Regression Model-Based Prediction for Building Energy Star Score of New York City

Fan Zhang, Baiyun Chen*, Fan Wu,
Faria Brishti, Sameeruddin Mohammed
Computer Science Department
Tuskegee University
Tuskegee, USA
e-mail: {fzhang9458;bchen;fwu;
fbrishti7995;smohammed8703}@tuskegee.edu

Ling Bai
School of Engineering
The University of British Columbia
Kelowna, Canada
e-mail: ling.bai@ubc.ca

***Abstract*—Machine learning algorithms have recently shown promise in predicting Energy Star Scores for buildings, outperforming traditional forecasting methods. While previous studies have focused on specific building types, this comprehensive research expands the scope to analyze and predict Energy Star Scores across four diverse building categories in New York City: residential, educational, commercial, and lodging structures. Our study employs rigorous feature engineering and selection to develop nine distinct regression models applied to these four building types. We compare various machine learning algorithms to identify the most effective predictive model for each category. The Gradient Boosting Regressor (GBR) consistently emerges as the top performer across building types, demonstrating superior accuracy and stability in predictions. We provide a detailed analysis of feature importance for each building category, offering insights into the key factors influencing energy efficiency across different sectors. By extending the analysis to multiple building types and employing a range of regression models, this study contributes to a more comprehensive understanding of urban energy efficiency and provides tailored strategies for improving energy performance across New York City's diverse building stock for urban planners, building managers, and policymakers.**

***Keywords*-Machine learning; *Regression*; *Data analysis*; *Model evaluation*.**

I. Introduction

This paper serves as an expansion of our initial study on the prediction of the Energy Star Score for residential buildings: A case study of New York City [1]. In this expanded version, we broaden the scope of our regression models to forecast Energy Star Scores across a diverse range of building types, encompassing not just residential structures but also educational buildings, commercial buildings, and lodging buildings in New York. All the buildings studied in this paper possess a common attribute, namely, that individuals spend substantial amounts of time within them.

As economic and social development has progressed, the consumption of energy and water resources by human behaviors has increased by an order of magnitude, leading to a rise in annual carbon dioxide emissions and a severe reduction of water resources [2]. This trend has significant implications for the sustainable development of human society. Buildings account for approximately 40% of global energy consumption, a percentage projected to increase in the coming decades [3]. This growth is attributed to two main factors: the frequent extreme temperature fluctuations caused by climate change [4], and rising human demands for housing and improved living standards [5]. Among various energy end uses in buildings, space heating typically consumes the largest share, accounting for over 30% of total energy use, which is followed by water heating, cooling, ventilation, and lighting, though the exact order can vary depending on the building type [6]. Notably, residential buildings are responsible for almost 70% of the energy consumption of the sector, mainly due to the usage for cooking and heating [7]. Fortunately, it illustrates a great potential to enhance the energy efficiency of buildings by analyzing the retrofit options or adjusting human activities in energy consumption.

The Energy Star Score for buildings was developed by the United States Environmental Protection Agency (EPA) in collaboration with the U.S. Department of Energy (DOE), evolving from the broader Energy Star Program launched in 1992 [8]. This 1-100 scoring system provides a standardized method for measuring and comparing energy efficiency across different types of buildings. A score of 100 indicates top performance, placing the building among the most energy-efficient nationwide, while a score of 1 represents the lowest performance [9]. The Energy Star Score is a crucial metric for assessing the energy efficiency of buildings, enabling stakeholders to evaluate and compare building performance objectively. Estimating this score is therefore essential for building owners, managers, and policymakers seeking to improve energy efficiency in the built environment.

However, the complexity of building energy consumption, influenced by numerous factors such as weather conditions, occupancy patterns, building characteristics, and operational schedules, etc. poses significant challenges to accurately predicting building energy consumption, which directly influences the Energy Star Score. A lot of efforts from academia, industry, and governments have originated multiple methods or tools for the estimation of buildings' energy consumption. The Building Energy Software Tools Directory [10] provides comprehensive information on building software tools for evaluating energy efficiency and sustainability in buildings. This directory also

*Corresponding author.

shows that efforts can be derived for different components to minimize energy consumption. With the widespread applic ation of machine learning techniques, a growing number of researchers have recently proposed to introduce regression models for predicting building energy consumption, offering a data-driven method to navigate this intricate landscape of variables and their interactions [11]–[13]. Linear regression model is the most basic model applied to predict building energy consumption, due to its simplicity, straightforward implementation, and computational efficiency [14]. However, linear regression often falls short in capturing the intricate, non-linear relationships between input variables and energy consumption outcomes. Thus, regression models capable of handling non-linear relationships are often necessary to achieve higher prediction accuracy in the multifaceted domain of building energy consumption.

Various advanced regression techniques have been proposed to address the limitations of linear regression in predicting building energy consumption. Jung et al. and Ma et al. suggested using support vector regression (SVR) due to its ability to handle complex non-linear relationships in data [15], [16]. Yu et al. proposed tree-like structures, particularly decision trees, to analyze building parameters and predict energy demand, allowing for the identification of key influencing factors [17]. To enhance the performance of single decision trees, the Random Forest method was introduced, which ensembles multiple trees [18]. This concept of ensembling improves predictive performance by combining multiple models together to leverage their collective strengths, reduce individual weaknesses, and capture diverse aspects of the data. Similar principles are employed in Gradient Boosting and extreme gradient boosting models, both of which have been applied to building energy consumption prediction [19], [20].

Artificial Neural Networks (ANN), inspired by biological neural networks, have gained popularity for their ability to solve non-linear problems associated with high-dimensional datasets [21]. Deep Learning, an advanced form of ANN, excels at capturing consumption patterns from historical data and discovering non-linear relationships between inputs and outputs [22]. Among the various types of neural networks, the Multilayer Perceptron (MLP) has emerged as a particularly effective tool for predicting building energy consumption, including heating and cooling loads [23]. This application represents a rapidly growing research area due to its potential to significantly enhance energy efficiency in building management systems. However, these methods, including Support Vector Machines, Decision Trees, Random Forest, and Artificial Neural Networks, often require significant computational resources for parameter optimization and model tuning. Deep Learning, in particular, demands not only substantial computing power but also high-quality, large-scale labeled datasets.

In contrast, some researchers have explored simpler methods like k-Nearest Neighbors (kNN) for building energy consumption prediction. kNN forecasts energy consumption by identifying similar past instances based on relevant factors such as weather conditions, appliance usage, and time of day [24]. This method's appeal lies in its simplicity, ease of interpretation, and minimal assumptions about data distribution, with only one parameter (k) to optimize.

This study focuses on urban buildings, given the high population density and concentrated energy consumption in metropolitan areas. Almost all the aforementioned regression models are employed to forecast the Energy Star Score of buildings using disclosed energy and water consumption data from New York City. The performance of these various approaches is then systematically compared. Moreover, by evaluating the significance of different features, this research identifies key factors that substantially influence energy consumption for each building type. These insights offer valuable guidance for future building design, retrofit strategies, and occupant behaviors related to heating and cooking. Ultimately, this work aims to support efforts to reduce emissions and conserve energy in urban environments, contributing to more sustainable and efficient city infrastructures.

The structure of the paper is as follows. Section II briefly introduces the five conventional regression methods utilized in this work. Section III depicts the modeling procedure and results for the residential building energy consumption data in New York, presenting and discussing the findings. We conclude with Section IV.

II. Methods

Regression approaches, one of the most popular types of machine learning algorithms, demonstrate superior predictability with promising results in various domains, including energy consumption [25], bankruptcy prediction [26], air pollution [27], epidemiology [28], and some other applications. This study introduces 9 typical regression methods, including k-Nearest Neighbor Regression [29], Linear Regression [30], Ridge Regression [31], Decision Tree Regression [32], Random Forest Regression [33], Support Vector Regression [34], and Gradient Boosting Regression [35], eXtreme Gradient Boosting Regression [36], and Multi-Layer Perceptron [37] to predict the Energy Star Score of residential buildings and investigates the prediction results using four metrics, i.e., MAE, SSE, R^2, Adjusted R^2 [38]. The coefficient of determination, R^2, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Adjusted R^2 is a modified version of R^2 that adjusts for the number of predictors in the model.

Mathematically, given a training dataset D with features X and target values Y, and a new data point $\boldsymbol{x}$ for which we want to predict the target value $\hat{y}$, we briefly introduce the nine regression models and calculate $\hat{y}$ in each regression model accordingly.

A. k-Nearest Neighbor Regression

kNN regression, or k-Nearest Neighbors regression, is a non-parametric regression method that predicts target values by averaging the observed values of the k nearest samples in the feature space [29]. Hence,

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i, \tag{1}$$

where y_i are the target values of the k nearest neighbors of $\boldsymbol{x}$. The nearest neighbors are typically determined based on a distance metric, such as Euclidean distance.

B. Linear Regression

Linear regression is a parametric regression technique that models the linear relationship between dependent and independent variables by minimizing the residual sum of squares [30]. The predict value $\hat{y}$ is calculated using (2):

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \tag{2}$$

where β_0, β_1, β_2, …, β_n are the estimated parameters for the linear regression model and x_1, x_2, …, x_n are the values of the independent variables for the new data point.

C. Ridge Regression

Ridge Regression is a regularized linear regression method that introduces an L2 penalty term to mitigate multicollinearity and reduce model variance [31]. The prediction $\hat{y}$ is calculated using:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \tag{3}$$

where β_0, β_1, β_2, …, β_n are the estimated parameters. These parameters are obtained by minimizing:

$$\sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} \beta_j^2, \tag{4}$$

where λ is the regularization parameter that controls the strength of the penalty.

D. Decision Tree Regression

Decision Tree Regression is a non-parametric model that predicts target values by recursively partitioning the feature space into regions with homogeneity and assigning predictions based on local sample averages [32]. The prediction $\hat{y}$ for a new data point $\boldsymbol{x}$ is given by:

$$\hat{y} = \sum_{m=1}^{M} c_m I(\boldsymbol{x} \in R_m), \tag{5}$$

where M is the number of leaf nodes, c_m is the predicted value in the m-th leaf node, R_m is the region of feature space corresponding to the m-th leaf node, and I is an indicator function that equals 1 if $\boldsymbol{x}$ is in region R_m and 0 otherwise.

E. Random Forest Regression

Random Forest Regression is an ensemble learning approach that combines predictions from multiple decision trees trained on bootstrapped samples to reduce variance and improve generalization [33]. $\hat{y}$ is predicted by (6):

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{x}), \tag{6}$$

where $f_i(\boldsymbol{x})$ is the prediction of the i^{th} decision tree for the new data point $\boldsymbol{x}$ and N is the total number of decision trees in the Random Forest.

F. Support Vector Regression

Support Vector Regression is a regression technique that seeks a hyperplane in a high-dimensional space to minimize prediction errors within a predefined tolerance, supported by a margin [34]. $\hat{y}$ is predicted by (7):

$$\hat{y} = \mathbf{w}^T \cdot \boldsymbol{x} + b, \tag{7}$$

where $\mathbf{w}$ is the weight vector and b is the bias term.

G. Gradient Boosting Regression

Gradient Boosting Regression is a sequential ensemble method that optimizes a differentiable loss function by constructing regression trees in a stage-wise manner using gradient descent in the function space [35]. $\hat{y}$ is predicted by (8):

$$\hat{y} = \sum_{i=1}^{N} \gamma_i f_i(\boldsymbol{x}) \tag{8}$$

where γ_i is the learning rate that controls the contribution for each learner,$f_i(\boldsymbol{x})$ is the prediction of the i^{th} decision tree for the new data point $\boldsymbol{x}$ and N is the total number of decision trees in the Gradient Boosting model.

H. eXtreme Gradient Boosting Regression

XGBoost is a highly efficient gradient boosting implementation that integrates advanced regularization techniques and parallel processing to enhance computational performance [36]. The prediction $\hat{y}$ is given by:

$$\hat{y} = \sum_{k=1}^{k} f_k(\boldsymbol{x}), \tag{9}$$

where k is the number of trees, f_k represents the k-th tree. The objective function to be minimized is:

$$\mathcal{L} = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k), \tag{10}$$

where l is a differentiable convex loss function and Ω is the regularization term.

I. Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is a fully connected feed-forward neural network that approximates complex functions through layered processing and non-linear transformations [37]. For a MLP with L layers, the prediction $\hat{y}$ is calculated as:

$$\hat{y} = f_L(W_L \cdot f_{L-1}(W_{L-1} \cdot \ldots f_1(W_1 \cdot \boldsymbol{x} + b_1) \ldots + b_{L-1}) + b_L), \tag{11}$$

where W_l and b_l are the weight matrix and bias vector for layer l respectively, and f_l is the activation function for layer l. Common choices for f_l include ReLU, sigmoid, and tanh functions.

J. Performance Metrics

Four commonly used performance metrics are employed in this work. They are Mean Absolute Error (MAE), Sum of Squared Errors (SSE), Coefficient of Determination (R-squared, R^2), and Adjusted R^2. MAE measures the average absolute difference between the predicted values and the actual values; SSE measures the total squared difference between the predicted values and the actual values; R^2 can be interpreted as the percentage of the variance in the dependent variable that is explained by the independent variables; Adjusted R^2 provides a more accurate assessment, which penalizes the addition of unnecessary variables to the regression model [39].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{14}$$

$$\text{Adjusted } R^2 = 1 - \frac{\left(1 - R^2\right)\cdot(n-1)}{n-k-1} \tag{15}$$

These performance measures aid in evaluating the quality of fit and accuracy of regression models, facilitating the comparison and assessment of various models and their capacity for prediction.

III. Case Study

Predicting the Energy Star Score follows the standard machine learning workflow, which consists of four stages: data collection, data preprocessing, model training, and model testing, as shown in Figure 1 [11]. Data collection gathers crucial building and energy consumption data. Data preprocessing involves cleaning and preparing data for analysis. Model training consists of selecting algorithms, setting parameters, and training the models. Finally, in the model testing stage, it examines the models' ability to predict the Energy Star Score.

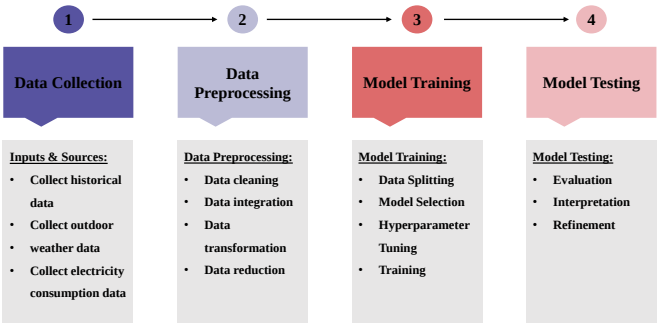


Figure 1. Workflow of predicting building Energy Star Score.

Data used for the regression prediction corresponds to the energy and water data disclosed for Local Law 84 of the New York City in the calendar year 2021 [40]. It encompasses a diverse range of building types, including residential buildings, educational buildings, commercial buildings, lodging buildings, factories, cultural institutions, and various other structures. In this study, we concentrated on four specific building types: multifamily residential buildings, K-12 schools, office buildings, and hotels. We chose these categories because they have significantly more available data compared to other types of buildings. A richer data set is advantageous in constructing a more robust predictive model and mitigating the uncertainty caused by limited data. After cleaning the dataset by removing rows with missing values and outliers, we extracted a total of 13,871 records from the original 22,479 rows, focusing on our four selected building types. This cleaned dataset comprises 10,802 records for multifamily residential buildings, 1,564 records for K-12 schools, 1,112 records for office buildings, and 393 records for hotel buildings. This substantial sample, representing about 62% of the original data, provides a robust foundation for our predictive models across these key urban building categories.

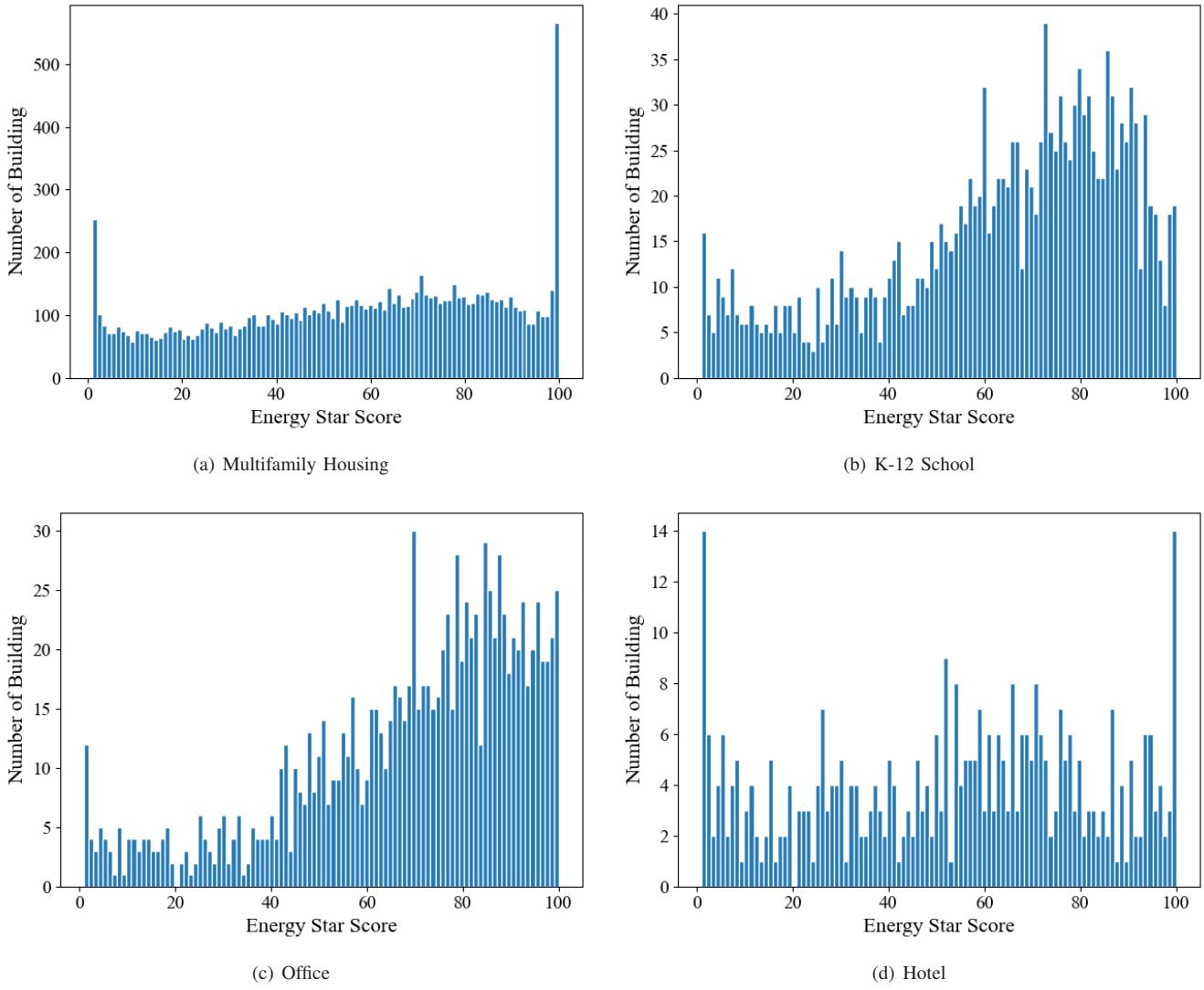
The original dataset comprises 249 columns, with the Energy Star Score column serving as the target variable for prediction. The score quantifies the property's performance relative to similar ones, rated on a scale of 1 to 100, where 1 denotes the poorest-performing buildings, and 100 indicates the best-performing ones. The remaining columns are considered as variables constituting the potential features in the regression model. A comprehensive explanation for each column can be found in the data dictionary [40].

Given that these four building types belong to distinct categories with varying energy consumption patterns, occupancy behaviors, and building functions, developing a single model to predict Energy Star Scores across all categories could lead to underfitting. The significant differences in sample sizes among the categories further complicate this issue. To address these challenges and to better capture the unique energy consumption characteristics of each building type, we opted to develop separate regression models for each category, which helps to maximize prediction accuracy by tailoring each model to the specific features and patterns of its respective building type.

A. Feature Statistics

Prior to constructing the predictive model for residential energy consumption, it is imperative to thoroughly explore the features within the original dataset. As it is known, each feature holds varying degrees of importance, with the Energy Star Score column being the most crucial as it serves as the target variable for prediction. Therefore, we first use a histogram to represent the distributions of this target variable, as shown in Figure 2.

Figure 2 illustrates the distribution of Energy Star Scores across four different building types: multifamily housing, K-12 schools, offices, and hotels. Each subfigure corresponds to one type separately. Notably, none of these distributions conform to either a uniform or a normal distribution. Multifamily housing in Figure 2(a) shows high frequencies at both ends with lower, uneven distribution in the middle. Both K-12 schools in Figure 2(b) and office buildings in Figure 2(c) exhibit



(a) Multifamily Housing

(b) K-12 School

(c) Office

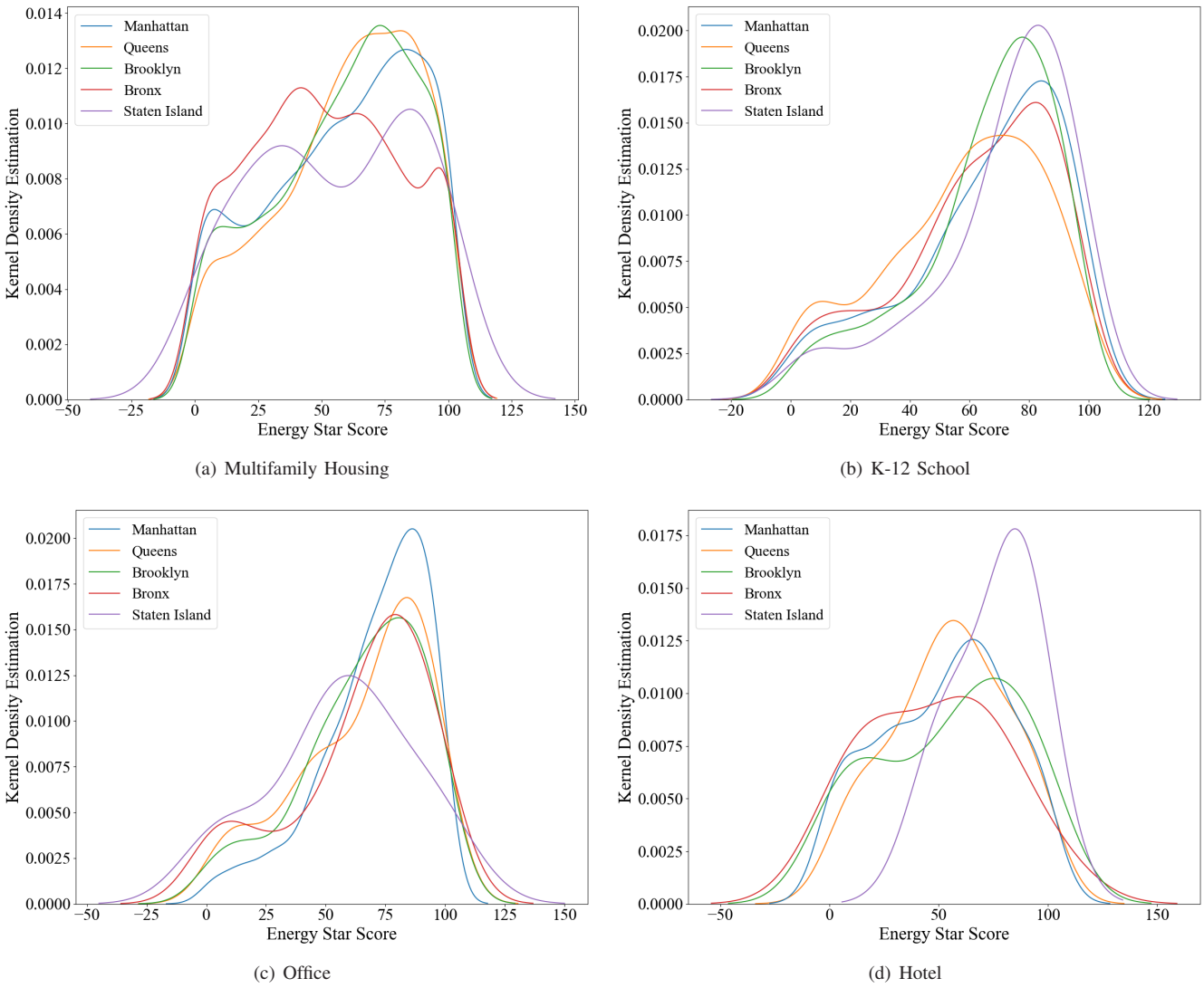
(d) Hotel

Figure 2. Distribution of Energy Star Score across four types of buildings.

similar patterns that can be described as slightly right-skewed bimodal distributions. They show a small peak in the lower score range (around 0-20), with scores gradually increasing towards the higher end, reaching a maximum peak near 100. This suggests that while a significant portion of these buildings achieve high energy efficiency, there's also a smaller group with lower efficiency. Hotels in Figure 2(d) present the most irregular pattern, with scores scattered across the range and multiple local peaks. These diverse, non-standard distributions across all building types underscore the complexity of energy performance in different sectors and highlight the necessity for advanced regression techniques rather than traditional statistical methods for accurate modeling and prediction of Energy Star Scores.

Next, we need to screen out the more important variables to the target variable for modeling from the 248 features, a step commonly known as feature selection. This process stands as one of the pivotal stages in the entire machine learning workflow. The efficacy of a machine learning model heavily relies on the predictive capability of the selected features. Even a simple linear model can showcase commendable performance if these features exhibit strong predictability. Conversely, the modeling process should exclude features with weaker predictive power. Their inclusion would not only increase model complexity but also compromise prediction accuracy.

In this study, we employ a non-parametric statistical technique, Kernel Density Estimation (KDE), to assess the effect of various variables on the distribution of the target variable. Variables demonstrating substantial fluctuations in the distribution of energy scores across different values are deemed significant, whereas those exhibiting minimal variation are deemed inconsequential. For example, we explore the impact of districts on the distribution of the Energy Star Score, as illustrated in Figure 3. We first categorize the datasets into different groups based on five districts in New York: Bronx,



(a) Multifamily Housing

(b) K-12 School

(c) Office

(d) Hotel

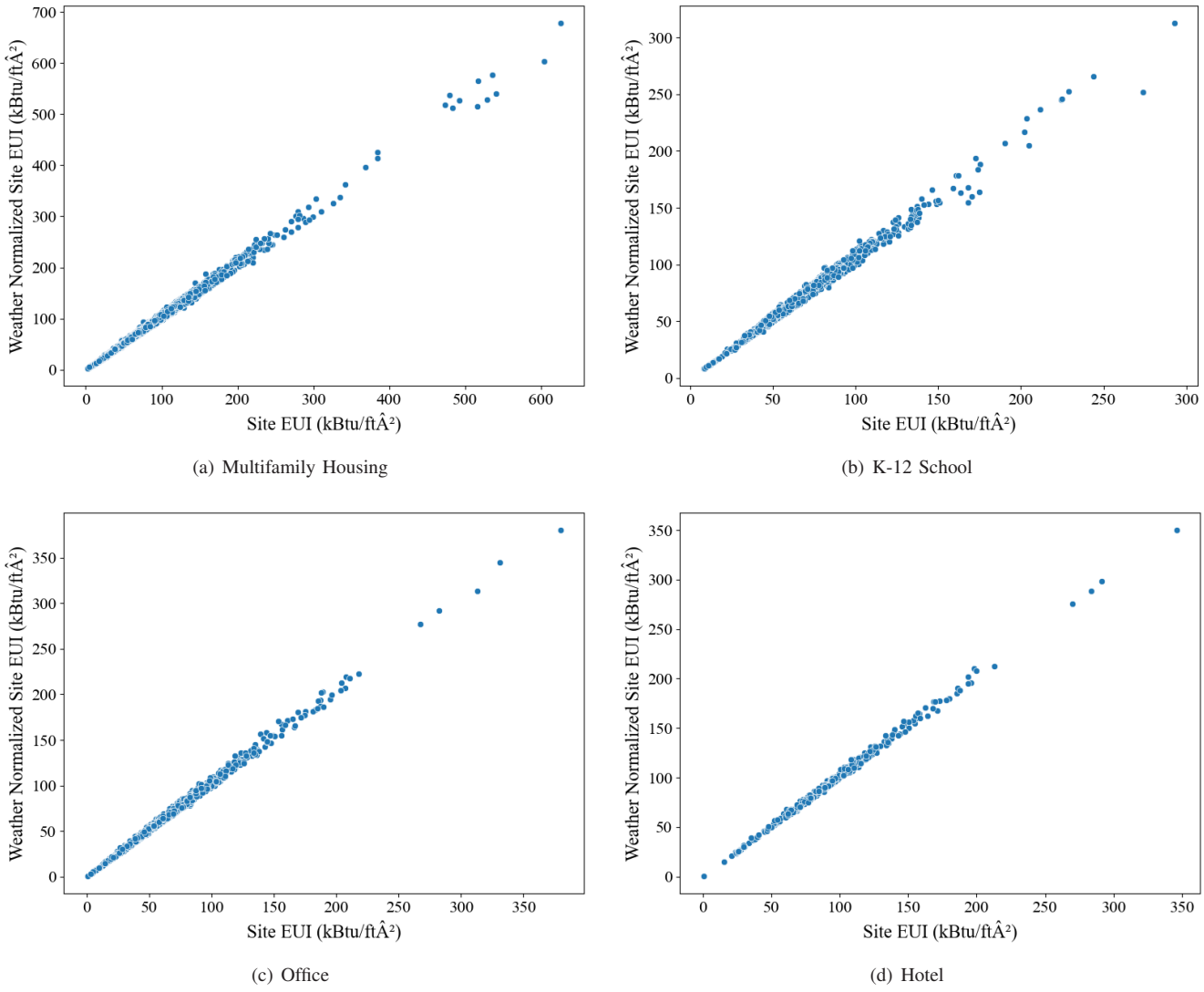
Figure 3. Distribution of Energy Star Score in different districts.

Manhattan, Brooklyn, Queens, and Staten Island, then we employ the Gaussian Kernel function to smooth the probability density estimation of different groups.

The KDE analysis across the four types reveals that the distribution of Energy Star Scores is generally consistent across the five districts, as shown in Figure 3. For all building types, the score distributions show similar patterns, with peaks around the same ranges. The analysis reveals that within each building category (multifamily, office, educational, and lodging), the Energy Star Score distributions show similar patterns across all five districts of New York City. This suggests that a building's geographical location within the city does not significantly influence its energy performance when compared to other buildings of the same type. Therefore, the district variable was not included as a predictor in our final models. Although Staten Island displays a somewhat distinct pattern for hotel buildings in Figure 3(d), this deviation is attributed to the fact that there are only three samples from this district, which is statistically insufficient to accurately represent the true distribution. Consequently, the district variable is not recommended for inclusion in the modeling process due to its limited contribution to predictive accuracy. This insight can help streamline future models and focus attention on variables that demonstrate greater discriminative power in the context of building energy efficiency.

Subsequently, we conduct correlation analysis to detect multicollinearity in two or more independent variables that are highly correlated with each other, possibly resulting in instability and inflated standard errors in regression models. By identifying and removing highly correlated variables, we can mitigate multicollinearity and improve the stability and interpretability of the model.

Figure 4 demonstrates the correlation analysis result of "Site EUI" and "Weather Normalized Site EUI" in the scatter diagram for four building types. EUI refers to the Energy Use Intensity, which measures the ratio of actual energy



(a) Multifamily Housing

(b) K-12 School

(c) Office

(d) Hotel

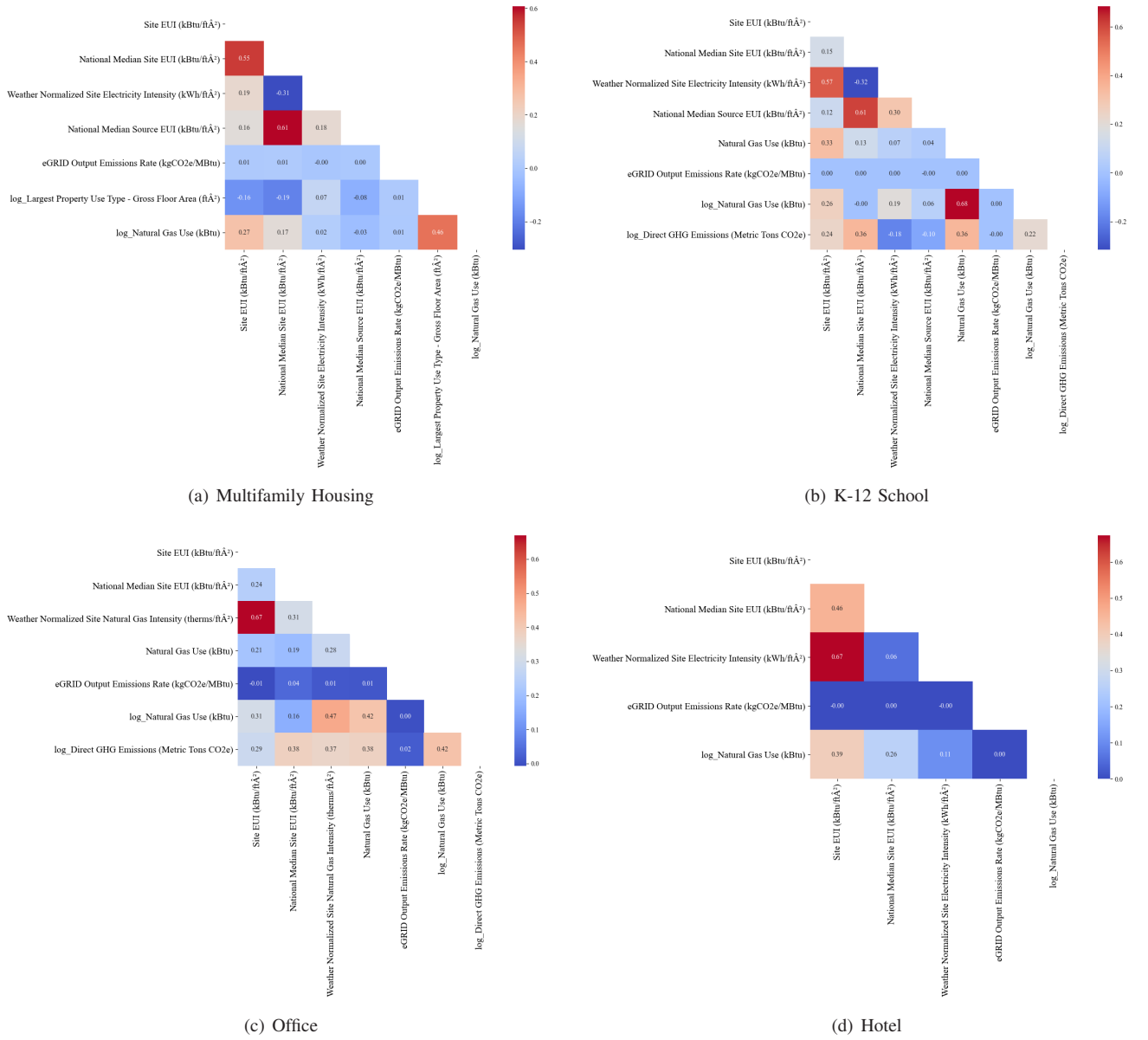
Figure 4. Distribution of correlation between “Site EUI (kBtu/ft²)” and “Weather Normalized Site EUI (kBtu/ft²)”.

consumption of a building or site to its area. Across all categories, an exceptionally strong positive linear relationship is observed, with correlation coefficients approaching 1. This near-perfect correlation is evident in the tight clustering of data points along the diagonal in each scatter plot. After checking the data dictionary, we find that “Site EUI" refers to the site energy use divided by the property square foot; the “Weather Normalized Site EUI" refers to the energy use one property would have consumed during 30-year average weather conditions [40]. Since the “Weather Normalized Site EUI" is calculated based on the “Site EUI", there is no doubt that there is such a high correlation between these two features. This high multicollinearity suggests that including both variables in predictive models would be redundant and potentially destabilizing. Therefore, only one of the features needs to be retained in the later modeling process. Given its more straightforward interpretation and direct measurement, we opt for keeping the “Site EUI" feature.

B. Feature Selection and Feature Engineering

Due to data measurement and collection challenges, we addressed missing data and potential multicollinearity by implementing a rigorous feature selection process. We removed features with substantial missing data and applied a correlation threshold of 0.7 to filter out highly correlated variables. This careful selection process yielded distinct sets of numeric features for each building type: 7 for multifamily housing, 8 for K-12 schools, 7 for offices, and 5 for hotel buildings. These selected features exhibit correlations below 0.7 with each other, as depicted in Figure 5, ensuring a balanced representation of predictors while minimizing redundancy.

During the feature selection stage, we also engage in feature engineering. Feature engineering entails the extraction or creation of new features from raw data, often involving the transformation of certain raw variables. This may include applying natural logarithm transformations to non-normally



(a) Multifamily Housing

(b) K-12 School

(c) Office

(d) Hotel

Figure 5. Correlation matrix of selected features.

distributed data or encoding categorical variables with one-hot codes to facilitate their inclusion in model training.

First, we apply the logarithms to the numeric features and add them to the original data. As we all know, most original data are not normally distributed. If we include this kind of data in the model directly, it might arise bias due to the skewed distribution of data. In Figure 5, the features starting with "log_" are the ones transformed by the logarithm functions.

Next, we apply the Min-Max normalization to the numerical features. Scaling these features to a comparable range helps mitigate bias toward features with larger scales, thereby fostering more accurate predictions and enhancing stability. With this step completed, our dataset is now fully prepared for the modeling phase.

C. Test Bench

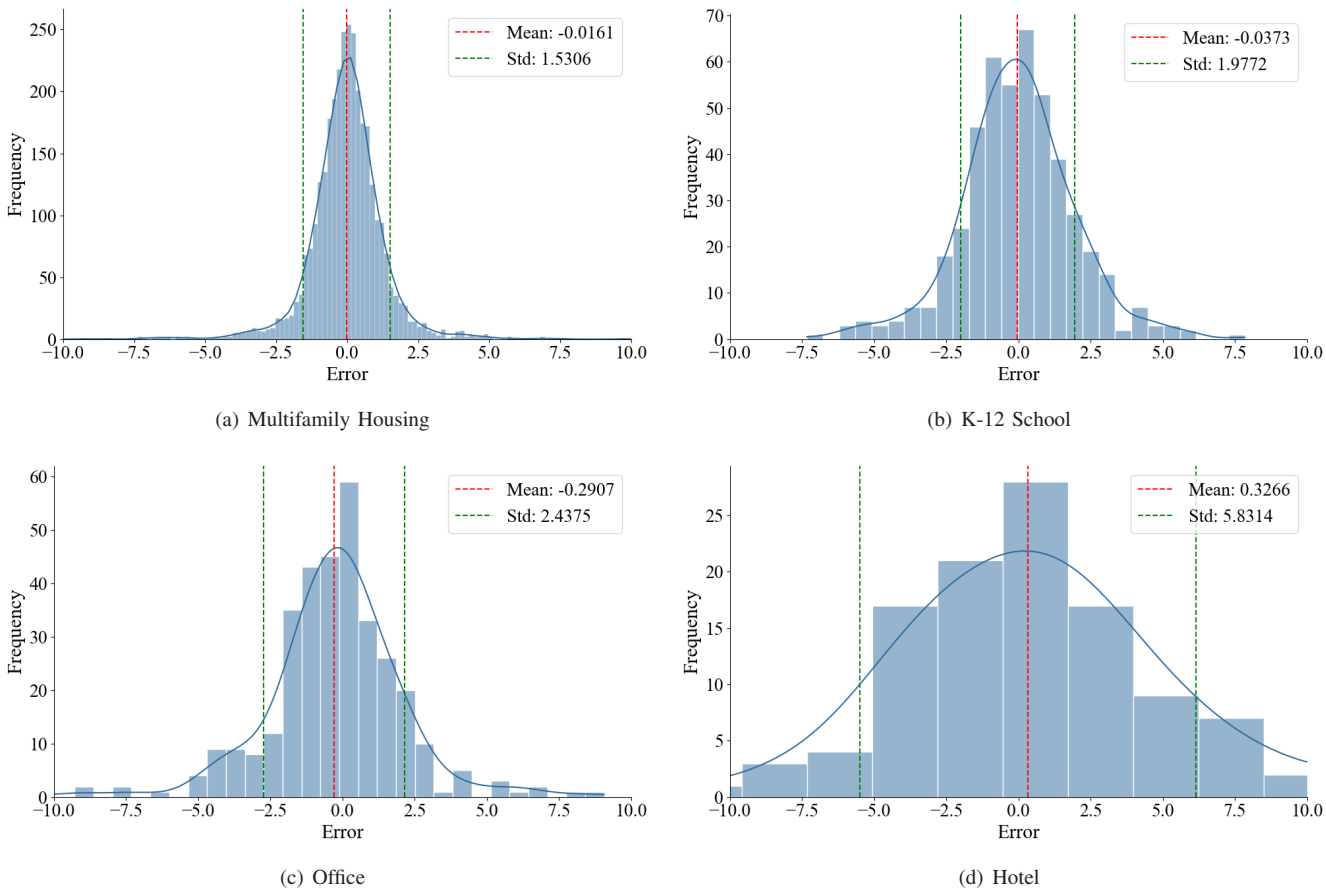
Our primary objective is to determine the model which best predicts the Energy Star Score of residential buildings. To achieve this goal, we split the dataset into two parts, 70% for training and 30% for testing. We enumerate a combination of different parameters and perform a 4-folds cross-validation to optimize each training model. The training model with the best performance under certain configuration will be used for the testing dataset. The entire experiment is repeated five times, and the average score and standard deviation are reported as the final results. Here, we list the parameters used for each model in the optimization process in Python 3.8.5:

- k-Nearest Neighbor Regression:
  - n_neighbors: [5, 10, 15, 20],

TABLE I
SUMMARY OF RESULTS OF THE CASE STUDY.

Building Type	Regressor	MAE	R-squared	Adjusted R-squared	SSE
Multifamily Housing	KNN	12.05±0.70	0.6718±0.0281	0.6655±0.0284	609722.05±52117.79
	Linear	10.96±0.24	0.6528±0.0638	0.6510±0.0642	955467.21±168765.45
	Ridge	11.49±0.33	0.6584±0.0478	0.6566±0.0480	940514.86±124643.94
	DT	1.41±0.05	0.9923±0.0011	0.9923±0.0011	21124.00±2978.79
	RF	2.49±2.07	0.9739±0.0276	0.9722±0.0282	48549.10±51378.32
	SV	6.73±1.82	0.8320±0.0354	0.8288±0.0360	312705.89±68675.44
	GB	**0.89±0.08**	**0.9967±0.0004**	**0.9966±0.0004**	**6199.90±806.99**
	XGB	1.16±0.04	0.9961±0.0006	0.9961±0.0006	10615.36±1691.00
	MLP	1.16±0.48	0.9937±0.0033	0.9937±0.0033	17512.53±9576.36
K-12 School	KNN	11.62±0.21	0.6450±0.0282	0.6308±0.0293	110224.45±8782.34
	Linear	6.46±0.14	0.8814±0.0125	0.8767±0.0130	36829.02±4025.72
	Ridge	7.16±0.16	0.8705±0.0092	0.8654±0.0095	40199.84±2926.95
	DT	2.75±0.20	0.9683±0.0078	0.9671±0.0082	9814.60±2335.98
	RF	1.82±0.08	0.9891±0.0010	0.9887±0.0010	3385.71±303.33
	SVR	17.31±0.27	0.2683±0.0240	0.2391±0.0249	227165.67±7852.88
	GB	**1.44±0.11**	**0.9909±0.0014**	**0.9906±0.0014**	**2812.97±436.87**
	XGB	1.76±0.08	0.9889±0.0014	0.9884±0.0014	3452.75±423.18
	MLP	3.40±0.26	0.9615±0.0063	0.9599±0.0065	11958.46±1934.40
Office	KNN	13.17±0.69	0.4940±0.0462	0.4668±0.0487	103329.18±9109.76
	Linear	7.00±0.26	0.7791±0.0587	0.7672±0.0619	45638.35±14317.48
	Ridge	7.97±0.49	0.7781±0.0260	0.7661±0.0274	45655.37±7580.89
	DT	2.99±0.16	0.9630±0.0053	0.9610±0.0056	7574.80±1146.03
	RF	**1.69±0.14**	0.9854±0.0022	0.9847±0.0023	2984.32±491.18
	SVR	16.65±0.66	0.1861±0.0508	0.1423±0.0535	166443.00±12696.20
	GB	1.72±0.15	**0.9894±0.0021**	**0.9888±0.0022**	**2169.33±408.58**
	XGB	1.90±0.16	0.9852±0.0042	0.9845±0.0044	2998.26±789.87
	MLP	6.99±0.78	0.8110±0.0437	0.8009±0.0461	39047.51±10999.97
Hotel	KNN	18.08±1.45	0.4294±0.0794	0.3455±0.0910	56948.41±9831.27
	Linear	10.18±0.54	0.6932±0.0959	0.6481±0.1100	30186.06±8709.30
	Ridge	12.82±0.96	0.6482±0.0618	0.5965±0.0709	34813.92±5025.98
	DT	6.86±0.31	0.8733±0.0154	0.8547±0.0177	12622.80±1825.87
	RF	4.96±0.46	0.9337±0.0189	0.9239±0.0216	6596.78±1943.27
	SVR	23.65±1.03	0.0729±0.0330	-0.0634±0.0378	92344.71±7476.52
	GB	**4.22±0.35**	**0.9483±0.0162**	**0.9407±0.0185**	**5150.55±1687.13**
	XGB	4.44±0.33	0.9398±0.0257	0.9309±0.0295	5978.02±2633.96
	MLP	15.62±1.15	0.5626±0.0335	0.4982±0.0384	43513.61±3876.85

  - weights: [‘uniform’, ‘distance’],
  - algorithm: [‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’],
  - leaf_size: [30, 40, 50]
- Ridge Regression:
  - alpha: [0.1, 1, 10, 100, 1000],
  - solver: [‘auto’, ‘svd’, ‘cholesky’, ‘lsqr’, ‘sparse_cg’],
- Decision Tree Regressor:
  - criterion: [‘squared_error’, ‘absolute_error’, ‘poisson’],
  - max_depth: [None, 2, 5, 10, 15],
  - min_samples_split: [2, 5, 10, 15],
  - min_samples_leaf: [1, 2, 4, 6],
  - max_features: [None, ‘sqrt’, ‘log2’]
- Random Forest Regression:
  - n_estimators: [100, 500, 900, 1100, 1500],
  - max_depth: [None, 2, 5, 10, 15],
  - min_samples_leaf: [1, 2, 4, 6, 8],
  - min_samples_split: [2, 4, 6, 10],
  - max_features: [‘sqrt’, None, 1]
- Support Vector Regression:
  - C: [0.1, 1, 10, 100],
  - kernel: [‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’],
  - gamma: [‘scale’, ‘auto’]
- Gradient Boosting Regression:
  - loss: [‘squared_error’, ‘absolute_error’, ‘huber’],



(a) Multifamily Housing (b) K-12 School

(c) Office (d) Hotel

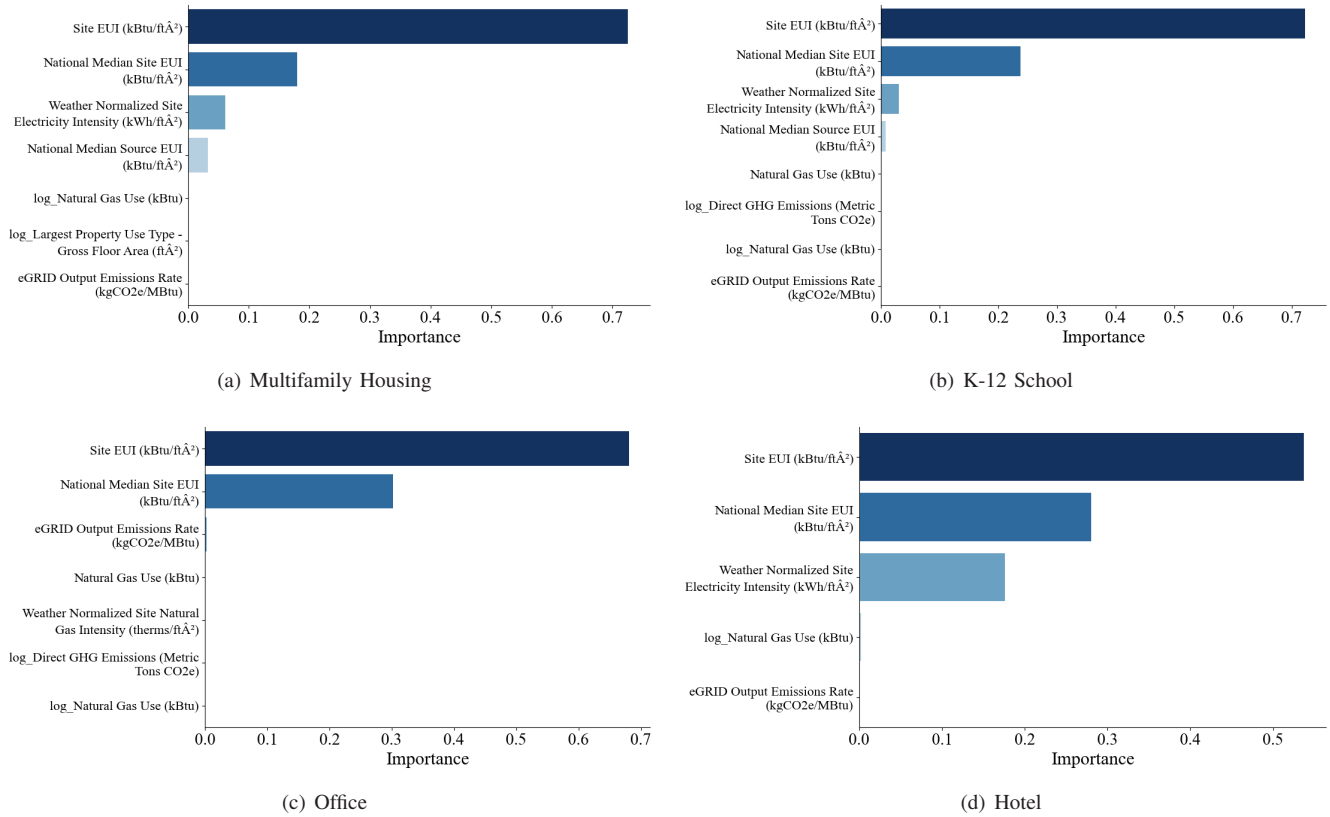
Figure 6. Distribution of residuals.

  - n_estimators: [100, 500, 900, 1100, 1500],
  - max_depth: [None, 2, 5, 10, 15],
  - min_samples_leaf: [1, 2, 4, 6, 8],
  - min_samples_split: [2, 4, 6, 10],
  - max_features: [‘sqrt’, None, 1]
- XGBRegressor:
  - n_estimators: [100, 200, 500, 1000],
  - max_depth: [3, 5, 7, 10],
  - learning_rate: [0.01, 0.1, 0.2, 0.3],
  - subsample: [0.8, 0.9, 1.0],
  - colsample_bytree: [0.8, 0.9, 1.0],
  - gamma: [0, 0.1, 0.2, 0.5]
- MLPRegression:
  - hidden_layer_sizes: [(50,), (100,), (100, 50), (100, 100)],
  - activation: [‘identity’, ‘logistic’, ‘tanh’, ‘relu’],
  - solver: [‘lbfgs’, ‘sgd’, ‘adam’],
  - alpha: [0.0001, 0.001, 0.01, 0.1],
  - learning_rate: [‘constant’, ‘invscaling’, ‘adaptive’],
  - max_iter: [200, 500, 1000]

Note that, there are no hyperparameters in Linear Regression, since its model parameters are determined directly by minimizing the least squares loss function. All machine learning models were implemented using Python with the Scikit-learn library, and the development environment was PyCharm Community Edition. Scikit-learn is a widely-used, open-source machine learning library that provides simple and efficient tools for data mining and data analysis. Detailed documentation and source code can be found on the official website [41].

D. Results

The analysis of Energy Star Score predictions across four building types in New York City consistently demonstrates the superiority of Gradient Boosting Regression (GBR), which achieves the lowest Mean Absolute Error (MAE) and Sum of Squared Errors (SSE), with R-squared values closest to 1 across all categories. GBR excels in multifamily housing (MAE: 0.89, R-squared: 0.9967), K-12 schools (MAE: 1.44, R-squared: 0.9909), offices (MAE: 1.72, R-squared: 0.9894), and hotels (MAE: 4.22, R-squared: 0.9483). Random Forest (RF) consistently ranks second, performing strongly in K-12 schools (MAE: 1.82, R-squared: 0.9891) and offices (MAE: 1.69, R-squared: 0.9854), while Extreme Gradient Boosting (XGB) follows closely. Support Vector Regression (SVR) shows inconsistent performance, ranging from poor in hotel (MAE: 23.65, R-squared: 0.0729) to moderate in multifamily housing (MAE: 6.73, R-squared: 0.8320). Simpler models like



(a) Multifamily Housing (b) K-12 School

(c) Office (d) Hotel

Figure 7. Distribution of importance ranking for the selected features.

KNN, Linear Regression, and Ridge Regression consistently underperform, with KNN showing particularly high MAEs across all types, especially in hotels (MAE: 18.08, R-squared: 0.4294). These results indicate the effectiveness of ensemble and boosting methods in accurately predicting Energy Star Scores for diverse urban building types while highlighting the limitations of simpler models in capturing the complex, non-linear relationships in building energy performance.

Given that the GBR model yielded the best performance across four types of buildings, we analyzed the residual distributions produced by GBR, as shown in Figure 6. Multifamily housing, with the largest dataset of 13,871 samples, shows the best performance with a mean error of -0.0161 and the lowest standard deviation of 1.5306, indicating highly precise and unbiased predictions. K-12 schools follow with a mean error of -0.0373 and a standard deviation of 1.9772. Office buildings show slightly less precision with a mean error of -0.2907 and a standard deviation of 2.4375. Hotels, with the smallest dataset of 393 samples, exhibit the highest variability with a mean error of 0.3266 and a standard deviation of 5.8314. The increasing standard deviations from multifamily housing to hotels directly correspond to the decreasing sample sizes, ranging from 13,871 to 393. Despite the differences in standard deviations, all distributions approximately follow a normal curve centered near zero, indicating that the regression models provide generally reliable predictions across all building types. Overall, the predictive performance is satisfactory and can offer valuable reference information for decision-makers in energy management and building efficiency across different types.

The feature importance analysis across all four building types reveals consistent patterns with some notable variations. For all building categories, “Site EUI" emerges as the most critical factor, with importance values ranging from approximately 0.5 to 0.7. “National Median Site EUI" consistently ranks second in importance across all types, though its influence varies, being particularly strong for offices and hotels. Hotels demonstrate a unique pattern with “Weather Normalized Site Electricity Intensity" having a notably higher importance, ranking third and showing more significance compared to other building types. For offices, the first two factors, “Site EUI" and “National Median Site EUI", significantly influence the model, with other factors showing much less importance. Multifamily housing shows a more balanced distribution of importance among secondary factors, with “National Median Source EUI" ranking fourth and contributing noticeably to the model. Across all building types, factors related to natural gas use and emissions generally show lower importance, though their rankings vary slightly between categories. This analysis highlights that while energy use intensity metrics are universally crucial for predicting Energy Star Scores, the relative importance of secondary factors can differ based on the specific building type, reflecting the unique energy consumption

patterns and characteristics of each category.

The importance values below 0.01 for the remaining features suggest that they have minimal influence on the model's predictions and can be considered less critical in explaining the variability in the Energy Star Score.

IV. CONCLUSION AND FUTURE WORK

Regression methods have been successfully applied to analyze and model Energy Star Scores across residential, educational, commercial, and lodging structures in New York City. Our comprehensive study, employing nine distinct regression models for these four building types, consistently demonstrates the superiority of the Gradient Boosting Regression (GBR) model. GBR outperforms other methods, achieving the best predictions with minimum errors and variances across all building types. Furthermore, the analysis highlights the universal importance of energy use intensity metrics, particularly "Site EUI" and "National Median Site EUI", while revealing varying influences of secondary factors specific to each building category. Moreover, accurately predicting building energy scores across various types will provide decision-makers with crucial information for retrofitting existing buildings and designing new, energy-efficient structures, ultimately contributing to reduced energy consumption, lower carbon emissions, and more sustainable urban development. Notably, our findings also indicate that the quantity of available data could impact model's stability, with larger datasets for multifamily housing buildings yielding less standard deviations compared to smaller datasets of hotels. Future research will focus on real-time energy emissions analysis and detailed energy usage distribution patterns to further refine energy conservation strategies across various building types.

V. ACKNOWLEDGMENTS

The work is partially supported by the National Science Foundation under NSF Awards Nos. 2234911, 2209637, 2100134. Any opinions, findings, or recommendations, expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] F. Zhang, B. Chen, F. Wu, and L. Bai, "Prediction of residential building energy star score: A case study of new york city", in *2024 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*, Porto, Portugal, Jul. 2024, pp. 180–186, ISBN: 978-1-68558-180-0.

[2] J. Syvitski *et al.*, "Extraordinary human energy consumption and resultant geological impacts beginning around 1950 ce initiated the proposed anthropocene epoch", *Communications Earth & Environment*, vol. 1, no. 1, p. 32, 2020.

[3] P. Nejat, F. Jomehzadeh, M. M. Taheri, M. Gohari, and M. Z. A. Majid, "A global review of energy consumption, co2 emissions and policy in the residential sector (with an overview of the top ten co2 emitting countries)", *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 843–862, 2015.

[4] K. Binita and M. Ruth, "Estimation and projection of institutional building electricity consumption", *Energy and Buildings*, vol. 143, pp. 43–52, 2017.

[5] W. Feist and J. Schnieders, "Energy efficiency–a key to sustainable housing", *The European Physical Journal Special Topics*, vol. 176, no. 1, pp. 141–153, 2009.

[6] *An assessment of energy technologies and research opportunities*, https://www.energy.gov/sites/prod/files/2015/09/f26/Quadrennial-Technology-Review-2015_0.pdf, Accessed: Aug 22, 2024, 2015.

[7] M. Santamouris and K. Vasilakopoulou, "Present and future energy consumption of buildings: Challenges and opportunities towards decarbonisation", *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 1, p. 100 002, 2021.

[8] *Our history about energy star*, https://www.energystar.gov/about/how-energy-star-works/history, Accessed: Aug 11, 2024, 2024.

[9] T. W. Hicks and B. Von Neida, "US national energy performance rating system and energy star building certification program", in *Proceedings of the 2004 Improving Energy Efficiency of Commercial Buildings Conference*, 2004, pp. 1–9.

[10] D. B. Crawley, "Building energy tools directory", *Proceedings of Building Simulation'97*, vol. 1, pp. 63–64, 1997.

[11] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies", *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018.

[12] S. Fathi, R. Srinivasan, A. Fenner, and S. Fathi, "Machine learning applications in urban building energy performance forecasting: A systematic review", *Renewable and Sustainable Energy Reviews*, vol. 133, p. 110 287, 2020.

[13] Q. Qiao, A. Yunusa-Kaltungo, and R. E. Edwards, "Towards developing a systematic knowledge trend for building energy consumption prediction", *Journal of Building Engineering*, vol. 35, p. 101 967, 2021.

[14] T. Ahmad *et al.*, "Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment", *Energy*, vol. 158, pp. 17–32, 2018.

[15] H. C. Jung, J. S. Kim, and H. Heo, "Prediction of building energy consumption using an improved real coded genetic algorithm based least squares support vector machine approach", *Energy and Buildings*, vol. 90, pp. 76–84, 2015.

[16] Z. Ma, C. Ye, and W. Ma, "Support vector regression for predicting building energy consumption in southern china", *Energy Procedia*, vol. 158, pp. 3433–3438, 2019.

[17] Z. Yu, F. Haghighat, B. C. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling", *Energy and Buildings*, vol. 42, no. 10, pp. 1637–1646, 2010.

[18] Y. Liu, H. Chen, L. Zhang, and Z. Feng, "Enhancing building energy efficiency using a random forest model: A hybrid prediction approach", *Energy Reports*, vol. 7, pp. 5003–5012, 2021.

[19] S. Touzani, J. Granderson, and S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings", *Energy and Buildings*, vol. 158, pp. 1533–1543, 2018.

[20] H. Lu, F. Cheng, X. Ma, and G. Hu, "Short-term prediction of building energy consumption employing an improved extreme gradient boosting model: A case study of an intake tower", *Energy*, vol. 203, p. 117 756, 2020.

[21] G. S. Georgiou, P. Christodoulides, and S. A. Kalogirou, "Implementing artificial neural networks in energy building applications—a review", in *2018 IEEE International Energy Conference (ENERGYCON)*, IEEE, 2018, pp. 1–6.

[22] N. Somu, G. R. MR, and K. Ramamritham, "A deep learning framework for building energy consumption forecast", *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110 591, 2021.

[23] S. Afzal, B. M. Ziapour, A. Shokri, H. Shakibi, and B. Sobhani, "Building energy consumption prediction using multilayer perceptron neural network-assisted models; comparison of different optimization algorithms", *Energy*, vol. 282, p. 128 446, 2023.

[24] F. Wahid, D. Kim, *et al.*, "A prediction approach for demand analysis of energy consumption using k-nearest neighbor in residential buildings", *International Journal of Smart Home*, vol. 10, no. 2, pp. 97–108, 2016.

[25] N. Fumo and M. R. Biswas, "Regression analysis for prediction of residential energy consumption", *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332–343, 2015.

[26] E. K. Laitinen and T. Laitinen, "Bankruptcy prediction: Application of the taylor's expansion in logistic regression", *International Review of Financial Analysis*, vol. 9, no. 4, pp. 327–349, 2000.

[27] D. J. Briggs *et al.*, "A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments", *Science of the Total Environment*, vol. 253, no. 1-3, pp. 151–167, 2000.

[28] E. Suárez, C. M. Pérez, R. Rivera, and M. N. Martínez, *Applications of Regression Models in Epidemiology*. John Wiley & Sons, 2017.

[29] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression", *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[30] J. Groß, *Linear regression*. Springer Science & Business Media, 2003, vol. 175.

[31] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[32] W.-Y. Loh, "Classification and regression trees", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[33] M. R. Segal, "Machine learning benchmarks and random forest regression", *UCSF: Center for Bioinformatics and Molecular Biostatistics*, 2004.

[34] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines", *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161, 1996.

[35] N. Duffy and D. Helmbold, "Boosting methods for regression", *Machine Learning*, vol. 47, pp. 153–200, 2002.

[36] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[37] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences", *Atmospheric Environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[38] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models", in *8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022)*, European Community on Computational Methods in Applied Sciences, 2022, pp. 1–25.

[39] A. V. Tatachar, "Comparative assessment of regression models based on model evaluation metrics", *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 09, pp. 2395–0056, 2021.

[40] *Energy and Water Data Disclosure for Local Law 84 2022 (Data for Calendar Year 2021)*, https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/7x5e-2fxh, [Online; retrieved: May,2024].

[41] *Scikit-learn*, https://scikit-learn.org, Accessed: June 22, 2024.