for home i ; $\text{cost}_{\text{elec},t}$ and $\text{cost}_{\text{gas},t}$ are the social costs of electricity and natural gas in year t ; r is a discount rate; TotalCost_i represents the total costs of the retrofits for home i ; and T_i denotes the expected lifespan of the retrofits installed in home i . Similarly, we calculate

¹⁷An alternate approach would be to predict measure-specific savings and optimize program cost-effectiveness by reallocating funds not only among homes but across measures performed in each home, as would be appropriate for a program such as WAP that aims to provide whatever cost-effective measures are appropriate for a low-income household. That exercise would require causal estimates of measure-specific effects which we do not have, and which are not currently available. We note that there are complex interactions among measures, which are also not randomly assigned across homes. We therefore view our exercise of allocation across homes as a lower bound on the potential improvements in cost-effectiveness. Better measure-specific ex-ante estimates of savings could potentially yield even larger improvements.

¹⁸State-level energy prices (reported in Appendix E) were obtained from the Energy Information Administration (EIA, 2017b), and were adjusted based on Borenstein and Bushnell (2018); Davis and Muehlegger (2010). Appendix E.4 shows that our results also hold when using retail energy prices without adjustment (i.e. private benefits).

benefit-cost ratios (BCR) by dividing total benefits by total costs. Total retrofit costs for each project include the labor and materials costs reported in the IHWAP administrative data set. The home-specific lifespans T_i are calculated based on weighted averages, depending on expenditures across different types of retrofits. Resulting lifespans are close to 30 years, on average. We assume a baseline a discount rate of 3%, which is recommended by the Department of Energy for evaluation of public programs. In Appendix E.3 we test the sensitivity of our estimates to lifetime and discount rate assumptions.

To assess the effect of prediction accuracy on cost-effectiveness, we compare a targeting exercise based on our ex-ante predictions of the effects of retrofits to two kinds of estimates: (1) ex-ante projections from the engineering model that currently guides decisions in the program and (2) from an ex-post evaluation. The first comparison estimates whether and how much the proposed method improves upon current estimates. The second comparison is akin to comparing the information available at the moment of decisions on retrofits to more complete information available years later. Since we cannot observe “true” savings, the ex-post estimates serve as a benchmark that is based on the more complete information set. State-of-the-art ex-post evaluation techniques are designed to account for unobserved factors (to the researcher) that may affect energy consumption patterns, such as changes in weather patterns, in consumer behavior, and in other economic factors that may occur simultaneously with treatment. We use the ex-post estimates from Christensen et al. (2021) as the benchmark for both sets of ex-ante predictions (see Appendix D for details).

In Figure 4, we report results from a simulation where we rank homes by net present benefits according to each of the models considered. The graph in the upper panel reports the cumulative social NPB from treating homes ranked highest to lowest. The lower panel reports the corresponding benefit-cost ratio. The ex-post model, represented by the blue line, serves as a benchmark for the maximum possible NPB at every point along the x-axis. The first 42% of homes have positive social NPB, after which the cumulative NPB declines. This demonstrates the potential for accurate ex-ante predictions to improve program cost-effectiveness. Targeting investments to the top 42% of projects would

maximize the program’s energy-related returns with the fully informed ex-post model. This corresponds to a possible increase in energy-related benefits from \$0.93 to \$1.36 for every dollar invested in efficiency retrofits.

We assess the potential gains from ex-ante predictions by looking at the orange and green lines of Figure 4. The orange and green lines illustrate the cumulative net benefits—as calculated using the benchmark value for each home—when ranking homes according to either the ML or engineering ex-ante estimates of those benefits. Those are compared against the blue line (ex-post estimates) and the red line, which reflects a random ordering of homes treated by the program. While both sets of ex-ante predictions exhibit better performance than a ranking of homes at random, the machine learning model (orange line) yields predictions that more closely approximate the ranking from the ex-post evaluation. This simulation illustrates that targeting investments using ML-based predictions can dramatically improve cost-effectiveness relative to models that currently guide funding allocations. In this sample, a ML-based targeting strategy would allocate funds to the top 43% of projects, increasing the social net benefits of a dollar invested from 0.93 to 1.23. In Appendix E.3, we examine the sensitivity of this finding to assumptions about the underlying lifespan and discount rate parameters and find that the gains from targeting according to the ML-based ranking are substantial across all scenarios considered.¹⁹

IV Conclusion

This study demonstrates that better ex-ante predictions could be important for optimizing investments in energy efficiency programs. Results indicate that: (1) ML-based predictions coupled with household billing data can outperform comparable predictions from engineering models, and (2) targeted investments based on this method could result in substantial increases in the cost-effectiveness of retrofit programs. While the sample analyzed in this study represents a single retrofit program with specific goals, our approach

¹⁹We consider scenarios with lower and higher insulation lifespans, resulting in average home-specific lifespans of 20 and 40 years, respectively. We also consider alternative discount rates of 2% and 4%. In Table E.1, we show that the gains from targeting according to the ML-based ranking range from \$0.27 to \$0.33 in net present benefits per dollar spent in the program.

could provide value to a wide range of programs that rely on similar engineering models. The International Energy Agency has promulgated an increase in worldwide investments in energy efficiency from \$140 Billion (currently expected) to \$220 Billion per year by 2025 (IEA, 2018). Extrapolating the 21% increase in benefits from the Illinois sample to worldwide investments yields an order-of-magnitude estimate of \$46 Billion in annual savings during the 5-year period. These savings would further increase as expenditures continue to ramp up on the basis of 2030 targets.

We identify several limitations in the present analysis to be addressed in future research. While the present study focuses on cost-effectiveness at the home level, the approach could be extended to applications that target measures within a given home. Second, the study relies on ex-ante billing data for metered fuels (electricity and natural gas), while retrofits to a smaller fraction of properties using delivered fuels such as propane, fuel oil, and wood may have outsized effects on greenhouse gas emissions.

We are not aware of any energy efficiency programs that combine household-level energy billing data with ML-based strategies to project savings or to select among candidate retrofits. However, rapid technological advances are creating opportunities to do so (Gagliano and Thomas, 2015). Energy efficiency programs are often sponsored by utilities that have recently developed the data infrastructure to store, query, and serve household billing data (EIA, 2017a). In programs that are not sponsored by utilities, government-led initiatives are making it easier for consumers to access and share their consumption data (The Green Button Alliance, 2020). Therefore software like Weatherworks could readily be updated to include household billing data. Likewise, it would be straightforward to integrate predictions from ML models, like those developed here. These models need only be run periodically, perhaps by consulting or academic groups, and the resulting predictions could be fed into the back end of the engineering software.

References

- Allcott, Hunt and Michael Greenstone (May 2017). “Measuring the Welfare Effects of Residential Energy Efficiency Programs”. *NBER Working Paper*. Working Paper Series(23386).
- Athey, Susan and Stefan Wager (2021). “Policy learning with observational data”. *Econometrica* 89(1), pp. 133–161.
- Auffhammer, Maximilian (2021). “Retrofit This”. *Energy Institute Blog, Haas School of Business, University of California Berkeley*.
- Barbose, Galen L, Charles A Goldman, Ian M Hoffman, and Megan Billingsley (2013). “The future of utility customer-funded energy efficiency programs in the USA: projected spending and savings to 2025”. *Energy Efficiency* 6(3), pp. 475–493.
- Berry, Linda G. and Michael B. Gettings (1998). “Realization Rates of the National Energy Audit”. *Thermal Performance of the Exterior Envelopes of Buildings VI Conference*.
- Borenstein, Severin and James B Bushnell (June 2018). “Do Two Electricity Pricing Wrongs Make a Right? Cost Recovery, Externalities, and Efficiency”. *NBER Working Paper*(24756).
- Bruegge, Chris, Tatyana Deryugina, and Erica Myers (2019). “The distributional effects of building energy codes”. *Journal of the Association of Environmental and Resource Economists* 6(S1), S95–S127.
- Burlig, Fiona, Christopher Knittel, David Rapson, Mar Reguant, and Catherine Wolfram (2020). “Machine Learning from Schools about Energy Efficiency”. *Journal of the Association of Environmental and Resource Economists* 7(6), pp. 1181–1217.
- California Air Resources Board (2017). “California’s 2017 Climate Change Scoping Plan”.
- Chaisemartin, Clément de and Xavier D’Haultfoeuille (Sept. 2020). “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects”. *American Economic Review* 110(9), pp. 2964–96.
- Christensen, Peter, Paul Francisco, Erica Myers, and Mateus Souza (2021). “Decomposing the Wedge Between Projected and Realized Returns in Energy Efficiency Programs”. *The Review of Economics and Statistics*. (Forthcoming).
- Dalhoff, Gregory K. (1997). “An Evaluation of the Performance of the NEAT Audit for the Iowa Low-Income Weatherization Program”. *1997 Energy Evaluation Conference*.
- Davis, Jonathan MV and Sara B Heller (2020). “Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs”. *Review of Economics and Statistics* 102(4), pp. 664–677.
- Davis, Lucas W., Alan Fuchs, and Paul Gertler (Nov. 2014). “Cash for Coolers: Evaluating a Large-Scale Appliance Replacement Program in Mexico”. *American Economic Journal: Economic Policy* 6(4), pp. 207–38.
- Davis, Lucas W. and Erich Muehlegger (2010). “Do Americans consume too little natural gas? An empirical test of marginal cost pricing”. *The RAND Journal of Economics* 41(4), pp. 791–810.
- Davis, Lucas W, Sebastian Martinez, and Bibiana Taboada (2020). “How effective is energy-efficient housing? Evidence from a field trial in Mexico”. *Journal of Development Economics* 143, p. 102390.

- Deshpande, Manasi and Yue Li (Nov. 2019). “Who Is Screened Out? Application Costs and the Targeting of Disability Programs”. *American Economic Journal: Economic Policy* 11(4), pp. 213–48.
- Edwards, J., D. Bohac, C. Nelson, and I. Smith (2013). “Field Assessment of Energy Audit Tools for Retrofit Programs”. *Technical Report, National Renewable Energy Laboratory*.
- Erel, Isil, Léa H Stern, Chenhao Tan, and Michael S Weisbach (2018). “Selecting directors using machine learning”. *NBER Working Paper* 24435.
- European Commission (2020). “A Renovation Wave for Europe - greening our buildings, creating jobs, improving lives.” Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions”.
- European Parliament (2012). “Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency”.
- Executive Office of Energy and Environmental Affairs (2018). ““Massachusetts Global Warming Solutions Act: 10-Year Progress Report””. [Online, accessed in 2020: <https://www.mass.gov/doc/gwsa-10-year-progress-report/download>].
- Farrell, Max H., Tengyuan Liang, and Sanjog Misra (2021a). “Deep Learning for Individual Heterogeneity: An Automatic Inference Framework”. *arXiv:2010.14694*.
- Farrell, Max H., Tengyuan Liang, and Sanjog Misra (2021b). “Deep Neural Networks for Estimation and Inference”. *Econometrica* 89(1), pp. 181–213.
- Finkelstein, Amy and Matthew J Notowidigdo (May 2019). “Take-Up and Targeting: Experimental Evidence from SNAP”. *The Quarterly Journal of Economics* 134(3), pp. 1505–1556.
- Fowlie, Meredith (2020). “The Search for Good Green Stimulus”. *Energy Institute Blog, Haas School of Business, University of California Berkeley*.
- Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram (2018). “Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program”. *The Quarterly Journal of Economics* 133(3), pp. 1597–1644.
- Gagliano, Jerone and Greg Thomas (2015). “NYSERDA Home Performance with Energy Star Realization Rate Attribution Study”. *Tech. Rep., New York State Energy Research and Development Authority*.
- Gerarden, Todd D. and Muxi Yang (2021). “Using Targeting to Optimize Program Design: Evidence from an Energy Conservation Experiment”. *Working Paper*.
- Gillingham, Kenneth and James H Stock (2018). “The cost of reducing greenhouse gas emissions”. *Journal of Economic Perspectives* 32(4), pp. 53–72.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press. Chap. 6 Deep Feedforward Networks.
- Goodman-Bacon, Andrew (2021). “Difference-in-Differences with Variation in Treatment Timing”. *Journal of Econometrics*. (Forthcoming).
- Google (2018). “Google Maps Platform: Geolocation API”. [Online, accessed in 2018: <https://developers.google.com/maps/documentation>].
- Hepburn, Cameron, Brian O’Callaghan, Nicholas Stern, Joseph Stiglitz, and Dimitri Zenghelis (May 2020). “Will COVID-19 fiscal recovery packages accelerate or retard progress on climate change?” *Oxford Review of Economic Policy* 36, S359–S381.

- Houde, Sébastien and Joseph E. Aldy (2014). “Belt and Suspenders and More: The Incremental Impact of Energy Efficiency Subsidies in the Presence of Existing Policy Instruments”. *NBER Working Paper* 20541.
- International Energy Agency (2018). “Energy Efficiency 2018: Analysis and outlooks to 2040”.
- International Energy Agency (2019). “World Energy Outlook 2019”.
- Johnson, Matthew S, David I Levine, and Michael W Toffel (2019). “Improving regulatory effectiveness through better targeting: Evidence from OSHA”. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 20-019.
- Knittel, Christopher R and Samuel Stolper (Dec. 2019). “Using Machine Learning to Target Treatment: The Case of Household Energy Use”. *NBER Working Paper* 26531.
- Kono, Jun, Yutaka Goto, York Ostermeyer, Rolf Frischknecht, and Holger Wallbaum (Feb. 2016). “Factors for Eco-Efficiency Improvement of Thermal Insulation Materials”. *Key Engineering Materials* 678, pp. 1–13.
- Levinson, Arik (Oct. 2016). “How Much Energy Do Building Energy Codes Save? Evidence from California Houses”. *American Economic Review* 106(10), pp. 2867–94.
- Lieber, Ethan M. J. and Lee M. Lockwood (Apr. 2019). “Targeting with In-Kind Transfers: Evidence from Medicaid Home Care”. *American Economic Review* 109(4), pp. 1461–85.
- Martín Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.
- McKenzie, David and Dario Sansone (2019). “Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria”. *Journal of Development Economics* 141, p. 102369.
- McKinsey & Co (2009). ““Pathways to a low-carbon economy: Version 2 of the global greenhouse gas abatement cost curve”. *Tech. Rep., McKinsey & Company*, pp. 1–165.
- Pigg, S., D. Cautley, and P. W. Francisco (2018). “Impacts of weatherization on indoor air quality: A field study of 514 homes”. *Indoor Air* 28(2), pp. 307–317.
- PRISM Climate Group, Oregon State University (2018). *PRISM Climate Data*. [Online, accessed in 2018: <https://prism.oregonstate.edu/>].
- Rushing, Amy S., Joshua D. Kneifel, and Barbara C. Lippiatt (2012). “Energy Price Indices and Discount Factors for Life-Cycle Cost Analysis – 2012: Annual Supplement to NIST Handbook 135 and NBS Special Publication 709”. *NIST Interagency/Internal Report (NISTIR)* 15(n29).
- Sentech Inc (2010). “Review of Selected Home Energy Auditing Tools: In Support of the Development of a National Building Performance Assessment and Rating Program”. *Tech. Rep., DOE’s Office of Energy Efficiency and Renewable Energy*.
- Sharp, T.R. (1994). “The North Carolina Field Test: Field Performance of the Preliminary Version of an Advanced Weatherization Audit for the Department of Energy’s Weatherization Assistance Program”. *Tech. Rep. ORNL/CON-362, Oak Ridge National Laboratory*.
- Souza, Mateus (2019). “Predictive Counterfactuals for Treatment Effect Heterogeneity in Event Studies with Staggered Adoption”. *SSRN Working Paper* (3484635).

- Stiglitz, Joseph E. (2019). “Addressing climate change through price and non-price interventions”. *European Economic Review* 119, pp. 594–612.
- The Green Button Alliance (2020). “Green Button Data”.
- The White House (2021). “FACT SHEET: The American Jobs Plan”.
- Tieleman, T. and G. Hinton (2012). “Lecture 6.5–RMSprop: Divide the gradient by a running average of its recent magnitude”. *Neural Networks for Machine Learning*.
- Tonn, B., E. Rose, and B. Hawkins (2018). “Evaluation of the U.S. Department of Energy’s Weatherization Assistance Program: Impact results”. *Energy Policy* 118, pp. 279–290.
- US Census Bureau (2013). “American Housing Survey for the United States: 2011”. [Online, accessed in 2018: <https://www.census.gov/programs-surveys/ahs/data.2011.html>].
- US Code (1964). “Weatherization Assistance for Low-Income Persons: Congressional findings and purpose”. *42 USC § 6861*. [Online, accessed in 2021: [http://uscode.house.gov/view.xhtml?req=\(title:42%20section:6861%20edition:prelim\)](http://uscode.house.gov/view.xhtml?req=(title:42%20section:6861%20edition:prelim))].
- US Department of Health & Human Services (2020). “U.S. Federal Poverty Guidelines Used to Determine Financial Eligibility for Certain Federal Programs”. [Online, accessed in 2020: <https://aspe.hhs.gov/poverty-guidelines>].
- US Energy Information Administration (2017a). “Nearly half of all U.S. electricity customers have smart meters”.
- US Energy Information Administration (2017b). “Residential Electricity and Natural Gas Prices”.
- US Environmental Protection Agency (1998). “AP 42, Fifth Edition Compilation of Air Pollutant Emissions Factors, Volume 1: Stationary Point and Area Sources”. *Tech. Rep., EPA*.
- Varma, Sudhir and Richard Simon (Feb. 2006). “Bias in error estimation when using cross-validation for model selection”. *BMC Bioinformatics* 7(1), p. 91.
- Wager, Stefan and Susan Athey (2018). “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. *Journal of the American Statistical Association* 113(523), pp. 1228–1242.
- Zivin, Joshua G. and Kevin Novan (2016). “Upgrading Efficiency and Behavior: Electricity Savings from Residential Weatherization Programs”. *The Energy Journal* 37(4).

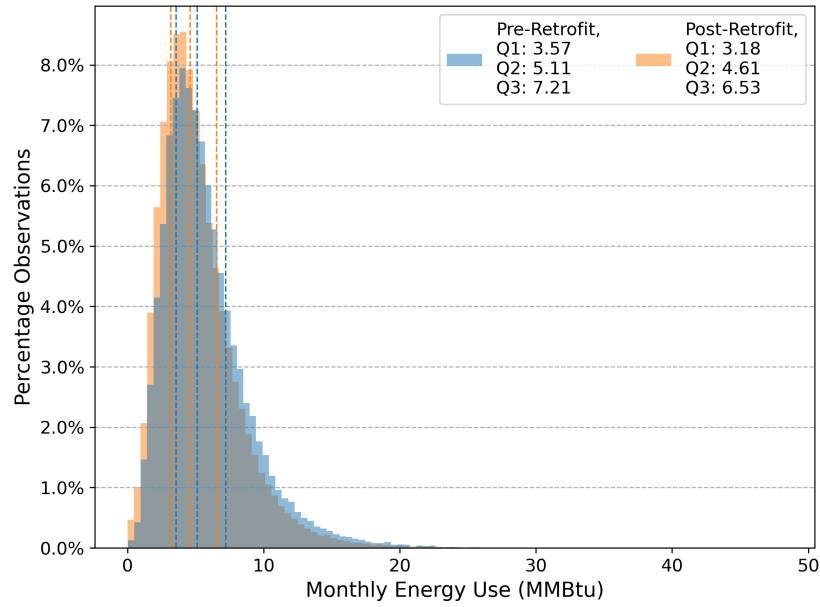
Table 1: Descriptive Statistics for Main Control Variables in the Study

	Average	Standard Deviation	Min	Max
<i>Demographics</i>				
Family Income (\$)	16,754.27	10,091.63	0	51,196.27
Family Size	2.68	1.65	1	9
Female Householder (%)	0.68	0.47	0	1
Householder Age	53.15	15.82	22	89
Renter (%)	0.06	0.24	0	1
County ID (Categorical)	43.95	26.04	1	86
<i>Housing Structure</i>				
Attic R-Value	11.43	10.96	0	40
Floor Area (sqft)	1450.3	622.8	0	4320
Pre-Retrofit Blower Door (CFM50)	3,648.79	1,786.18	980	13,662
Main Heat Type (Categorical)	2.25	1.15	1	10
Main Heat Age	19.44	14.6	0	100
Main Heat Size (BTU)	76,735.14	41,939.71	0	750,000
Main Heat Operational (%)	0.83	0.38	0	1
Building Vintage (Categorical)	6	2.44	1	13
Has Air-Conditioning (%)	0.01	0.11	0	1
Has Attic (%)	0.7	0.46	0	1
Has Multiple Stories (%)	0.32	0.46	0	1
Num. Bedrooms	2.76	0.98	1.00	6.00
Num. Windows	15.12	5.4	2	26
Shielding Class (Categorical)	1.85	0.87	1	5
Operational Water Heater	0.99	0.12	0	1
Water Heater Setting (Categorical)	2.02	0.4	1	3
<i>Administrative Variables</i>				
Audit Month	6	3.4	1	12
Audit Year	2010	2.29	2005	2016
Retrofit Year	2011	2.21	2006	2016
<i>Costs (\$) per Retrofit Categories</i>				
Air Conditioning	6.8	90.14	0	1,827
Air Sealing	296.78	287.45	0	1,981.28
Attic	930.71	714.49	0	3,451.65
Baseload	175.65	232.23	0	1,099.57
Door	341.58	360.11	0	2,082.22
Foundation	300.73	500.35	0	2,988.15
Furnace	1,352.84	1,179.08	0	4,907.51
General	99.3	488.31	0	5,451.54
Health and Safety	486.67	334.03	0	1,725.11
Wall Insulation	274.75	622.03	0	3,467
Window	668.82	890.98	0	4,760.76
Water Heater	138.02	229.82	0	1,553.93
Number of Homes in Sample	13,638	-	-	-

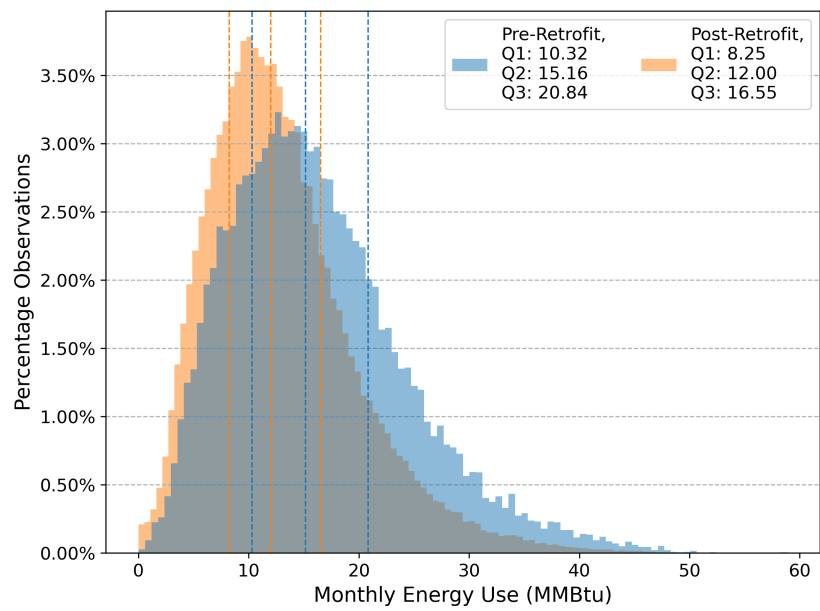
Notes: This table presents descriptive statistics for the main control variables used in the statistical analyses. All monetary values have been inflation-adjusted, by converting to US dollars in 2017. We also consider transformations of floor area (squared and log), and winsorized main heat size.

Figures

Figure 1: Monthly Energy Use Before/After Energy Efficiency Retrofits



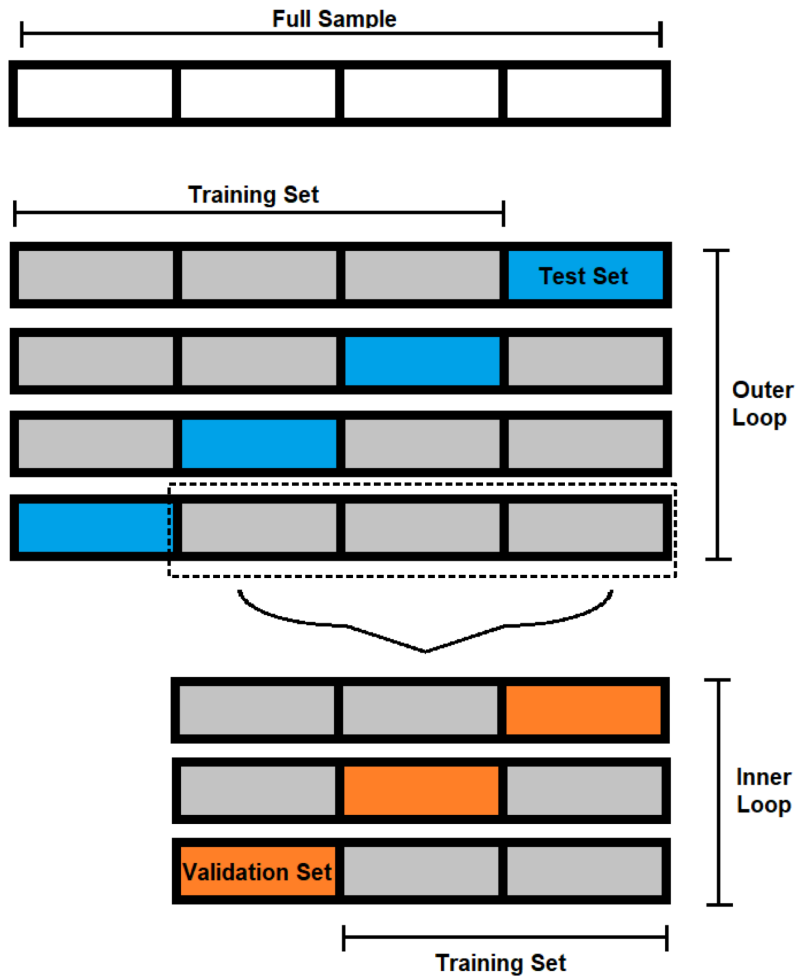
(a) Non-Winter Months



(b) Winter Months

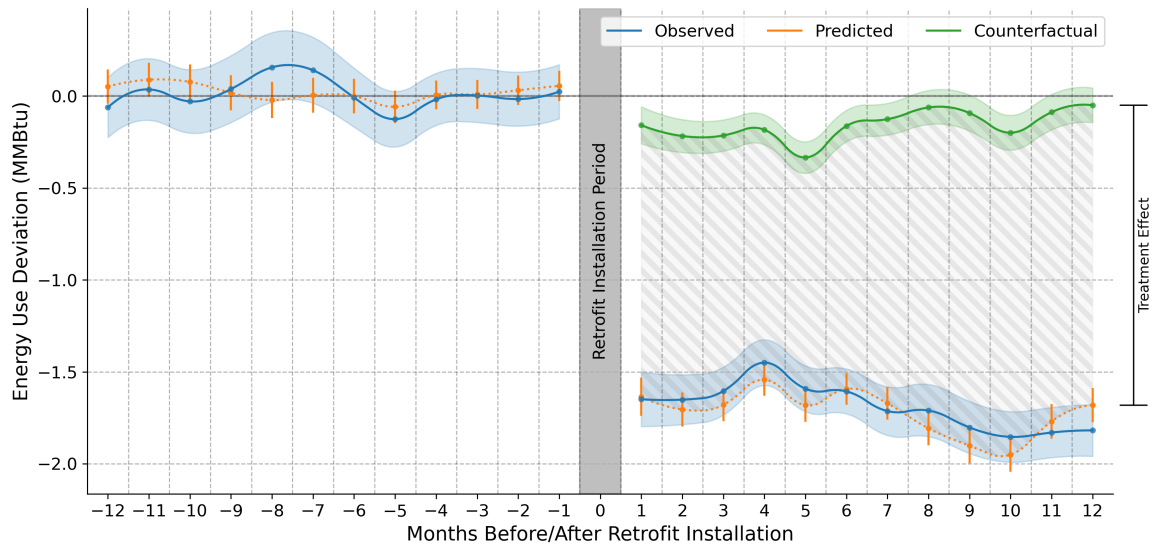
Notes: The figure compares pre-retrofit (blue) and post-retrofit (orange) monthly energy use for homes served by the energy efficiency program. Winter months are defined as November through March. The lower (Q1), middle (Q2), and upper (Q3) quartiles are represented by vertical dashed lines.

Figure 2: Nested Cross-Validation Design



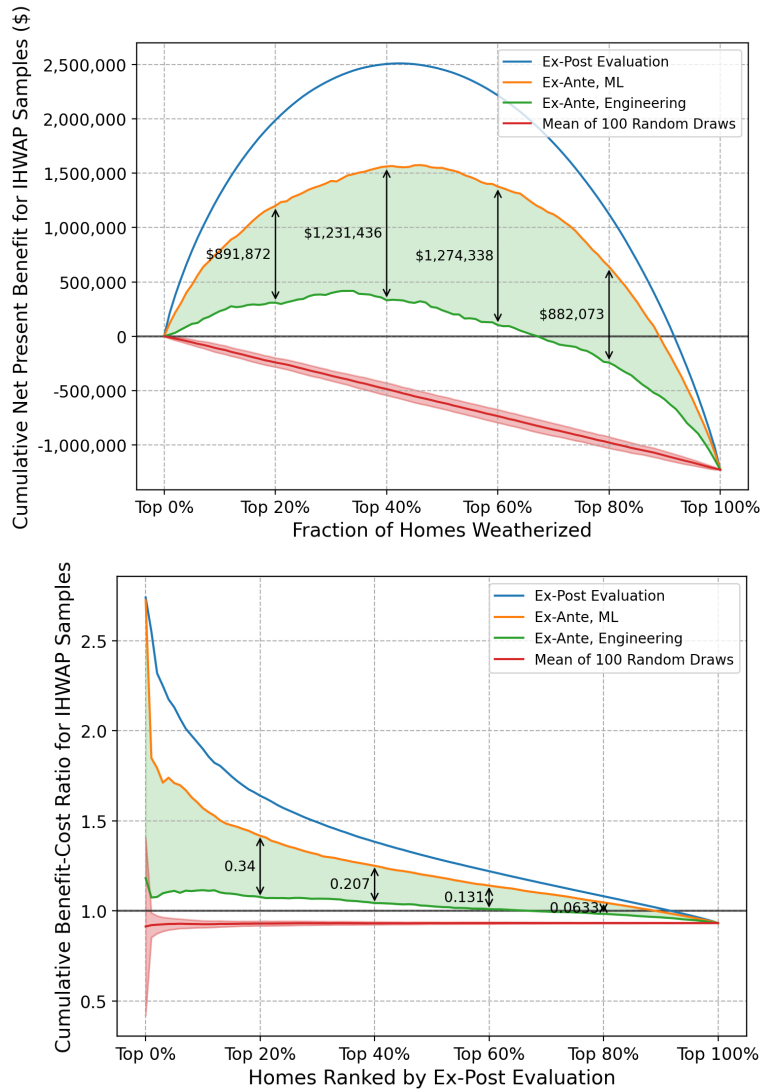
Notes: This figure illustrates our nested cross-validation (CV) design. The top panel represents the outer CV loop. Subsamples in blue are the test set which, for each iteration, are completely unseen by the model and are used to generate out-of-sample estimates. Subsamples in gray are used for model training. The bottom panel represents the inner CV loops. In that case, subsamples in orange represent the validation set, which are used to obtain proxy out-of-sample errors, thus to guide hyperparameter tuning.

Figure 3: Predicted Effects of Energy Efficiency Retrofits



Notes: The figure reports averages and 95% confidence intervals for observed (blue), predicted (orange), and counterfactual (green) energy use in retrofitted homes. Treatment effects are the difference between counterfactual and predicted use, $\hat{\beta}_{it}^{EA}$ from Eq. 1, illustrated by diagonal shading. The x-axis plots the number of months relative to retrofit installation. Data were normalized to account for seasonality (Appendix A).

Figure 4: Net Present Benefits from ML-based Targeting



Notes: The figure reports estimates of cost-effectiveness when ranking homes based on the ex-ante predicted net present benefits generated by the ML model (orange), structural engineering model (green), and the mean of 100 iterations of random selections (red) with std. dev. as the red shaded region. The upper panel reports cost-effectiveness as cumulative NPB and the lower panel in terms of benefit-cost ratios. The y-axis plots cumulative NPB according to ex-post estimates but following the ranking from each ex-ante approach. NPB calculations use home-specific retrofit lifespans (30 years on average), a 3% discount rate, and account for the social cost of carbon. ML models were trained using over thirteen thousand homes. Results were assessed in a restricted sample of 3,913 homes for which a complete year of post-retrofit data was available. The total expenditures for this subset of homes was close to \$18 million.

Appendix – For Online Publication

A Observed and Projected Weather

Geocoded addresses were used to match all homes in the sample with daily minimum outdoor temperature, maximum outdoor temperature, and precipitation from the PRISM Climate Group (2018). The Group compiles and validates observations from monitoring stations across the US, which are then interpolated based on climate models to produce fine resolution (4km grid cell) estimates of weather variation. We calculate the average daily maximum temperature, minimum temperature, heating degree days (with bases $60F$ and $65F$), cooling degree days (with base $75F$), and precipitation during all homes' (monthly) energy billing cycles using daily weather data from 2003 to 2017 from PRISM (thus covering the full sample period). Table A.1 Panel A presents descriptive statistics for the observed weather data, while Panel B presents descriptives for projected weather data.

We further provide evidence that the main results from this research are not highly sensitive to the use of projected instead of observed weather. Figure A.1 plots energy predictions and treatment effects obtained with projected weather (panel a) and observed weather (panel b). Energy use data in this figure were normalized to take into account seasonality. First, the pre-retrofit mean energy usage was calculated for each month in the sample. Then the monthly energy usage of a given home is subtracted by the mean pre-retrofit usage in each corresponding month. The y-axis thus represents deviations from the mean pre-retrofit monthly usage.

As expected, predictions are better aligned with actual energy consumption when the model uses observed weather. Nevertheless, the differences between predicted and real post-treatment usage are not economically or statistically significant when comparing a model with weather projections (as used in the paper) versus with realized weather observed ex-post. This suggests that projected weather yields savings predictions that are highly similar to those produced by a model with actual weather data. The shaded areas of the graphs reveal minimal differences in the estimated treatment effects from both

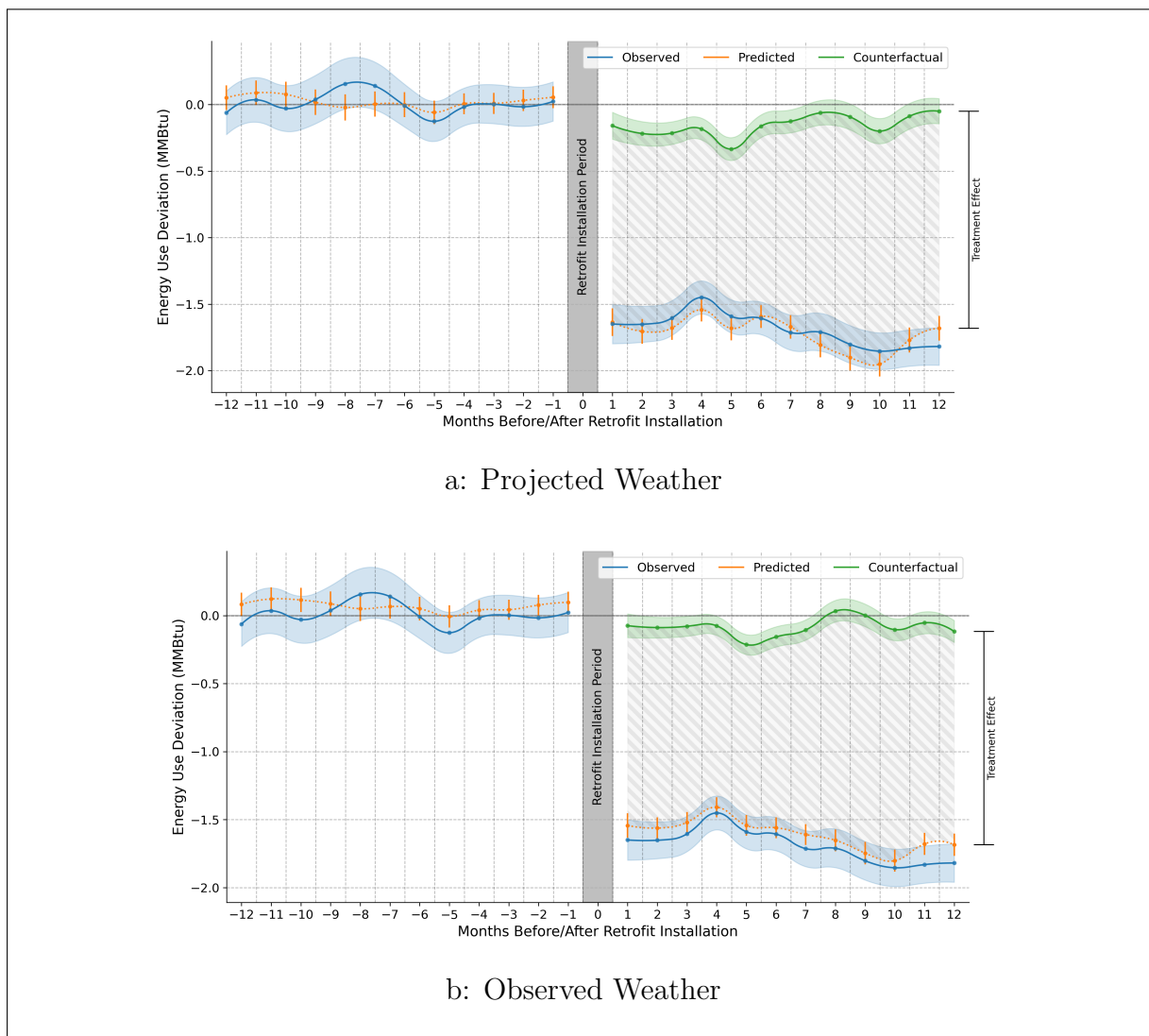
Table A.1: Descriptive Statistics for Weather Variables

Panel A: Observed Variation				
	Average	Standard Deviation	Min	Max
Cooling Degree Days 75F	18.76	37.6	0	426.64
Heating Degree Days 60F	351.94	388.41	0	2577.13
Heating Degree Days 65F	443.15	442.4	0	2862.13
Precipitation (cm)	3.02	1.89	0	18.36
Max Temperature (C)	17.57	10.13	-7.05	37.1
Min Temperature (C)	6.73	9.44	-19.83	24.37
Number of Obs.	457,224	-	-	-
Panel B: Projected Variation				
	Average	Standard Deviation	Min	Max
Cooling Degree Days 75F	17.08	30.85	0	332.95
Heating Degree Days 60F	362.79	388.32	0	2403.6
Heating Degree Days 65F	455.14	442.28	0	2713.6
Precipitation (cm)	2.91	1.07	0	16.67
Max Temperature (C)	17.32	10.13	-5.84	36.65
Min Temperature (C)	6.31	9.24	-18.91	24.37
Number of Obs.	457,205	-	-	-

Notes: This table presents descriptive statistics for the observed and projected weather variables used in the analyses.

approaches. Predictions from both approaches exhibit similar seasonal patterns. Finally, the home cost-effectiveness ranks produced with observed (not reported) or projected weather are also similar. Comparing the ranks with a Kendall rank correlation coefficient hypothesis test yields a p-value of 3.81×10^{-15} , thus rejecting the null hypothesis that the ranks produced by both approaches are independent from each other.

Figure A.1: Machine Learning Predictions; Projected Versus Observed Weather



Notes: This figure presents energy usage prediction results from the machine learning approach, comparing a model that uses projected weather (a) versus one that uses observed weather (b).

B IHWAP’s Model for Projecting Energy Savings

Program management for IHWAP is aided by a software called WeatherWorks. Within WeatherWorks, embedded engineering equations are used to project energy savings for an audited house, and to project savings-to-investment ratios (SIR). The current formula for the whole-house SIR is defined as:

$$SIR_{ov} = (\$Heat_{sav} + \$AC_{sav} + \$Base_{sav} + \$WH_{sav}) / (TotalCost) \quad (B.1)$$

where SIR_{ov} represents the overall SIR for a given home; $\$Heat_{sav}$ are heating savings; $\$AC_{sav}$ are air-conditioning savings; $\$Base_{sav}$ are baseload savings (i.e. from refrigerators, lightbulbs, and other electric appliances); $\$WH_{sav}$ are water heater savings; and $TotalCost$ are the total costs of the retrofits.

Each element from the numerator of SIR_{ov} is estimated with complex formulas based on assumed relationships between heat exchange and energy consumption within homes. Once energy savings are obtained, they are transformed into monetary terms by multiplying by fuel costs, discounted with a 3% annual rate, and assuming different types of retrofits have different expected lifespans. For example, the assumed lifespans for a few of the major retrofits are: 25 years for insulation; 20 for air sealing; 20 for furnace replacement; 15 for central ACs; 10 for window ACs; 15 for water heater replacements; 15 for refrigerators; and 5 for fluorescent light bulbs. Further details on each element from equation B.1 are presented in the “WeatherWorks General Design” document, which is available upon request.

For the purpose of this research, SIR_{ov} , $TotalCost$, and the combined whole-house WeatherWorks projected savings have been provided directly to the authors. Section E presents comparisons between SIR_{ov} from WeatherWorks and benefit-cost ratios estimated according to alternative models. Further, with whole-house projected savings it is possible to obtain monetized WeatherWorks projected benefits for each home, which can then be subtracted by $TotalCost$ to obtain net present benefits. Comparisons of NPV across models are also presented in Appendix E.

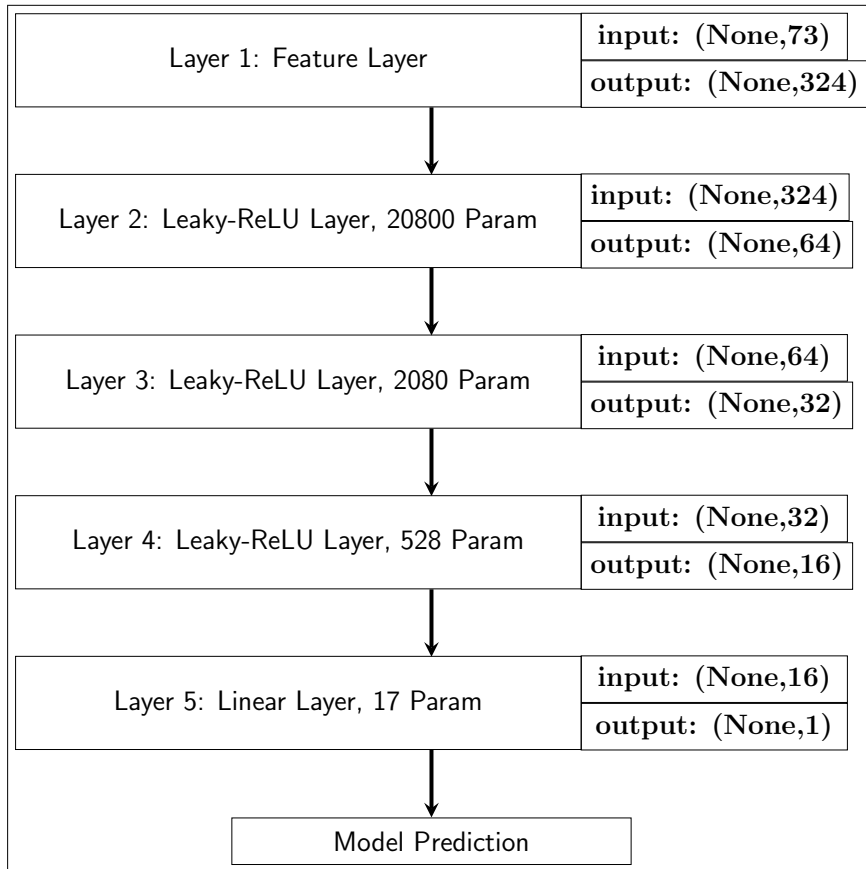
C Machine Learning Algorithms

C.1 Neural Networks

In this paper, neural networks are used for predictive tasks as an intermediate step for ex-ante estimation of program savings. Specifically, we use feedforward neural networks. Feedforward neural networks take features (or covariates) \mathbf{X} as inputs, and map them onto functions to generate predictions of a given outcome y (energy consumption). Importantly, these functions are connected in a chain. Following Goodfellow, Bengio, and Courville (2016), let f^1 , f^2 , f^3 , f^4 , and f^5 denote the layers or functions of a given neural network. The complete neural network can then be expressed as: $f(\mathbf{X}) = f^5(f^4(f^3(f^2(f^1(\mathbf{X}))))))$. Often, the outputs of some of the layers are not known to the researcher, making them “hidden layers.”

Further, each layer is itself a combination of many neurons (or functions), where the output of all neurons will be stacked to generate the output of each layer. Figure C.1 illustrates the neural networks used in this study. The first layer is the feature layer, for which each numeric feature is normalized and each categorical feature is one-hot encoded (separate binary features for each category). Temperature variables were split into 5 bins to allow for non-linear effects. An indicator for winter months was also added, defined as November through March. The feature layer is followed by three hidden layers, constituted of leaky-ReLU activation functions (described below). The first leaky-ReLU layer contains 64 neurons, the second contains 32 neurons, and the last one contains 16 neurons. The final (5th) layer uses a simple linear activation function to generate the model predictions.

Figure C.1: Neural Network Layers



Notes: This figure illustrates the configuration of the preferred neural network used in this study, selected based on validation-set predictive performance.

Each neuron in a layer is a (linear or non-linear) function that takes a vector of inputs, multiplies it by a weight vector and outputs a transformed outcome. Each neuron can therefore be defined as:

$$y = f(\beta \mathbf{X}),$$

where \mathbf{X} is the input vector, β is the learnable weight vector and y is the output of the neuron. For the linear layer, the $f(\cdot)$ function is simply a linear function: $y = \beta \mathbf{X}$.

For the leaky-ReLU layers, the $f(\cdot)$ function is non-linear. Specifically, the the

output of each neuron in the leaky-ReLU layer is:

$$y = f(\beta \mathbf{X}) = \begin{cases} \beta \mathbf{X} & \text{if } \beta \cdot \mathbf{X} \geq 0 \\ \alpha * \beta \mathbf{X} & \text{otherwise.} \end{cases}$$

where α is a hyperparameter defined by the researcher. The algorithm described above was trained using the TensorFlow library (Martín Abadi et al., 2015). For this study, a default value of $\alpha = 0.3$ was used. Also, an L1 regularization penalty is applied to each neuron, with varying regularizers depending on the model being considered (Table C.1). An RMSprop optimizer with learning rate equal to 0.00009 is used to find the optimal parameters of the neural network (Tieleman and Hinton, 2012), using mean squared error as the loss function.

The model was trained on the complete training set first for 10 epochs, to learn general patterns. The model was then further trained for 30 epochs on pre-retrofit data and post-retrofit data separately, to learn specific patterns for each. Monthly energy use in the counterfactual condition was predicted using the pre-retrofit model and by changing the value of the treatment indicator to zero. Post-retrofit usage was predicted by the model trained on post-retrofit data. Observed weather data was used for pre-retrofit predictions, while projected weather was used for post-retrofit and counterfactual predictions. Finally, the ex-ante predicted treatment effect was obtained by computing the difference between the post-retrofit predictions and the counterfactuals.

C.2 Algorithm Selection and Hyperparameter Tuning

Model selection and hyperparameter tuning were implemented via nested cross-validation (CV). While other cross-validation approaches, such as k-fold cross-validation, may be biased for out-of-sample errors, nested CV has been shown to significantly reduce such bias (Varma and Simon, 2006). Also, nested CV is more desirable in the context of this paper because it is better aligned with a social planner’s problem: that is, the social planner must make decisions ex-ante, based on models trained without information

about the homes that are candidates for targeting. Our nested CV design is as follows. Prior to training the candidate algorithms, the full sample of observations from this study was randomly split into four equally-sized subsamples, stratifying by home such that all monthly observations of a given home were allocated to only one of the subsamples. Our nested CV design, illustrated by Figure 2, thus has an outer loop with four iterations, and inner loops with three iterations each. As shown in Figure 2, for each iteration of the outer loop, the subsamples colored in blue represent the test set which are used to obtain out-of-sample estimates reported in the main text. Subsamples in gray are used to train the models. For the inner loops, subsamples in orange are the validation set, used to assess prediction errors and for hyperparameter tuning. Specifically, the best models were selected based on the Mean Squared Error (MSE) for predicting annual home-specific energy reductions (i.e. the program’s Treatment Effect).

Table C.1 presents configurations and prediction accuracy metrics for several neural networks considered. All models have three hidden layers with 64, 32, and 16 neurons, respectively (as illustrated in Figure C.1), with varying regularizers. We present treatment effect MSE for all CV subsamples separately, both for when they served as the validation set and when they served as the test set. The first three columns, for example, present results for the iteration with Fold 1 as the test set. The first column presents the settings for the regularizers, the second column presents the validation set MSE, and the third column presents the test set MSE. The rows in gray highlight the best performing models, selected based on validation set MSE.

Comparing columns two and three from Table C.1, for example, it can be noted that validation and test set MSE are similar throughout, suggesting that our nested CV design is unlikely to produce biased out-of-sample errors. For some folds, the test set MSE are even slightly smaller than validation set MSE. Overall, the MSE range from 37 to 45, resulting in Root Mean Squared Errors (RMSE) ranging from 6 to 6.7. These represent approximately 34% to 38% of the average per-home annual savings from the program (around 17.7 MMBtu according to the ex-post ML method). These errors are not negligible. Nevertheless, the ex-ante model still results in substantially more accurate

estimates than the status quo model currently in place, as discussed in the main text.

Table C.2 presents cross-validation results for other candidate machine learning algorithms. For this Table, we implemented 5-fold cross-validation (i.e. not nested cross-validation, which is substantially more computationally demanding). This Table was produced at an earlier stage of the project, after which MSE of different models were compared. As can be noted, MSE for other types of algorithms are substantially higher compared to those from neural networks. Since neural networks resulted in substantially lower validation set MSE, we chose to focus on those types of models, thus implementing nested CV only for those.

Table C.1: Nested CV Results and Hyperparameter Tuning For Neural Networks

Regularizer	Fold 1 Test Set			Fold 2 Test Set			Fold 3 Test Set			Fold 4 Test Set					
	Validation Set MSE	Test Set MSE	Regularizer	Validation Set MSE	Test Set MSE	Regularizer	Validation Set MSE	Test Set MSE	Regularizer	Validation Set MSE	Test Set MSE	Regularizer	Validation Set MSE	Test Set MSE	Regularizer
1, 0.6, 0.8	41.0267	38.4655	1, 0.9, 0.8	41.2119	38.9581	1, 0.9, 0.8	40.2521	38.2661	1, 0.9, 0.8	41.2605	41.9979	1, 0.9, 0.8	41.2605	38.2661	1, 0.9, 0.8
1.2, 0.6, 0.8	41.4340	39.1346	0.9, 0.9, 0.8	40.7673	39.6166	1.1, 0.9, 0.8	40.8743	38.5912	1.1, 0.9, 0.8	41.6940	42.0887	1.1, 0.9, 0.8	41.6940	38.5912	1.1, 0.9, 0.8
0.9, 0.6, 0.8	40.7797	38.1490	0.8, 0.9, 0.8	40.4860	37.0120	0.9, 0.9, 0.8	39.8543	38.0942	0.9, 0.9, 0.8	40.9128	42.9775	0.9, 0.9, 0.8	40.9128	38.0942	0.9, 0.9, 0.8
0.8, 0.6, 0.8	41.1950	37.4924	0.7, 0.9, 0.8	40.3389	37.0238	0.8, 0.9, 0.8	39.5619	37.7234	0.8, 0.9, 0.8	40.8355	44.0094	0.8, 0.9, 0.8	40.8355	37.7234	0.8, 0.9, 0.8
0.9, 0.5, 0.8	40.9162	38.2768	0.6, 0.9, 0.8	40.8131	39.0279	0.7, 0.9, 0.8	39.3125	37.6872	0.7, 0.9, 0.8	40.7608	43.9808	0.7, 0.9, 0.8	40.7608	37.6872	0.7, 0.9, 0.8
0.9, 0.7, 0.8	40.6887	37.9356	0.7, 0.8, 0.8	40.3027	37.2739	0.6, 0.9, 0.8	39.1242	38.0720	0.6, 0.9, 0.8	40.8901	44.9226	0.6, 0.9, 0.8	40.8901	38.0720	0.6, 0.9, 0.8
0.9, 0.8, 0.8	40.5172	37.7860	0.7, 0.7, 0.8	40.5172	37.7411	0.5, 0.9, 0.8	39.7495	38.8418	0.7, 1.0, 0.8	41.4424	43.7979	0.7, 1.0, 0.8	41.4424	38.8418	0.7, 1.0, 0.8
0.9, 0.9, 0.8	40.3097	37.6383	0.7, 0.9, 0.8	40.3389	37.0238	0.6, 0.8, 0.8	39.4183	37.9601	0.7, 0.8, 0.8	40.8598	44.3152	0.7, 0.8, 0.8	40.8598	37.9601	0.7, 0.8, 0.8
0.9, 1.0, 0.8	40.3184	37.6173	0.7, 0.8, 0.9	40.7823	38.4516	0.6, 1.0, 0.8	39.1401	38.1964	0.7, 0.7, 0.8	40.9488	44.7300	0.7, 0.7, 0.8	40.9488	38.1964	0.7, 0.7, 0.8
0.9, 0.9, 0.7	40.5537	37.5104	0.7, 0.8, 0.7	40.1696	36.4351	0.6, 0.9, 0.9	40.1263	38.6208	0.7, 0.9, 0.9	41.4700	43.7533	0.7, 0.9, 0.9	41.4700	38.6208	0.7, 0.9, 0.9
0.9, 0.9, 0.9	40.7879	37.6167	0.7, 0.8, 0.6	40.2189	36.7940	0.6, 0.9, 0.7	38.6978	37.7130	0.7, 0.9, 0.7	40.2058	44.5816	0.7, 0.9, 0.7	40.2058	37.7130	0.7, 0.9, 0.7
0.7, 0.9, 0.7	41.4070	37.0017				0.6, 0.9, 0.6	38.7996	37.7797	0.7, 0.9, 0.6	40.3276	45.5583	0.7, 0.9, 0.6	40.3276	37.7797	0.7, 0.9, 0.6

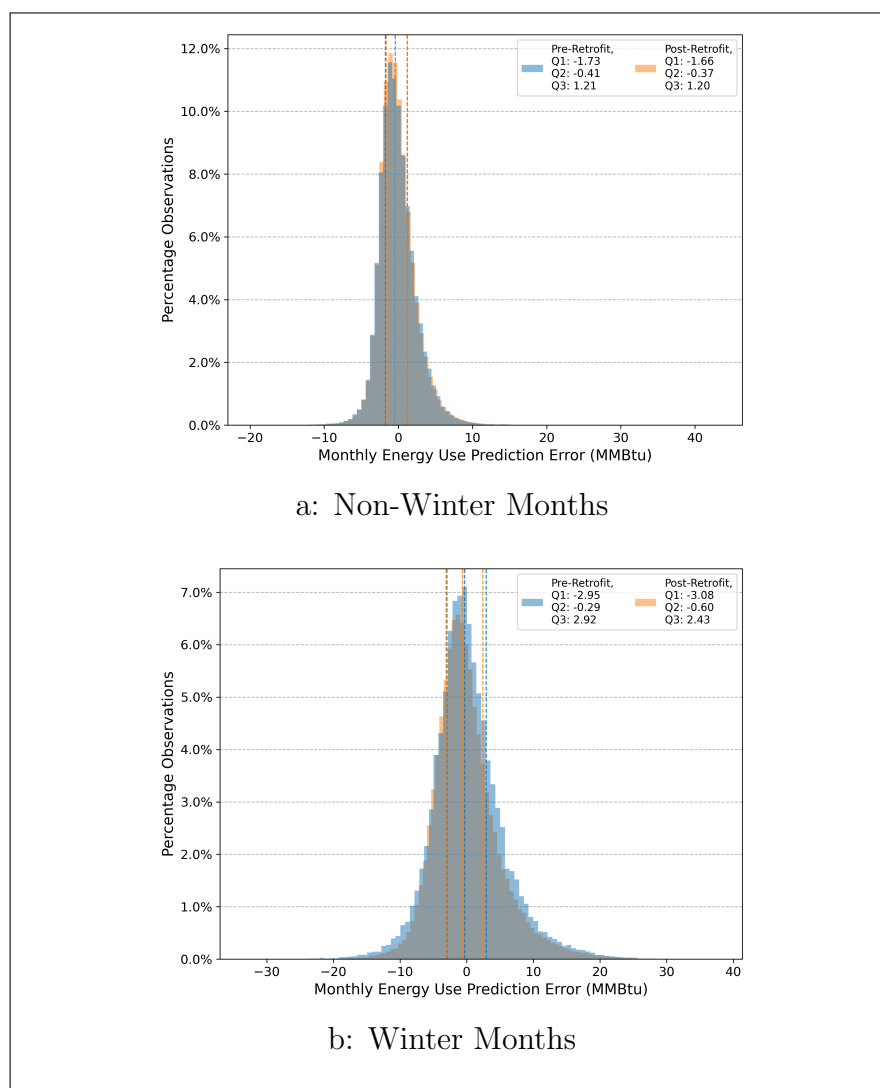
Notes: This table presents results from the nested cross-validation procedure for neural network models. The first three columns present results for the iteration with Fold 1 as the test set (i.e. for that iteration, no data from Fold 1 was used for training the models). Columns 4 through 6 present results for when Fold 2 is the test set, and so on. We present both Validation Set Mean Squared Error (MSE) and Test Set MSE for estimating the treatment effect (i.e. comparing the benchmark treatment effect from the ex-post model with treatment effects according to the ex-ante models). All models have three hidden layers with 64, 32, and 16 neurons, respectively (as illustrated in Figure C.1). The first column from each corresponding fold shows the regularizer that was used in each layer. The rows in gray highlight the model that was selected for each iteration of the nested cross-validation procedure, based on the validation set MSE.

Table C.2: Cross-Validation Results for Other Algorithms

Panel C: Performance of Other Algorithms						
Model ID	Model Type	Model Parameter	MSE (Treatment Effect)	MSE (MMBtu)	MSE (Pre, MMBtu)	MSE (Post, MMBtu)
1	GradientBoosting	boosting stages = 100, subsample = 0.8	47.308	12.652	13.804	12.383
2	GradientBoosting	boosting stages = 120, subsample = 0.8	44.685	12.526	13.639	12.266
3	RandomForest	number of trees = 20	72.722	13.323	14.336	13.086
4	RandomForest	number of trees = 30, max_depth = 4	279.469	16.171	19.468	15.400
5	RandomForest	number of trees = 30	67.992	13.122	14.078	12.898
6	Lasso	alpha = 1	365.241	19.466	25.483	18.059
7	Lasso	alpha = 0.1	109.867	15.717	18.252	15.124
8	Lasso	alpha = 0.01	58.700	14.610	16.949	14.063
9	Lasso	alpha = 0.005	59.786	14.508	16.838	13.963

Notes: This table presents results from 5-fold cross-validation for several machine learning algorithms considered in this study.

Figure C.2: Distribution of Pre- and Post-Retrofit Prediction Errors



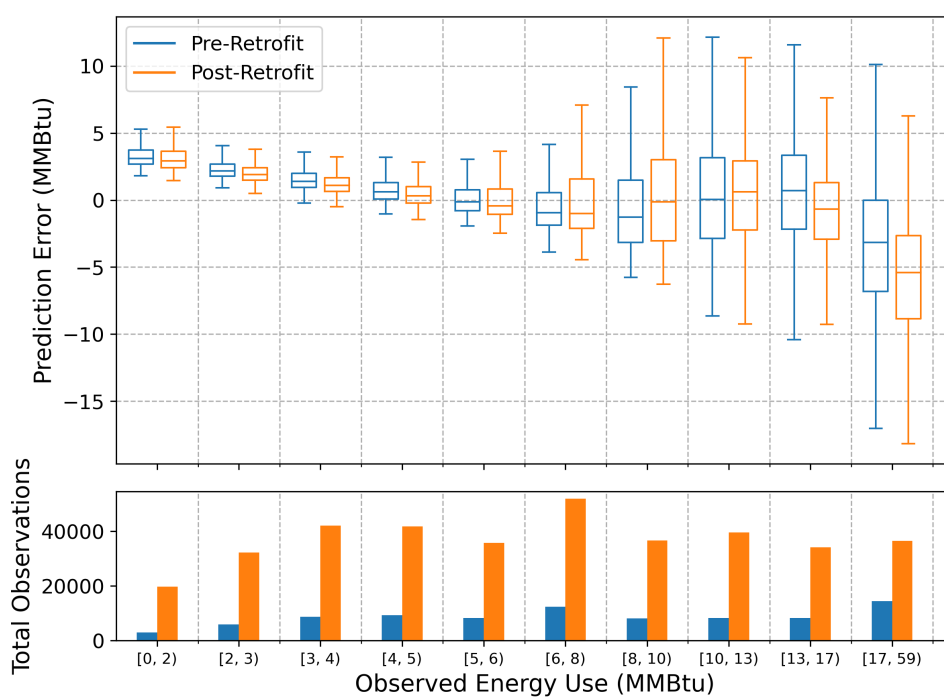
Notes: This figure presents validation set prediction errors based on household by month observations. Winter months are defined as November through March.

C.3 Analysis of Out-of-Sample Prediction Errors

Figure C.2 presents the distributions of errors for predicting energy usage both before and after the retrofits. It is clear that the errors are centered around zero, both for winter and non-winter months, such that biases along that dimension are unlikely. Absolute errors are less than 2 MMBtu for more than 75% of non-winter months, and less than 3 MMBtu for more than 75% of winter months.

Figure C.3 plots prediction errors by bins of observed usage. This allows for a more precise identification of regions of the sample that may have biased predictions. The top panel presents the errors for each bin, while the bottom panel presents the number of observations in each corresponding bin. The selected algorithm overestimates energy usage for months at the lower end (i.e. months with actual usage lower than 3 MMBtu). That is expected, as those could constitute outlier months for which households were not occupying their residences (e.g. travel or vacation). The selected algorithm should not be accurate for predicting outlier observations, which would suggest overfitting. Also, bias operates in the same directions for both pre- and post-retrofit observations, such that they cancel and reduce bias in the residuals of estimated treatment effects.

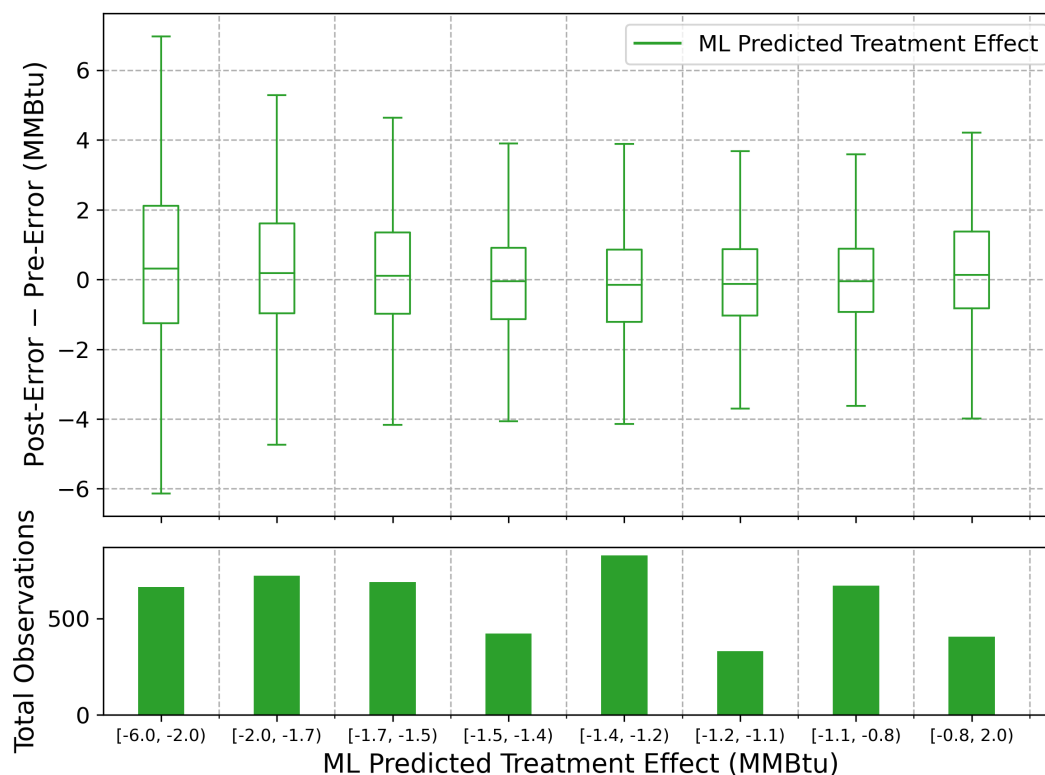
Figure C.3: Usage Prediction Errors (MMBtu) by Bins of Observed Usage



Notes: The top panel presents validation-set prediction errors, based on household by month observations. The x-axis represents bins of observed energy use. The bottom panel presents the number of observations each corresponding bin.

To further assess potential bias, Figure C.4 plots the per-home differences between pre- and post-retrofit prediction errors. Those are plotted across several bins of estimated treatment effects. The figure reveals that the pre- and post-retrofit predictions errors are highly correlated, such that their difference is close to zero, on average. That result holds across all bins of estimated treatment effects.

Figure C.4: Treatment Effect Prediction Errors (MMBtu) by Bins



Notes: The top panel presents the difference between pre- and post-retrofit prediction errors in the validation folds, based on household level observations. The x-axis represents bins of predicted treatment effect. The bottom panel presents the number of households each corresponding bin.

D Ex-Post Evaluation Method

For estimation of ex-post program savings, this paper implements a machine learning-based event study approach. That is a frontier method which has been used for recent impact evaluations, especially of energy efficiency programs (Burlig et al., 2020; Christensen et al., 2021). The advantage of that method, compared to traditional regressions, is that it allows more precise estimation of the program effects for each treated home in the sample. Traditional approaches, such as fixed effects regressions, were designed to estimate average effects, thus fail to capture heterogeneity. It has also been shown that while fixed effects regressions can result in short-term biased estimates in the presence of heterogeneity (Chaisemartin and D’Haultfoeuille, 2020; Goodman-Bacon, 2021), the machine learning-based approach does not (Souza, 2019).

The first step of the approach is to predict counterfactual energy consumption in absence of the retrofits. Then, similar to equation (1) from the main text, energy savings are estimated as:

$$b_{it}^{EP} = Y_{it}(1) - \hat{Y}_{it}(0) , \quad (\text{D.1})$$

where b_{it}^{EP} is an *ex-post* estimation of the reduction in energy use resulting from energy efficiency retrofits for home i in month t , $Y_{it}(1)$ is *observed* energy use in the presence of treatment, and $\hat{Y}_{it}(0)$ is a *counterfactual* prediction of energy use in the absence of treatment. Note that since post-treatment data are available, in this case it is only necessary to predict counterfactual energy consumption $\hat{Y}_{it}(0)$ in absence of treatment.

The ex-post savings used in this paper are taken directly from Christensen et al. (2021), who employ Gradient Boosted Trees for counterfactual predictions. Once initial estimates of b_{it}^{EP} are obtained, it is then possible to add structure based on knowledge on how the program operates and on which type of retrofits were implemented in each home. For that, a second-step regression is implemented as follows:

$$b_{it}^{EP} = \boldsymbol{\gamma}\mathbf{X}_{it} + \varepsilon_{it} , \quad (\text{D.2})$$

where \mathbf{X}_{it} is a vector including information about demographics, housing structure, pro-

gram costs, a constant, and an idiosyncratic error term ε_{it} . Simulations demonstrate the improved performance of implementing that two-step approach (Souza, 2019). Finally, predictions obtained from the model described in equation (D.2) are aggregated in order to represent a home’s annual energy savings attributable to the retrofits:

$$\hat{b}_i^{EP} = \sum_{t=jan}^{dec} \hat{b}_{it}^{EP} , \quad (D.3)$$

where \hat{b}_{it}^{EP} represents the average predicted savings for home i in a given month of the year (January through December); and \hat{b}_i^{EP} is the ex-post estimate of annual energy savings for home i .

Note that ex-post savings obtained in this way are robust to potential confounders, such as changes in weather patterns or the macroeconomic context before and after the retrofits. Other ex-post methods that rely on simple comparisons between pre- and post-retrofit usage and that do not incorporate a causal framework are likely to produce biased estimates of savings for several homes, thus producing an inaccurate ranking of cost-effectiveness. This paper therefore focuses on the frontier ex-post model described above as the most valid benchmark for ex-ante models.

E Details on Net Present Benefits

E.1 Assumptions

The ML approach described in the main text provides estimates of annual combined energy savings $\hat{\beta}_i$, rather than gas and electricity savings separately. We disaggregated savings by assuming that 17% of those are attributable to electricity consumption and 83% to natural gas, based on prior literature (Christensen et al., 2021). An annual discount rate of $r = 3\%$ is assumed throughout, as recommended by Department of Energy (DOE) for evaluation of governmental programs (Rushing, Kneifel, and Lippiatt, 2012).

Results from the main text incorporate energy costs that account for the social costs of carbon emissions in the energy sector (Borenstein and Bushnell, 2018; Davis and Muehlegger, 2010), although the program that we used for this analysis uses private benefits due to the nature of program goals. Following prior literature, citygate natural gas prices were used for the marginal private costs of gas (to which social costs of carbon were added) based on emissions factors from the US Environmental Protection Agency (EPA, 1998). Estimates assume a price of \$40 per ton for CO₂ emissions. Based on a similar approach, prior literature provides estimates of the social marginal costs of electricity for the state of Illinois (Borenstein and Bushnell, 2018). The resulting energy prices that incorporate costs of carbon were \$6.74 and \$33.95 per MMBtu for natural gas and electricity, respectively.

Appendix E.4 evaluates results using an alternative approach that calculated energy costs using the average residential energy prices for Illinois over 2009-2016. Those were obtained from the US Energy Information Administration (EIA, 2017b), resulting in \$8.32 and \$34.26 per MMBtu for natural gas and electricity, respectively. Price escalation was applied to project future energy prices, also based on recommendations from the DOE (Rushing, Kneifel, and Lippiatt, 2012). Results from this approach reflect the private benefits to consumers that face reduced energy bills thanks to the retrofits. Note that those prices are slightly higher than the ones that incorporate social costs of carbon as a result of energy taxes in Illinois. Figures from Appendix E.4 show that results from this

study are robust to assumptions regarding energy prices.

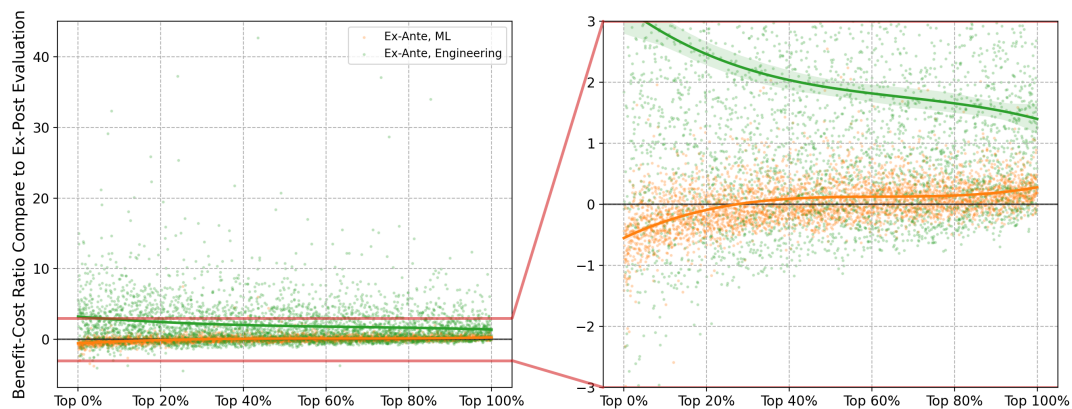
Different types of retrofits installed by the program might have different expected lifespans. To account for these differences, the per-home expected lifespan T_i is calculated as a weighted average based on expenditures across retrofits. For example, a home with expenditures predominantly on wall insulation will have a relatively higher expected lifespan (closer to 25 years) than a home with expenditures predominantly on the water heater (closer to 15 years). Retrofit-specific lifespan recommendations from WeatherWorks documentation suggest home-specific lifespans are around 20 years on average. However, updated estimates from recent engineering literature suggest that some insulation measures may have up to 50-year lifespans (Kono et al., 2016). Retrofit lifespans used in preferred estimates account for longer longevity of insulation measures, resulting in home-specific lifespans of approximately 30 years. Figures from E.3 show that results from this study are robust to assumptions regarding retrofit lifespans and discount rates.

E.2 Comparison of Cost-Effectiveness Predictions Across Models

This section compares ex-ante estimated benefit-cost ratios with those obtained from the ex-post approach. Figure E.1 plots benefit-cost ratio prediction errors for the ex-ante ML method (orange) and for the ex-ante engineering approach (green). Each point represents the error for a given home, while the lines represent cubic fits. Errors are sorted based on the cost-effectiveness ranking from the ex-post model. The panel on the right zooms in to illustrate the most relevant regions for the sample. The errors from the ML approach are substantially lower than those from the engineering model. The majority of ML absolute errors are lower than 1 and the cubic fit curve ranges from -0.5 to 0.2. Errors from the engineering approach are several times larger, and could be over 30 for some homes. Errors from the engineering model systematically *overestimate* savings. This stark discrepancy may partially explain the poor performance of traditional engineering approaches for targeting funding (Figure 4 from the main text). The mean squared relative error of the benefit-cost ratios is 272.1 for the ML approach, and 686.6

(more than 2 times larger) for the engineering approach.

Figure E.1: Household Benefit-Cost Ratio Prediction Errors Ranked by Ex-Post Evaluation



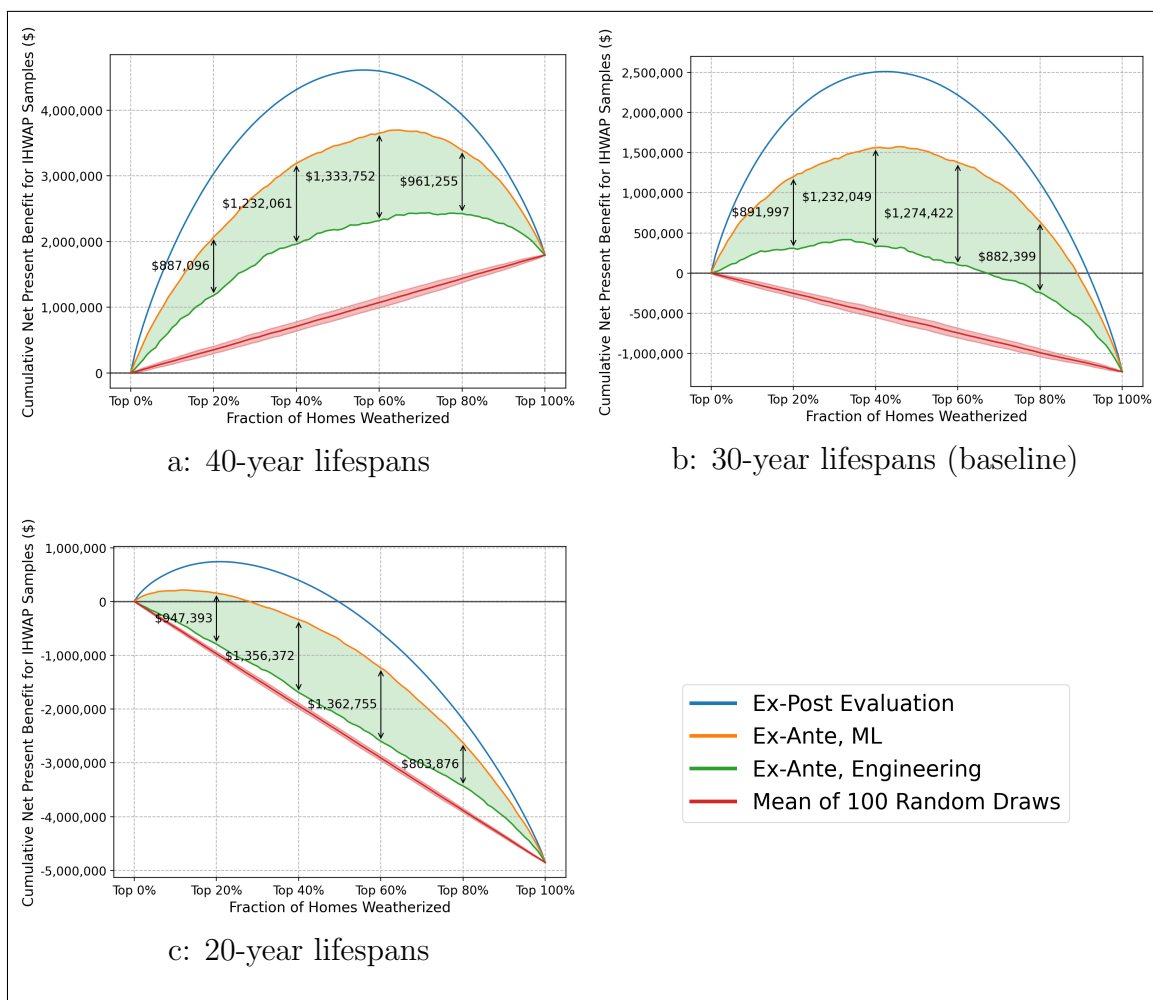
Notes: This figure compares errors in benefit-cost ratios (BCR) generated by the ex-ante ML and engineering approaches. The errors are sorted by the ex-post evaluation rank. BCR are calculated using 30-year retrofit lifespans, a 3% discount rate, and incorporating the social cost of carbon. The dots represent errors for a given home, while the lines represent cubic fits.

E.3 Sensitivity to Lifespan and Discount Rate Assumptions

This section presents results for cumulative net present benefits and benefit-cost ratios, with varying assumptions regarding retrofit lifespans and discount rates. The objective is to analyze the sensitivity of the main findings reported in the study to these parameter assumptions. In the main text, the assumed lifespan is close to 30 years and the discount rate is 3%. Figure E.2 presents results with varying retrofit lifespans, holding the discount rate at 3%. Panel (a) suggests that the program results in almost \$2 million in net benefits if the retrofits are assumed to have 40-year lifespans. However, with 20-year lifespans (Panel c) the program is associated with almost \$5 million in losses. Nevertheless, the ranking of homes remains stable across lifespan assumptions. Importantly, note that the gains from the ex-ante ML ranking (compared to the engineering approach) are substantial in all three panels, ranging from \$1.23 to \$1.36 million for the top 40% homes. The implication is that targeting with ML methods will have similarly beneficial impacts even if overall program cost-effectiveness is low.

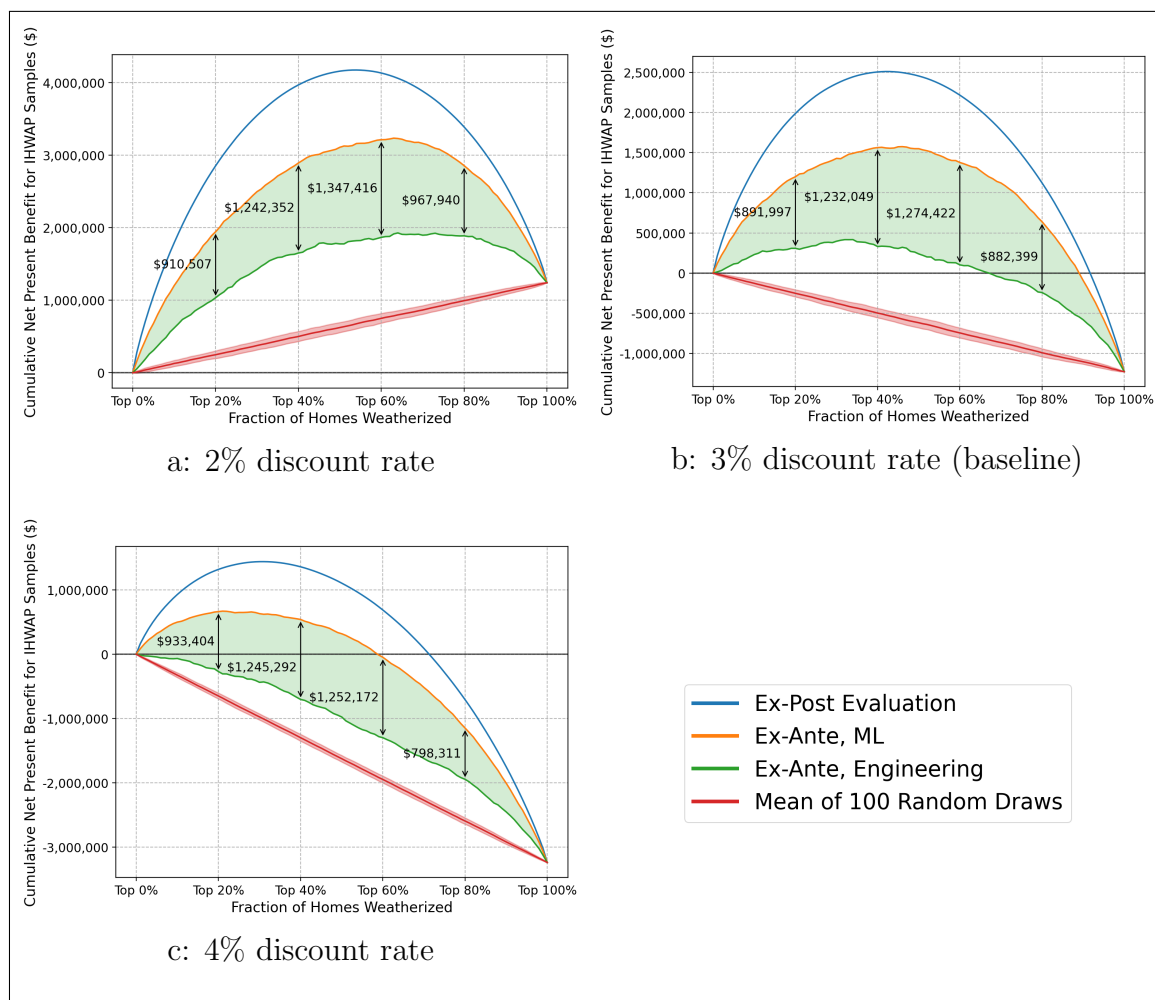
Similarly, Figure E.3 presents results with varying discounts rates, but holding lifespans at 30 years. Again, the program's overall cost-effectiveness varies significantly, depending on assumed discount rates. Cumulative net benefits are reduced by almost \$4.5 million when moving from a 2% to a 4% discount rate. Nevertheless, the conclusions from the main text still hold: rank-distributions produced by the ex-ante ML approach lead to substantial gains. The gains from ML modeling are strikingly similar regardless of discount rate assumptions.

Figure E.2: Cumulative Net Present Benefits, Varying Retrofit Lifespans



Notes: This figure compares presents cumulative net present benefits according to the different models, and with varying retrofit lifespan assumptions. Panel (b) are the results with baseline assumptions (30-year lifespans, and 3% discount rate). Panel (a) presents results with lifespans increased to 40 years, while Panel (c) is for reduced lifespans of 20 years.

Figure E.3: Cumulative Net Present Benefits, Varying Discount Rates



Notes: This figure compares presents cumulative net present benefits according to the different models, and with varying discount rates. Panel (b) are the results with baseline assumptions (30-year lifespan, and 3% discount rate). Panel (a) presents results with the discount rate reduced to 2%, while Panel (c) is for an increased discount rate of 4%.

Table E.1: Max Benefit-Cost Ratios, Varying Lifespan and Discount Rate Assumptions

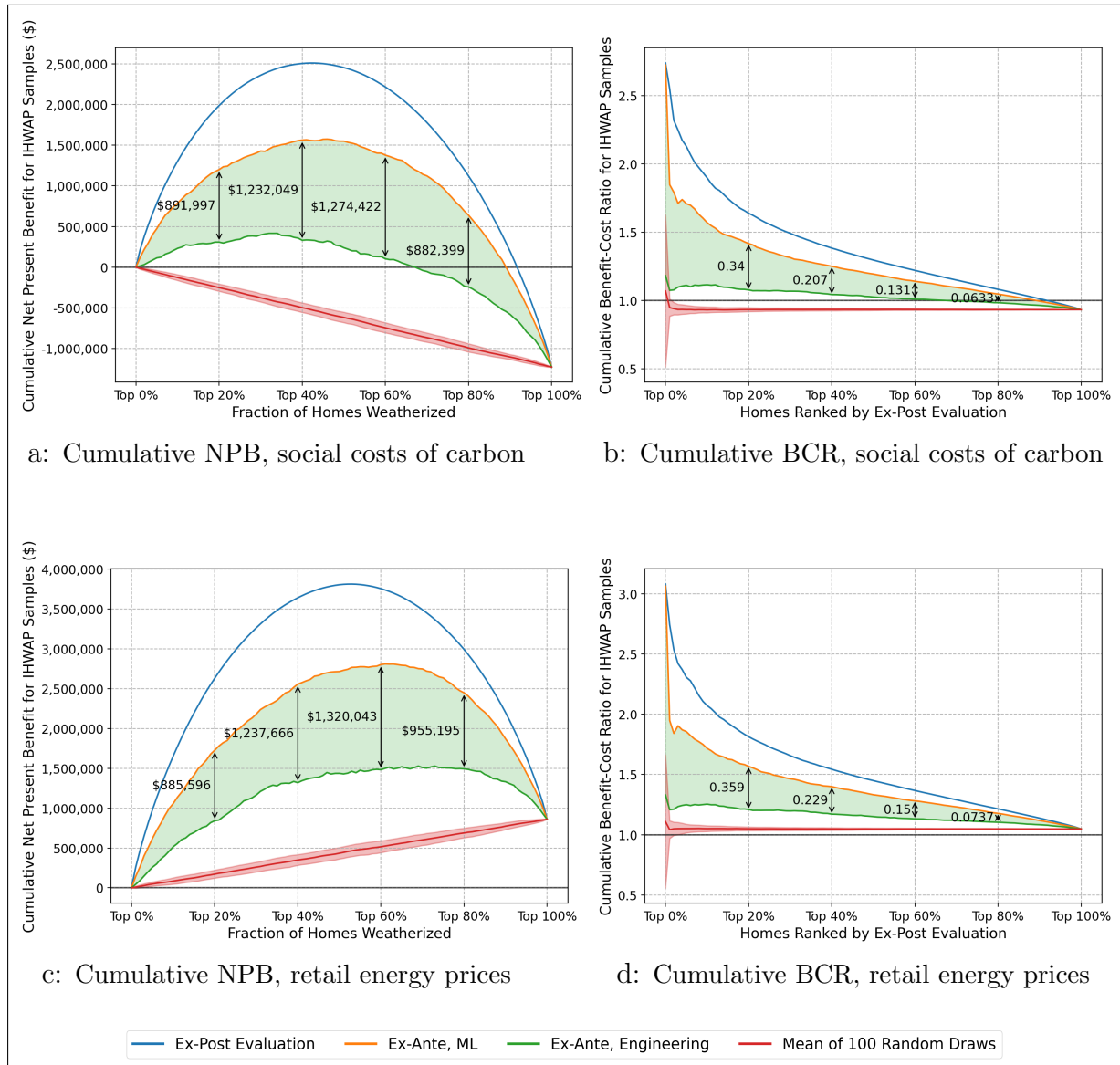
	30-year lifespan, varying discount rates			3% discount rate, varying lifespans		
	2%	3%	4%	40 years	30 years	20 years
Full sample BCR	1.07	0.93	0.82	1.10	0.93	0.73
Max BCR, ex-post approach	1.46	1.36	1.31	1.48	1.36	1.25
Gains from ex-post targeting	0.39	0.43	0.49	0.38	0.43	0.52
Max BCR, ex-ante ML approach	1.34	1.23	1.14	1.37	1.23	1.06
Gains from ex-ante targeting	0.28	0.30	0.32	0.27	0.30	0.33

Notes: This table presents the sensitivity of estimates of benefit-cost ratios (BCR) to varying assumptions of retrofit lifespans and discount rates. We present BCR for treating homes in order of net benefits, up to the maximum points (i.e. when marginal benefits equal marginal costs) according to each approach and across scenarios. For the final two rows, the ex-ante ML approach serves to produce a ranking of homes, while “true” net benefits are assessed based on the ex-post estimates.

E.4 Private versus Social Benefits

Results from the main text incorporate the social cost of carbon to the net present benefit calculations, as described in detail in the Methods section. An alternative approach is to consider only the private benefits to the households served by the program, in which case retail energy prices should be used. Results with retail energy prices are presented in Figure E.4, panels (c) and (d). Results with baseline assumptions (panels a and b) are presented for ease of comparison. Note that the overall program cost-effectiveness increases when considering only the program's private benefits to consumers. That is because retail energy prices in the state of Illinois are higher than prices when accounting for marginal production costs plus the social costs of carbon. That is most likely due to the energy taxation policies in the state. Nevertheless, the gains from using the ML approach remain similar comparing social versus private benefits: for the top 40% homes, the gains are \$1.232 versus \$1.237 million in net present benefits; 0.207 versus 0.229 in benefit-cost ratios.

Figure E.4: Social Cost compare to Retail Energy Prices



Notes: This figure compares presents cumulative cost-effectiveness according to the different models, and with varying assumptions regarding energy prices. Panels (a) and (b) present the same results as in Figure 4 from the main text, incorporating the social cost of carbon. Panels (c) and (d) present cumulative net present benefits and cumulative benefit-cost ratios using retail energy prices, thus representing private benefits to the consumers. Note that retail energy prices in Illinois are higher than prices constituted of marginal production costs plus the social costs of carbon.