PREDICTING OPERATIONAL ENERGY CONSUMPTION IN LA BUILDINGS

CS109A - FALL 2023

Helen Huang	hah079@g.harvard.edu
Yiwei Lyu	yiweilyu@gsd.harvard.edu
Dominika Randle	drandle@hbs.edu
Sang Won Kang	sangwonkang@gsd.harvard.edu

INDEX

MILESTONE 3

MILESTONE 4

Summary of the Data

The dataset was generated using a building energy simulation tool called 'Honeybee', which is a plugin tool that works via a CAD software called Rhino. The data targets operational energy simulations for buildings located in Los Angeles, California, providing a clear overview of pure operational performance. The pseudo designs are generated using a Monte Carlo approach with random building attributes.

Some of the key attributes of the data that we filtered are:

PREDICTORS

Name	Metric	Data Type
Orientation	Degrees	Quantitative
nonMassWallR	m^2-K/W	Quantitative
MassWallR	m^2-K/W	Quantitative
RoofR	m^2-K/W	Quantitative
ExteriorFloorR	m^2-K/W	Quantitative
WWRNorth	R-value	Quantitative
WWRWest	R-value	Quantitative
WWREast	R-value	Quantitative
WWRSouth	R-value	Quantitative
SHGC	ratio(percentage)	Quantitative
WindowR	R-value	Quantitative
numFloor	count(integer)	Quantitative
AspectRatio	ratio(percentage)	Quantitative
VolumetoFacadeRatio	ratio(percentage)	Quantitative
Equipment (0-5)	One-hot-encoded	Categorical
Program(0-9)	One-hot-encoded	Categorical
WallType(0-3)	One-hot-encoded	Categorical

RESPONSE

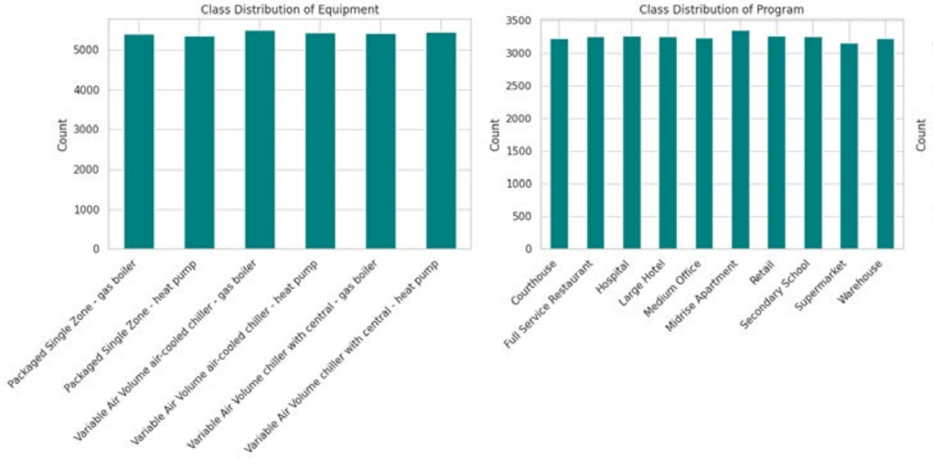
Name	Metric	Data Type
Operational Energy(OE)	kWh/m^2	Quantitative

Please refer to the attached ipynb notebook for more details regarding summary of features and purpose of the project (written in detail)

Deeper Understanding

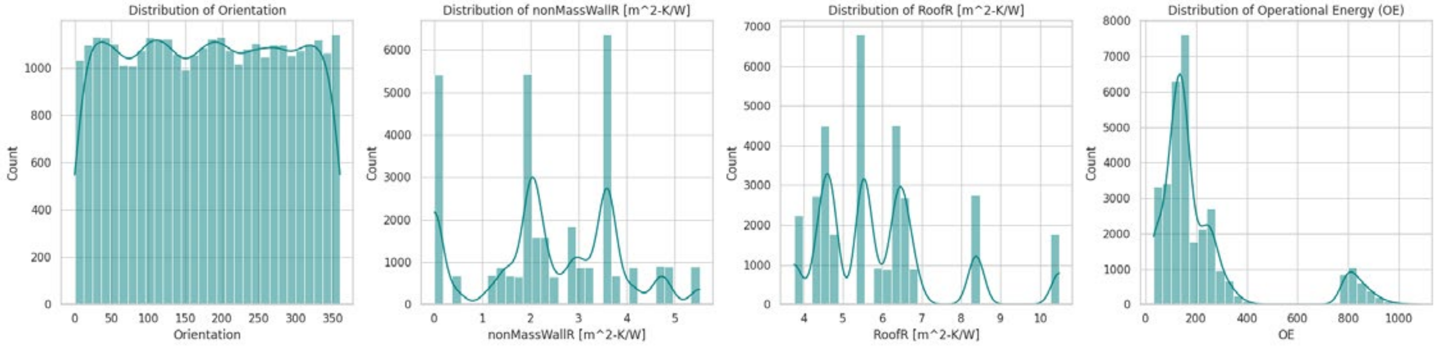
In order to understand our data, we sought for primarily patterns, trends, and relationships of variables, in addition to confirming that the data is good enough to work with and identify inconsistencies or any erroneous/troublesome values.

Data Imbalances



All programs, wall types and equipment are well balanced in terms of data counts, confirming that there are no imbalances.

Distributions

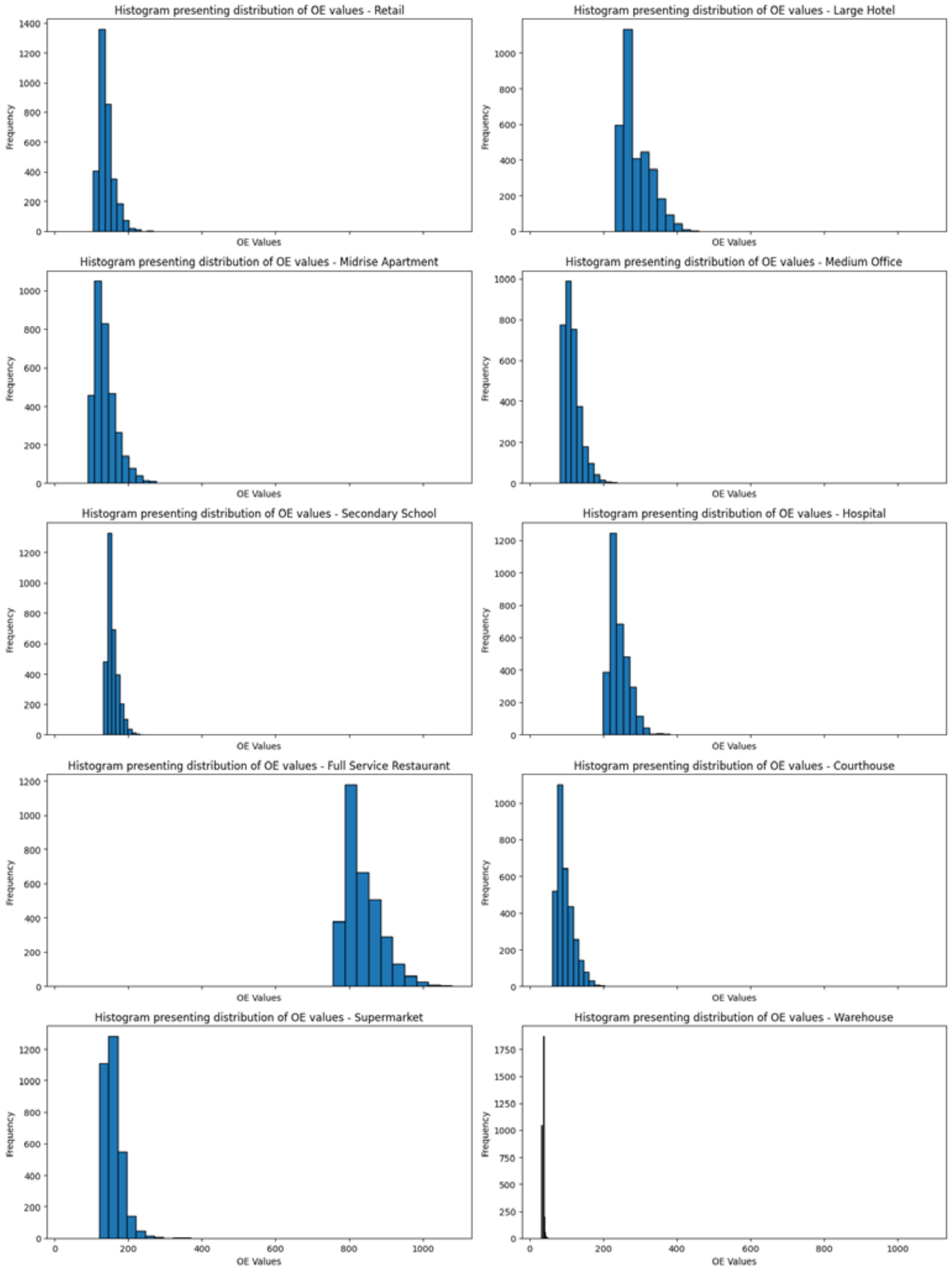


Select few predictors plotted to observe distributions.

The summary statistics for the predictors reveal a variety of scales and distributions. For instance, the orientation seems uniformly distributed between 0 and 360 degrees, non-mass wall R-values range between 0 and 5.5, and roof R-values range between 3.76 and 10.49. The response variable, operational energy (OE), has a mean of approximately 221 with a wide standard deviation of about 217, indicating significant variation in energy consumption across the dataset.

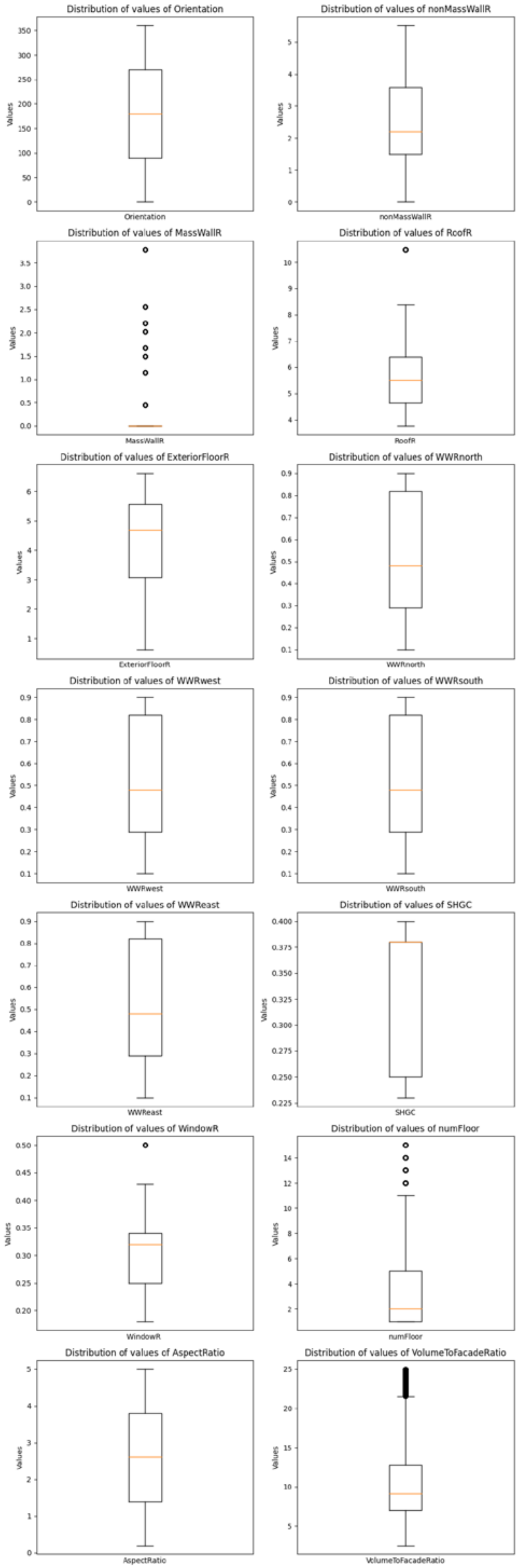
The histograms for selected predictors and the operational energy suggest the following:

The Orientation histogram is approximately uniform, suggesting no particular orientation bias in the dataset.
nonMassWallR is non-uniform, with varying frequency of values
RoofR shows a slight right skew.
The operational energy distribution is right-skewed, with a few values significantly higher than the rest, which could be potential outliers or simply represent high-energy-consuming buildings.

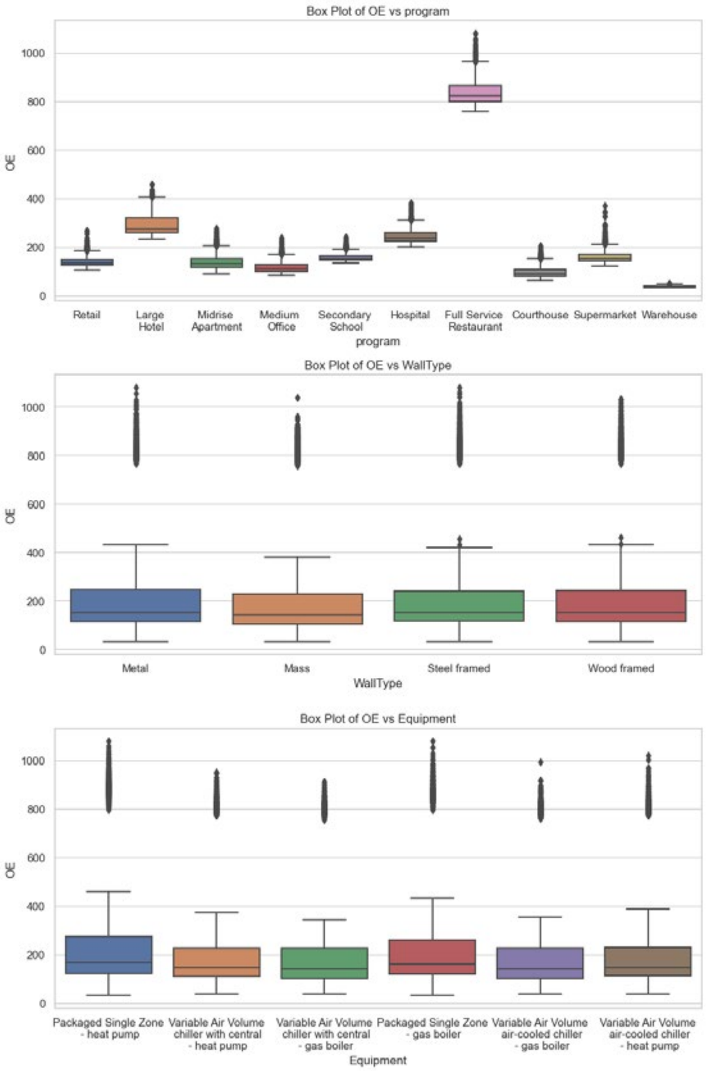


Distribution of OE per programs(labeled)

The distributions of Operational energy consumption per program is very helpful to determine how buildings of different types differ in energy usage and at what frequency. In the plots above, Warehouse types for instance use very low evergy as they often have low numbers of embedded mechanical systems and equipment loads. In contrast, a building like a restaurant tend to operate actively throughout 24 hrs, some operating beyond 10 - 12 hours per day, with a lot of loads from cooking equipment, lighting and ventilation systems. The plot displays high fidelity with common assumptions of building usages.



Distribution of values per Predictor

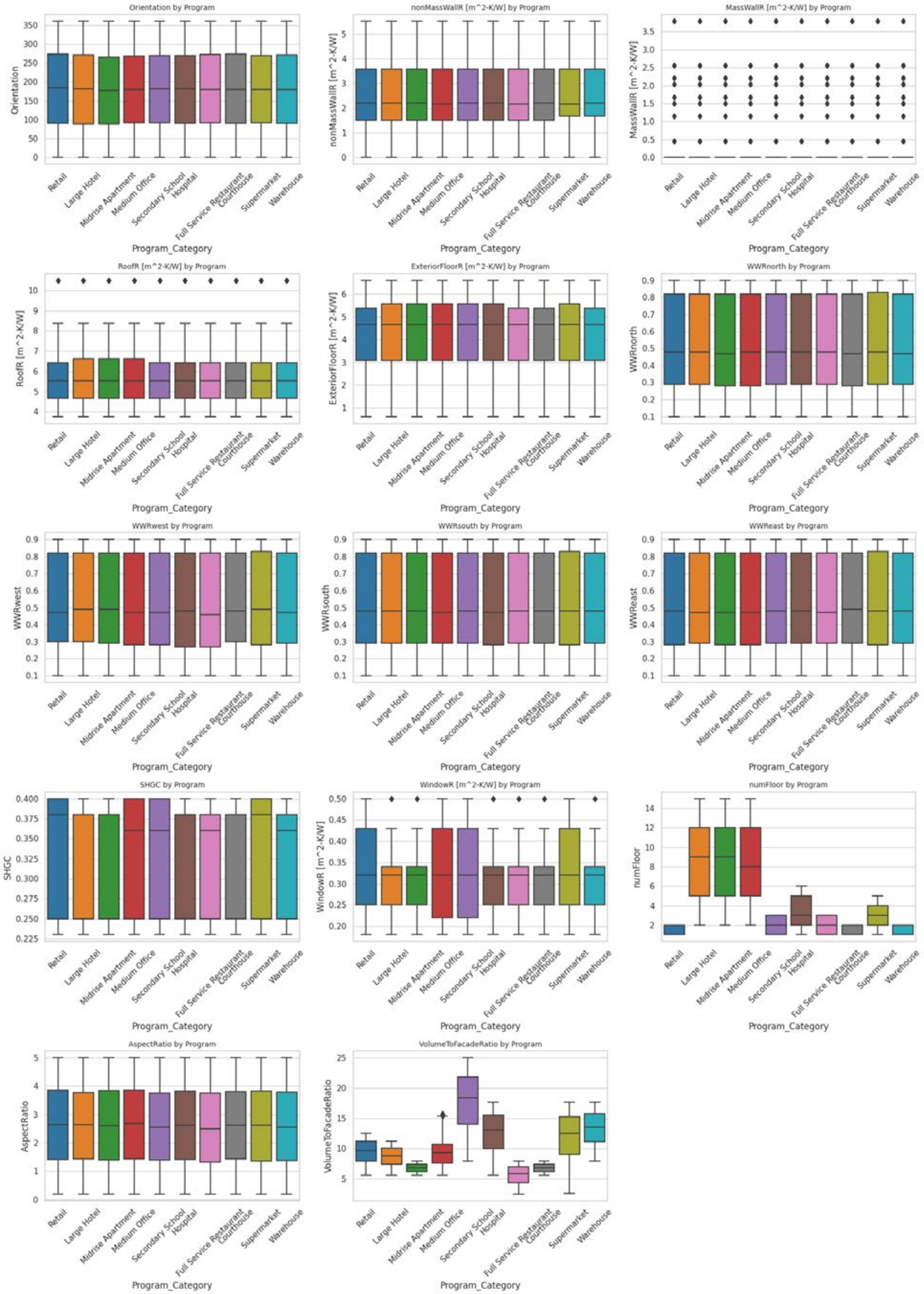


OE vs Categorical Variables

In this section of research, we are seeking for outliers and trying to understand where these may come from. In the left plots for distribution values per predictor we can easily find outliers of values per predictor, particularly Window R values, mass wall R values, number of floors and volume to facade ratio having the most noticeable outliers. To further understand why, we plotted the OE vs the categorical variables. The plot above contradicts our initial belief that perhaps the outliers came from full service restaurants which used the most OE, and instead we find outliers in all building types, which means the outliers are nearly well distributed across all types, being general outliers rather than a particular type's outliers.

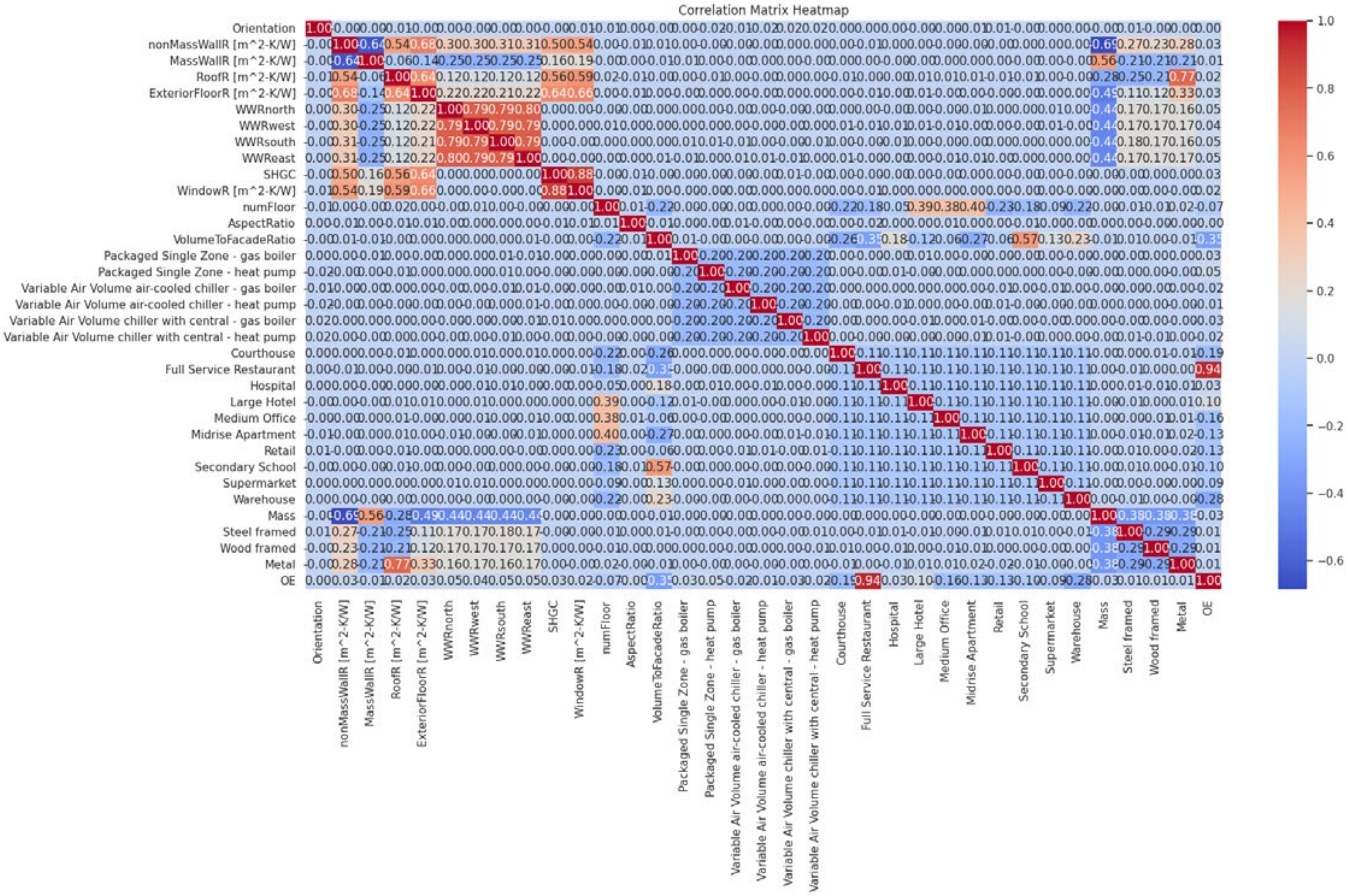
In the next page, we analyze this even further by plotting each individual predictor vs building programs, which further strengthen our findings so far.

Refer to plot in pg5



Predictors vs Programs

Trends and Relationships



A correlation matrix is made and visualized in a heat map using seaborn to show relationships and correlation coefficients between variables. In this process we are trying to identify which variables have strong relationship with each other.

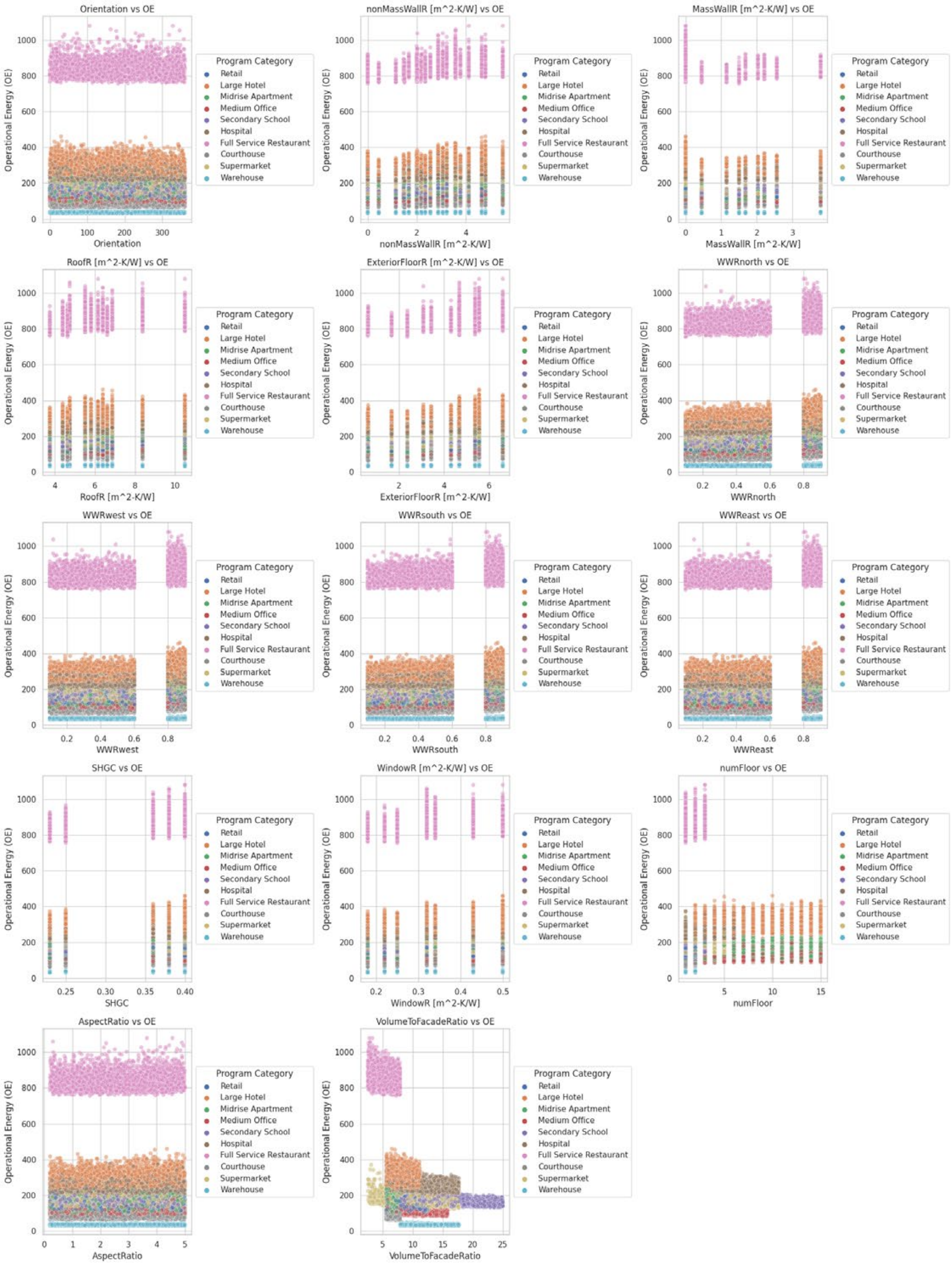
In the visualization, most interesting aspect is the VolumeToFacadeRatio having a negative correlation with Operational Energy (OE) of -0.35, suggesting that buildings with high volume to facade ratio tend to have lower operational energy consumptions, holding all variables constant. This is interesting in the context of building technology, as one may often assume that larger buildings would consume more energy, but as we are holding all other variables constant, we need to look into how the other variables influence the output.

As we move to analyzing relationships further by observing trends when plotting each predictor vs OE, we can see a few important aspects to remember and carefully consider: (plot in page 7)

1. Some predictors such as window to wall ratios per orientation have gaps between 0.6 - 0.8 consistently across all 4, this is because the simulations take the situation concerning the glass storefront as well.

2. All predictors, when observed plotted against the response variable, have non-linear relationships, this is crucial to observe in order to determine what model to choose to implement for predictions. For milestone 2, we roughly started with a Random Forest Regression model, which this proves why the model we chose initially could be potentially effective.

3. Across all plots we have a trend of restaurants having the highest energy consumption when considering a predictor holding all others constant.

Orientation vs OE
Operational Energy (OE)
0
200
400
600
800
1000
Orientation
0
100
200
300
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
nonMassWallR [m^2-K/W] vs OE
Operational Energy (OE)
nonMassWallR [m^2-K/W]
0
2
4
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
MassWallR [m^2-K/W] vs OE
Operational Energy (OE)
MassWallR [m^2-K/W]
0
1
2
3
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
RoofR [m^2-K/W] vs OE
Operational Energy (OE)
RoofR [m^2-K/W]
4
6
8
10
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
ExteriorFloorR [m^2-K/W] vs OE
Operational Energy (OE)
ExteriorFloorR [m^2-K/W]
2
4
6
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
WWRnorth vs OE
Operational Energy (OE)
WWRnorth
0.2
0.4
0.6
0.8
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
WWRwest vs OE
Operational Energy (OE)
WWRwest
0.2
0.4
0.6
0.8
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
WWRsouth vs OE
Operational Energy (OE)
WWRsouth
0.2
0.4
0.6
0.8
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
WWReast vs OE
Operational Energy (OE)
WWReast
0.2
0.4
0.6
0.8
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
SHGC vs OE
Operational Energy (OE)
SHGC
0.25
0.30
0.35
0.40
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
WindowR [m^2-K/W] vs OE
Operational Energy (OE)
WindowR [m^2-K/W]
0.2
0.3
0.4
0.5
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
numFloor vs OE
Operational Energy (OE)
numFloor
5
10
15
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
AspectRatio vs OE
Operational Energy (OE)
AspectRatio
0
1
2
3
4
5
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse
VolumeToFacadeRatio vs OE
Operational Energy (OE)
VolumeToFacadeRatio
5
10
15
20
25
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse

Project Questions & Plans

How can we make an effective predictive model based on the fact that the data has non-linear relationships, relatively high dimensionality, and correlations between predictors?

How do we address and handle gaps? are they critical in determining the accuracy and effectiveness of our model?

Is there a safe range of resulting Operational Energy consumption that we can use to roughly determine before proceeding to validation strategies that our model is yielding accurate predictions?

Our baseline model has been implemented already in Milestone 2 (Random Forest Regression).
We plan to test and compare other models such as:

- Decision Trees
- Polynomial regression
- Lasso|ridge regression
- Gradient Boosting (pending discussion)

The decision for such models comes from the data analysis. We have carefully observed the non-linear relationships and also a relatively high level of complex interactions between variables. Decision Trees and Random Forest regression are models that handle non-linearity effectively without explicitly defining it, while polynomial regression is effective in handling non-linearity between independent and dependent variables by adding polynomial terms to the model. A lasso and ridge regression will also be a good model to test, again because it is effective in handling non-linearity as an extension to polynomial regression.

***For more information on the work in progress please refer to the github link below, where all our notebooks and data can be found:

https://github.com/snwnkang/CS109A_Final/tree/main

Problem Statement

How can we predict the Operational Energy consumption in buildings when we are given certain variables that are not entirely obvious in assuming how they may influence it? We can roughly estimate the total energy consumption based on specific metrics such as HVAC, and other quantitative variables that directly state numbers that add to a total energy consumption, but when categorical variables such as program, equipment, or wall types are introduced, the relationships and predictions are more difficult to predict and less obvious in making assumptions. Throughout the progress of our project, we have obtained the data, explored relationships and patterns, and have manipulated the data in ways we can fit and develop predictive models.

In this milestone, we present different models and tests of each model one by one, enabling easy comparisons and eventually determine which model(s) perform best.

Note: We have, in addition, included a method that was not part of what we have learned, a neural network that has proved incredibly effective, we have used it to set an ideal goal and compare our models to get the best performing model possible.

Models
-Overview-

Linear Regression

Baseline model, assuming a linear relationship between predictors and response.

Polynomial Regression + Lasso Regularization

Efficient in modeling non-linear relationships (most real world data) and Lasso to aid in feature selection and regularization to work with high dimensionality

PCA Linear Regression

Another efficient method to reduce dimensionality, to capture most variance in the data and improve performance.

Decision Tree Regression

Effective for non linear relationships and interactions without the need for feature engineering. Additionally handles categorical variables very well.

Random Forest Regression

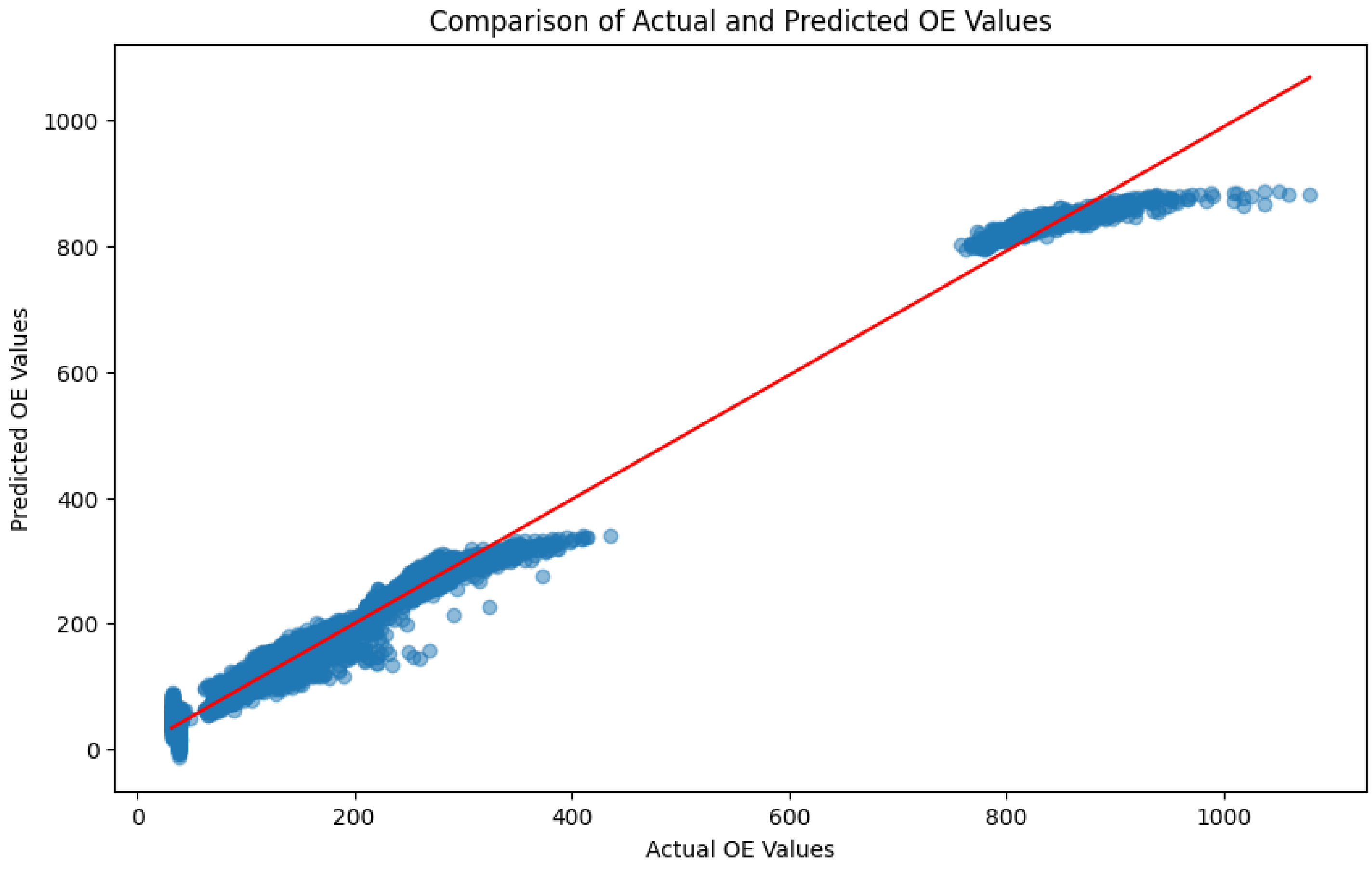
Ensemble of decision trees that improve predictive performance and robustness over a single tree by reducing overfitting

Extreme Gradient Boosting

Capable of building trees one at a time, where each new tree helps to correct errors made by the previous trees. Very effective in handling a combination of numerical and categorical variables.

*Neural Network

Highly flexible, capable of handling high complexity and non-linear relationships. (*ideal model)



All the models were split into 80% train and 20% test splits.

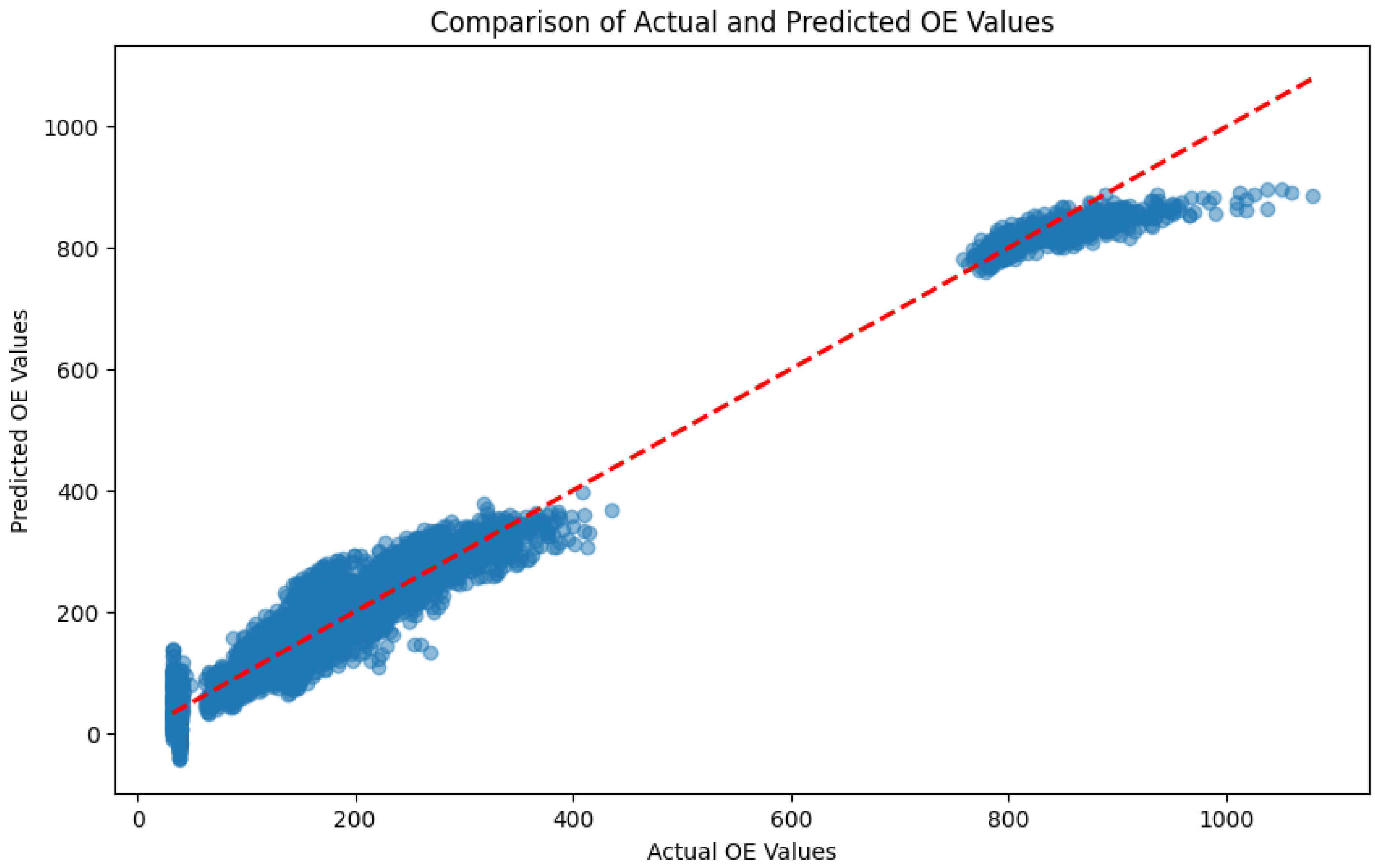
The linear model, as expected is not efficient in making accurate predictions, as the variables have non linear relationships and a linear model would not be able to accurately capture the predictions well. The plot below is self-explanatory (please refer to the ipynb notebook for details on the model). What is promising is that the plot below demonstrates that a linear model has some predictive capability, until the points begin diverging from the test line, particularly for high Operational Energy values, suggesting the model could be improved.

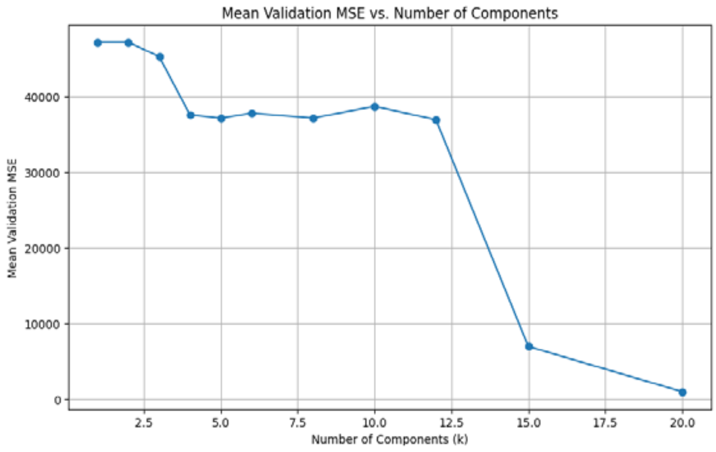
Mean Squared Error: 380.887786

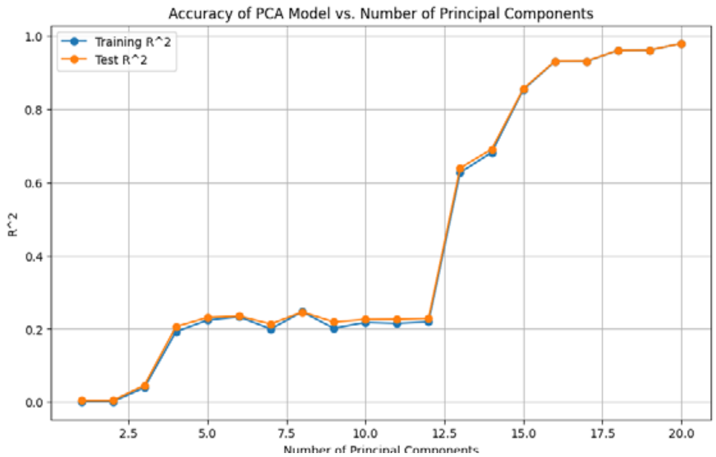
Root Mean Squared Error: 19.516347

R-Squared: 0.992138

Principal Component Analysis







From the plots above, it is evident that as the number of principal components used in the model increases, both train and test R squared values increase, suggesting that adding more principal components do in fact enhance the model's ability to explain variance in the data, up to around 16 components or somewhere close. Beyond this point, new information do not contribute in the predictive ability of the model.

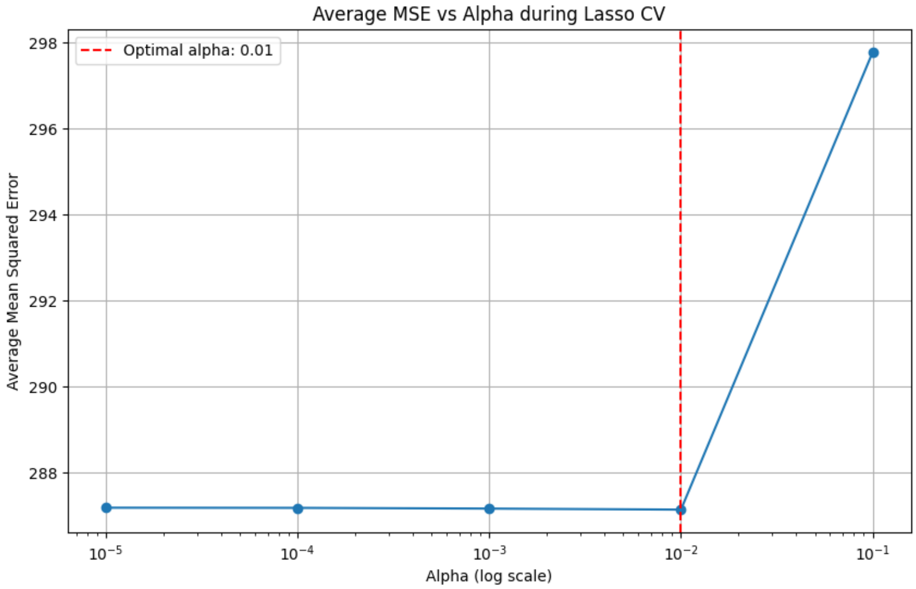
Yet, PCA is not performing comparatively better than the linear regression model, yet increasing the number of components reduces the MSE significantly, up to a point where it perhaps begins to overfit.

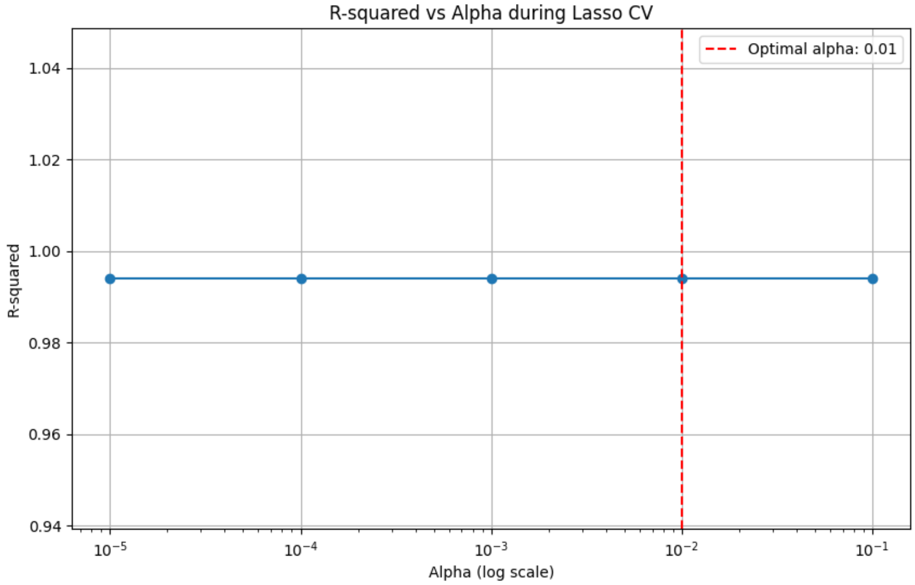
Mean Squared Error: 1009

Root Mean Squared Error: 31.777421

R-Squared: 0.979157

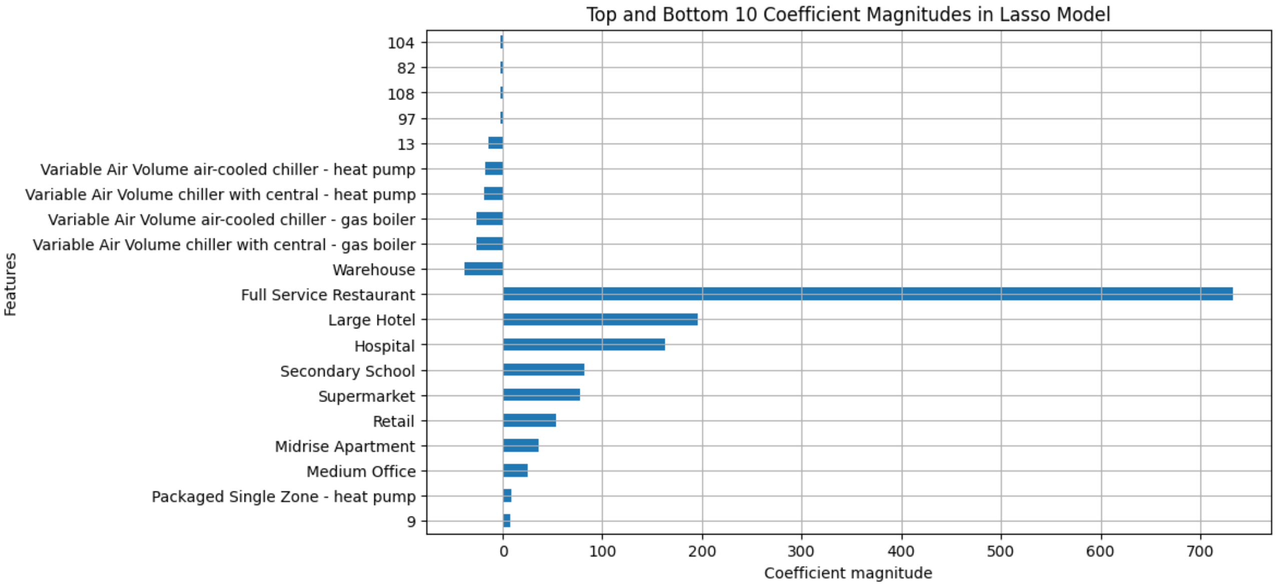
Polynomial Regression

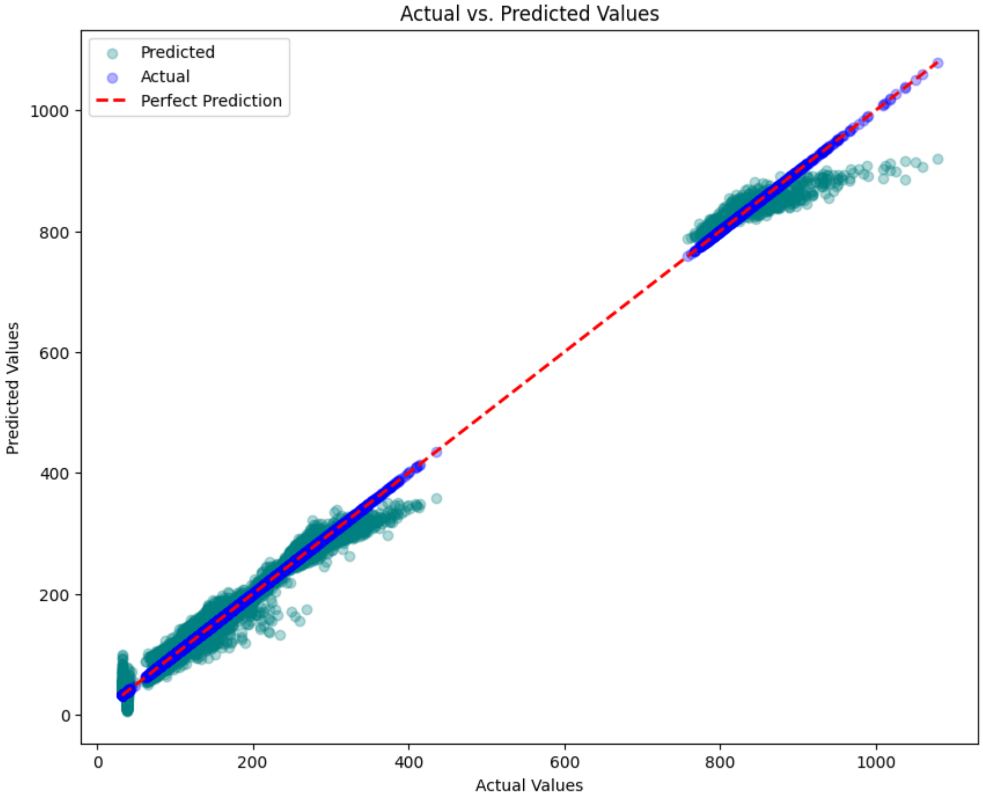




In polynomial regression, the first plot shows the relationship between different alpha values and the average MSE across cross validation folds. The red dashed line marks the optimal alpha value of 0.01 which is the regularization strength that resulted in the lowest MSE. Increasing the alpha beyond this point drastically increases the MSE, indicating too much regularization will weaken the model.

In the R squared vs Alpha plot, we can see that the R squared value remains fairly consistent across different alpha values, suggesting that the model's power is robust to regularization strength, up until the optimal alpha. If the plot was expanded to show more alpha values, we could assume that the R squared value would decrease perhaps further beyond what the current plot displays.





The magnitudes of the top and bottom 10 coefficients in the Lasso regularization shows that there are negative coefficients as well as positive coefficients. This visualization helps in understanding which features are most influential in the model, and the direction of the relationship shown is also important in knowing how features influence the model (negative or positive)

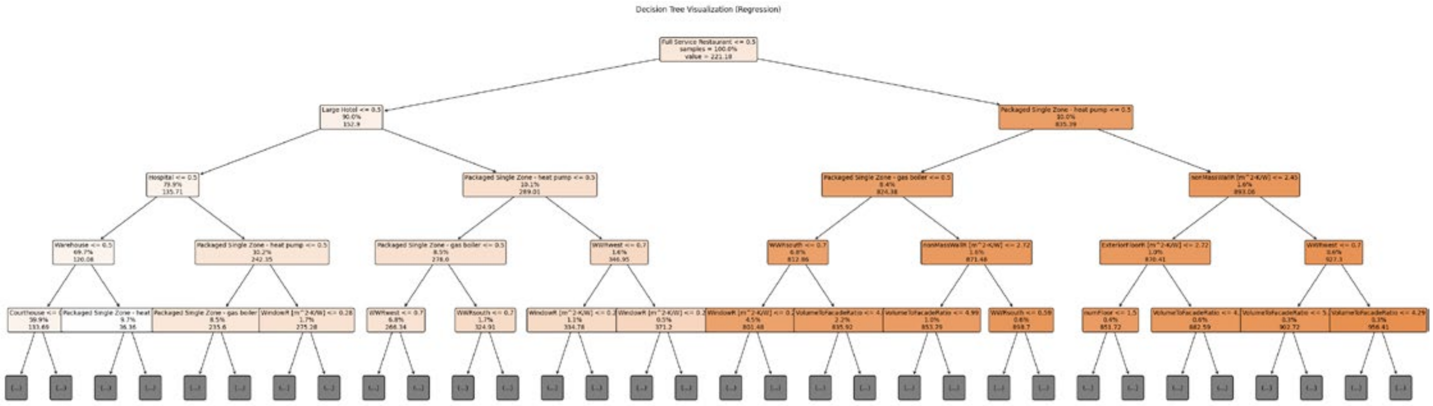
In comparison to the linear base model and PCA analysis, the polynomial regression model shows a slight improvement, with more alignment to the perfect prediction line. The performance of this model in comparison to the base model is also evident by looking at the MSE, RMSE and R2, which show that this model is predicting slightly more accurately. Yet this model shows that there could be a better model that could perform better.

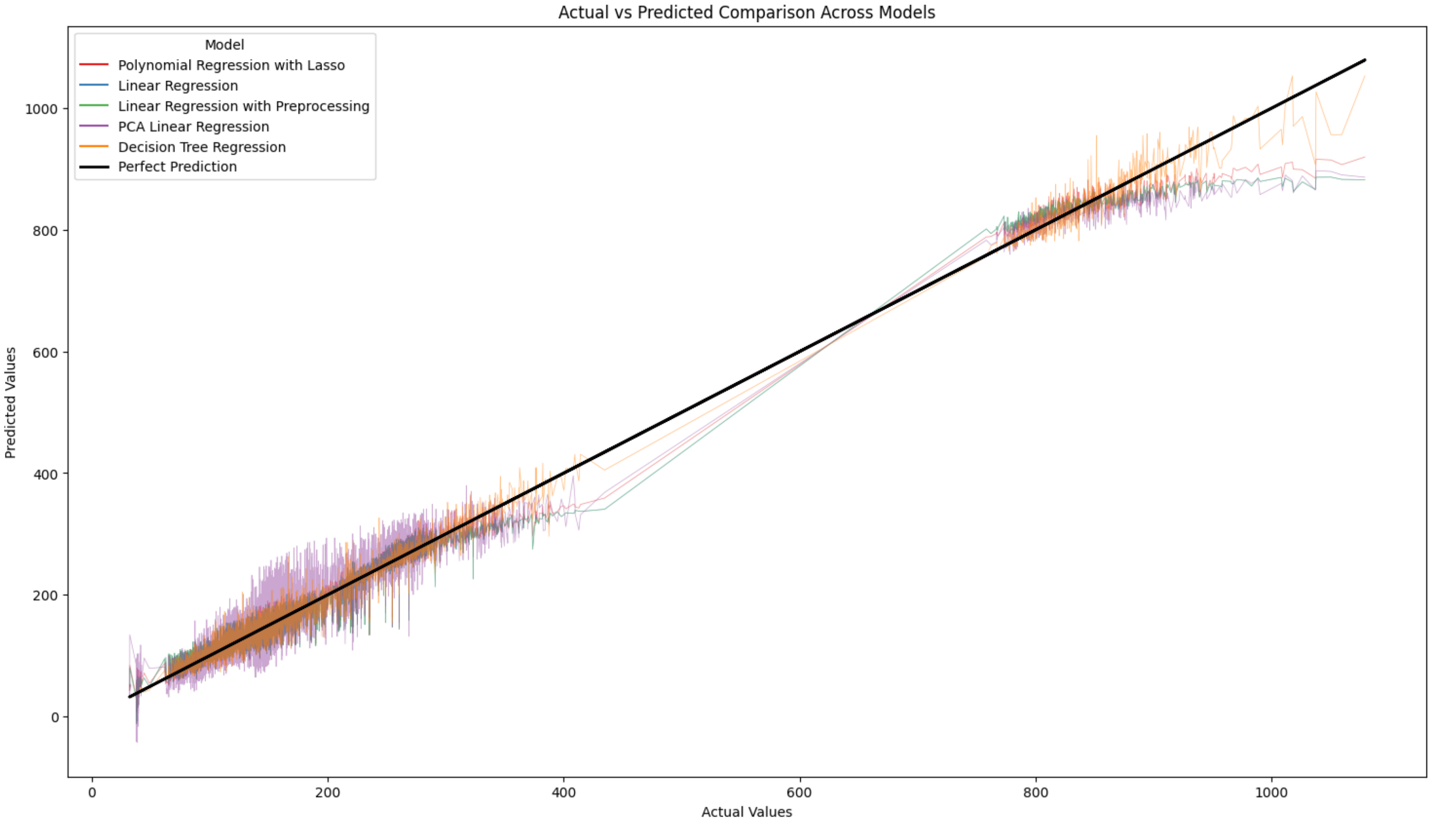
Mean Squared Error: 302.581275

Root Mean Squared Error: 17.394863

R-Squared: 0.993755

Decision Tree





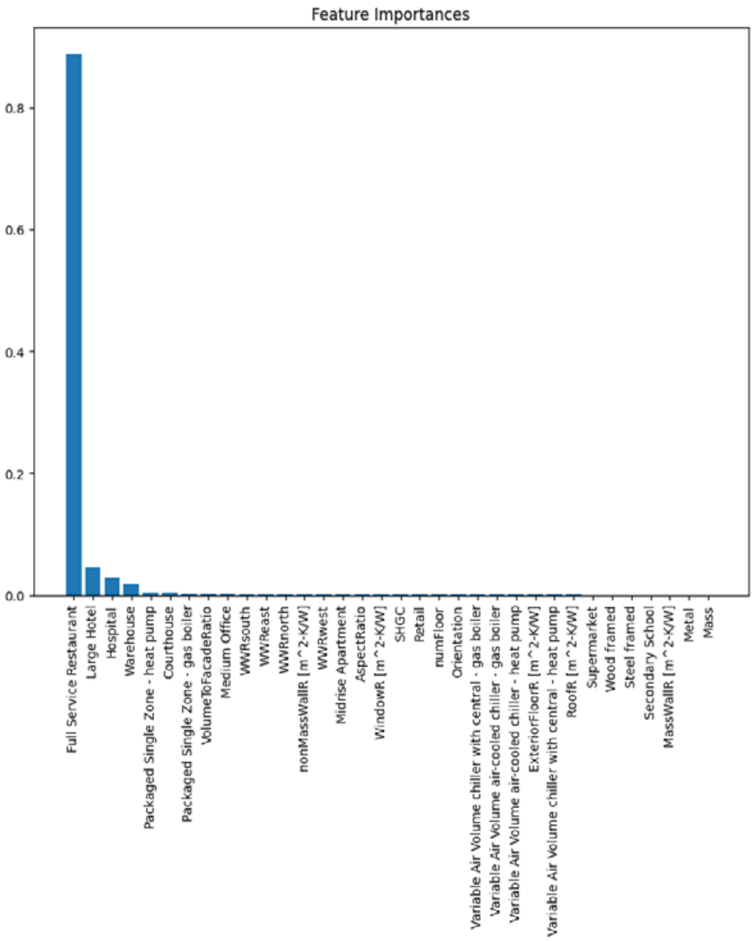
The decision tree model was successful in outperforming the previous models. The plot comparison of all models so far clear shows that the Decision Tree model is closer to the prediction line, which proves its effective predictive power.Yielding the lowest MS, RMSE and higher R Squared value, the model explains a lot of the variance in the response. The tree's structure suggests a relatively complex model that has learned disinctions in the data based on the values of various features. Unfortunately the interpretability of the tree is compromised by the complexity of the tree structure, which extends far beyond the visualization above. So we move on to a Random Forest Regression model which we hope could perform even better and could be more interpretable.

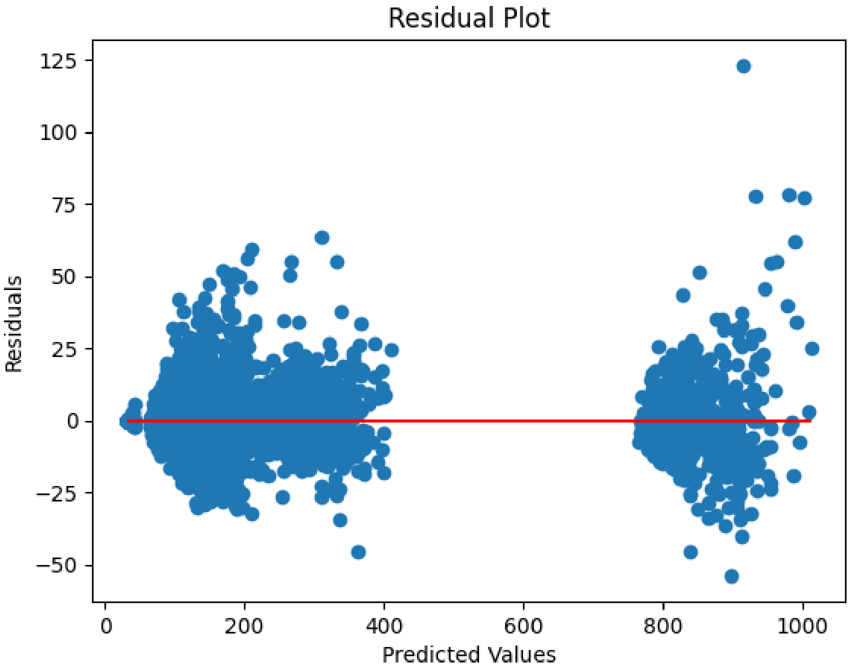
Mean Squared Error: 164.727118

Root Mean Squared Error: 12.834606

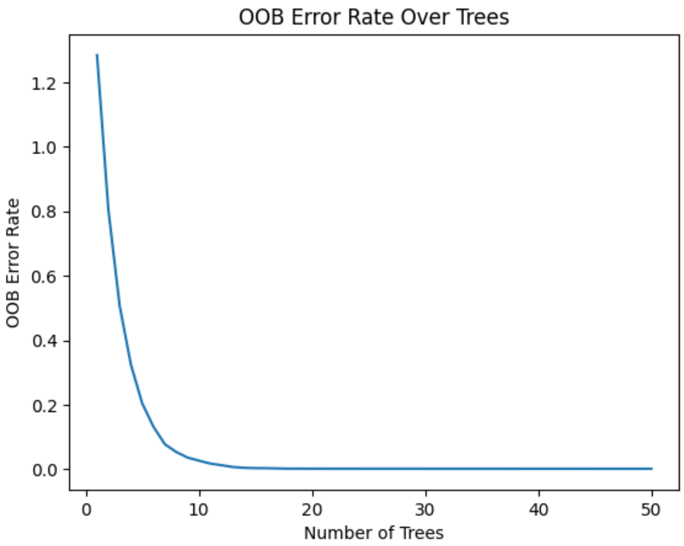
R-Squared: 0.996600

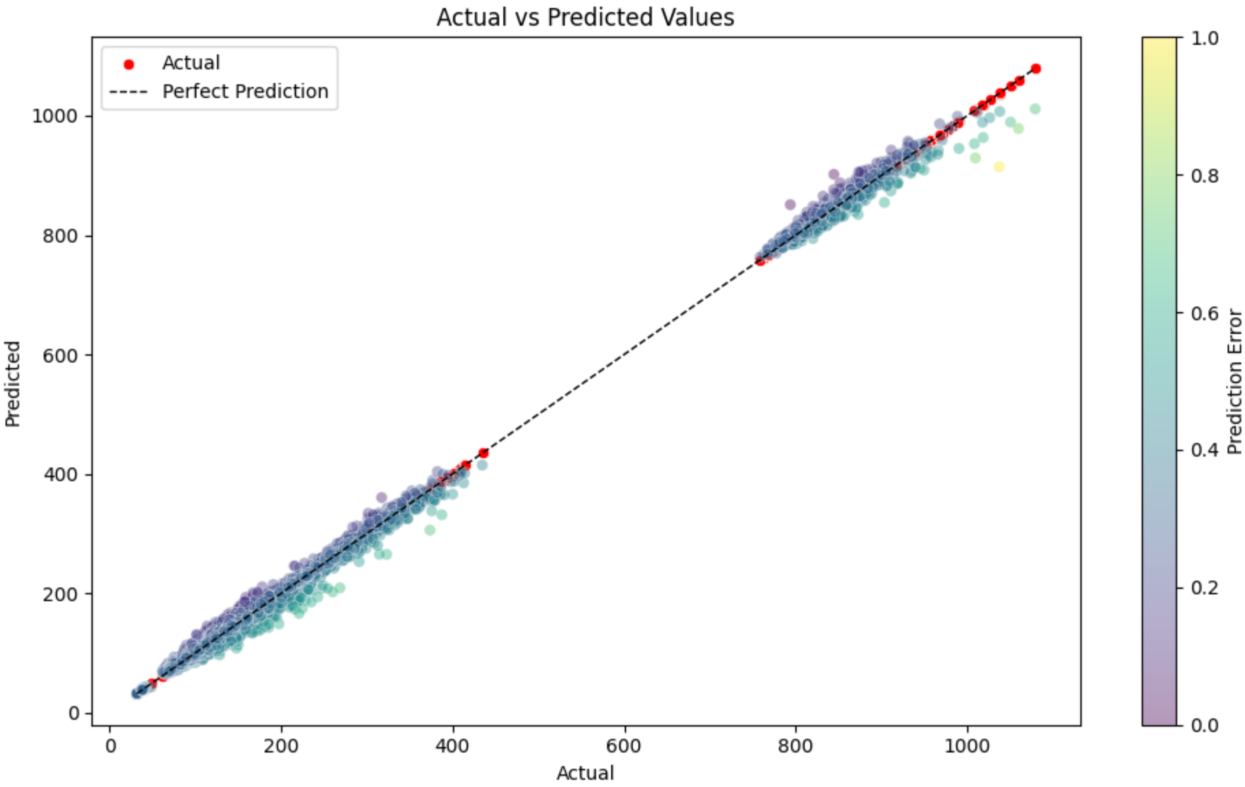
Random Forest Regression





For this model, feature importance analysis shows which features the model found most predictive. The dominance of some of the categorical variables suggest that these have strong relationships with the target variable and the model may rely heavily on these for predicting. The residual plot shows a relatively even spread of residuals (homoscedasticity), increasing in variance as the predicted values increase, which suggests that there may be heteroscedasticity where the model's performance is not consistent across all levels of the target variable.





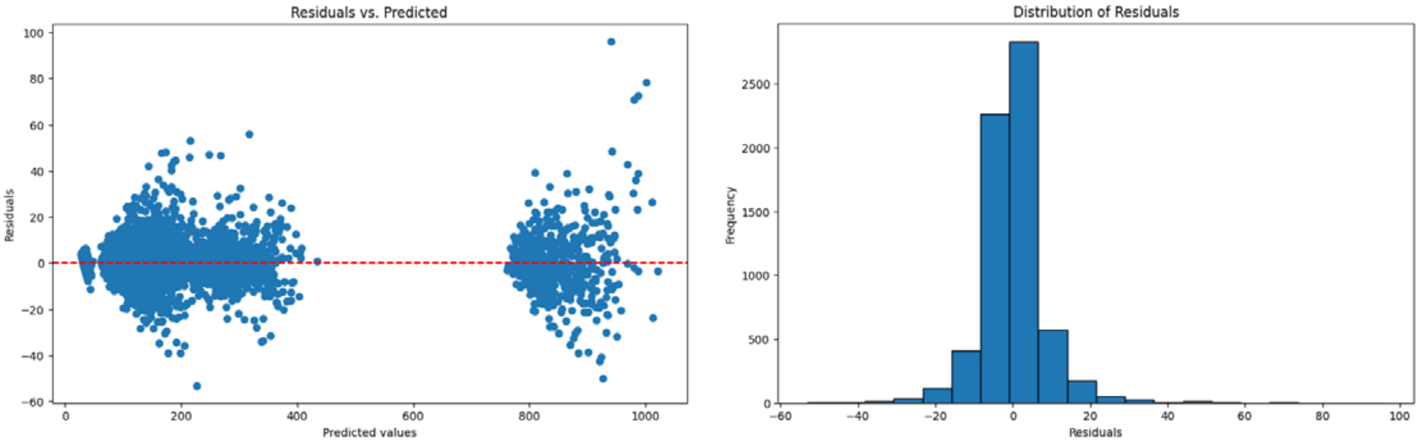
The Out of Bag error rate over trees decreases as more trees are added, up to a certain point, suggesting that beyond that optimal point in between 9-10 may not benefit the model. The out of bag score also shows that the model can generalize well to unseen data.

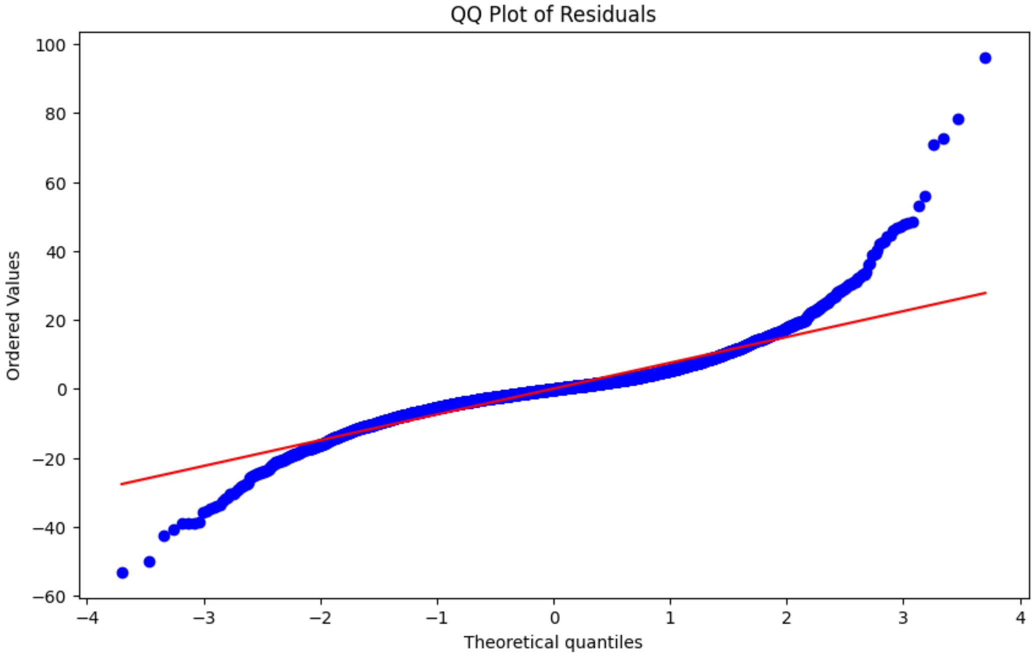
In conclusion, the random forest regression model has significantly outperformed all previous models, to the point of being all close to the prediction line, unlike the linear model or the polynomial model, and perhaps even better than the decision tree model (looking at the numbers below).

Please note that the random forest regression model was tuned (please refer to the notebook for details), improving the model only slightly but yet clear that it performs the best so far.

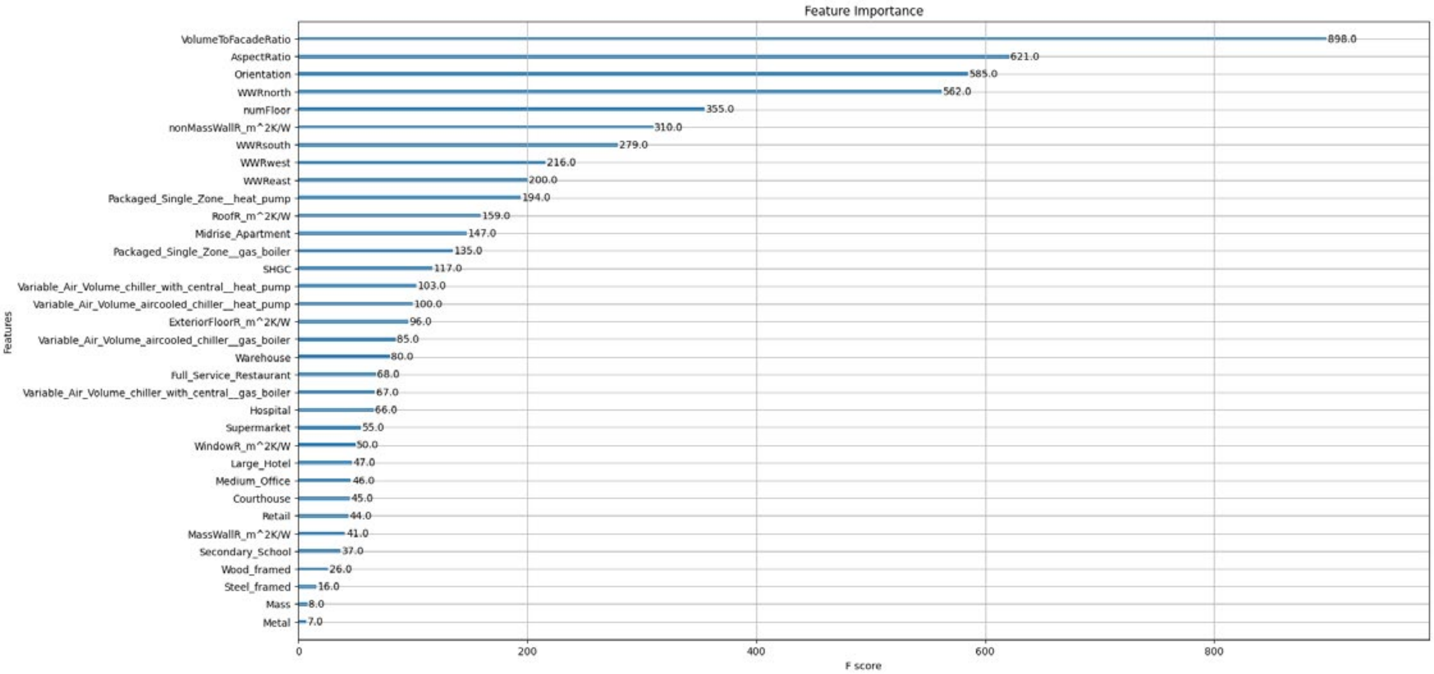
Pre Tuning	Post Tuning
Mean Squared Error: 80.288517	Mean Squared Error: 80.574747
Root Mean Squared Error: 8.960386	Root Mean Squared Error: 8.976344
R-Squared: 0.996600	R-Squared: 0.998337

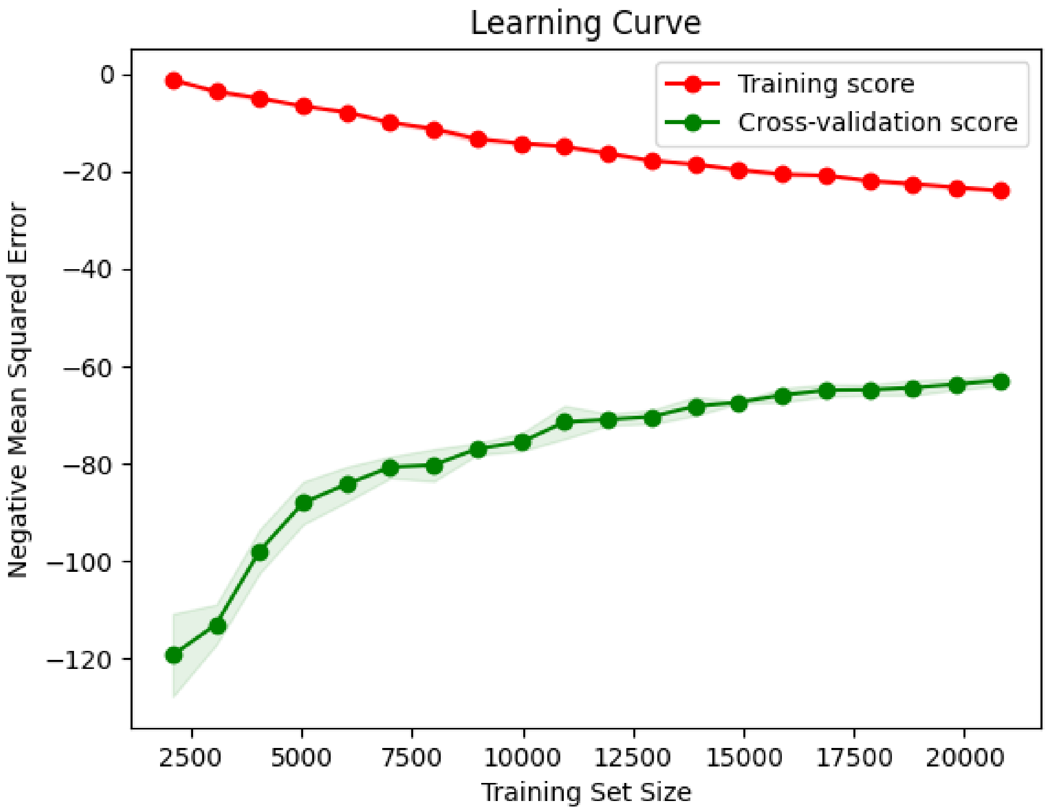
Extreme Gradient Boosting





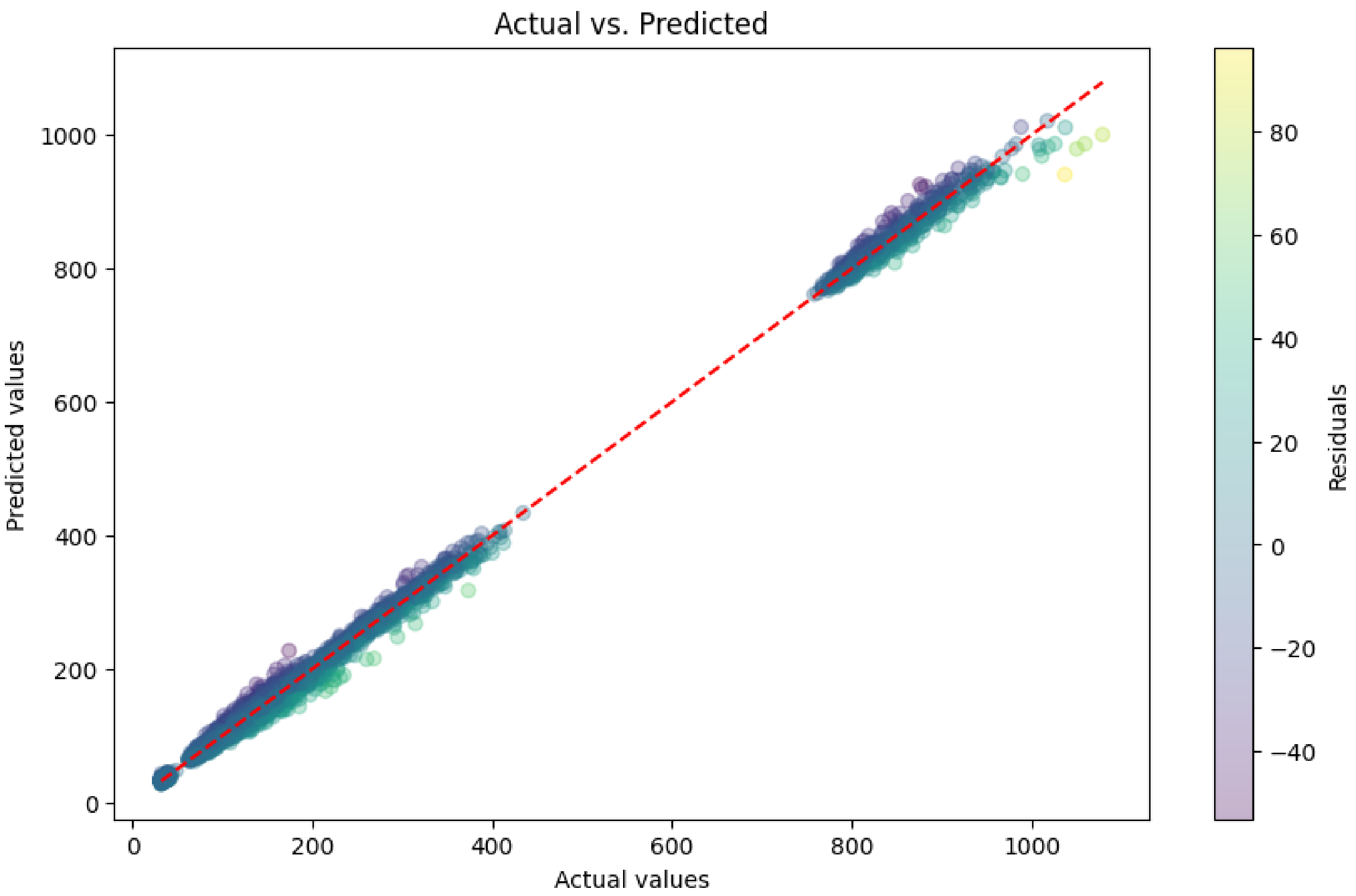
The residual plots remain the same as displayed in the previous models. Most residuals however are close to zero, which is ideal. However, the QQ plot shows that the residuals are not normally distributed, indicating potential issues with outlier predictions.

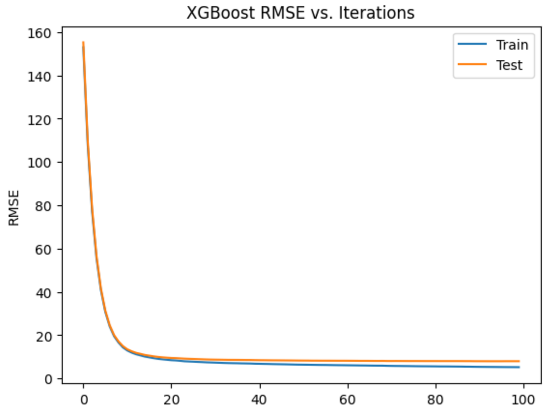


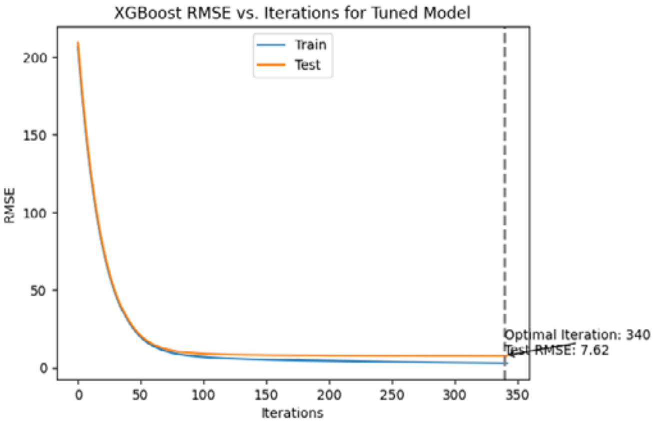


We plot another feature importance for this model, which now show different important variables, especially those that are numerical variables.

The learning curve is crucial in this model as it shows how the model's performance improves with the addition of more training data. The cross validation score is converging towards the training score, which suggest that adding more data could eventually bring these two lines closer together, helping the model generalize better.





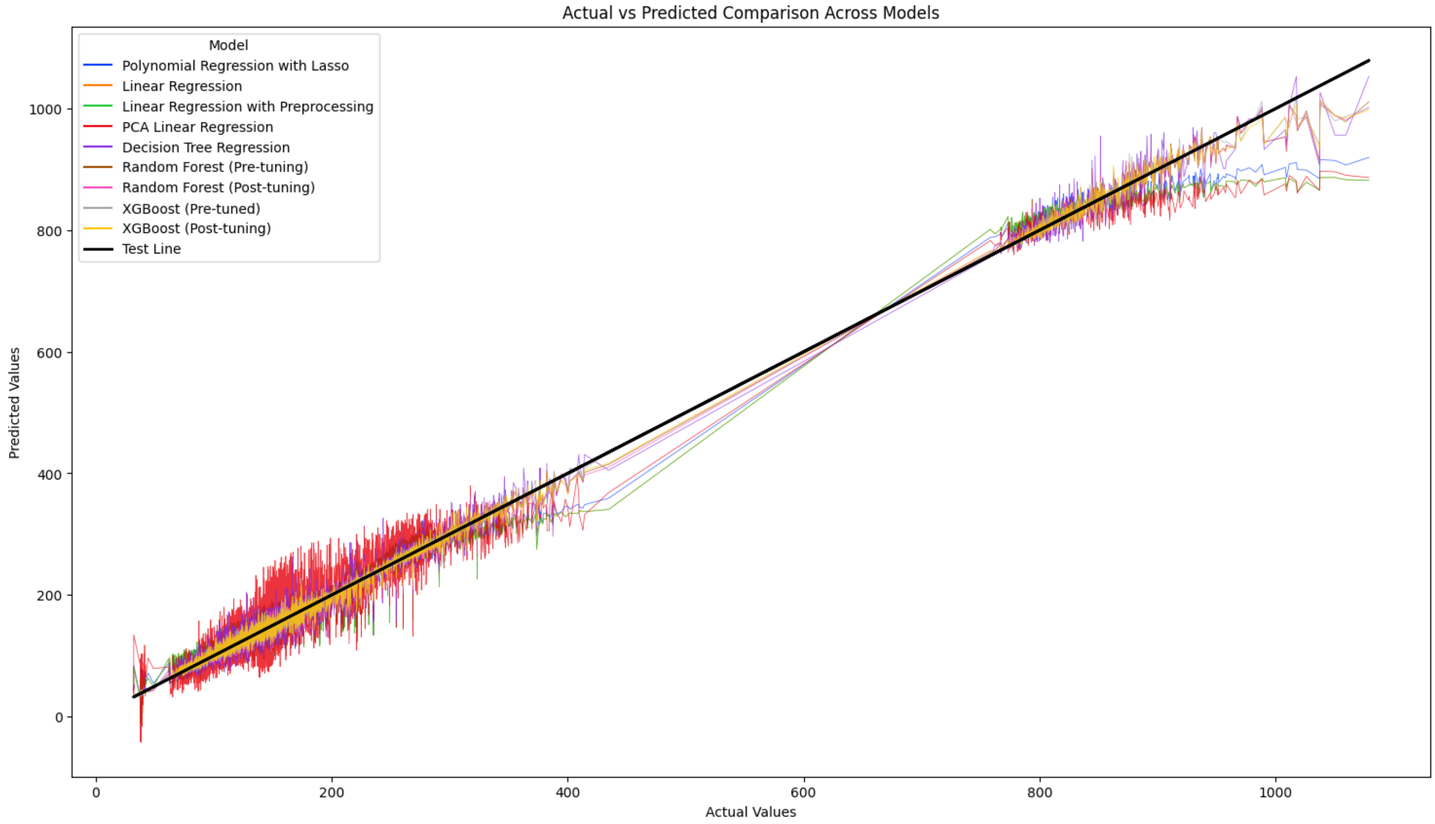


The model has outperformed all previous models (please refer to the plot and the numbers).

The model was fit, then tuned, yielding slightly better results. This is more evident in the RMSE vs iterations plot comparisons (pre tune and post tune), in which the RMSE starts much lower before tuning and decreases much faster, as for after tuning, the RMSE starts much higher but decreases more smoothly and continues improving across more iterations, to an optimal iteration at 340 with RMSE of 7.62. The extreme gradient boosting has proved to be significantly effective in predicting our target variable.

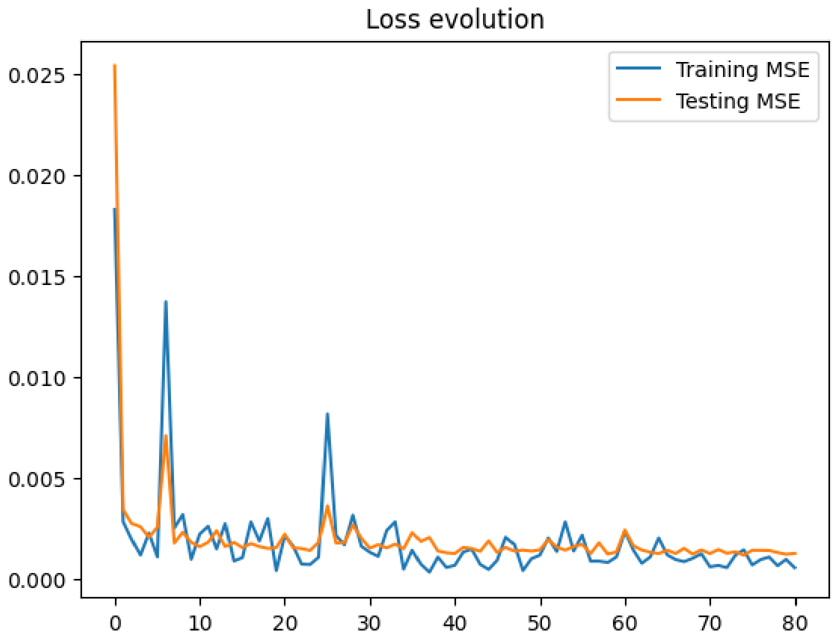
Pre Tuning	Post Tuning
Mean Squared Error: 63.260267	Mean Squared Error: 58.036042
Root Mean Squared Error: 7.953632	Root Mean Squared Error: 7.618139
R-Squared: 0.998694	R-Squared: 0.998802

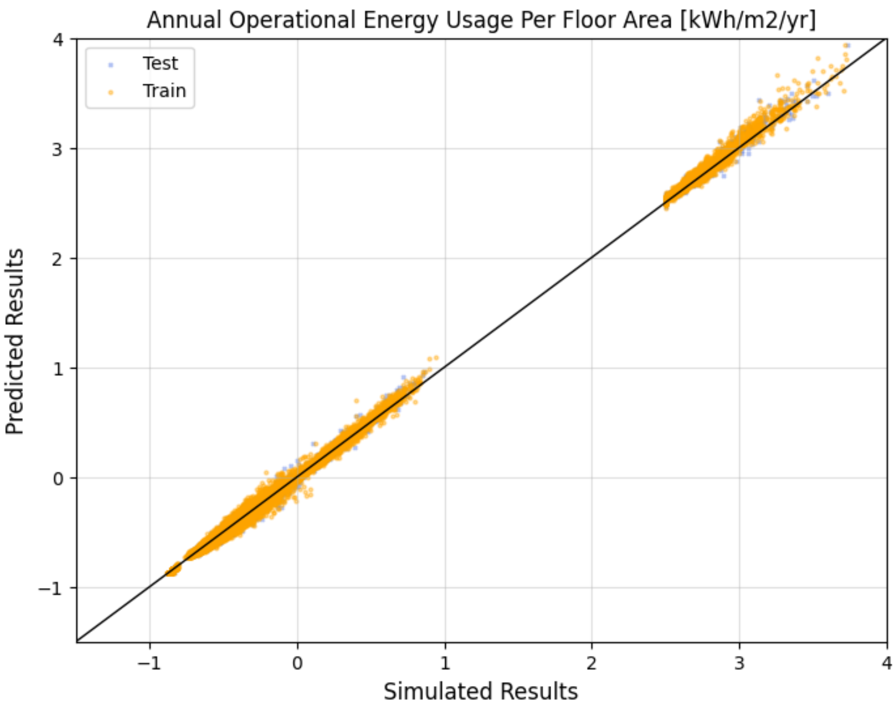
Overall Comparison of all models tested in terms of performance



Model	MSE	RMSE	R²
Polynomial Regression with Lasso	302.581275	17.394863	0.993755
Linear Regression	380.887786	19.516347	0.992138
Linear Regression with Preprocessing	380.883383	19.516234	0.992138
PCA Linear Regression	1009.804470	31.777421	0.979157
Decision Tree Regression	164.727118	12.834606	0.996600
Random Forest (Pre-tuning)	80.288517	8.960386	0.998343
Random Forest (Post-tuning)	80.574747	8.976344	0.998337
XGBoost (Pre-tuned)	63.260267	7.953632	0.998694
XGBoost (Post-tuning)	58.036042	7.618139	0.998802

Optimal Model: Neural Network





Optimal model because of its incredible ability to improve significantly. The loss evolution graph makes evident that as epochs increase, the loss stabilizes. This is characteristic of a neural network training where early iterations substantially improve and subsequently training refines the model. It's ability of learning is evident by observing the plot as well; the spikes initially suggest that the model is encountering new patterns, and as epochs increase, the model gradually learns the patterns and stabilizes.

The monitoring of the loss is crucial to stop the training, as the model would continue learning and eventually overfit.

Please refer to the neural network ipynb for more details.