Prediction of Residential Building Energy Star Score: A Case Study of New York City

Fan Zhang
Computer Science Department
Tuskegee University
Tuskegee, USA
e-mail: vfun.zhang@gmail.com

Baiyun Chen*
Computer Science Department
Tuskegee University
Tuskegee, USA
e-mail: bchen@tuskegee.edu

Fan Wu
Computer Science Department
Tuskegee University
Tuskegee, USA
e-mail: fwu@tuskegee.edu

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

***Abstract*—Over the past few years, machine learning algorithms have garnered widespread attention in predicting the Energy Star Score of residential buildings. Traditional forecasting models, relying on software and statistical methods, failed to deliver accurate predictions owing to the intricacies of factors, non-linear relationships, and the noise of the data utilized in prediction. In this paper, we propose to use machine learning algorithms to enhance the performance of the Energy Star Score of residential buildings, due to their capability to capture the complex relationships between various kinds of features. We carefully choose the essential features to construct a regression model capable of accurately predicting the score through feature engineering and selection procedures. Additionally, we unveil the significant factors by ranking the importance of various features. Furthermore, we compare the performances of different machine learning algorithms in prediction and identify the optimal model, Gradient Boosting Regressor (GBR), as the best forecaster of Energy Star Scores for residential buildings in New York City. GBR outperforms all other methods, exhibiting the lowest Mean Absolute Error (MAE) of 0.89 and Sum of Squared Errors (SSE) of 6199.90, as well as R^2 of 0.9967 and adjusted R^2 of 0.9966. The variances for all the metrics in the GBR model are also minimized. Our study results not only enhance the prediction performance of energy scores but also provide valuable insights for decision-makers involved in retrofitting or constructing similar residential buildings with energy-saving considerations.**

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

I. Introduction

As economic and social development has progressed, the consumption of energy and water resources by human behaviors have increased by an order of magnitude, leading to a rise in annual carbon dioxide emissions and a severe reduction of water resources [1]. This trend has significant implications for the sustainable development of human society. Buildings account for approximately 40% of the global energy consumption and this percentage will continue to grow in the coming decades [2]. Notably, residential buildings are responsible for almost 70% of the energy consumption of the sector, mainly due to the usage for cooking and heating [3]. Fortunately, it illustrates a great potential to enhance the energy efficiency of residential buildings by analyzing the retrofit options or adjusting human activities in energy consumption. Hence, it is necessary to estimate the Energy Star Score of residential buildings, which is designed to assess the energy efficiency of buildings or appliances by the U.S. Environmental Protection Agency (EPA) and the U.S. Department of Energy (DOE) [4].

The primary challenge lies in accurately predicting residential 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 residential buildings energy consumption. The Building Energy Software Tools Directory [5] provides comprehensive information on building software tools for evaluating energy efficiency and sustainability in buildings. It also shows that efforts can be derived for different components to minimize energy consumption. With the widespread application of machine learning techniques, a growing number of researchers have recently proposed to introduce machine learning algorithms in predicting residential building energy consumption. In [6], Allard et al. compare the traditional calculation methods for energy performance analysis in Nordic countries, highly depending on the definition of energy performance in various countries. In [7], a neural network is trained for modeling and estimating the hourly energy consumption for a typical residential building in Athens. In [8], Kialashaki and Reisel employ both artificial neural networks and regression models to model the energy demand in the residential sector of the United States, forecasting energy demand in the residential sector until 2030. In [9], Swan and Ugursal provide a review of the multiple techniques used for modeling residential sector energy consumption, where regression and neural networks are utilized to identify the impactors of end-use energy consumption. Although these researches concentrate on distinct areas, they all offer a basis for forecasting residential energy consumption.

This paper focuses on urban residential areas due to the high population density and energy consumption in metropolitan areas. Five representative regression models are used to forecast the Energy Star Score of residential buildings employing the

*: Baiyun Chen (bchen@tuskegee.edu) is the corresponding author.

disclosed energy and water consumption data from New York City, and the performances of various approaches are compared. Furthermore, by evaluating the significance of various features, this study identifies factors that significantly affect energy consumption, offering guidance for future home design or retrofit as well as human activities in heating and cooking to support the attempts to reduce emissions and conserve energy.

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 V.

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 [10], bankruptcy prediction[11], air pollution [12], epidemiology [13], and some other applications. This study introduces 5 typical regression methods, including k-Nearest Neighbor Regression [14], Linear Regression [15], Random Forest Regression [16], Support Vector Regression [17], and Gradient Boosting Regression [18] 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 [19]. 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 five 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 technique used for estimating the value of a continuous target variable. In kNN regression, the predicted value for a given data point is determined by averaging the target values of its k nearest neighbors [14]. 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 linear approach to model the relationship between a dependent variable y and one or more independent variables $\boldsymbol{x}$ [15]. 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. Random Forest Regression

Random Forest Regression is an ensemble learning method that combines multiple decision trees to make predictions. Each tree in the forest independently predicts the target variable, and the final prediction is the average value of all the predictions from individual trees [16]. $\hat{y}$ is predicted by (3):

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{x}), \tag{3}$$

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.

D. Support Vector Regression

Support Vector Regression uses support vector machines to search for the best-fitting hyperplane to predict the target variable. It aims to minimize margin violations while ensuring that deviations from the predicted values (the errors) are within a predefined margin [17]. $\hat{y}$ is predicted by (4):

$$\hat{y} = \mathbf{w}^T \cdot \boldsymbol{x} + b, \tag{4}$$

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

E. Gradient Boosting Regression

Gradient Boosting Regression also builds a sequence of decision trees. The difference lies in that each tree corrects the errors made by the previous ones. It minimizes the loss function by adding trees sequentially in a greedy manner [18]. $\hat{y}$ is predicted by (5):

$$\hat{y} = \sum_{i=1}^{N} \gamma_i f_i(\boldsymbol{x}) \tag{5}$$

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.

F. 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 [20].

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{6}$$

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

$$\text{Adjusted } R^2 = 1 - \frac{\left(1 - R^2\right) \cdot (n-1)}{n-k-1} \tag{9}$$

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

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 [21]. It encompasses a diverse range of building types, including schools, banks, hospitals, factories, multifamily residences, and various other structures. To focus on the residential buildings energy consumption, we utilize the subset of the multifamily energy and water data. Excluding the rows containing the missing values and outliers, we extract 7888 tuples from the entire 22479 rows.

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 [21].

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 1.

From Figure 1, we can see that the distribution of the Energy Star Score does not conform to either a uniform or a normal distribution. Instead, it exhibits high frequency at both ends with a relatively lower and uneven distribution in the middle. Due to its non-uniform and non-normal distribution, traditional statistical methods are inappropriate for modeling the score distribution. Instead, regression prediction emerges as a feasible solution.

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.

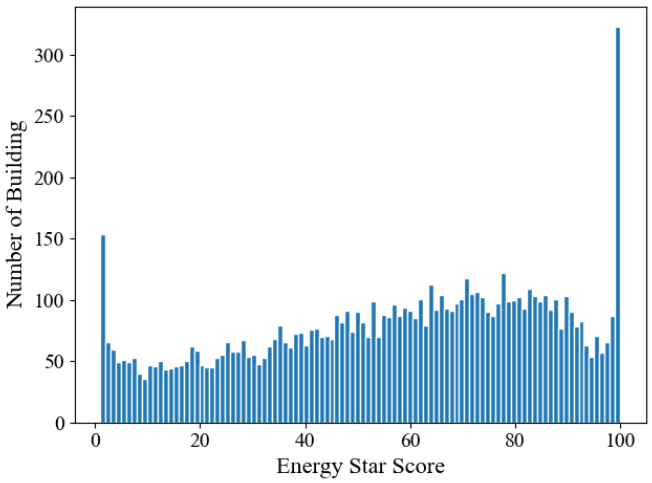

Figure 1. Distribution of energy star score.

In this study, we initially 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 2. We first categorize the datasets into different groups based on their different districts, then we employ the Gaussian Kernel function to smooth the probability density estimation of different groups.

From Figure 2, it can be found that the different groups show similar Energy Star Score distribution, implying that it lacks sufficient discriminative power to distinguish the target variable. Hence, this feature is not suggested to be maintained in the modeling process.

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 3 demonstrates the correlation analysis result of "Site EUI" and "Weather Normalized Site EUI" in the scatter diagram. EUI is the Energy Use Intensity, which measures the ratio of actual energy consumption of a building or site to its area. The correlation coefficient between the two features is

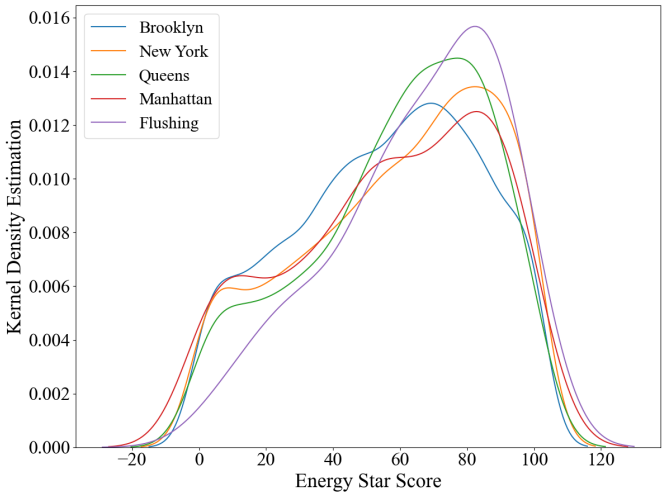

Figure 2. Distribution of energy star score in different districts.

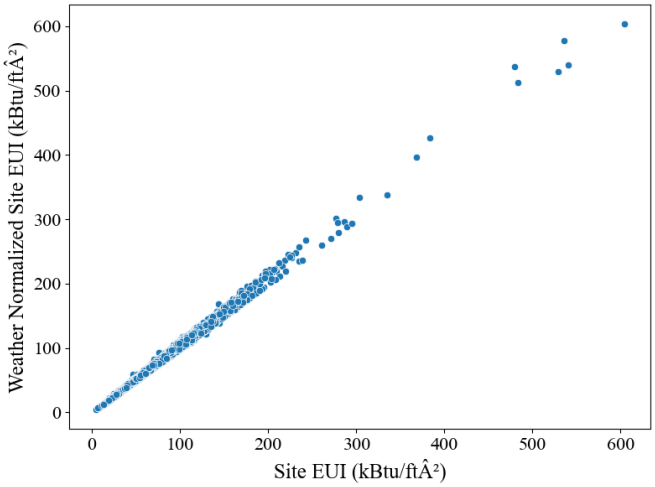

Figure 3. Distribution of correlation between “Site EUI (kBtu/ft²)” and “Weather Normalized Site EUI (kBtu/ft²)”.

up to 0.9969, indicating an extremely strong positive linear relationship between them. 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. 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. In this case, only one of the features needs to be retained in the later modeling process. To enhance the interpretability of the model, we opt for keeping the “Site EUI" feature.

B. Feature Selection and Feature Engineering

Due to data measurement and collection challenges, numerous features in the original dataset contain missing or unavailable data. After removing these features and those exhibiting highly correlated features, we identified a total of 11 numeric features to construct the regression model. The selected features exhibit correlations of less than 0.7 between each other, as depicted in Figure 4.

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 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 4, the features starting with “log_” are the ones transformed by the logarithm functions.

Secondly, we utilize one-hot encoding to transform the categorical variables into numerical representations. One-hot encoding is a widely used technique for handling categorical variables. In this study, we apply one-hot encoding to the "district" feature. However, as illustrated in Figure 2, this feature demonstrates limited predictive power. Therefore, we exclude it from the regression model construction.

The last step in data preprocessing involves applying 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.

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],
  - weights: [‘uniform’, ‘distance’],
  - algorithm: [‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’],
  - leaf_size: [30, 40, 50]
- 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],

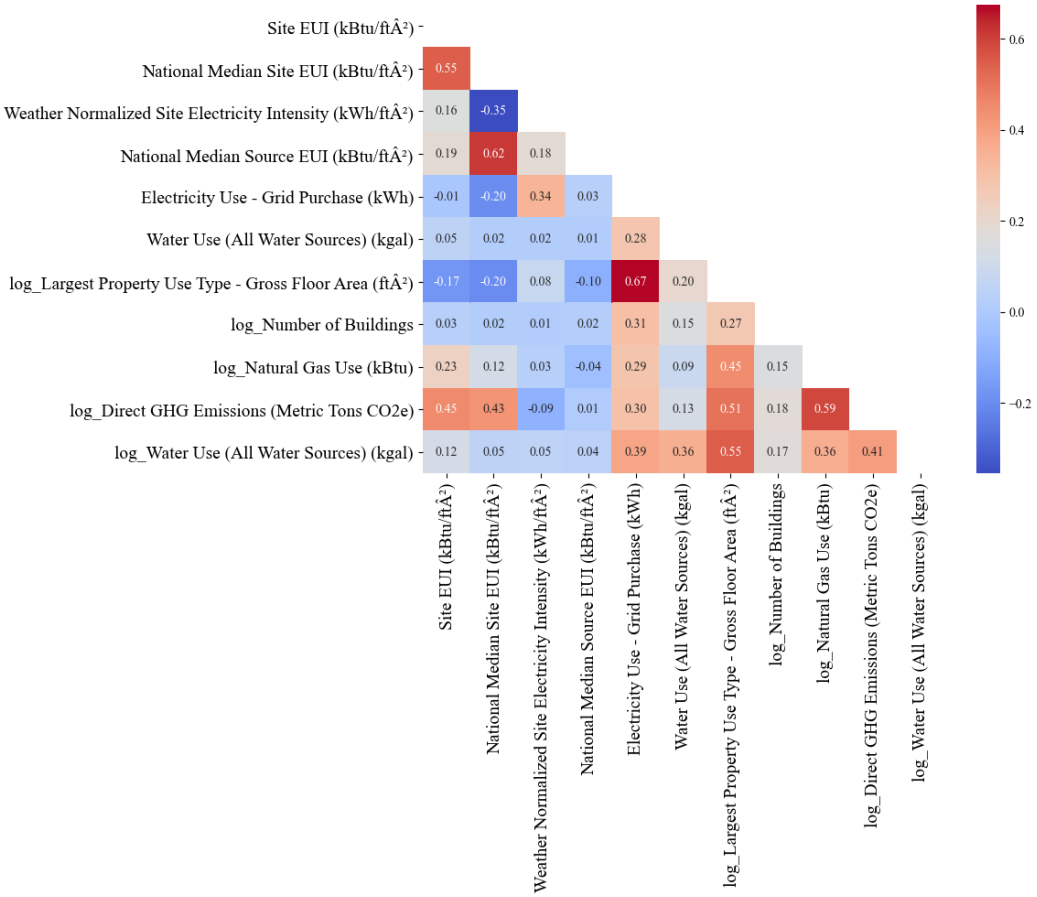

Figure 4. Correlation matrix of selected features.

TABLE I
SUMMARY OF RESULTS OF THE CASE STUDY.

Regressor	MAE	R^2	Adjusted R^2	SSE
GradientBoosting	**0.89±0.08**	**0.9967±0.0004**	**0.9966±0.0004**	**6199.90±806.99**
RandomForest	2.49±2.07	0.9739±0.0276	0.9722±0.0282	48549.10±51378.32
SVR	6.73±1.82	0.8320±0.0354	0.8288±0.0360	312705.89±68675.44
Kneighbors	12.05±0.70	0.6718±0.0281	0.6655±0.0284	609722.05±52117.79
Linear	9.28±0.16	0.7469±0.0287	0.7420±0.0292	470894.70±58898.11

  - kernel: [‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’],
  - gamma: [‘scale’, ‘auto’]
- Gradient Boosting Regression:
  - loss: [‘squared_error’, ‘absolute_error’, ‘huber’],
  - 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]

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 [22].

D. Results

The prediction results of the Energy Star Score of residential buildings in New York City are reported in Table I. Gradient Boosting Regression (GBR) outperforms all other methods, exhibiting the lowest Mean Absolute Error (MAE) of 0.89 and Sum of Squared Errors (SSE) of 6199.90, as well as the values

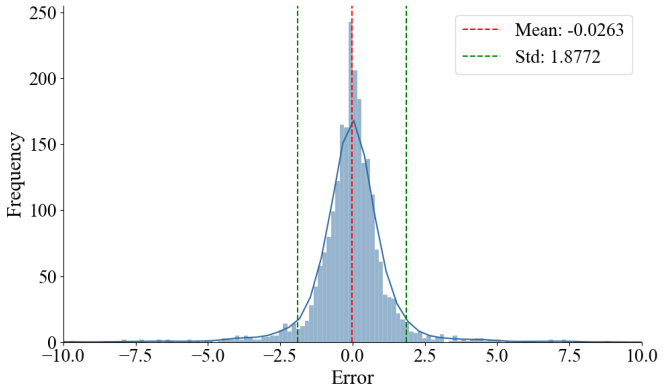

Figure 5. Distribution of residuals.

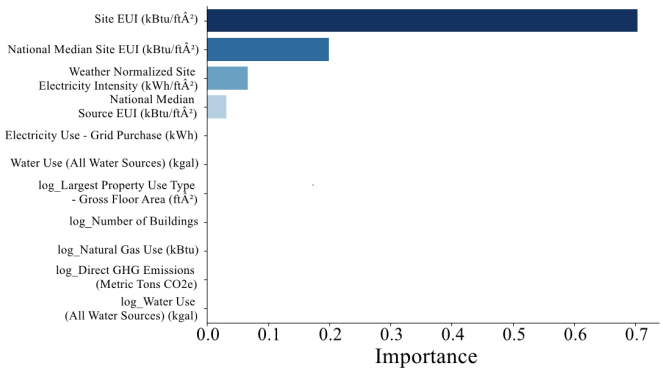

Figure 6. Distribution of importance ranking for the selected features.

closest to 1 for both R^2 of 0.9967 and adjusted R^2 of 0.9966. Besides, the variances for all the metrics in the GBR model are minimized. The promising outcome in Table I also shows the potential to exploit machine learning techniques to predict the Energy Star Score for residential buildings in urban areas. These results empower decision-makers to pinpoint necessary updates for retrofitting or constructing similar buildings, particularly for reducing energy consumption. Random Forest regression also exhibits a notable prediction performance regarding the MAE, R^2, and adjusted R^2 metrics. However, the remaining three models demonstrate weaker predictability across all metrics, indicating their limited applicability to this dataset.

Given that the GBR model yielded the best performance, it obtains a mean error of 0.0467 and a standard deviation of 1.8396. The corresponding residual histogram of predicting the Energy Star Score closely aligns with a normal distribution, as shown in Figure 5. It indicates that the residuals are distributed with a narrow dispersion around the mean, implying that the model's predictions are unbiased and reliable.

Taking a closer look at the GBR model, we focused our analysis on the features with the greatest impact on predicting the Energy Star Score. Figure 6 illustrates the rank of importance values for each numeric feature. "Site EUI" has the highest importance value of 0.703892849, suggesting that it has the most significant impact on the predicted score. "National Median Site EUI" follows with an importance value of 0.215082687, indicating that it also plays a notable role in the predictions, albeit to a lesser extent than "Site EUI". "Weather Normalized Site Electricity Intensity" and "National Median Source EUI" have importance values of 0.052931567 and 0.027565661, respectively. While these features contribute to the model's predictions, their impact is comparatively smaller compared to the previous two features.

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, one of the most used machine learning techniques, are used to analyze and model the Energy Star Score of residential buildings in New York City. The result shows that the Gradient Boosting Regression model exceeds all other methods, achieving the best prediction with the minimum errors and variances. These findings have important ramifications for modeling and analysis of predicting energy use trends in the future. The regression model can also be broadened to forecast energy ratings for many other buildings, such as business, medical, and educational buildings. Furthermore, accurately predicting building energy scores will aid decision-makers in retrofitting or constructing similar buildings, which is crucial for reducing energy consumption and carbon emissions, and promoting sustainable development. In the future, we will further investigate real-time residential energy emissions and conduct detailed research on the distribution of residential energy usage to guide users in energy conservation and emission reduction efforts.

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] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] I. Allard, T. Olofsson, and O. A. Hassan, "Methods for energy analysis of residential buildings in nordic countries", *Renewable and Sustainable Energy Reviews*, vol. 22, pp. 306–318, 2013.

[7] G. Mihalakakou, M. Santamouris, and A. Tsangrassoulis, "On the energy consumption in residential buildings", *Energy and buildings*, vol. 34, no. 7, pp. 727–736, 2002.

[8] A. Kialashaki and J. R. Reisel, "Modeling of the energy demand of the residential sector in the united states using regression models and artificial neural networks", *Applied Energy*, vol. 108, pp. 271–280, 2013.

[9] L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques", *Renewable and sustainable energy reviews*, vol. 13, no. 8, pp. 1819–1835, 2009.

[10] 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.

[11] 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.

[12] 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.

[13] E. Suárez, C. M. Pérez, R. Rivera, and M. N. Martínez, *Applications of regression models in epidemiology*. John Wiley & Sons, 2017.

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

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

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

[17] 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.

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

[19] 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.

[20] 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.

[21] *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].

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