A Comparative Study of Machine Learning Models in Predicting Commercial Building Electricity Consumption

Xiaohan Liu

1 Introduction

Predicting electricity consumption in commercial buildings is essential for energy management and policy planning, as it aids in understanding energy demands and supports the development of sustainable energy strategies. This project employs data from the 2018 Commercial Buildings Energy Consumption Survey (CBECS), focusing specifically on the dataset's electricity-related features.

The primary objective of this study is to develop predictive models using linear regression, random forest, and XGBoost techniques to forecast annual electricity consumption (kWh) in commercial buildings. The project compares the performance of these three methods, aiming to identify the key factors influencing electricity consumption the most effective approach in predicting and analyzing electricity consumption patterns.

Through this comparative analysis, the study aims to enhance the understanding of electricity consumption in commercial environments. The insights gained may inform future energy policies and strategies for building design and sustainability. This report is structured to walk through the model settings, result analysis, and a discussion in terms of future research direction.

2 Methods

2.1 Dataset

The dataset used in this study is derived from 2018 CBECS Survey, which comprises a comprehensive set of 1249 variables. However, not all these variables are relevant to the study's focus on electricity consumption, so a feature selection process was undertaken.

The primary response variable of interest for this analysis is 'ELCNS', representing the annual electricity consumption in kilowatt-hours (kWh). Alongside this, variables specifically related to electricity usage were selected to serve as predictors. During the selection process, it was noted that a majority of these variables had over 50% missing values. Given the extensive missing data, imputation methods are considered impractical, as they could introduce significant biases and inaccuracy. Consequently, the predictors with high levels of missing data were excluded from the analysis. An exception was the variable 'DAYLTP', indicating the percentage of daylight. 'DAYLTP' had approximately 279 missing entries. In this case, median imputation was employed to fill the missing values. For the response variable 'ELCNS', only 79 missing values were observed. Imputing missing values for the response can be particularly challenging, as it directly affects the outcome of the predictive models. Therefore, to ensure the reliability of the analysis, observations with missing response values were removed from the dataset.

The final dataset comprised 6357 observations with 15 predictors. These predictors consisted of 3 continuous variables, 11 categorical variables, and 1 ordinal variable. The variables covered a broad range of aspects, such as building size, building age, building type, geographic location, and the usage of electricity for various functions within the building, like air conditioning and cooking.

2.2 Model Setup

In the development of the predictive models, the project constructed three distinct types of models, including linear regression, random forest, and XGBoost.

The process involved a training/testing split process, dividing the dataset into a training set comprising 80% of the data and a testing set with the remaining 20%. This split ensures that the models are trained on a substantial portion of the data, while still having a separate dataset to evaluate their performance.

The modeling process began with a multiple linear regression model using the entire set of predictors in the training dataset. The formula are expressed as

$$\begin{aligned} ELCNS \sim & SQFT + WLCNS + RFCNS + ELHT1 + ELCOOL + \\ & ELWATR + ELCOOK + ELMANU + CAPGEN + \\ & BLDSHP + YRCONC + PBA + CENDIV + PUBCLIM + DAYLTP \end{aligned} \tag{1}$$

For the random forest model, the parameter 'mtry' representing the number of variables to randomly sample as candidates at each split was tuned through 10-fold cross validation. A grid search was conducted by varying 'mtry' over a range from 2 to the number of features, in steps of 3. After determining the optimal parameter, the final random forest model was trained with 500 trees. Predictions were then made on the testing set to evaluate the model's performance.

For the XGBoost model, I applied one-hot encoding to the categorical variables, converting them into binary format. Following the data preparation, the XGBoost model requires a detailed tuning of its parameters, achieved also through a grid search approach and 5-fold cross validation. The parameters and their respective ranges were set as

- number of boosting rounds: 50, 100, 150
- maximum depth of a tree: 3, 5, 7
- learning rate: 0.01, 0.05, 0.1
- minimum loss reduction required for further partition: 0, 0.1, 0.2
- fraction of features used per tree: 0.5, 0.7, 0.9
- minimum sum of instance weight needed in a child: 1, 3, 5
- subsample ratio of the training instances: 0.5, 0.7, 0.9

The final XGBoost model was trained with the optimal combination of parameters obtained.

3 Results

3.1 Parameter Tuning Results

For the Random Forest model, the optimal number of variables to sample as candidates at each split ('mtry') was determined to be 14. In the case of the XGBoost model, the grid search process identified an optimal set of parameters that reached the least RMSE. The best-performing model was achieved with 100 boosting rounds ('nrounds'). The trees were allowed a maximum depth ('max_depth') of 3. The learning rate ('eta') was set at 0.05, balancing the speed of convergence. A 'gamma' value of 0.2 helped to control the model's tendency to overfit by requiring a minimum loss reduction to make additional partitionings on a tree's leaf nodes. The 'colsample_bytree' was optimized at 0.7, indicating that 70% of the features were considered for building each tree, providing a good compromise between model robustness and prediction accuracy. The 'min_child_weight' was set at 3, which is a regularization parameter that can help to prevent overfitting by making the algorithm more conservative. Finally, the 'subsample' ratio was chosen to be 0.9, suggesting that a high percentage of the data was used to train each individual tree.

3.2 Model Comparison

Three models are evaluated through the performance metrics including the coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE). As summarized in Table 1, the random forest model outperforms the other two models with an R^2 value of 0.7323, indicating that approximately 73.23% of the variance in the electricity consumption can be explained by the model. This model also achieves the lowest RMSE and MAE, with values of 292,852 and 100,013 respectively, suggesting a strong predictive capability with the least deviation from the actual values. XGBoost follows closely, with an R^2 of 0.7266. However, it exhibits a slightly higher RMSE and MAE than that of the random forest model, indicating marginally less precise predictions on average. The linear regression model, while still demonstrating moderate predictive power with an R^2 of 0.6328, has higher RMSE and

MAE values of 343,489 and 146,477 respectively. This indicates that the linear regression model does not predict values as close to the actual values as the more complex models. The results highlight the effectiveness of ensemble methods such as Random Forest and XGBoost given the complexity of the dataset. The betther performance of these models suggest that they are suitable for capturing nonlinear relationships and interactions between variables that may exist in the data.

	R^2	RMSE	MAE
Linear Regression	0.6328	3439489	1467477
Random Forest	**0.7323**	**2925852**	**1005013**
XGBoost	0.7266	2933880	1161357

Table 1: R^2, RMSE, MAE of three models

The residual plots in Figure 1 provide a visual representation of the differences between the actual and predicted values for each model. The residuals of linear regression model increase with the increase of electricity consumption. Although the overall performance is not as good as the other two models, the linear regression model is better at predicting extreme values than the remaining two. For the random forest model, the residuals are more densely clustered around the zero line, suggesting that the model has a better predictive performance. However, there are still notable deviations for higher values of electricity consumption, which may indicate that while the model captures the general trend well, it may still struggle with certain complex patterns in the data. The XGBoost model presents a slightly higher concentration of data points close to the zero line of residuals. It also shows some challenges with extreme values, which could be attributed to outliers that are not well-represented in the training data.

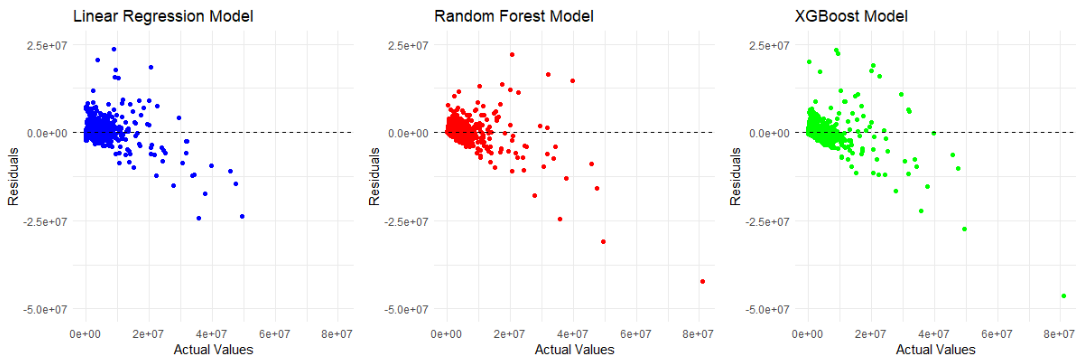

Figure 1: Residual plots of three models

4 Conclusion

This study developed predictive models for annual electricity consumption in commercial buildings using a dataset from the 2018 CBECS survey. Through the application of linear regression, random forest, and XGBoost methods, I sought to uncover patterns and predictors that significantly influence electricity consumption.

The random forest model performs the best, demonstrating the highest R^2 value and the lowest errors, indicating its robustness in capturing the complex relationships within the data. The XGBoost model follows closely. The linear regression model, despite the simplicity, provides a moderate level of predictive power, but does not perform as well as ensemble methods when handling larger-scale consumption data. Residual plots reveal all models faced challenges with extreme values. These insights highlight the potential for further refining techniques.

In conclusion, the findings of this study affirm the potential of advanced machine learning techniques in predicting electricity consumption. The successful application of these models can guide energy management strategies, inform policy planning, and support the drive towards more sustainable energy use in commercial buildings. For future research, the incorporation of additional data, such as temporal patterns and more granular usage details, could further enhance model accuracy and provide deeper insights into energy consumption behaviors.