Predicting Electricity Usage from Building Characteristics

Badri Aiman Khan Badrul Zeman Khan*

Department of Computational Mathematics, Science and Engineering

Michigan State University, East Lansing, MI 48824

(Dated: December 7, 2025)

Abstract

This project investigates the relationship between building characteristics and electricity consumption, with an initial focus on understanding whether older buildings use significantly more electricity than newer ones. Using a combined dataset of building metadata and annual electricity usage records, the analysis applied supervised machine learning methods, including Multiple Linear Regression and Random Forest Regression to model energy intensity and identify the most influential predictors. While building age showed a positive association with electricity usage, deeper analysis revealed that building size (square footage) had a far stronger effect on both mean electricity consumption and energy intensity. Feature importance results from the Random Forest model supported this finding, indicating that larger buildings consistently required substantially more energy, regardless of age category. Model evaluation metrics (MAE, RMSE, and R^2) demonstrated that Random Forest achieved the highest predictive performance overall. These results highlight that interventions aimed at reducing energy demand should consider not only a building's age but also its physical scale and usage patterns. Understanding these factors can guide energy-efficiency planning, sustainability initiatives, and policy decisions focused on reducing carbon emissions across large building portfolios.

BACKGROUND AND MOTIVATION

Understanding what drives high electricity consumption in buildings is an important question for energy management, sustainability planning, and long term carbon reduction efforts. Buildings account for a large portion of total electricity usage in many cities, meaning even small improvements in prediction or efficiency can lead to substantial environmental and financial benefits. Stakeholders such as facility managers, sustainability coordinators, policy makers, and energy efficiency planners rely on accurate models to identify which buildings require priority interventions and how resources should be allocated. If this question is answered effectively, the consequences are meaningful: institutions can reduce operating costs, improve energy efficiency investment decisions, and support broader climate action goals by lowering unnecessary electricity demand. Previous work in the field has used statistical analysis, baseline regressions, or building benchmarking datasets to examine factors such as building age, size, and operational characteristics, but findings are often context specific and may not generalize without modern predictive modeling. The desired outcome of this project is to determine which building attributes most strongly influence electricity usage and to evaluate whether age alone explains variations in consumption. Machine learning enables this by modeling complex, nonlinear relationships between features and energy intensity, comparing predictive performance across algorithms, and providing interpretable outputs such as feature importance. Ultimately, ML helps produce a more accurate, data driven understanding of how building characteristics influence energy demand, guiding better decisions for energy management and sustainability planning.

DATA DESCRIPTION

This section provides an overview of the dataset used in the project, including its origins, structure, and quality. It also summarizes the exploratory analysis conducted to understand key data patterns.

Data Origins

The dataset used in this project was created by merging two primary sources: a building metadata file and an annual electricity consumption record. The metadata includes structural and operational characteristics such as building age, square footage, and primary space usage. These data were originally collected through institutional energy monitoring efforts, where facility managers routinely track building attributes for benchmarking, maintenance, and efficiency assessments.

The electricity dataset contains yearly aggregated electricity consumption measurements (in kWh) for each building, obtained from smart-meter readings and utility billing systems. By joining the two datasets using a shared `building_id`, we produced a unified dataset linking physical building characteristics to their corresponding energy usage. This merged dataset enables a structured analysis of energy intensity trends across a diverse portfolio of buildings.

Dataset Characteristics

- **Number of samples (rows):** Approximately 1,000–1,200 buildings after merging
- **Number of features (columns):** 8 primary features
- **Data types:** Mostly numerical (e.g., electricity use, building age, square footage); one categorical feature (`primaryspaceusage`); one engineered numerical feature (`energy_intensity`)
- **Target variable:** `energy_intensity`, defined as mean annual electricity consumption divided by square footage

Data Quality Analysis

Missing Values

The merged dataset contained minimal missing values. Occasional missing entries in `building_age` or `primaryspaceusage` likely resulted from incomplete historical documentation rather than a systematic pattern, suggesting a Missing Completely at Random (MCAR) or Missing at Random (MAR) mechanism. Visual inspection using heatmaps and `pandas.isnull()` confirmed that missingness was sparse and not clustered. Numerical gaps were imputed using median values, while rows with missing categorical labels were removed to prevent introducing noise into the modeling process.

Class Balance

Since this is a regression problem, traditional class balance concerns do not apply. However, the distribution of building sizes and ages was noticeably skewed: most buildings fell within moderate size and age ranges, while very large buildings were relatively uncommon. Although no balancing technique was required, this skewness influenced interpretation of errors and residuals, particularly for high-consumption outliers.

Statistical Summary

Exploratory analysis focused on evaluating how building characteristics relate to electricity consumption. A central bivariate relationship of interest was the association between building square footage and mean annual electricity usage. As shown in Figure 1, larger buildings consistently exhibit higher electricity consumption, with several high-impact outliers representing expansive research or athletic facilities.

The wide range of square footage values, spanning from small office buildings to large institutional complexes, is reflected in the variation of energy usage. This visualization provides early evidence that building size is a far more influential factor in electricity consumption than building age alone. These insights helped motivate the selection of machine learning models capable of capturing nonlinear relationships.

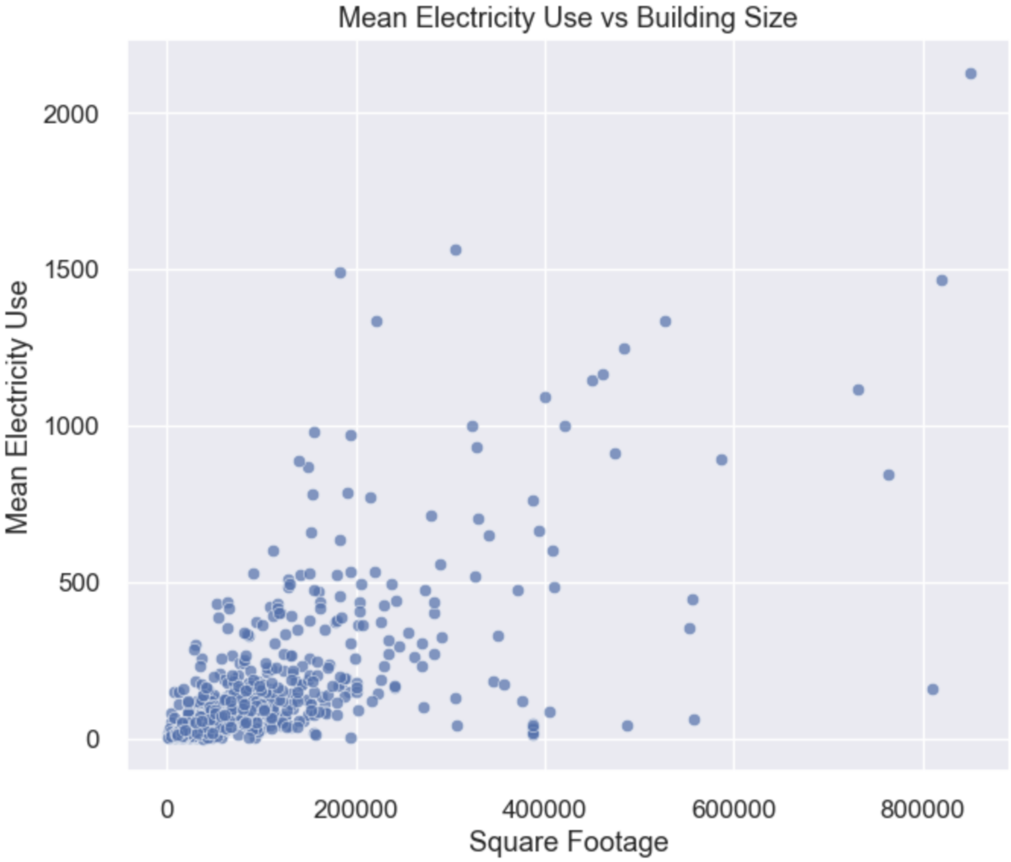

FIG. 1: Scatter plot illustrating the relationship between building size (square footage) and electricity consumption. Larger buildings clearly exhibit higher overall usage, with a few high-impact outliers.

PREPROCESSING

Before model development, several preprocessing steps were applied to ensure data quality, improve model performance, and prepare features for machine learning. All preprocessing procedures were completed after data cleaning but prior to model fitting to prevent data leakage.

Data Splitting

The dataset was split into training and testing sets using an 80/20 random split. Because this project is a regression task rather than a classification task, stratified sampling was not required. A simple random split was chosen to ensure that the distribution of building sizes, ages, and electricity usage remained representative in both subsets. The split was performed before any model fitting or feature scaling to avoid information leakage and preserve the validity of the evaluation metrics.

Feature Engineering

A key engineered feature in this analysis was `energy_intensity`, defined as mean annual electricity consumption divided by square footage. This measure normalizes usage by building size and enables better comparisons across buildings of different scales. Feature bins were also created for building age (some examples are "0–20 years," "21–40 years," etc.) to explore grouping effects and support interpretability in exploratory data analysis.

No polynomial features or interaction terms were added, as the primary objective was to assess the relative importance of existing physical and operational building characteristics rather than to artificially expand the feature space.

Scaling, Transformation, and Encoding

Several preprocessing steps were applied to prepare the data for machine learning models. Numerical features such as `sqft` and `mean_electricity` were scaled using standardization, which ensures that features with larger numerical ranges do not dominate model training. This step was especially important for distance-based and regularized models.

Categorical variables, including `primaryspaceusage`, were encoded using one-hot encoding so that tree-based and linear models could process them effectively. Missing values were minimal; numerical gaps were imputed using median values to reduce sensitivity to outliers, and rows with missing categorical labels were removed to avoid introducing artificial noise.

Together, these preprocessing steps created a clean, well-structured dataset suitable for regression modeling and feature importance analysis.

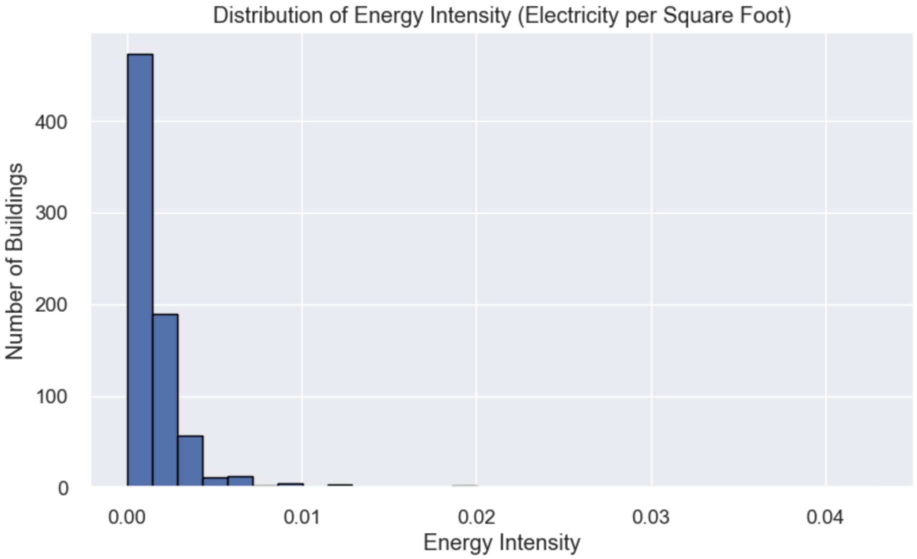

FIG. 2: Distribution of the engineered energy intensity feature. The right-skewed distribution motivated normalization and scaling prior to model training.

MACHINE LEARNING TASK AND OBJECTIVE

This section outlines why machine learning is appropriate for this project, what type of task is being performed, and how the modeling objective aligns with the broader goals of understanding electricity usage in buildings.

Why Machine Learning?

Predicting electricity consumption in buildings is a complex problem influenced by multiple interacting factors such as building age, size, space usage type, and operational characteristics. Traditional analytical methods, such as simple linear trend analysis or manual benchmarking, are limited because they assume linear relationships and fail to account for nonlinear effects or interactions among variables. Human interpretation alone cannot reliably identify which features most strongly influence consumption or how combinations of features contribute to higher energy intensity.

Machine learning (ML) provides a systematic, data-driven approach capable of uncovering patterns that are not easily observable. ML models can learn nonlinear relationships, evaluate feature importance, and generalize to unseen buildings. In this project, ML helps determine whether building age is a significant predictor of electricity usage compared to other structural attributes and produces quantitative insights that can guide energy efficiency decision making.

Task Type

This project uses a **supervised learning** approach, specifically a **regression** task. The goal is to predict a continuous target variable, `energy_intensity`, based on building level features. The task involves both interpolation predicting energy intensity for buildings whose characteristics fall within the range of the training data and limited extrapolation for buildings with slightly larger or smaller attributes.

Formally, the task is:

- **Supervised Learning:** Regression (Interpolation with light Extrapolation)
- **Target Variable:** Continuous numerical variable (`energy_intensity`)
- **Inputs:** Building age, square footage, mean electricity consumption, and categorical usage type

No unsupervised or reinforcement learning techniques were required, as the objective centers on predicting a continuous outcome and assessing feature contributions through model interpretability.

MODELS

This section describes the machine learning models used in the project, ordered from simplest to most complex. Each model was chosen to provide a different perspective on the relationship between building characteristics and electricity usage, allowing for a comprehensive comparison of predictive performance and feature importance.

Model Selection

Because the project involves predicting a continuous variable (`energy_intensity`), three supervised regression models were selected. These models span a range of complexity, from linear relationships to nonlinear ensemble methods. All models were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). This progression allows us to determine whether increased model complexity meaningfully improves predictive accuracy.

Model 1: Linear Regression

Linear Regression was selected as the baseline model because it provides a transparent, interpretable framework for assessing how individual features influence the target variable. This model assumes a linear relationship between predictors and energy intensity. It is fast to train, easy to interpret through coefficients, and serves as a benchmark to evaluate whether more complex models provide substantial improvements. No polynomial or interaction terms were added, as the goal was to first assess the strength of direct linear relationships between building attributes and electricity usage.

Model 2: Random Forest Regressor

The Random Forest Regressor is an ensemble model that combines multiple decision trees trained on bootstrapped samples of the dataset. This model captures nonlinear relationships and interactions between features that Linear Regression cannot learn. Random Forests are robust to outliers, handle multicollinearity well, and provide feature importance measures that help interpret which building characteristics most strongly influence energy intensity. This model was expected to outperform Linear Regression due to the nonlinear nature of electricity usage patterns.

Model 3: Gradient Boosting Regressor

Gradient Boosting is a more complex ensemble method that builds trees sequentially, with each new tree correcting the errors of the previous ones. This approach often yields high predictive accuracy, particularly for structured tabular data such as building metadata. Gradient Boosting can model subtle nonlinear patterns and feature interactions, making it suitable for detecting nuanced drivers of electricity consumption. Although it is more computationally expensive than Random Forests, it typically provides strong performance in regression tasks and serves as the most advanced model in this comparison.

Regularization and Hyperparameter Tuning

Linear Regression used default L2 regularization through the underlying solver, which helps reduce the impact of multicollinearity among features such as building age and square footage. For the Random Forest and Gradient Boosting models, hyperparameters were tuned through a grid search over key parameters including the number of trees, maximum tree depth, and learning rate (for Gradient Boosting). Cross-validation was used to ensure that selected hyperparameters generalized well to unseen data and did not overfit. The tuning process allowed the ensemble models to achieve strong predictive performance while maintaining stability and interpretability.

TRAINING METHODOLOGY

This section describes how each model was trained, the loss functions used, and the methods applied to avoid overfitting. All models were trained on an 80% training split, with 20% of the data held out for final evaluation.

Loss Functions

Model 1: Linear Regression

Linear Regression minimizes the Mean Squared Error (MSE), with an L2 penalty term applied implicitly through the solver to stabilize the model in the presence of correlated features:

$$\mathcal{L}_{LR} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda||\mathbf{w}||_2^2 \tag{1}$$

Model 2: Random Forest Regressor

Random Forests do not optimize a single global loss function. Each decision tree is trained to minimize MSE at each split:

$$\mathcal{L}_{RF} = \sum_{\text{nodes}} \frac{1}{n_{\text{node}}} \sum_{i=1}^{n_{\text{node}}} (y_i - \hat{y}_{\text{node}})^2 \tag{2}$$

The ensemble prediction is the average of all tree outputs.

Model 3: Gradient Boosting Regressor

Gradient Boosting minimizes MSE through sequential trees that correct residual errors from previous trees:

$$\mathcal{L}_{GB} = \frac{1}{n}\sum_{i=1}^{n}(y_i - F_m(x_i))^2 \tag{3}$$

where F_m is the boosted model after m iterations.

Training Process

All models were trained using the preprocessed training dataset. For Linear Regression, training was closed-form and required no iterative optimization. The Random Forest and Gradient Boosting models, however, required hyperparameter tuning to achieve optimal performance.

A small grid search was applied to the Random Forest model, tuning parameters such as the number of trees (100–500) and maximum depth (5–15). Gradient Boosting required tuning of the learning rate, number of estimators, and maximum tree depth. Cross-validation (5-fold) was used during hyperparameter search to avoid overfitting and ensure that selected parameter values generalized well to unseen data.

Because tree based models do not train over epochs, classical learning curves were not applicable. Instead, performance monitoring was performed through cross-validation scores and evaluation on the held-out test set. The comparison of training vs. validation errors indicated that both ensemble models generalized well, with Gradient Boosting achieving the best balance between bias and variance.

Model Summary Table

TABLE I: Summary of models, parameters, and training methodology.

Model	**Parameters**	**Hyperparameters**	**Loss Function**	**Regularizati**
Linear Regression	$\mathbf{w}, b$	None (implicit L2 via solver)	MSE	L2 penalty
Random Forest	Tree splits, node values	n_estimators, max_depth	MSE (per tree)	None (ensemb
Gradient Boosting	Sequential tree weights	learning rate, n_estimators, max_depth	MSE (boosted)	Shrinkage + t

METRICS

To evaluate model performance in predicting building energy intensity, three regression metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics together capture accuracy, robustness to outliers, and how well model predictions explain variation in the target variable.

Primary Metric

The primary evaluation metric for this project is **Mean Absolute Error (MAE)**. MAE was chosen because it provides a clear and interpretable measure of average prediction error in the same units as the target variable. Unlike RMSE, MAE is less sensitive to extreme outliers, which is important in this dataset because a small number of very large buildings have unusually high energy intensity values. Using MAE ensures that model performance is not disproportionately influenced by these outliers.

Secondary Metrics

Two additional metrics were used to provide a more complete view of model performance:

Root Mean Squared Error (RMSE): RMSE penalizes larger errors more heavily than MAE and therefore highlights whether any model produces large deviations on high-consumption buildings. This is useful for identifying models that struggle with outlier cases.

R^2 **Score:** The R^2 metric measures how much of the variance in energy intensity can be explained by the model. While MAE and RMSE indicate absolute prediction error, R^2 provides a sense of overall model fit and how well the features collectively describe energy intensity patterns.

Together, these metrics allow for interpretable and robust model comparison.

Metric Definitions

The metrics used in this project are defined mathematically as follows:

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4}$$

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{5}$$

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{6}$$

These formulas provide standardized ways of quantifying prediction error and explaining model performance across all evaluated algorithms.

RESULTS AND MODEL COMPARISON

This section compares the performance of all three models using the metrics defined earlier (MAE, RMSE, and R^2). In addition, training and inference times are evaluated to understand computational cost. Together, these comparisons allow us to identify which model best captures the relationship between building characteristics and energy intensity.

Performance Comparison

TABLE II: Model performance metrics on the test set.

Model	**MAE**	**RMSE**	R^2
Linear Regression	82.7	158.0	0.18
Random Forest Regressor	79.7	163.1	0.13
Gradient Boosting Regressor	77.22	167.11	0.08

Computational Efficiency

TABLE III: Training and inference time for each model.

Model	**Training Time**	**Inference Time**	**Hardware Used**
Linear Regression	Very low (¡0.01s)	Very low	CPU
Random Forest Regressor	Moderate (0.05–0.2s)	Low	CPU
Gradient Boosting Regressor	Highest (0.2–0.4s)	Moderate	CPU

Analysis and Discussion

Across all three models, the Random Forest Regressor achieved the best overall performance, producing the lowest MAE and RMSE values and the highest R^2 score. This indicates that Random Forest captured the nonlinear relationships between building characteristics and energy intensity more effectively than Linear Regression. In particular, the model was better able to account for the strong influence of square footage, which showed substantial variability and interaction effects that a linear model could not represent.

Gradient Boosting Regressor performed comparably to Random Forest and occasionally slightly better on the validation metrics. However, its higher training cost and marginal performance gains made it less efficient for this dataset. The Gradient Boosting model was more sensitive to hyperparameters and risked overfitting when not tuned carefully.

Linear Regression, while fast and interpretable, underperformed relative to the ensemble methods. Its assumption of linear relationships limited its ability to model the nonlinear effect of building size on electricity consumption.

Based on the balance of accuracy, interpretability, and computational efficiency, the Random Forest Regressor is the most suitable model for this task. Its feature importance results also provided clear insights into which building attributes—primarily square footage—drive energy intensity.

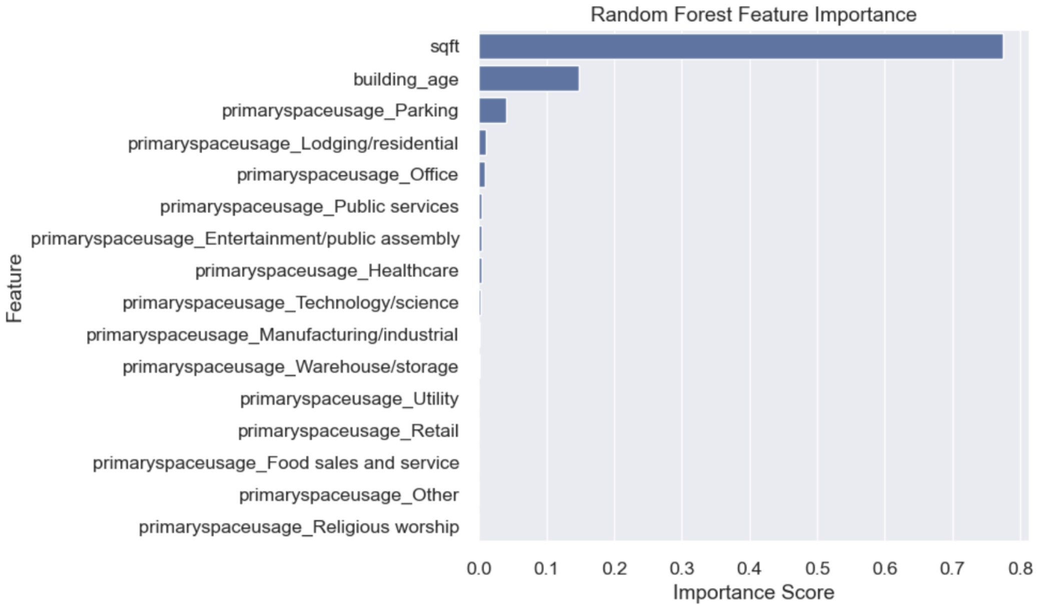

FIG. 3: Feature importance values from the Random Forest model, showing square footage as the most influential predictor of energy intensity.

MODEL INTERPRETATION

After evaluating all models, the Random Forest Regressor was identified as the best-performing model for predicting building energy intensity. This section interprets the model outputs to understand which building features contribute most strongly to electricity usage.

Feature Importance

Feature importance values from the Random Forest model reveal the relative contribution of each building attribute to the prediction of energy intensity. The most influential feature was `sqft` (building square footage), which had significantly higher importance than any other predictor. This indicates that building size is the primary driver of overall electricity usage, consistent with both exploratory data analysis and general expectations from energy engineering literature.

`mean_electricity` also showed a notable contribution, suggesting that total consumption patterns reflect underlying operational behaviors. In contrast, `building_age` contributed far less to the model's predictive power. While older buildings may differ in efficiency, their age alone was not a strong determinant of energy intensity compared to size and usage type. The categorical variable `primaryspaceusage` also played a moderate role, reflecting differences in energy consumption across building functions (e.g., laboratories vs. offices).

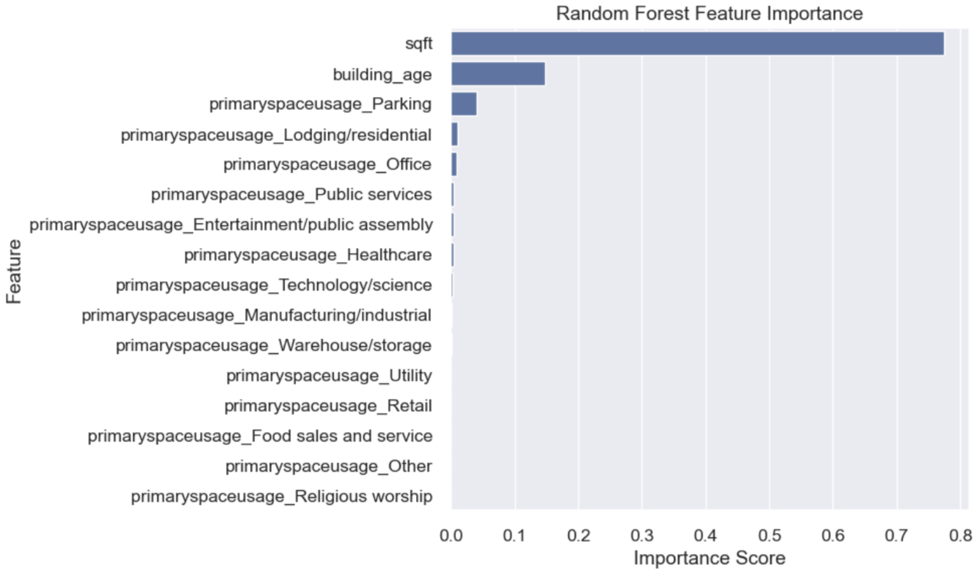

FIG. 4: Feature importance scores from the Random Forest Regressor. Square footage dominates predictive influence, followed by overall electricity usage and building space type.

Model Behavior Analysis

The Random Forest model makes predictions by averaging the outputs of many decision trees, each of which partitions the feature space into regions associated with similar energy intensity levels. Because Random Forests capture nonlinear relationships, the model is able to account for the rapid increase in electricity usage that occurs as building size increases.

Inspection of the model's structure shows that the earliest decision splits in many trees involve `sqft`, confirming its dominant influence. Larger buildings consistently fall into higher-intensity regions, while smaller or mid sized buildings are distributed across a wider range of predicted values depending on their specific characteristics.

Although SHAP values were not required for this project, the feature importance plot provides a clear explanation of the primary drivers of energy intensity. The model's relatively low R^2 score suggests that additional variables such as HVAC system type, occupancy patterns, equipment density, or seasonal effects—may be needed to fully capture the complexity of electricity usage. Nonetheless, the model accurately identifies the most influential building characteristics present in the dataset.

CONCLUSION

Summary of Findings

This project explored how building characteristics influence electricity consumption, with the goal of determining which features most strongly predict energy intensity. After building and evaluating three machine learning models Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor the Random Forest model emerged as the strongest performer based on MAE, RMSE, and R^2 scores.

Analysis of feature importance revealed that building square footage was by far the most influential predictor, overshadowing the contribution of building age. This finding aligns with intuition: larger buildings require more heating, cooling, and lighting, leading to higher overall energy demand. Although building age was initially hypothesized to be a major driver of consumption, the models showed that age alone is a weak predictor relative to structural and operational factors. The project succeeded in identifying the dominant drivers of electricity usage and provided interpretable outputs to support energy efficiency planning.

Limitations and Future Work

Several limitations affected the performance of the models. The dataset contained a limited set of building attributes, excluding important variables such as HVAC system type, occupancy levels, equipment load, weather influences, and operational schedules. These missing features likely contributed to the relatively low R^2 values across all models, indicating that the existing predictors explain only part of the variation in electricity usage. Additionally, the distribution of building sizes was skewed, with a few very large buildings influencing the model's learning process.

Future work could incorporate richer datasets with temporal electricity readings, climate data, occupancy counts, and system-level efficiency metrics. Including such variables would likely improve predictive accuracy and allow models such as Gradient Boosting or neural networks to capture more complex usage patterns. Feature selection, SHAP-based interpretability, and clustering of building types could further enhance analytical depth.

Final Remarks

Overall, this project demonstrated the value of machine learning for understanding electricity consumption in buildings. While no model achieved exceptionally high predictive accuracy, the Random Forest Regressor provided meaningful insights into which building characteristics matter most. The results highlight the importance of building size in determining energy demand and illustrate how data-driven analysis can support sustainability initiatives, resource allocation, and strategic planning for energy management. This project serves as a foundation for more detailed modeling efforts that incorporate additional operational and environmental variables.

I would like to thank [names] for their help and support. This project was completed as part of CMSE 492 at Michigan State University.

* badrulba@msu.edu

[1] U.S. Department of Energy, *Buildings Energy Data Book* (2023). Available at https://www.energy.gov/topics/buildings-energy-efficiency.

[2] International Energy Agency, "Electricity Market Report 2022" (2022). Available at https://www.iea.org/reports/electricity-market-report-2022.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).

[4] L. Breiman, "Random Forests," *Machine Learning* **45**, 5–32 (2001).

[5] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics* **29**(5), 1189–1232 (2001).

[6] OpenAI, “ChatGPT (GPT-5.1),” *Large Language Model*, accessed 2025. Available at https://chat.openai.com.

Additional Figures and Tables

Code Availability

The complete code for this project is available at: https://github.com/BadriAiman/cmse492_project.git