Seattle EnergyStar Prediction Analysis

Problem Statement

The level of decarbonization necessary to meet the goals set by the Paris Climate Agreement will require the curtailment of emissions in numerous sectors. Residential and commercial buildings are a key such sector; in 2019, these buildings were responsible for 28% of global energy-related CO2 emissions. The first step toward reducing these emissions is establishing a benchmark of how much carbon dioxide buildings currently emit. Once this is understood, building developers, owners, and operators can implement efficiency solutions and rigorously track their progress toward meeting broader emissions targets.

Fortunately, through the EPA's Energy Star program, building managers can collapse the environmental qualities of their buildings into a single metric that can allow them to compare their building's impact with that of others. This can be done through a proprietary program on the EPA's website; however, through recreating the tool with regression modeling, managers can develop a better grasp of which building attributes have a pronounced impact on Energy Star performance ratings and focus their capital investment in efficiency upgrades that can deliver the greatest return on investment.

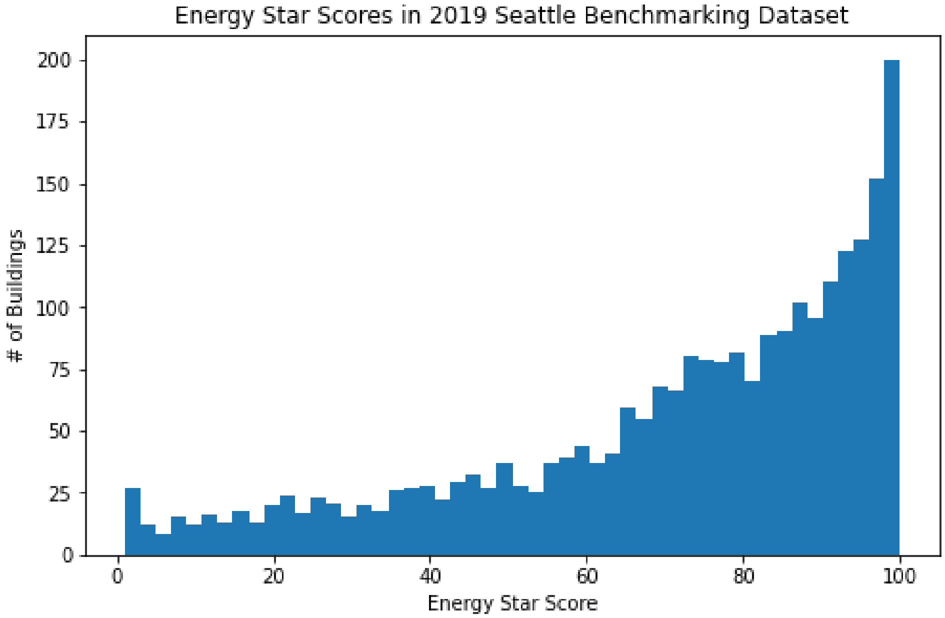
Using benchmarking data on buildings in the Seattle area, I created a model that would allow developers to predict building Energy Star ratings based on existing building information readily available for public access. The data reflects various attributes of these buildings including number of floors, floor area, location in the city, and energy usage. Through the usage of random forest regression, the model was able to achieve a r^2 of 0.927 on the training set and 0.596 on the test set. The corresponding mean absolute error was 4.45 on the training set and 10.4 on the test set. When using cross-validation, the mean test set r^2 decreased to 0.480, and the mean absolute error increased to 11.9.

Data Wrangling

The raw dataset used in this analysis was initially provided by the City of Seattle on the Open Data Portal website. The data was formatted as a csv file with 3,582 rows and 42 columns. There were several steps taken to tailor the file into the format necessary to begin analysis.

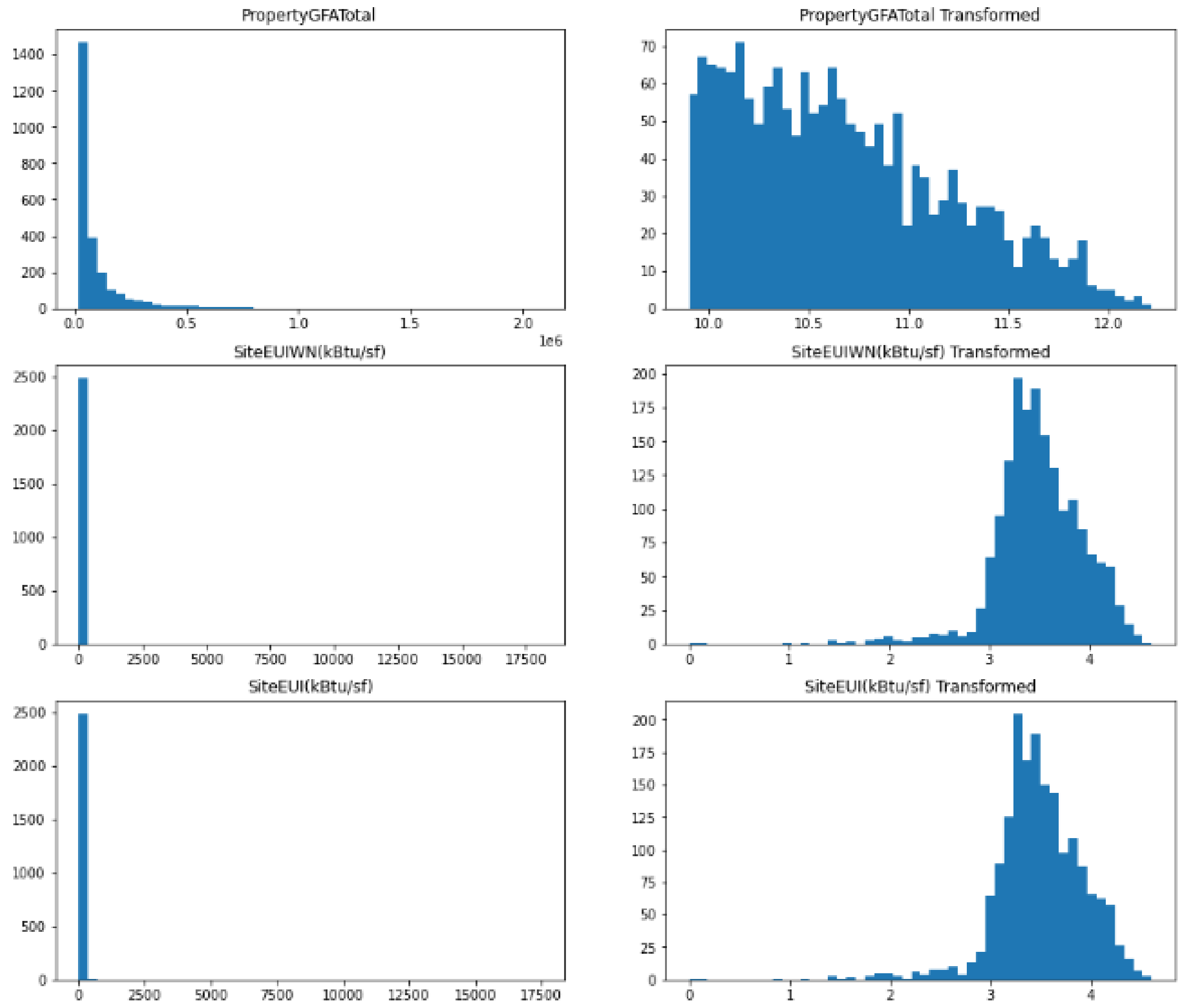
First, I removed any redundant columns in the file, namely the City and State columns because all of the buildings are located in Seattle. Next, I removed all rows that contained a null value in the ENERGYSTARScore column, as this is our target variable for the regression analysis. Finally, after setting the null floor areas to zero, there were very few rows left (less than 3% of those remaining) that had null values, so I removed these rows as well given their exclusion would not have a material impact on the analysis.

At this stage, the dataset had been reduced to 2,497 rows and 40 columns. Below is a snapshot of the Energy Star ratings in the dataset at this stage.



Exploratory Data Analysis

After removing most of the null values, I removed values in certain columns containing numerical data that were over three standard deviations from the mean so as to limit the outliers in the analysis. I then performed a log transformation on these columns to make the distribution of data points in each of these columns more representative of a normal distribution. Performing this transformation drove extensive changes in the visual representation of these numerical columns, greatly reducing the tail size for each. Below is a visual comparison of select columns (representing the total floor area of each building and the site's energy use intensity, both with and without weather normalization) prior to and after transformation.

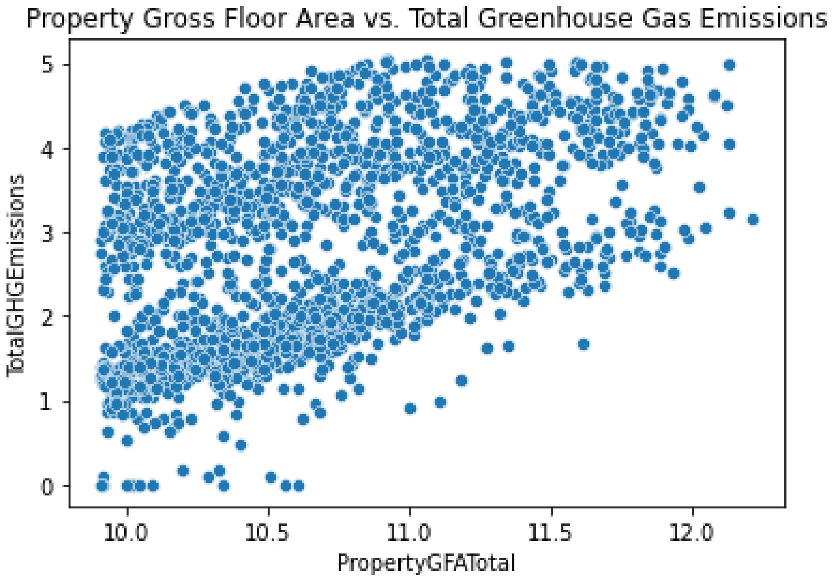


The transformations improved our interpretability of the data in these columns by reducing the range of possible values.

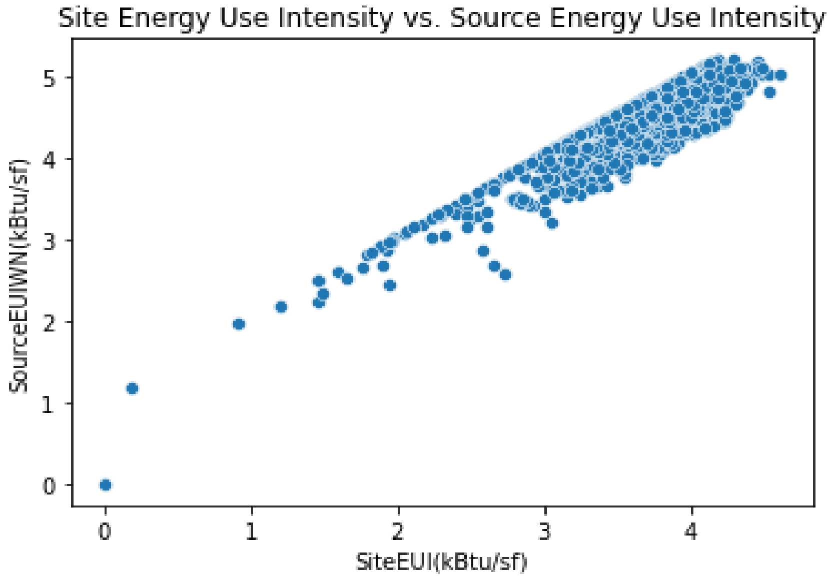
After removing outliers and transforming the numerical data, we were effectively left with 1,766 rows in the dataset. Specific columns of interest in at this stage included:

- PropertyGFATotal, total building and parking gross floor area
- SiteEUI, the annual amount of energy consumed by the property from all sources of energy, divided by gross floor area
- SourceEUI, the annual amount of energy used to operate the property, including losses from generation, transmission, and distribution, divided by gross floor area
- Electricity, the amount of electricity consumed by the property
- NaturalGas, the amount of utility-supplied natural gas consumed by the property
- TotalGHGEmissions, the total amount of greenhouse gas emissions, including carbon dioxide, methane, and nitrous oxide gases released into the atmosphere as a result of energy consumption at the property

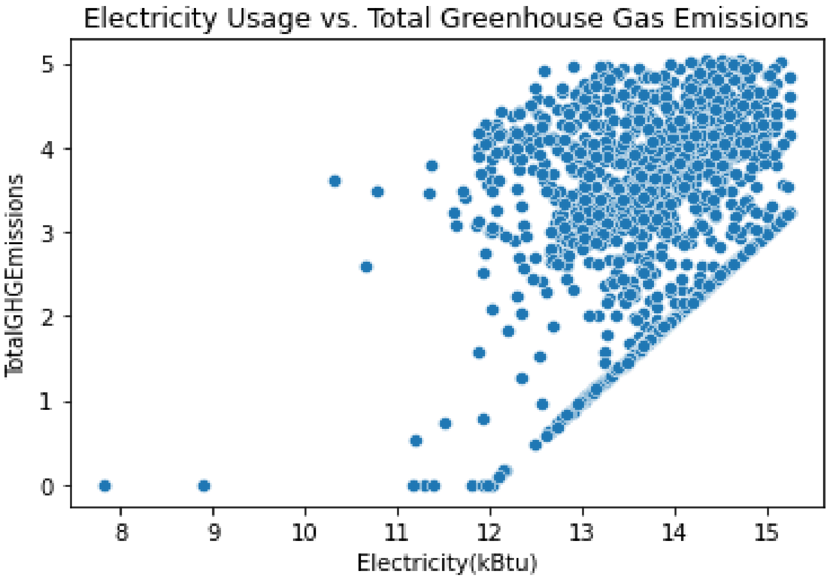
Using the transformed data, I tried to get a better understanding of how different columns compared against one another. The below graph examines the relationship between the gross floor area and the total greenhouse gas emissions.



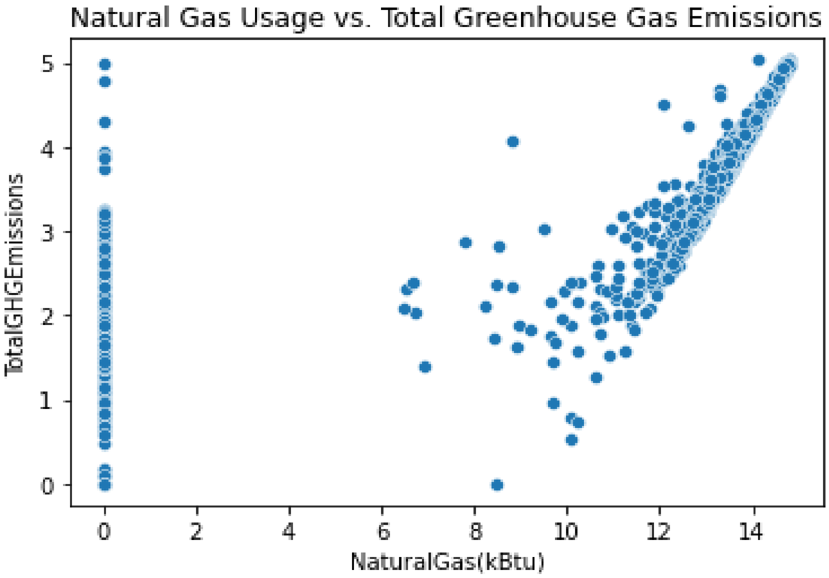
While there does seem to be a positive correlation between the two variables, this relationship does not appear to be strong upon visual examination.



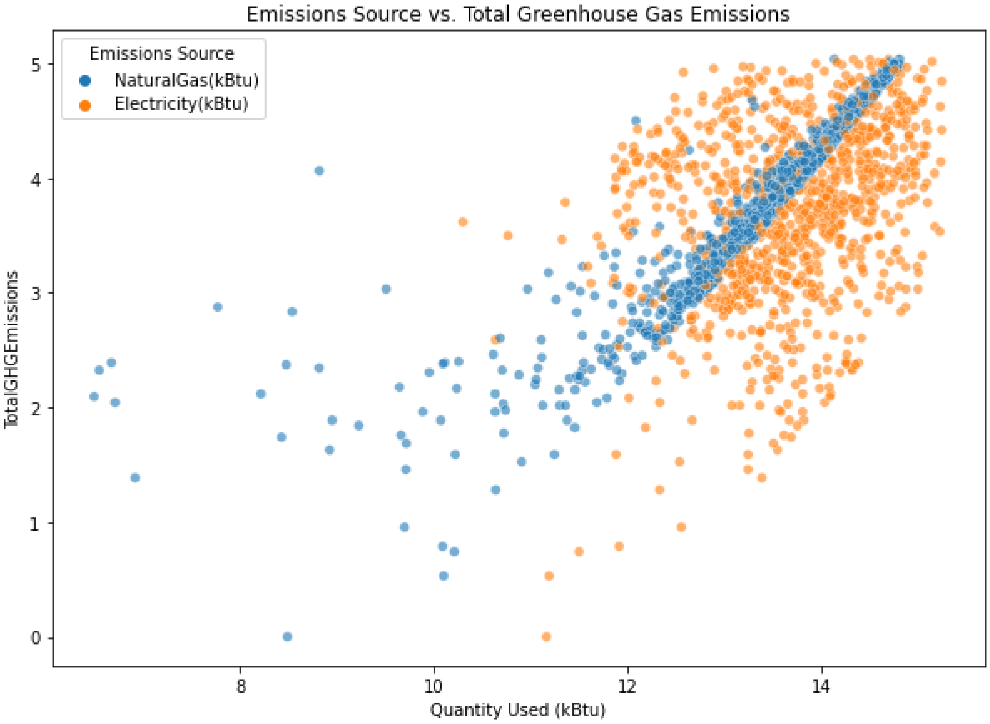
Site Energy Use Intensity and Source Energy Use Intensity, however, seem to be more strongly correlated. This is in line with my intuition because these variables should nearly be the same, but for differences driven by Source EUI’s accounting of transmission and distribution losses.



There does seem to be a positive correlation between electricity consumption and total greenhouse gas emissions, as anticipated.



Except for the outlier values on the left (nearly zero after transformation because certain properties did not consume natural gas), there is generally a positive correlation between natural gas consumption and total greenhouse gas emissions as well. The below graph removes these outliers and compares the total greenhouse gas emissions values against both natural gas and electricity consumption.



Here the natural gas distribution seems to have a stronger positive correlation with total greenhouse gas emissions than electricity consumption does. This raises the question of what mix of sources of electricity generation is supplying power to the buildings in the dataset. Lower carbon sources such as renewable power generation could be contributing to the relationship shown above.

Feature Engineering

After obtaining a more firm understanding of the relationships between variables, I created dummy variables for certain columns containing categorical data to facilitate modeling. Two columns were selected: Building and Neighborhood. To avoid making too many additional columns, I tried to consolidate the values in the original Building and Neighborhood columns such that each value within would appear at least 10 times in the dataset. At this stage, there were 57 columns in the dataset.

Model Selection and Results

Four models were compared: a linear regression model and three random forest regression models. 67% of the remaining data was used to train each of these models, while 33% was used as a control group.

The linear model returned the least impressive results, yielding a training set r^2 of 0.42 and a mean absolute error between the training set actual results and predictions of 12.9. The corresponding test set r^2 and mean absolute error were 0.44 and 12.4, respectively. This suggests the model’s Energy Star rating predictions would deviate from their actual levels by 12 to 13 points, on average.

I expected performance to improve by moving from a linear regression model to a random forest regression model. Using the default hyperparameters, the regression model returned a training set r² of 0.93 and a test set r² of 0.60. Furthermore, the mean absolute error on the training data and predictions decreased to 4.4, while the mean absolute error on the testing data decreased to 10.0. This represented a material improvement over the linear regression model.

To further improve results, I explored the impact of tuning the random forest model’s hyperparameters using a grid search. This involved first performing a random search to obtain a general sense of the bounds in which to perform a grid search. I focused on three hyperparameters: n_estimators, max_features, and max_depth. The random grid returned 200 estimators, 6 columns, and 40 nodes of depth for these fields. Subsequently using a grid search further refined these results, adjusting the maximum depth to 30 nodes instead. Building a model using these hyperparameters kept the training set r² at 0.93 and the test set r² at 0.60; there were negligible changes to each. This suggests that the hyperparameter tuning had minimal impact on model performance.

The final iteration of the random forest regression model assessed featured cross-validation. Upon incorporating cross-validation into the random search, the grid search, and the model creation, I obtained an average test set r² of 0.48 and average mean absolute error of 11.9. The reduction in performance suggests that the other random forest regression models used were overfitting on the data provided, and the five-fold cross-validation results were more representative of how the model would perform with other data.

Takeaways

Ultimately, the models created did not have the best performance. The deviation between actuals and predictions highlighted this, as these variations often exceeded 10 points, while the Energy Star rating scale is bound between 0 and 100. More sophisticated modeling techniques could have increased the level of overfitting observed, and the relative lack of change in results between linear regression and random forest modeling suggests that further refinement may not have yielded much improvement. That said, additional feature selection could potentially yield benefits, as the 57 columns could have caused the analysis to suffer the effects of the curse of dimensionality. Perhaps refining the dataset by eliminating variables with low correlation with Energy Star ratings could yield a more robust model.

It would be interesting to see how these results compare with those generated by using datasets from other cities. Taking the analysis a step further, it would be interesting to recreate this model with more information on building composition. For example, if I had access to additional data on building materials, we could focus on their relationships with greenhouse gas emissions. This information could help city governments pinpoint which materials in the building supply chain are most responsible for pollutants and potentially help incentivize developers to find alternatives.