Data-Driven Decision Support for Building Retrofit Planning

Milan Jain, Chitra Nambiar, Xuechen Lei, Pacific Northwest National Laboratory
Supriya Goel, West Monroe

ABSTRACT

For building owners and facility managers, existing buildings represent a near-term opportunity to cut energy use and operating costs through targeted efficiency retrofits—often on the order of ~10–30% depending on building type and measures. For utilities and program implementers, these upgrades can also reduce peak demand and help manage growing electricity loads. However, retrofit decision-making requires tools that are not only accurate but also transparent and interpretable. Studies have shown that advanced machine learning (ML) methods can identify retrofit opportunities, yet their "black box" nature often limits adoption among building owners and program implementers.

This study presents an interpretable ML framework that combines tree-based models with explainable artificial intelligence (XAI) to support building-level and portfolio-level retrofit screening. Using the ComStock dataset and a wide range of building attributes, the models predict energy use intensity (EUI) and apply Shapley additive explanations (SHAP)-based explanations to identify key performance drivers. Envelope characteristics and cooling system characteristics emerge as dominant factors, with effects that vary across commercial building types. Counterfactual simulations of LED lighting and window upgrades demonstrate the framework's ability to rapidly estimate retrofit impacts, showing strong directional agreement with physics-based upgrade simulations.

The framework is designed to complement—not replace—detailed engineering models by enabling fast, scalable early-stage prioritization of retrofit measures. Although current limitations include reliance on learned correlations, sensitivity in feature attributions, lack of integrated cost metrics, and unconstrained counterfactual assumptions, the approach advances transparent, data-driven decision support. By linking accurate predictions to clear explanations, this work provides a portfolio-scale screening framework that uses building-stock data to rank retrofit opportunities, quantifies expected energy savings from adopting those retrofits, thereby providing a replicable, practical model for strategic energy retrofit planning.

Introduction

Rising energy prices and aging equipment are increasing pressure on existing-building owners to find opportunities to reduce operating costs. Energy efficiency programs recommend identifying energy retrofit opportunities by beginning with benchmarking against others of a similar type as an initial step to assess the performance of existing buildings. In this process, a building's energy consumption is measured and compared either to its historical use or to that of similar buildings. This process facilitates peer comparisons, uncovers operational inefficiencies, and guides the development of specific energy-reduction strategies. Data-driven benchmarking approaches, including statistical modeling and advanced analytics techniques, offer robust tools for characterizing building energy performance across multiple temporal scales (annual, monthly, daily, and hourly) and at increasing spatial resolutions, from individual buildings to portfolios and campus-scale systems (Wei et al. 2018; Li et al. 2024).

However, while benchmarking is essential for performance assessment and peer comparison, it represents only the initial step toward achieving realized energy and cost savings. Translating benchmarking insights into measurable reductions requires detailed building assessments aligned with energy savings goals, cost savings goals, and long-term capital investment strategies. These assessments are typically conducted through on-site energy audits which are inherently time- and resource-intensive- limiting their scalability across large building portfolios and constraining widespread implementation.

Machine learning (ML) is increasingly gaining traction as an intermediate step between benchmarking and on-site energy audits. ML algorithms can reduce the cost and effort of detailed assessments by leveraging real-world or synthetic datasets to identify retrofit opportunities and estimate energy savings. These data-driven techniques complement traditional physics-based simulation methods by providing an early-stage screening layer for portfolio-level benchmarking. In doing so, they identify buildings with the greatest energy-saving potential, thereby helping prioritize buildings for more detailed physics-based analyses and enabling more efficient allocation of audit resources. Additionally, at the individual building level, ML-based frameworks can help identify and prioritize retrofit measures with the greatest energy-saving potential.

This paper builds on the current state of the art in ML applications for retrofit assessment and proposes a novel capability that integrates building characteristics with ML and Explainable Artificial Intelligence (XAI) techniques to identify retrofit potential and quantify associated energy savings. Building on prior work (Jain, Goel, and Lei 2026), which used energy audit data submitted through the U.S. Department of Energy's Audit Template tool (Goel et al. 2022), this study extends the methodology by applying it to large-scale building stock data derived from the U.S. DOE Comstock dataset (Parker et al. 2023). Comstock (Commercial Building Stock) is a publicly available, physics-based, synthetic dataset developed by the U.S. Department of Energy that represents the energy performance of the U.S. commercial building stock across diverse building types, vintages, locations, and operational characteristics. It provides high-resolution simulation outputs and building attribute data that enables large-scale analysis of energy use patterns, technology impacts, and retrofit pathways.

We apply four machine learning algorithms to the publicly available ComStock data and evaluate their ability to build reliable predictive models. The best (trained) model is then used to assess pre-retrofit building energy performance. The XAI layer provides post-hoc explanations of these predictions, highlighting the key factors driving energy consumption at the individual-building level. Next, the model is applied to simulate post-retrofit scenarios, allowing for a comparative assessment of energy performance before and after the retrofits.

Key contributions of this paper are:

- Model development: ML methods to utilize ComStock data for identifying energy retrofit opportunities in existing building stock.
- From Explanation to Quantification: Integrating SHAP with counterfactual simulations to directly estimate energy savings from targeted retrofits, showcasing an extension of its application from simple feature importance prediction.
- Portfolio-Level Decision Support: Demonstrating how XAI frameworks can support energy efficiency objectives by identifying retrofit potential across diverse building stocks.

Background

Benchmarking and energy prediction play distinct but complementary roles in advancing high-performance commercial buildings. Benchmarking provides a comparative framework, situating a building's energy use relative to peers, historical performance, or modeled expectations after normalizing factors such as weather, occupancy, and operating characteristics. Building energy usage prediction, in contrast, estimates expected future consumption and is typically used to inform new construction, major renovations, incentive program participation, and other market activities.

Benchmarking has scaled through platforms such as the ENERGY STAR® Portfolio Manager, which uses multivariate regression models, with weather and operational normalization, to generate 1–100 performance scores for millions of commercial buildings nationwide (EPA 2023). While traditional linear regression and physics-based simulation models remain dominant, recent research and industry efforts have explored more flexible, data-driven approaches. ML methods—including gradient boosted trees—have demonstrated improved predictive accuracy when paired with interpretability techniques that enhance transparency and user trust (Arjunan, Poolla, and Miller 2020). Emerging trends emphasize automation and the integration of feature attribution methods to translate benchmarking results into actionable insights for building owners and supporting industry stakeholders.

In parallel, retrofit assessment has expanded but remains largely focused on energy performance evaluations. A recent systematic review of envelope retrofit strategies found that most studies evaluate interventions primarily through modeled energy savings and thermal performance outcomes (Dabous and Hosny 2025). Other scholars have noted the importance of incorporating life-cycle economic analysis and broader decision dimensions, noting that differences in system boundaries, time horizons, and data availability substantially influence reported results and limit cross-study comparability (Toosi et al. 2020). Reviews of residential retrofit decision-support tools similarly conclude that most platforms prioritize energy and financial analyses, frequently relying on scenario modeling and simulation-based approaches (Gonzalez-Caceres, Rabani, and Martínez 2019).

Jain, Goel, and Lei (2026) used audit records from U.S. DOE's Audit Template tool to link building attributes to energy use and to estimate savings from efficiency upgrades using SHAP integrated with counterfactual analysis in retrofit planning. However, data limitations and potential data-entry errors in the audit template dataset reinforced the need to expand the approach to larger feature-rich datasets. The current work, using the ComStock dataset, addresses these limitations observed in the previous study. The paper presents the framework we developed using ComStock data to identify the contributions of building-level attributes to energy performance and post-retrofit savings, enabling the adoption of holistic approaches to retrofit decision planning.

Methodology

This study utilizes a building-level modeling framework on the ComStock 2025_3 dataset (Parker et al. 2023), a simulated representation of the U.S. commercial building stock spanning multiple locations and building types. The ComStock dataset includes detailed assumptions about building characteristics, including envelope properties, HVAC systems, lighting technologies, service hot water systems, and annual energy consumption by fuel type and end use. The nationwide ComStock data consists of 149,272 buildings, each assigned a weight representing the number of real-world buildings corresponding to that configuration. Each record is associated with a unique building identifier (bldg_id) that reflects the corresponding simulation model. For ML model development, we randomly divided the datasets into 80% for training, 10% for validation, and 10% for testing. All models were trained to minimize mean squared error (MSE), and hyperparameters were optimized using Optuna (Akiba et al. 2019) on the training and validation datasets. Model performance was subsequently evaluated on the held-out test data using root-mean-squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2).

The framework proposed in this study is intended to enable an understanding of how building-level attributes contribute to energy performance and the potential for retrofit energy savings. The modeling framework involves three key stages: (i.) predicting building-level annual site energy use intensity (EUI); (ii.) identifying the most influential characteristics of driving energy intensity performance; and (iii.) estimating the energy impact of plausible retrofit scenarios. Unlike traditional benchmarking approaches that focus solely on energy performance, this methodology couples ML modeling and interpretability with counterfactual modeling to quantify the expected energy impacts of targeted retrofits. Adding predictive analytics empowers stakeholders to rapidly evaluate and prioritize retrofit pathways before committing to complex and detailed physical simulations.

I. Predicting building annual site energy use intensity (EUI)

In this stage, we train supervised machine learning models to estimate annual building EUI from building characteristics. Prior to model development, we conducted a structured review of the ComStock data dictionary and variable distributions to identify features most strongly associated with energy consumption and most relevant from a retrofit perspective (see Table 1).

The ComStock dataset consists largely of discrete building descriptors; therefore, we focused on tree-based algorithms that natively handle non-linear relationships and interactions in structured tabular data. Specifically, we evaluated gradient-boosted decision tree methods, including extreme gradient boosting (XGBoost) (Chen and Guestrin 2016), categorical boosting (CatBoost) (Prokhorenkova et al. 2018), and light gradient boosting machine (LightGBM) (Ke et al. 2017).

Prior to model training, categorical variables are encoded using ordinal encoding to convert discrete categories into numeric representations compatible with the learning algorithms. Continuous variables were normalized to place them on comparable scales, and missing values were imputed to ensure consistent model inputs. Hyperparameters governing tree depth, learning rate, and ensemble size were tuned using Optuna, a Bayesian optimization framework that efficiently searches the hyperparameter space to minimize prediction error. We then evaluate model performance using out-of-sample error metrics to ensure that predictive accuracy generalizes across building types.

Table 1. Extracted ComStock features for modeling and counterfactual analysis

Group	Feature
Energy Goal	Building EUI (kWh/ft^2)

Lighting	Lighting technology
Wall	Construction type, Avg. U-value (Btu/ft^2-hr), Exterior wall area (m^2)
Window	Window type, Exterior window area (m^2), Window-to-wall ratio
Roof	Roof absorptance, Avg. U-value (Btu/ft^2-hr), Exterior roof area (m^2)
Cooling	Chiller design COP, DX cooling capacity (tons), DX cooling design COP
Heating	DX heating min. operating temp (C), DX heating capacity (kBtu/hr), DX heating design COP, Hot water loop boiler fraction, Hot water loop load (Joules)

II. Identifying key drivers of energy use intensity

To determine which building characteristics most strongly influence predicted EUI, we applied SHAP to the best-performing model. SHAP (Lundberg and Lee 2017) is a post hoc interpretability method grounded in cooperative game theory that attributes a model's prediction to individual input features. For each building, SHAP quantifies how much each characteristic increases or decreases the predicted EUI relative to the dataset average, thus producing an additive breakdown of the model output.

This framework supports both stock and building-level analysis. At the stock level, features are ranked by the mean absolute SHAP value across all buildings, reflecting their overall contribution to model predictions. At the building level, SHAP values explain why a specific feature is predicted to have relatively high or low energy intensity by identifying the features driving that deviation.

It is important to note that SHAP explains model behavior, not physical causality. The identified drivers represent influential predictors within the trained statistical model and are used to prioritize candidate retrofit levers.

III. Counterfactual retrofit simulation

For features identified as influential, we construct counterfactual retrofit scenarios by modifying input variables within realistic engineering ranges. These modifications represent plausible efficiency upgrades, such as increasing HVAC system efficiency to high-performance levels, reducing infiltration rates, converting lighting systems to LED technologies, and improving envelope thermal performance, among others. The ranges for these adjustments were grounded in energy code-based or best-practice retrofit values observed in commercial buildings. Further details regarding upgrades are available in the ComStock documentation (Parker et al. 2023).

For each scenario, the modified building characteristics are fed into the trained model to predict the post-retrofit EUI. Potential energy savings are then computed as the difference between the baseline-predicted EUI and the post-retrofit predicted EUI.

We use this counterfactual framework to conduct building-level and stock-level analysis. At the building level, it quantifies the expected energy savings from each retrofit for an individual property, allowing identification of the most effective measures given its current configuration. At the stock level, it evaluates how a uniform retrofit program applied across a portfolio would influence the energy performance of the building stock. By aggregating predicted savings across the ComStock population, we assess how consistent upgrade strategies perform at scale and how their impacts vary across building types.

Results

Table 2 depicts the prediction performance of the trained models on the held-out test set. Among the evaluated approaches, CatBoost achieved the lowest RMSE (10.34) and the highest R^2 value (0.93). XGBoost achieved the lowest MAE (3.65) and reached R^2 of 0.93. LightGBM performed slightly lower, with an RMSE of 11.26, MAE of 4.55, and R^2 of 0.92.

Sensitivity analysis showed marginal improvements in predictive accuracy from switching from ordinal to one-hot encoding and from log-transforming the target variable. The results presented here are based on one-hot encoding with the log-normal target variable.

Table 2. Prediction performance of the trained models

Model Type	RMSE	MAE	R^2
XGBoost	10.56	3.65	0.93
CatBoost	10.34	4.78	0.93
LightGBM	11.26	4.55	0.92

The relatively consistent performance across tree-based models suggests that predictive accuracy is not sensitive to a specific algorithm. An R^2 value close to 0.9 indicates that the model accounts for a large share of differences in EUI across buildings using the feature set (mentioned in Table 1). This level of explanatory power also suggests that the chosen predictors capture much of the systematic variation in energy use intensity. Next, we discuss SHAP analysis, followed by counterfactual studies.

Key drivers of commercial building energy use

Figure 1 presents mean SHAP values for key features across major building types. Bar color indicates direction: red bars represent positive contributions to predicted EUI (i.e., an increase in energy use), while blue bars represent negative contributions (i.e., a decrease in energy use). Each bar shows whether that feature is associated with higher or lower energy use for that specific building type, and by how much on average. Overall, the results show that both envelope characteristics and cooling system parameters are dominant features that impact building EUI, with magnitudes and signs of effect varying across building types.

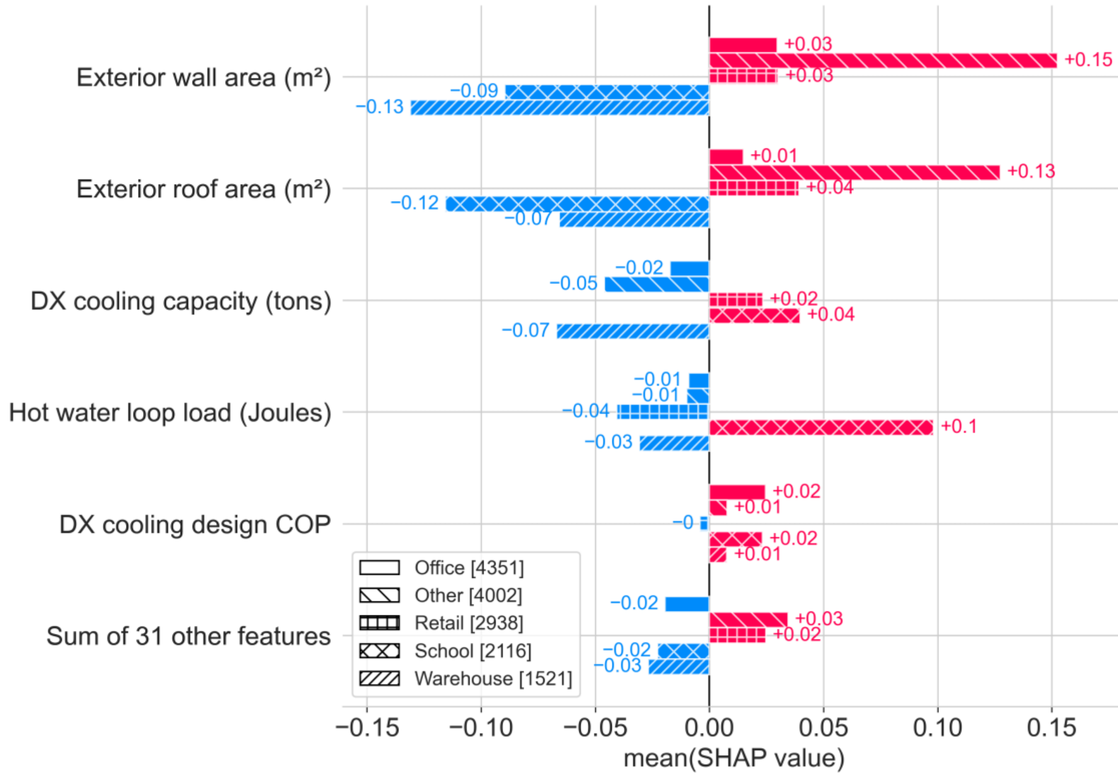

Figure 1: Mean SHAP values by building type for the best model

One of the primary drivers is envelope characteristics, including exterior walls and roofs. The plot shows that the exterior wall area has a positive mean SHAP value for Offices (+0.15) and Retail (+0.03), and negative values for Warehouses (-0.13) and Schools (-0.09). Akin to that, the exterior roof area is a strong positive driver for Offices (+0.13) and Retail (+0.04), and a negative driver for Schools (-0.12) and Warehouses (-0.07). These findings suggest that total exposed surface area, and by extension building geometry and form, play a dominant role in explaining building EUI. However, the impact is highly dependent on building use.

Cooling system characteristics are the second most important factor in predicting building EUI, although their impact varies across building types. For most building types, the plot shows that a higher cooling efficiency (COP) correlates with higher energy intensity. This counterintuitive trend may occur because high-performance equipment is most often found in complex, cooling-intensive commercial buildings that have high energy needs due to their size and internal workloads. However, the impact of cooling capacity follows a different pattern: it contributes to higher energy use in Retail (+0.04) and School (+0.02) but shows a negative association for Warehouses (-0.07) and Offices (-0.02). The different patterns likely reflect how these buildings are used. Retail stores and schools have high occupancy density, frequent door openings, and higher ventilation needs. These factors significantly impact cooling demand, and buildings with larger cooling systems use more energy. In contrast, warehouses with fewer occupants and limited air-conditioning needs (relatively to retail stores and schools), and offices operating on more regular schedules, are less impacted. This does not imply that greater capacity or higher efficiency causes increased energy use; it merely reflects correlation in building and system characteristics within the dataset.

Lastly, the hot water loop load also affects building EUI. While this factor negatively influences the energy intensity of Offices and Warehouses, it positively impacts the School EUI (+0.1), which likely reflects the high hot water demands unique to schools, such as in high-volume cafeterias, gym locker rooms, showers, and for specialized laboratory requirements. In these buildings, the hot water system is a primary contributor to the total energy footprint, whereas in other sectors this load can be negligible.

Counterfactual retrofit analysis

In this stage, we apply the trained machine learning model to evaluate the energy impacts of selected retrofit measures using a counterfactual framework, focusing on LED lighting upgrades and window replacements. For each retrofit, relevant building attributes are modified to represent the upgraded condition while all other characteristics are held constant. The updated configuration is then processed through the same preprocessing and prediction pipeline used during model training. The difference between the predicted baseline and retrofit EUI represents the estimated impact of the measure.

LED lighting

In the ComStock dataset, interior lighting is characterized by a generation-based categorical variable that reflects technology tiers ranging from legacy Gen1 systems to high-efficiency Gen4-5 light-emitting diode (LED) systems. The LED lighting retrofit measure represents a categorical technology upgrade for this variable, elevating it to the high-efficiency LED tier. Buildings in the baseline already classified as Gen-4 or Gen-5 are excluded from the retrofit intervention.

After the categorical update, the modified dataset is passed through the same preprocessing pipeline used during model training. Predicted site energy use intensity (EUI) is then computed for both baseline and upgraded configurations, and the difference represents the modeled energy impact of the LED retrofit.

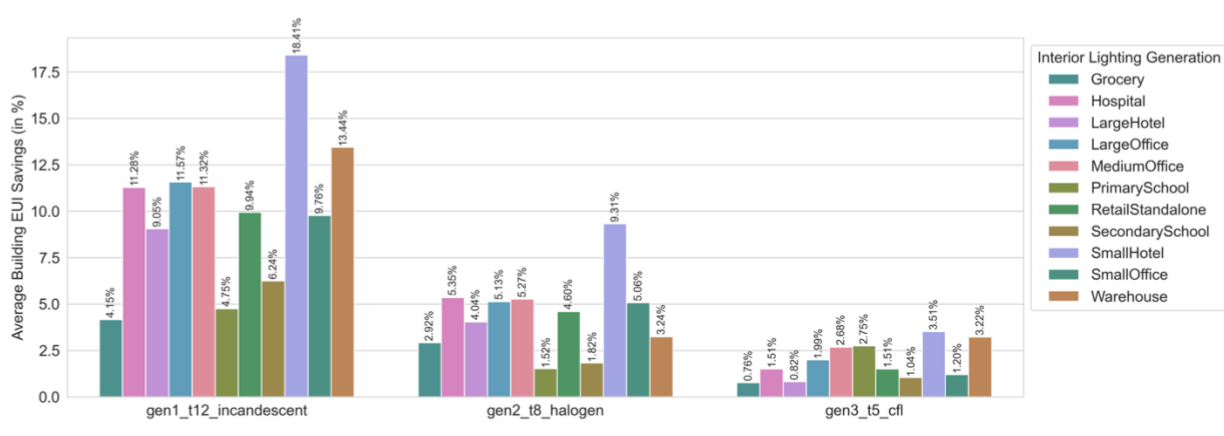

Figure 2: Average percentage reduction in site energy use intensity (EUI) from LED lighting upgrades by baseline lighting generation and building type.

Figure 2 presents the average percentage reduction in building EUI resulting from the LED upgrade, disaggregated by baseline lighting generation and building type. Results are shown for three baseline technologies: gen1_t12_incandescent, gen2_t8_halogen, and gen3_t5_cfl, across multiple commercial building categories.

Across all building types, the magnitude of savings is strongly dependent on the baseline lighting. Buildings with first-generation incandescent systems (gen1_t12_incandescent) achieve the largest energy savings, ranging from approximately 4% (Grocery) to 18% (SmallHotel), with most office, hospital, and hotel categories achieving 9–13%. For second-generation T8/halogen systems (gen2_t8_halogen), savings are more moderate, generally between 2% and 9%, with SmallHotel again exhibiting the highest impact. For third-generation T5/CFL systems (gen3_t5_cfl), savings are comparatively small, typically 0.8%-–3.5%. This monotonic decline in savings from Gen1 to Gen3 reflects diminishing incremental benefits as baseline lighting efficiency improves, since more efficient starting technologies leave less room for additional gains from upgrading to high-efficiency LED systems.

The results also highlight heterogeneity in retrofit impact across building types. Hotels consistently exhibit larger relative savings, likely due to higher lighting loads and longer operating hours, which amplify the effect of lighting efficiency upgrades. In contrast, grocery and certain retail categories show smaller relative reductions, suggesting that a larger portion of their total energy use intensity is driven by refrigeration, plug loads, or other process loads rather than lighting.

New windows

The window replacement retrofit measure represents a categorical upgrade applied to building envelope systems. In the ComStock dataset, window performance is characterized by the variable in.window_type, which encodes glazing type, coating, and framing system. The retrofit replaces existing window assemblies with location-specific high-performance configurations aligned with the Advanced Energy Design Guide (ASHRAE 2014). Buildings already using high-performance triple-glazed systems are not modified.

After the categorical update, the modified dataset is passed through the same preprocessing pipeline. The predicted EUI is then computed for both the baseline and upgraded configurations, and the difference represents the modeled energy impact of the window replacement retrofit.

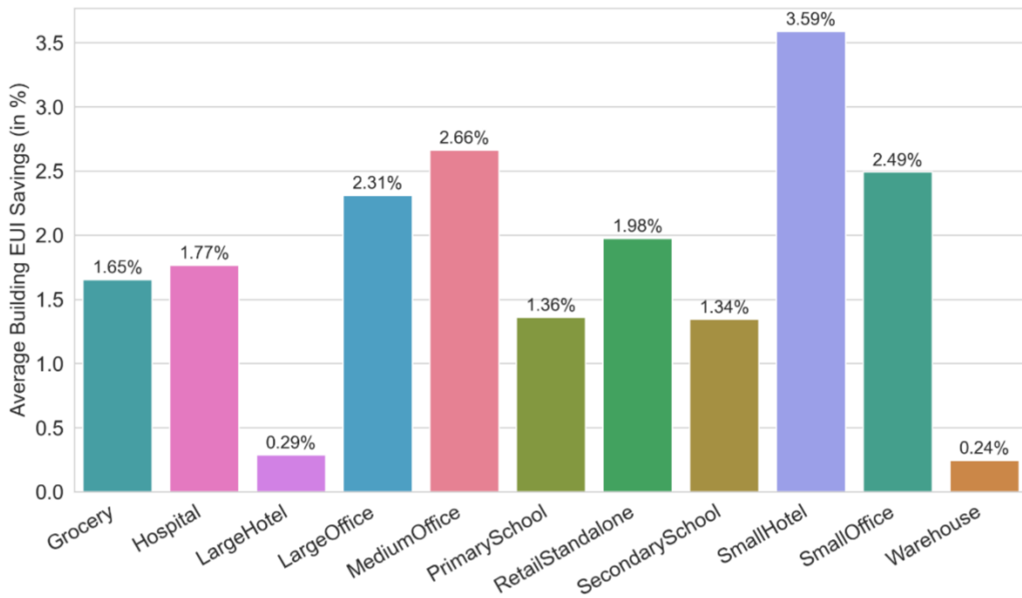

Figure 3: Average percentage EUI savings from window replacement across commercial building types.

Figure 3 shows the average reduction in building EUI from upgrading to high-performance window technologies, disaggregated by building type. The modeled savings from window replacement are lower than the lighting upgrades and vary across building types. SmallHotel shows the highest reduction at 3.59%, followed by MediumOffice (2.66%), SmallOffice (2.49%), and LargeOffice (2.31%). Our analysis estimates 2% savings for RetailStandalone, 1.6–1.8% for Grocery and Hospital buildings, 1.3–1.4% for schools, and 0.24-0.29% for LargeHotel and Warehouse.

The results indicate clear building-type heterogeneity. Office buildings are consistently more sensitive to window upgrades, likely due to larger conditioned floor areas, higher window-to-wall ratios, and extended operating hours. In contrast, warehouses demonstrate smaller relative impacts, suggesting that window-related heat transfer contributes less to total site EUI in these buildings.

Validation against ComStock upgrade simulations

ComStock provides physics-based upgrade simulations for a wide range of retrofit measures, including lighting and envelope improvements. These upgrade scenarios are generated by modifying building characteristics within the simulation framework and re-running EnergyPlus to estimate energy impacts. The tool developed in this study supports this process by offering a fast, data-driven way to estimate retrofit impacts across many buildings. It is not meant to replace detailed simulations, but to identify the most promising measures and building types for physics-based modeling. In this section, we evaluated the proposed framework against the ComStock upgrade simulation outputs to validate our framework.

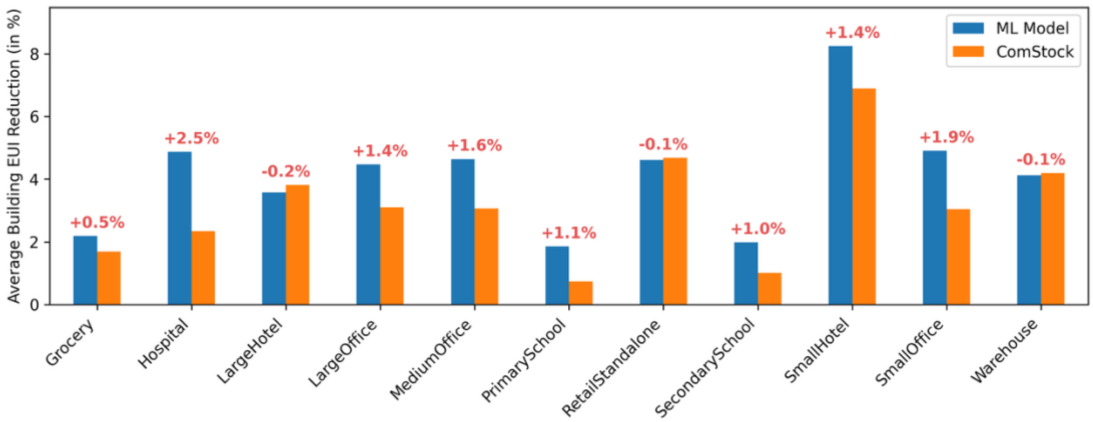

Figure 4: Comparison of ML Model Estimated v/s ComStock Average Building EUI Reduction for LED Replacement

Figure 4 compares the average percentage reduction in building EUI predicted by the proposed framework against published ComStock upgrade simulation outputs for the LED retrofit measure, disaggregated by building type.

Overall, the ML model's estimated savings align with those observed in the ComStock simulations across most building categories. For example, SmallHotel exhibits the highest savings in both approaches, with the model predicting approximately 8.2% reduction compared to 6.9% in ComStock. RetailStandalone shows strong agreement between the two methods, with both estimates near 4.6–4.7%. Grocery, SecondarySchool, and LargeHotel also demonstrate consistent ordering and similar magnitudes.

However, the model tends to overestimate savings relative to ComStock in several building types, including Hospital, LargeOffice, MediumOffice, PrimarySchool, and SmallOffice. The magnitude of overprediction ranges between approximately 0.5 and 2.5 percentage points, depending on the building type.

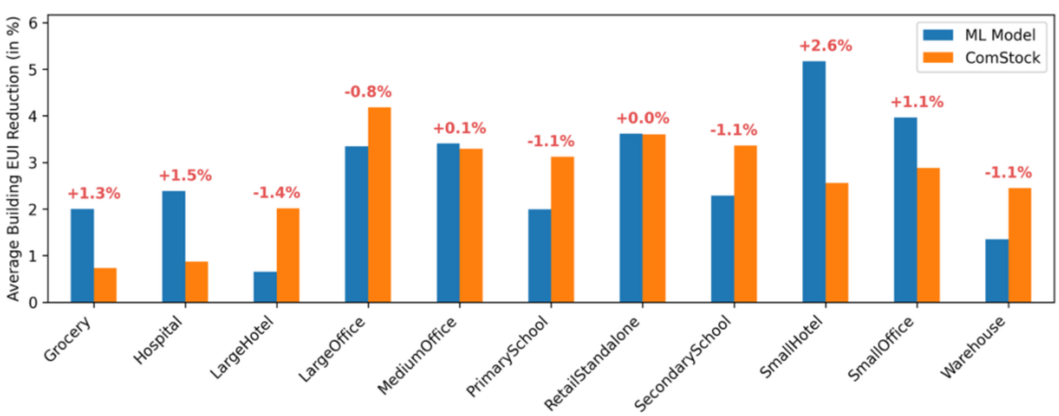

Figure 5: Comparison of ML Model Estimated v/s ComStock Average Building EUI Reduction for Window Replacement

Figure 5 compares the average percentage reduction in building EUI predicted by the counterfactual machine learning framework against the ComStock upgrade simulation results for the window replacement measure. Overall, the model captures the relative ordering of savings across building categories, but the agreement in magnitude is less consistent than in the LED retrofit case. RetailStandalone shows strong alignment between the two approaches, with nearly identical savings (~3.6%). MediumOffice and LargeOffice also exhibit reasonably close agreement, with differences within approximately one percentage point.

However, notable deviations appear in several building types. The model predicts substantially higher savings for SmallHotel (approximately 5.2%) compared to ComStock (~2.6%), and for SmallOffice (approximately 4.0% vs. ~2.9%). Conversely, the model underestimates savings in LargeHotel, SecondarySchool, and Warehouse relative to ComStock. Grocery and Hospital categories also show larger discrepancies, with the model predicting roughly two to three times the ComStock savings. These differences likely occur because window upgrades interact with heating and cooling in complex ways that physics-based simulations model directly, but the statistical model captures only indirectly. Window performance depends on solar heat gain, heat loss, weather region, glazing area, internal loads, and HVAC controls. Since the counterfactual approach relies on learned patterns rather than rerunning thermal simulations, some differences in the magnitude of savings are expected.

Despite differences in absolute magnitude, the overall directional consistency and preservation of building-type ranking across retrofit measures indicate that the counterfactual framework captures the primary structural drivers governing energy impacts. While envelope-related measures such as window replacement exhibit greater sensitivity to simulation-specific interactions, and lighting upgrades show modest deviations, the relative performance patterns remain largely consistent with the physics-based ComStock simulations. These results provide confidence that the counterfactual approach generates directionally reliable and program-relevant estimates of retrofit impact at both building and stock levels, making it suitable for comparative analysis.

Applications of the Proposed Framework

The proposed interpretable machine learning framework provides a capability that can be leveraged by building owners, large portfolio owners, energy service providers and energy efficiency program administrators to better understand existing building performance and identify improvement strategies with minimal data and investment. This framework has several practical applications across building energy management, program design, and planning.

Engineering and Energy Service Companies

Early-Stage Audit Prioritization: For engineering firms and energy service companies (ESCOs), the framework can function as a pre-audit diagnostic tool. By highlighting dominant drivers of EUI—such as envelope exposure or cooling system characteristics—it helps practitioners focus site assessments on systems most likely to yield measurable savings.

Building Owners, Portfolio Managers

Portfolio-Level Retrofit Screening: The primary application is screening retrofit opportunities across diverse building portfolios. Owners of commercial real estate, public agencies, and institutional campuses can use the framework to rapidly rank building features by their predicted energy-savings potential before investing in detailed audits or physics-based simulations. This would serve as a mechanism for prioritizing buildings based on potential for energy savings and would hence enable a more strategic allocation of engineering resources and capital planning efforts.

Benchmarking and Stakeholder Communication: The integration of explainable AI improves interpretability for non-technical stakeholders. Facility managers and building owners can better understand how specific characteristics influence performance and how proposed upgrades contribute to savings. This transparency can increase trust in data-driven recommendations and facilitate internal decision-making.

Utilities and Energy Efficiency Program Managers

Program administration decision support: Energy efficiency program administrators can apply the model to identify high impact retrofit measures across specific market segments. Because the framework provides transparent feature attributions, it supports defensible incentive design, targeted outreach strategies, and data-driven program evaluation. The ability to generate portfolio-level counterfactual scenarios also enables estimation of aggregate savings potential under different program or incentive structures.

Program and Scenario Analysis: At the regional or national level, the framework can support scenario analysis by estimating the potential energy impact of widespread adoption of specific measures (e.g., lighting upgrades or window retrofits). When combined with cost data, it could inform cost-effectiveness screening and long-term resource planning. Overall, the framework serves as a complementary layer between benchmarking and detailed simulation—bridging large-scale data analytics and actionable retrofit decision-making.

Limitations and Future Work

This study demonstrates that an interpretable machine learning framework can be used to conduct building-level and portfolio-level counterfactual analysis for retrofit screening. The goal of this work is not to replace physics-based simulation, but to provide a fast and scalable tool that supports early-stage prioritization of retrofit measures before detailed simulation or engineering analysis is performed. At the same time, the results show strong directional consistency with ComStock upgrade simulations, and several limitations frame opportunities for improvement.

First, model accuracy remains constrained by the limited representation of underlying physical behavior. The model is trained on structural and system-level attributes available in ComStock. However it does not explicitly encode thermodynamic processes, control logic, or dynamic interactions between envelope and HVAC systems. As a result, EUI predictions are based on learned correlations rather than mechanistic relationships. Incorporating richer operational features, temporal information, or physics-informed modeling approaches could improve the model's ability to capture true energy behavior and reduce residual bias in retrofit impact estimates.

Second, the robustness of SHAP-based explanations warrants careful consideration. SHAP assumes additive feature effects and can outweigh correlated variables, potentially affecting feature ranking stability and the interpretation of retrofit drivers. While SHAP provides transparency and interpretability, more robust explainability methods, such as interaction-aware attribution techniques or causal feature importance frameworks, could increase confidence in identifying the true drivers of EUI. Strengthening the reliability of explanations is especially important when counterfactual decisions are based on feature-level insights.

Third, the current framework evaluates energy savings but does not incorporate cost-effectiveness or economic feasibility. Retrofit decisions in practice depend not only on percentage EUI reduction but also on capital costs and lifecycle benefits. Integrating cost data and simple financial metrics would enable the tool to move from energy-impact screening to investment prioritization.

Finally, the counterfactual analysis assumes that modified features represent feasible building states and does not explicitly enforce system-level consistency. Single-variable interventions may generate combinations that are statistically valid but operationally unrealistic. Future work should explore constrained or joint counterfactual approaches that ensure internally consistent retrofit scenarios, improving the reliability of estimated impacts.

Addressing these limitations would strengthen the framework's predictive accuracy, interpretability, and decision relevance, enhancing its role as a complementary tool for retrofit screening and prioritization.

Conclusion

This study developed an interpretable machine learning framework to predict building EUI and evaluate retrofit impacts using counterfactual analysis on the ComStock dataset. Tree-based models combined with SHAP explanations identified envelope exposure and cooling system characteristics as dominant drivers of energy use, with effects that vary by commercial building type. Counterfactual simulations for LED lighting and window upgrades demonstrated that the framework can rapidly estimate building- and portfolio-level savings. Results show strong agreement when compared with ComStock physics-based upgrade simulations. The framework is not intended to replace detailed physics-based simulation or on-site building assessments, but rather to complement them by providing a fast, interpretable screening tool. This framework can be leveraged by ESCOs to enhance the efficiency and effectiveness of building energy audits, and by building owners and portfolio managers to prioritize assessments and allocate resources toward buildings with the greatest energy savings potential. It also provides energy efficiency program administrators with actionable insights into the characteristics and performance of their building stock, enabling them to identify high-impact retrofit measures suitable for incentive design and program implementation. By enabling rapid prioritization of retrofit opportunities across large portfolios, this approach supports more strategic allocation of audit and simulation resources. Though the current work focuses on the development of the framework, future work to convert this into an open-source platform will promote transparent, data-driven methods in retrofit decision-making.

References

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. "Optuna: A Next-generation Hyperparameter Optimization Framework." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining: 2623-2631. 2019 arxiv.org/abs/1907.10902.

Arjunan, P., K. Poolla, and C. Miller. 2020. "EnergyStar++: Towards more accurate and explanatory building energy benchmarking." Applied Energy 276: 115413.

ASHRAE. 2014. "Advanced Energy Design Guide for Small to Medium Office Buildings".

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785-794.

Dabous, S. A., and F. Hosny. 2025. "A review of building envelope retrofitting methods for improving energy efficiency, aesthetic, and indoor environmental quality." Energy Nexus: 100407.

EPA (U.S. Environmental Protection Agency). 2023. ENERGY STAR Portfolio Manager. www.energystar.gov/buildings.

Goel, S., M. Borkum, R. Fowler, S. Newman, H. Bergmann, D. Prahl, H. Berk, and P. K. Chui. 2022. "Audit Template: Facilitating Data-Driven Decision Making for Jurisdictions." In Proceedings of the 2022 ACEEE Summer Study on Energy Efficiency in Buildings. Washington, DC: American Council for an Energy-Efficient Economy (ACEEE).

Gonzalez-Caceres, A., M. Rabani, and P. W. Martínez. 2019. "A systematic review of retrofitting tools for residential buildings." In IOP Conference Series: Earth and Environmental Science 294: 012035.

Jain, M., S. Goel, and X. Lei. 2026. "Explainable AI and Machine Learning for Transparent Building Energy Benchmarking and Retrofit Decisions." Forthcoming in Proceedings of SimBuild 2026.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." Advances in neural information processing systems 30.

Lei, X., M. Kang, M. Bergés, and B. Akinci. 2015. "Assessment of Industry Foundation Classes (IFC) in Supporting Building Energy Benchmarking." In Proc. of the 32nd CIB W78 Conference.

Li, T., T. Liu, A. O. Sawyer, P. Tang, V. Loftness, Y. Lu, and J. Xie. 2024. "Generalized building energy and carbon emissions benchmarking with post-prediction analysis." Developments in the Built Environment 17: 100320.

Lundberg, S. M., and S.-I. Lee. 2017. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30.

Parker, A., Horsey, H., Dahlhausen, M., Praprost, M., CaraDonna, C., LeBar, A., & Klun, L. (2023). *ComStock Reference Documentation: Version 1* (NREL/TP-5500-83819). National Renewable Energy Laboratory. https://doi.org/10.2172/1967948

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. 2018. "CatBoost: unbiased boosting with categorical features." Advances in neural information processing systems 31.

Toosi, H. A., M. Lavagna, F. Leonforte, C. D. Pero, and N. Aste. 2020. "Life cycle sustainability assessment in building energy retrofitting; A review." Sustainable Cities and Society 60: 102248.

U.S. Department of Energy. 2016. Best Practices for Providing Whole-Building Energy Data: Guide for Utilities. betterbuildingssolutioncenter.energy.gov/sites/default/files/attachments/Best%20Practices%20for%20Providing%20Whole-Building%20Energy%20Data%20-%20Guide%20for%20Utilities.pdf.

DOE (U.S. Department of Energy). 2019. Benchmarking and Transparency: Resources for State and Local Leaders.

www.energy.gov/sites/default/files/2019/02/f59/Benchmarking_Transparency_Resource_PDF_Final_2.14.pdf.

Wei, Y., X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao. 2018. "A review of data-driven approaches for prediction and classification of building energy consumption." Renewable and Sustainable Energy Reviews 82: 1027-1047.