REPORT/WHITE PAPER

ENERGY CONSUMPTION PREDICTIVE MODEL

Joseph Choi | Jenny Overby | Daniel Meier | Serge Nane

DSC450 Applied Data Science
Spring 2024

TABLE OF CONTENTS

INTRODUCTION	3
BUSINESS PROBLEM/HYPOTHESIS	3
METHODS/ANALYSIS	4
RESULTS	6
RECOMMENDATIONS	8
CONCLUSION	9
REFERENCES	10
APPENDIX	12

INTRODUCTION

In today's world, energy sustainability has long been an important topic of discussion as it directly impacts the health of our planet and the quality of life for current and future generations. Therefore, optimizing and conserving energy consumption has been globally prioritized as an essential goal in combating climate change, preserving natural resources, and encouraging people to care for the environment for a more sustainable future (Zaharia et al., 2019). One way to contribute and support sustainability is by making changes in our own homes (Chen et al., 2023). For instance, reducing energy usage by turning off lights and appliances when not in use can make a difference in conserving resources and protecting the environment (Debebe et al., 2023). Taking these concerns into consideration, our project's objective is to address them using data analysis and data science methodologies.

To achieve this, we are using simulated historical data to anticipate future energy usage and conduct extensive analysis to comprehend the factors affecting energy consumption in a hypothetical setting. The dataset, sourced from Kaggle (Energy-consumption-prediction (kaggle.com)), consists of time-stamped records detailing logged energy consumption measures at a specific date and time, along with variables influencing energy usage. Here are the dataset's variables and their descriptions:

- **Timestamp**: Date and time when the data was recorded (1/1/22 - 2/11/22)
- **Temperature**: Degree of Celsius at the time of the recording
- **Humidity**: Amount of moisture in the air as a percentage
- **Square Footage**: Total area of the space, measured in square feet
- **Occupancy**: Number of people present in the space
- **HVAC Usage**: Indicates whether an HVAC system was used or not (On/Off)
- **Lighting Usage**: Indicates whether lighting systems were used or not (On/Off)
- **Renewable Energy**: Contribution of renewable energy to the total energy usage in percentage
- **Day Of Week**: Days ranging from Monday to Sunday
- **Holiday**: Indicates whether the data was recorded on a holiday or not (Yes/No)
- **Energy Consumption**: Measure that indicates total energy consumption

Pursued by Joseph Choi, Jenny Overby, Daniel Meier, and Serge Nane, our primary focus is to identify key drivers strongly correlated to energy consumption, explore different machine learning methodologies for accurate prediction, and share our findings.

BUSINESS PROBLEM/HYPOTHESIS

The core of our project revolves around two fundamental research questions:

1. What are the key indicators or factors influencing energy consumption?
2. Can we accurately predict future energy consumption in buildings using historical usage data?

Throughout the project, we focused on addressing these two main business problems, ensuring that all our data analysis and data science tasks stayed aligned with these objectives to avoid going out of scope.

Analyzing this data and extracting insights into these business problems can greatly impact building energy management practices as it can empower everyday energy users to make informed decisions and manage energy more effectively. Identifying and understanding the main influencing factors affecting energy consumption promotes efficiency, reduces costs, and supports sustainability efforts (Guo et al., 2018).

Based on our initial research and our preliminary investigation of our dataset, our hypothesis is as follows:

- We anticipate that variables like Temperature and HVAC usage will most likely significantly impact energy consumption. This is because of their direct influence on heating and cooling systems, which are widely known to be major contributors to overall energy usage. Our hypothesis is supported by our high energy bills during temperature fluctuations, such as in summer or winter, prompting our behavior to increase HVAC usage, consequently increasing energy consumption.
- Our assumption about the machine learning model that would project the best output and accuracy is the Random Forest Regressor. As highlighted by Sarswatula, Pugh, and Prabhu (2022), the Random Forest has a proven record of performing well in various predictive modeling scenarios, particularly involving non-linear relationships and interactions between multiple variables, which is the case with our dataset.

METHODS/ANALYSIS

Our approach encompasses a structured plan to arrive at our goal, which includes four phases. The description of each of these phases and the analysis we made are as follows:

Phase 1: Data Wrangling

To kickstart the project, our team explored and assessed the data quality of our dataset to understand its readiness for analysis and identify any inconsistencies or dirty data. From the initial quality assessment, we found a list of issues that needed to be addressed before proceeding to the next phase of the project. The data cleaning and transformation tasks we performed included:

- Correcting features with incorrect data types
- Handling missing values
- Standardizing values from categorical features
- Addressing potential miss-entered values from numerical and categorical features
- Converted temperature from Celsius to Fahrenheit

Phase 2: Exploratory Data Analysis

Once we confirmed that the dataset was ready for further analysis, we proceeded to exploratory data analysis (EDA). Our objective in the EDA phase was to comprehend three aspects. First, we wanted to understand the distribution of our features to identify any data skewness. Next, we tried to identify any outliers in the features and address them if needed. Lastly, we wanted to examine the correlations between our input and target variables. Here are the insights we extracted from our EDA phase:

- **Distribution**: Histograms and bar charts were plotted to grasp the distributions for each feature.
- **Outliers**: We built box plots to detect outliers for each feature. Our visualizations identified only one outlier point in the “EnergyConsumption” feature. After review, our team concluded that it should be retained as the recorded row is valid.
- **Correlation**: A correlation matrix (Fig. 1) was constructed to understand the relationship between each feature. We’ve identified three notable features with strong positive correlations from the computed correlation coefficient: Temp_F, HVACUsage, and Occupancy. Other features had little to no correlation with the target variable. In addition, many of the input

features indicated weak correlations with each other, indicating no significant dependencies among them.

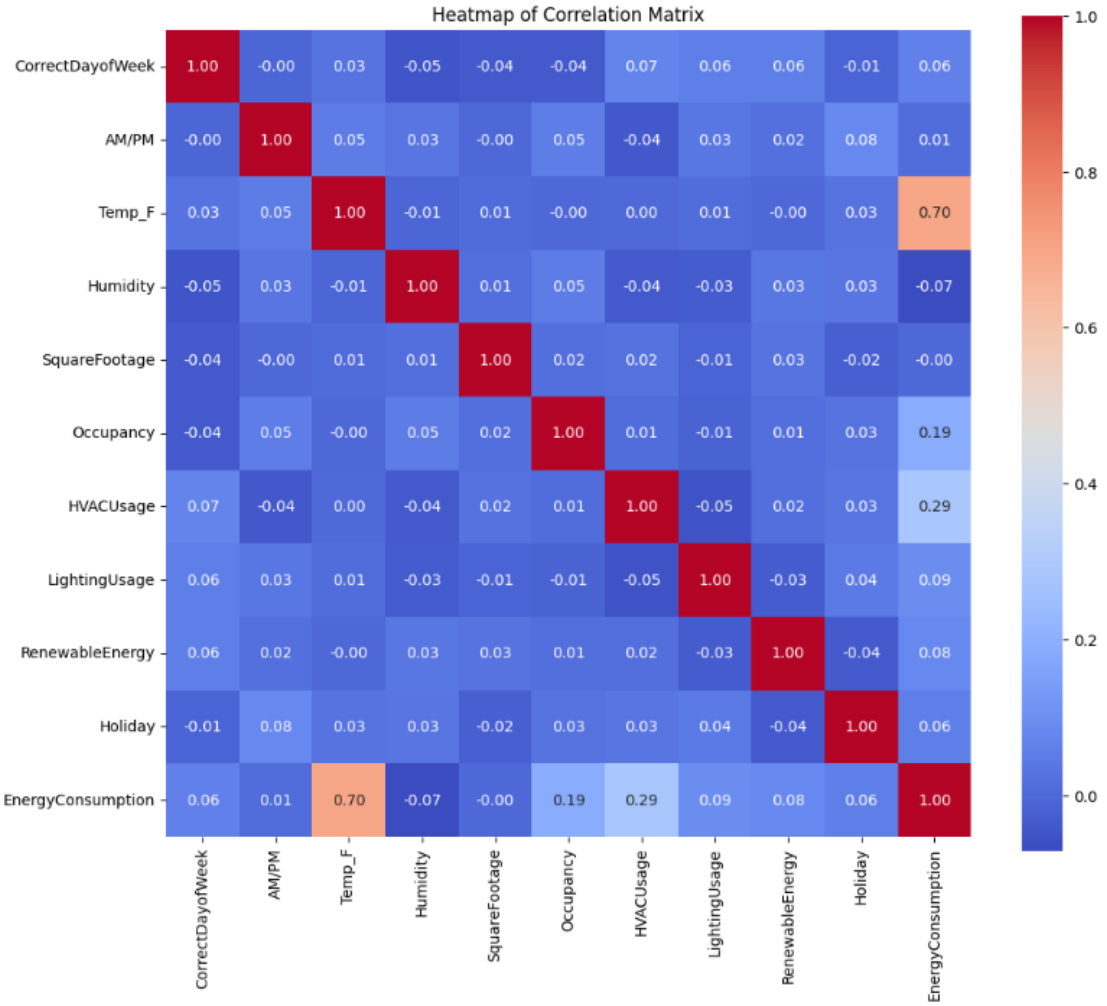

Fig. 1: Correlation Matrix Heatmap

Phase 3: Feature Engineering

After completing the EDA phase, we transitioned to the feature engineering stage of the project. The goal here was to prepare our data for model training by transforming it to ensure optimal learning and predicted output. Some of the feature engineering tasks we performed include:

- Separating the date and time from the “Timestamp” feature to extract AM or PM information, as unique columns like “Timestamp” do not directly help the learning process due to the lack of repeatable patterns
- Using MinMaxScaler to scale numerical features as this process ensures uniform scaling across all features. Scaling is performed to prevent features with more extensive ranges from dominating the model’s decision-making process.

- Encoding categorical features to convert them into a format that machine learning models can better understand and utilize. Depending on the nature of the feature, data was transformed into either numerical labels or binary variables.

Phase 4: Data Modeling

During the data modeling phase, our general workflow involved training and evaluating the model's performance metrics. Our approach followed an iterative process, experimenting with different model types (random forest regressor and linear regression), hyperparameters using GridSearch, and various combinations of input features. Our initial model included all the features to assess its performance. Based on the insights we gathered from the EDA, we carefully removed redundant and irrelevant features to enhance the model's performance. Details of the model's performance and its output will be discussed in the Result and Appendix section of the report.

RESULTS

Based on the insights derived from the correlation matrix and the feature importance visualization (Fig. 2), we have determined that building temperature is the primary factor impacting energy consumption. Factors like HVAC usage and occupancy have a low to moderate influence, while other features show minimal impact. Our intention is to incorporate these key factors into our model, ensuring it effectively learns and predicts outputs.

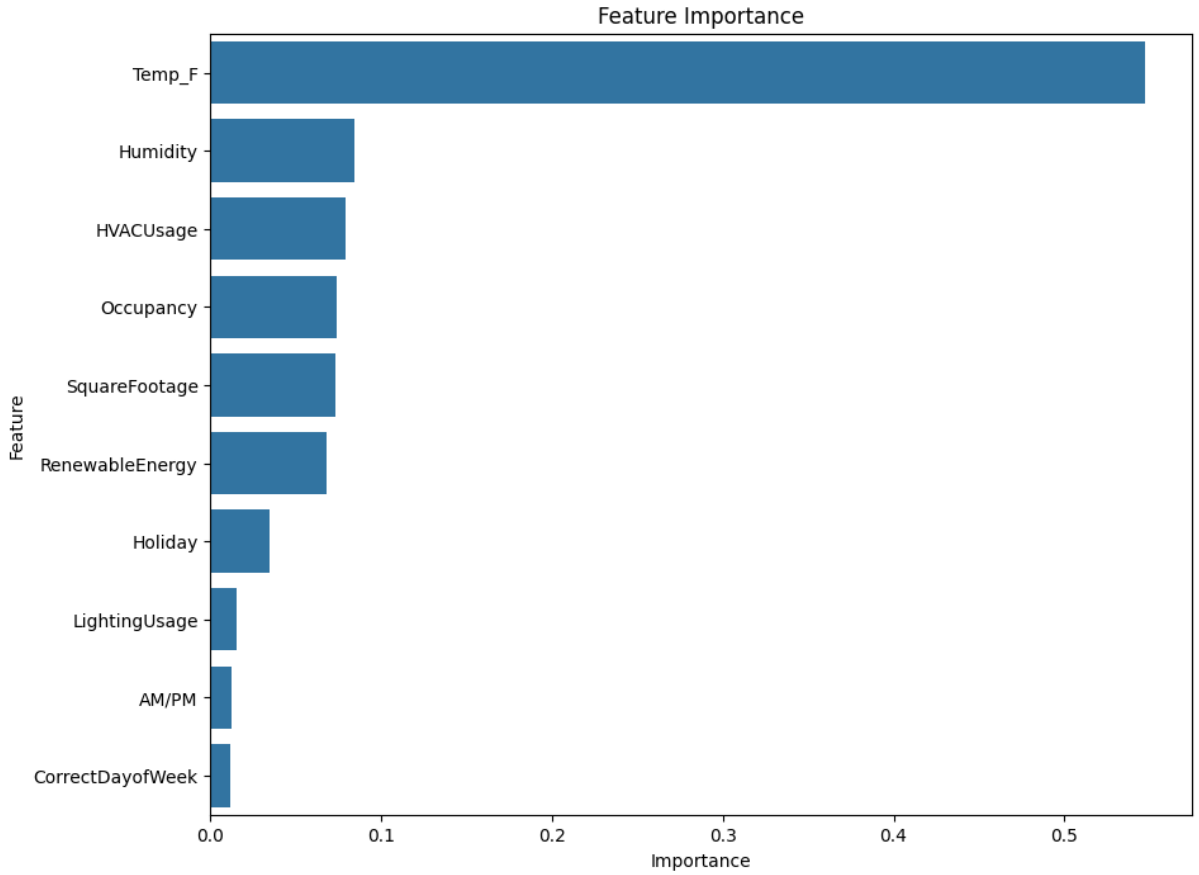

Fig. 2: Feature Importance Bar Graph

For our modeling process, we employed random forest regressor and linear regression to predict future energy consumption accurately using historical data. We took an iterative approach, experimenting with different features, training/testing ratios (percent split), and hyperparameters across various models. Our primary metric for evaluating model performance and accuracy was the R2 value. Contrary to our initial hypothesis, we concluded that the linear regression model performed best, achieving an R2 value of 0.64. The features used for the best linear regression model were the Temp_F, HVACUsage, Occupancy, and LightingUsage. This was in comparison to the Random Forest model's best R2 score of 0.60, which was returned using all of the features (Temp_F, Humidity, HVACUsage, Occupancy, SquareFootage, RenewableEnergy, Holiday, LightingUsage, AM/PM, CorrectDayofWeek). Detailed iterations and corresponding R2 values are documented in the Appendix.

There were a few reasons that we determined could be why the Linear Regression model performed better than the Random Forest. This could be explained by the fact that linear regression performs well on linear data and has more of an issue handling outliers in data. As demonstrated by the boxplots of our data, there was only one outlier, so outliers did not play a factor in how the model performed. Random Forest also does not perform as well if it is not tuned properly, which can lead to overfitting or underfitting of the data. It is possible there are other variations of hyperparameters that could be used in the tuning process in order to improve the Random Forest performance.

Below are data visualizations portraying our model's performance:

- **Actual vs. Predicted Energy Consumption Plot**: The dashed diagonal line represents a perfect prediction, which would mean that if the model predicted energy consumption perfectly, all of the scatter plot points would fall on that line. The scatter plot points represent each prediction the model made. Overall, our predictions align closely with the diagonal line, indicating reasonable accuracy. However, some points deviate from the line, indicating less accuracy prediction in certain areas (Fig. 3).
- **Predicted vs. Residual Energy Consumption Plot**: The dashed horizontal line represents a perfect scenario where the residuals are precisely zero. The scatter points represent individual predictions made by the model and their respective differences from the residuals. The points scattered above and below the horizontal line indicate instances where the model overestimated or underestimated the energy consumption. Overall, the even distribution of points above and below the horizontal line suggests that our model is unbiased. It has an equal chance of overestimating and underestimating energy consumption, which enhances model reliability (Fig. 4).

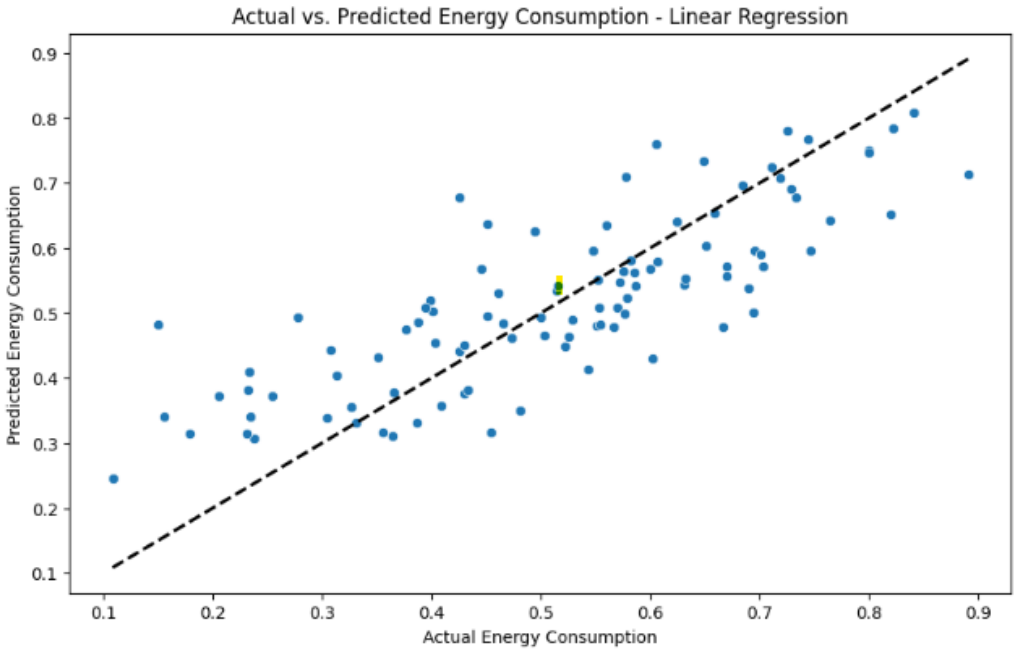

Fig. 3: Actual vs. Predicted Plot

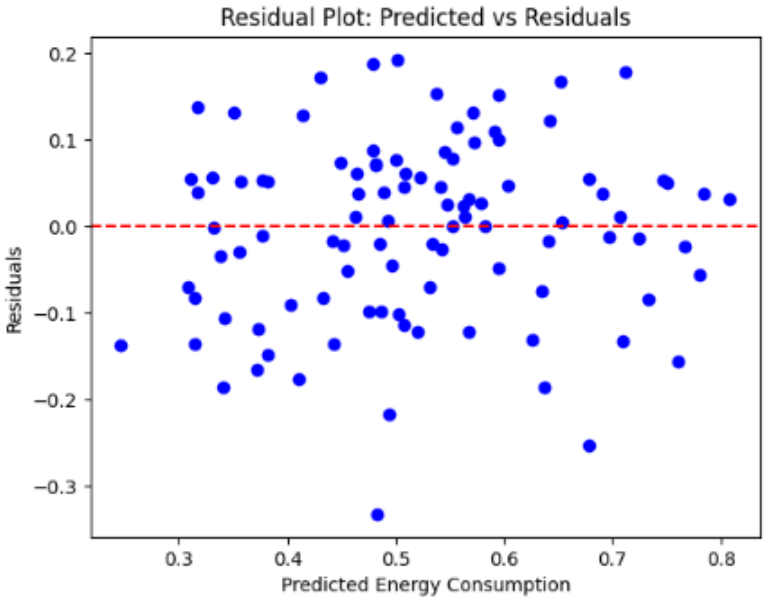

Fig. 4: Residual Plot

RECOMMENDATIONS

Based on our findings that temperature, HVAC usage, and occupancy are the top factors impacting energy consumption per our dataset, one recommendation to optimize energy usage is to install a

programmable or smart thermostat or HVAC unit that adjusts the temperature based on weather forecasts and occupancy patterns to optimize heating and cooling for actual needs and reduce energy waste in unoccupied areas.

To build out this smart device, our energy consumption model can be implemented as the predictive engine that informs the thermostat or HVAC unit when and how to adjust settings for optimal energy usage. Using historical energy consumption data in real-time, our model will enable the device to make effective energy management decisions, minimizing energy consumption and costs.

It is important to note that our current model is not yet ready for deployment due to its low R2 value. While this level of accuracy is not ideal, we have identified two main factors contributing to this:

- **Nature of Dataset**: It was artificially generated, which may not accurately represent the true relationships and correlations, preventing the model from learning and predicting output effectively
- **Model Selection**: The two selected models may not effectively capture interactions between features.

Moving forward, we plan to conduct further analysis to identify additional relevant features and perform more feature engineering. Then, we'd like to experiment with different model types and hyperparameters to find the best combinations that improve model performance.

CONCLUSION

Tying everything together, our project's main objective was to understand and predict energy consumption using key factors identified through data analysis. These key factors are temperature, HVAC usage, and occupancy. While our linear regression model showed potential by achieving an R2 value of 0.64, we ran into challenges training our models with a Kaggle synthetic dataset. Our recommendation focuses on using smart thermostats and HVAC systems guided by our predictive model to optimize energy use. For the next steps, we'd like to explore additional feature engineering techniques, more complex models, and real-world data to enhance our model's accuracy, driving toward more efficient energy management.

REFERENCES

Chen, Y., Li, Y., Jiang, H., & Huang, Z. (2023). Research on household energy demand patterns, data acquisition, and influencing factors: A review. *Sustainable Cities and Society*, *99*, 104916. https://doi.org/10.1016/j.scs.2023.104916

Debebe, B., Senbeta, F., Guta, D. D., Teferi, E., & Teketay, D. (2023). Determinants of household energy choice for domestic chores: Evidence from the Semien Mountains National Park and Adjacent Districts, Northwest Ethiopia. *Cleaner Energy Systems*, *4*, 100063. https://doi.org/10.1016/j.cles.2023.100063

Guo, Z., Zhou, K., Zhang, C., Lu, X., Chen, W., & Yang, S. (2018). Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies. *Renewable & Sustainable Energy Reviews*, *81*, 399–412. https://doi.org/10.1016/j.rser.2017.07.046

Piao, X., & Managi, S. (2023). Household energy-saving behavior, its consumption, and life satisfaction in 37 countries. *Scientific Reports (Nature Publishing Group)*, *13*(1). https://doi.org/10.1038/s41598-023-28368-8

Ramos, P. V. B., Villela, S. M., Silva, W. N., & Dias, B. H. (2023). Residential energy consumption forecasting using deep learning models. *Applied Energy*, *350*, 121705. https://doi.org/10.1016/j.apenergy.2023.121705

Sarswatula, S. A., Pugh, T., & Prabhu, V. V. (2022). Modeling energy consumption using machine learning. *Frontiers in Manufacturing Technology (Lausanne)*, *2*. https://doi.org/10.3389/fmtec.2022.855208

Tran, L. N., Cai, G., & Gao, W. (2023). Determinants and approaches of household energy consumption: A review. *Energy Reports*, *10*, 1833–1850. https://doi.org/10.1016/j.egyr.2023.08.026

Wang, T., Zhao, Q., Gao, W., & He, X. (2024). Research on energy consumption in household sector: a comprehensive review based on bibliometric analysis. *Frontiers in Energy Research*, *11*. https://doi.org/10.3389/fenrg.2023.1209290

Zaharia, A., Diaconeasa, M. C., Brad, L., Lădaru, G., & Ioanăş, C. (2019). Factors influencing energy consumption in the context of sustainable development. *Sustainability (Basel)*, *11*(15), 4147. https://doi.org/10.3390/su11154147

Zheng, J., Dang, Y., & Ullah, A. (2024). Household energy consumption, energy efficiency, and household income–Evidence from China. *Applied Energy*, *353*, 122074. https://doi.org/10.1016/j.apenergy.2023.122074

APPENDIX

Detailed model training iterations and corresponding R2 values:

Random Forest Model

1. Ran initial model with all features with a training/test ratio of 80/20 (**R2: 0.5417**)
2. Ran model again, adjusted training/test ratio of 90/10 (**R2: 0.5799**)
3. Hypertuned the model and ran again with updated parameters (**R2: 0.6023**)
4. Ran feature importance and proceeded to remove the 2 lowest-scoring features (AM/PM and Holiday). Re-ran the model with the base parameters for the Random Forest (**R2: 0.5974**)
5. Hypertuned the model again with the updated features. Ran the model with updated parameters (**R2: 0.5878**)
6. Ran a third iteration of the model with the top 3 features with base parameters (**R2: 0.5370**)
7. Hypertuned the model a third time with the top 3 features. Ran the model with updated parameters (**R2: 0.5612**)

Linear Regression Model

1. Ran initial model with a training/test ratio of 80/20 with all features (**R2: 0.5987**)
2. Ran model again, adjusting the training/test ratio of 90/10 (**R2: 0.6224**)
3. Ran with the bottom 2 features removed (**R2: 0.6261**)
4. Ran with top 3 features only (**R2: 0.5693**)
5. Ran with a combination of 4 features that produced the highest R2 score - Temp_F, HVACUsage, Occupancy, and LightingUsage (**R2: 0.6480**)