PREDICTING ENERGY STAR® SCORE FOR BUILDINGS BASED OFF 2017 NEW YORK CITY ENERGY AND WATER DATA DISCLOSURE

HARSHIL PATEL
ISE 4120 – REGRESSION PROJECT
NOVEMBER 29TH, 2019

THE OHIO STATE UNIVERSITY
1971 NEIL AVENUE – 210 BAKER SYSTEMS
COLUMBUS, OH 43210

ABSTRACT

Energy Star Score is a rating that provides information on the energy consumption of many types and usage of buildings. This report analyzes data related to energy efficiency of over 25,000 New York City Buildings with the primary output response of the Energy Star Score. The factors or features to include Site EUI (kBtu/ft²), Natural Gas Use Intensity (kBtu/ft²), and Electricity Use Intensity (kWh/ft²). The resulting prediction model for Energy Star Score indicate reducing energy intensity consumption is a very important consideration to gain the highest Energy Star Score along with assessing the value energy consumption. Furthermore, the model provides building owners and regulators an indication to invest and evaluate a building's energy efficiency rating.

1.) INTRODUCTION

Energy Star (trademarked ENERGY STAR) is a program run by the U.S. Environmental Protection Agency (EPA) and U.S. Department of Energy (DOE) that promotes energy efficiency, its label is commonly found on a variety of categories. The Energy Star program provides on the energy consumption of products, devices and infrastructure. This report is primary concerned with the energy consumption of infrastructure including but not limited to: housing,[36] bank/financial institutions, courthouses, hospitals (acute care and children's), hotels and motels, houses of worship, K-12 schools, medical offices, offices, residence halls/dormitories, retail stores, supermarkets, warehouses (refrigerated and non-refrigerated), data centers, senior care facilities, and wastewater facilities, commercial buildings and industrial plants.

This report explores the publicly available building energy data, or 2017 New York City (NYC) Energy and Water Data Disclosure dataset which includes 60 variables or features that correspond to the energy usage of 34, 686 buildings in the NYC Region. The main objective of the analysis conveyed in the report is to develop a defensible model using regression techniques to predict the Energy Star Score, or an aggregate rating of the overall efficiency of a building based of numerical and continuous factors. The Energy Star Score is a 1-100 rating system or percentile measure of a building's energy performance calculated from self-reported energy usage. Benchmarking is the process used to compute a rating and the Energy Star Score reveals how a building's energy consumption compares to similar buildings. On average, earning a 75 or higher rating for the Energy Star Score is the initial step in receiving a Energy Star for a building; furthermore, it is incorporated in earing LEED certification for green building standards. The objectives are laid out as the following:

- Identify predictors within the dataset for the Energy Star Score, based off numerical and continuous factors of a certain building.

- Build a regression model that can predict the Energy Star Score of building given the building's energy data.

- Interpret the results of the model and use the trained model to infer Energy Star Scores of new buildings.

Assessing the value of energy efficiency in a building through a model can be very vital as once a building is overhauled the new (lower) energy consumption is compared against modeled values for the original building or new buildings to calculate the savings from the retrofit. More accurate models could support better market incentives and enable lower cost financing.

2.) ENERGY STAR SCORE DATA FROM 2017 NYC ENERGY AND WATER DISCLOSURE

The public disclosure of benchmarking data for energy consumed in 2017 in NYC that is organized into a single spreadsheet is described in this section. The Energy Star Score is evaluated as the response/target variable (1-100 rating system) and it is a continuous variable. This report will further enhance upon a supervised regression task as both the features and the target are available. As stated earlier, the report's aim is to establish a model that can map between the features and response variables. This model's is to be developed in order to accurately predict the Energy Star Score, and defensible to justify the validity of those predictions to the true score.

2.1) Cleaning Data

The dataset contains over 32,000 buildings with 60 energy-focused variable known as features. Many columns contain a significant portion of missing values or anomalies. The first step was to remove all categorical variables that fall outside the scope of this study for simplicity as this report will study the numerical and continuous factors that affect Energy Star Scores or have no clear relationship to the Energy Scope Score within context to the data. Refer to the spreadsheet containing the dataset to follow along the data cleaning process.

Now, after removing all the categorical features, the dataset contains over 32,000 buildings with 31 energy-related features for each index. Many of the columns in the data after removing all categorical features contain a substantial percentage of missing values. The next step in the data cleaning process is to remove all buildings (rows in the dataset) that do not have an Energy Star Score, as these rows are not sufficient for using the dataset as training set to create a prediction model on. After this step, that dataset contains 25,455 buildings (rows in data). Subsequently, a table containing the percentage of missing values in each feature or column is build as shown in Table 2. The threshold for the maximum percentage of missing values is set at 30%, which is an arbitrary value to create a meaningful model while being cautious of removing information. Features that have numerous missing values are likely nor useful to our model, based on the provided raw dataset for a building's Energy Star Score in New York City. The features that do not meet this threshold are removed from the analysis provide in this study, reducing the number of features to 19. This concludes the step of cleansing the data.

Table 2: Percentage of Missing Values in Dataset Features

Field Name	% of Missing Values	Remove From Dataset
Self-Reported Gross Floor Area (ft²)	0%	NO
Largest Property Use Type – Gross Floor Area (ft²)	0%	NO
Year	0%	NO
Occupancy	0%	NO
ENERGY STAR Score	0%	NO
Source EUI (kBtu/ft²)	0%	NO
Weather Normalized Source EUI (kBtu/ft²)	11%	NO
Site EUI (kBtu/ft²)	0%	NO
Weather Normalized Site EUI (kBtu/ft²)	11%	NO
Weather Normalized Site Electricity Intensity (kWh/ft²)	3%	NO
Weather Normalized Site Natural Gas Intensity (therms/ft²)	11%	NO
Natural Gas Use (kBtu)	9%	NO
Weather Normalized Site Natural Gas Use (therms)	11%	NO
Electricity Use – Grid Purchase (kBtu)	2%	NO
Electricity Use – Grid Purchase (kWh)	2%	NO
Weather Normalized Site Electricity (kWh)	3%	NO
Total GHG Emissions (Metric Tons CO2e)	0%	NO
Direct GHG Emissions (Metric Tons CO2e)	0%	NO
Indirect GHG Emissions (Metric Tons CO2e)	0%	NO
2nd Largest Property Use – Gross Floor Area (ft²)	84%	YES
3rd Largest Property Use Type – Gross Floor Area (ft²)	95%	YES
Fuel Oil #1 Use (kBtu)	100%	YES
Fuel Oil #2 Use (kBtu)	82%	YES
Fuel Oil #4 Use (kBtu)	92%	YES
Fuel Oil #5 & 6 Use (kBtu)	99%	YES
Diesel #2 Use (kBtu)	100%	YES
Propane Use (kBtu)	100%	YES
District Steam Use (kBtu)	95%	YES
District Hot Water Use (kBtu)	100%	YES
District Chilled Water Use (kBtu)	100%	YES
Annual Maximum Demand (kW)	92%	YES
Annual Maximum Demand (MM/YYYY)	92%	YES
Water Use (All Water Sources) (kgal)	44%	YES
Water Use Intensity (All Water Sources) (gal/ft²)	44%	YES

2.2) Cleaned Data of 2017 NYC Energy and Water Disclosure

After the above step of cleansing the data cleansing process is still incomplete, there are now 19 features left to evaluate a model for the Energy Star Score. There are still many features that are redundant as they are features that are calculations based off other features. Through evaluating the definitions of the features, features were identified that are very likely to have high correlation through being collinear. By removing those features, there are 11 features that will be used in this report to build a model that is generalized and become more interpretable.

Table 2 provides a small sample of the data from the final dataset with its features for clarification of the next steps of exploratory data analysis. In addition, Table 3 substantiates the definition of all the available features for each building in the dataset. In the next section, it evaluates the cleaned dataset to search for a model based off trends, patterns, or relationships within the data.

Table 3: Energy Star Scores and Features for NYC Buildings

Building #	ENERGY STAR Score	Year Build	Occupancy (%)	Self-Reported Gross Floor Area (ft²)	Largest Property Use Type - Gross Floor Area (ft²)	Source EUI (kBtu/ft²)	Site EUI (kBtu/ ft²)	Natural Gas Use (kBtu)	Electricity Use - Grid Purchase (kWh)	Total GHG Emissions (Metric Tons CO2e)	Direct GHG Emissions (Metric Tons CO2e)	Indirect GHG Emissions (Metric Tons CO2e)
1	90	1909	95	169416	164754	138.4	53.8	1435755	1920103.6	732.4	76.3	656.1
3	100	1963	100	94380	94380	43.5	28.4	2068300	180640	164.5	109.9	54.6
4	83	1999	85	125000	125000	271.1	130.2	8245445	2354605.3	1150.2	438	712.3
5	27	1994	100	50000	50000	163	76.5	1848519	579335.2	273.4	98.2	175.3

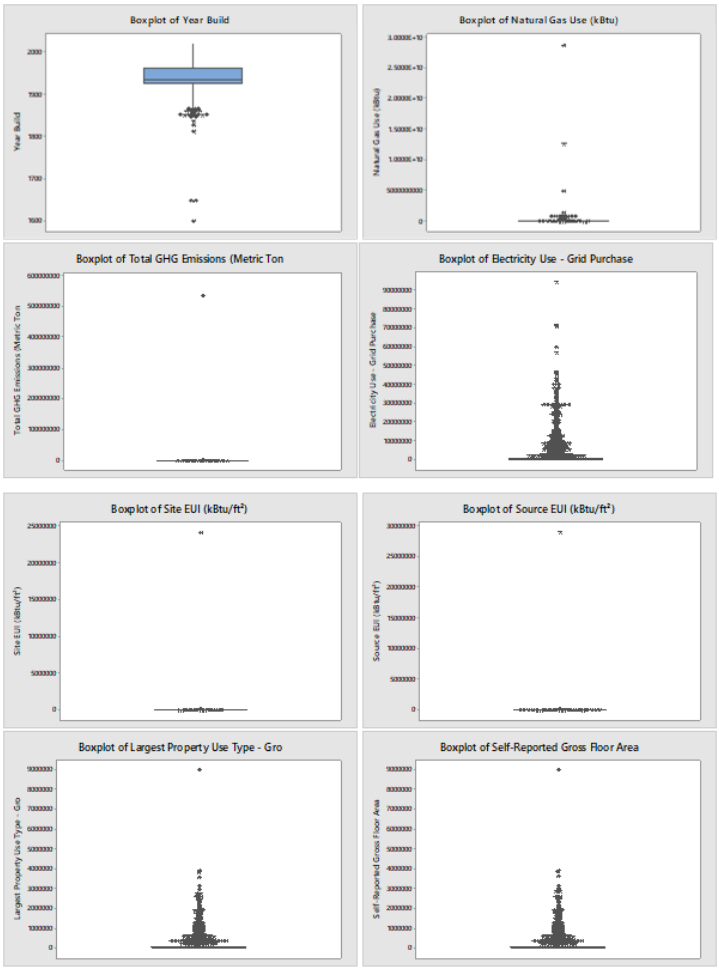
Table 4 : Definitions of Features for NYC Buildings

Field	Definition
Largest Property Use Type - Gross Floor Area (ft²)	The GFA for the largest Property Type.
Year Built	This is the year in which your property was constructed. If your property has undergone a complete renovation that included gutting and rebuilding the interior, then you can indicate the date of this renovation as the year built. If you don't know the exact year the property was built, enter an estimate.
Occupancy (%)	The percentage of your property's Gross Floor Area (GFA) that is occupied and operational.
ENERGY STAR Score	1-to-100 percentile ranking for specified building types, calculated in Portfolio Manager, based on self-reported energy usage for the reporting year.
Site EUI (kBtu/ft²)	Energy use intensity as calculated by Portfolio Manager at the property site in kBtus per gross square foot (kBtu/ft2), for the reporting year.
Natural Gas Use (kBtu)	Energy Use by Type is a summary of the annual consumption of an individual type of energy. Annual totals are available for Natural Gas.
Electricity Use - Grid Purchase (kWh)	Energy Use by Type is a summary of the annual consumption of an individual type of energy. Annual totals are available for Electricity Use - Grid Purchase.
Total GHG Emissions (Metric Tons CO2e)	The total direct and indirect greenhouse gases emitted by the property, reported in metric tons of carbon dioxide equivalent (MtCO2e) for the reporting year.
Direct GHG Emissions (Metric Tons CO2e)	The total direct greenhouse gases emitted by the property, reported in metric tons of carbon dioxide equivalent (MtCO2e) for the reporting year.
Indirect GHG Emissions (Metric Tons CO2e)	The total indirect greenhouse gases emitted by the property, reported in metric tons of carbon dioxide equivalent (MtCO2e) for the reporting year.
Property GFA - Self-Reported (ft²)	Self-reported total gross square footage (ft2) of the property.
Source EUI (kBtu/ft²)	Energy use intensity as calculated by Portfolio Manager at the source of energy generation in kBtus per gross square foot (kBtu/ft2), for the reporting year.

2.3) Final Cleaned Data of 2017 NYC Energy and Water Disclosure

From an initial inspection of the dataset after the previous sections cleansing, it seems apparent that there are major extreme outliers to many features. As this is an actual dataset that was collected on NYC buildings, there is likely to be errors from data entry, disparities in units, or could possibly be legitimate values that are extreme. For the benefit of a more cohesive model on predicting the Energy Star Score of a building, the final cleaned dataset will remove all outliers. Figure 1 displays a boxplot of each feature.

Figure 1: Boxplots of the Features



The extreme outliers removed will be based on all data points or values that fall below of the first quartile -3* interquartile range (IQR) and above the third quartile + 3 ∗ interquartile range. Minitab software created a new dataset that excludes these outliers. Even though these outliers may represent meaningful data, the dataset will contain large portion of data to make a defensible model of predicting Energy Star Score. Through data manipulation in Excel software, these extreme outliers were identified and those corresponding building were removed from consideration using the following formula:

=IF(OR(C2<(QUARTILE($C$2:$C$24632,1)-(3*(QUARTILE($C$2:$C$24632,3)-QUARTILE($C$2:$C$24632,1)))),C2>(QUARTILE($C$2:$C$24632,1)+(3*(QUARTILE($C$2:$C$24632,3)-QUARTILE($C$2:$C$24632,1))))),TRUE,FALSE)

At the end of this data cleaning process, there over 20,000 buildings with 11 defining features. The cleaned dataset and the raw dataset with the process of cleaning the data can be inspected for verification.

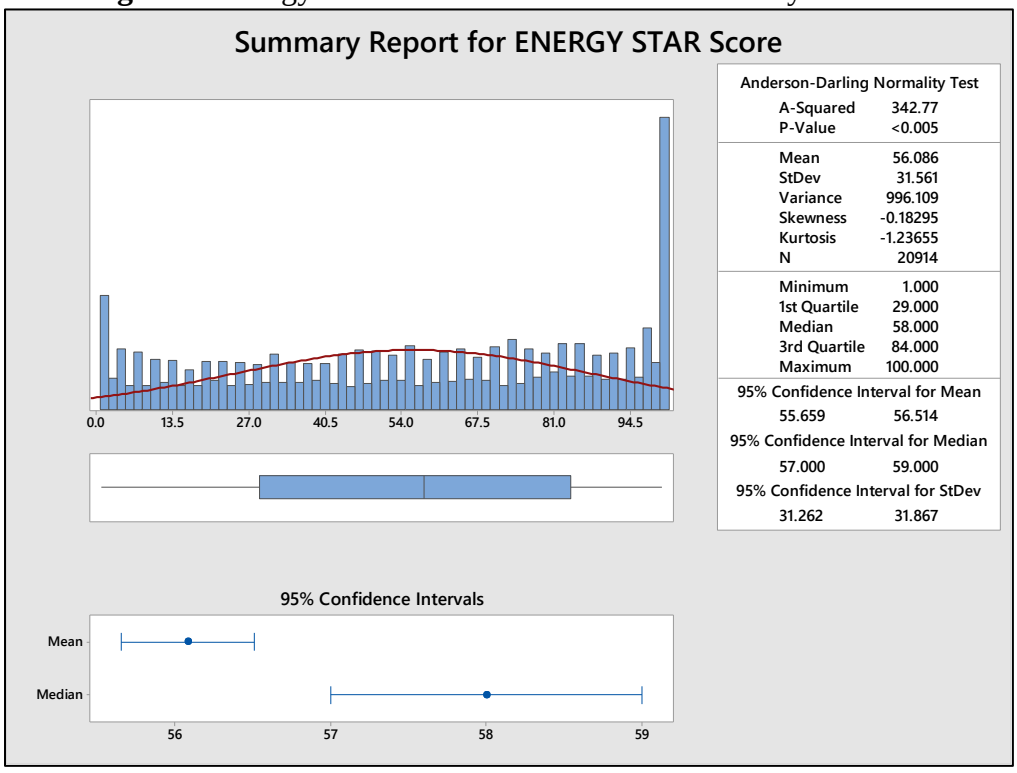
3.) EXPLORATION AND DATA ANALYSIS

In this section, the cleaned dataset will be analyzed in accordance to determine and visualize the data's statistics, and evaluate trends, anomalies, patterns, or relationships between the features and response variables. This section will further allow the study of the data and the initial knowledge that can be gained from inspecting it. This inspection process narrows to specific areas as the data is understood. By conducting this process, features are studied that can be used to inform our modeling choice. This section is aided through a Minitab notebook to create visualizations.

3.1) Distribution of Energy Star Scores

The target or response variable is the Energy Star Score, Figure 2 conveys the distribution of the Energy Star Score across the dataset of buildings in the NYC region.

Figure 2: Energy Star Score Distribution and Summary Statistics

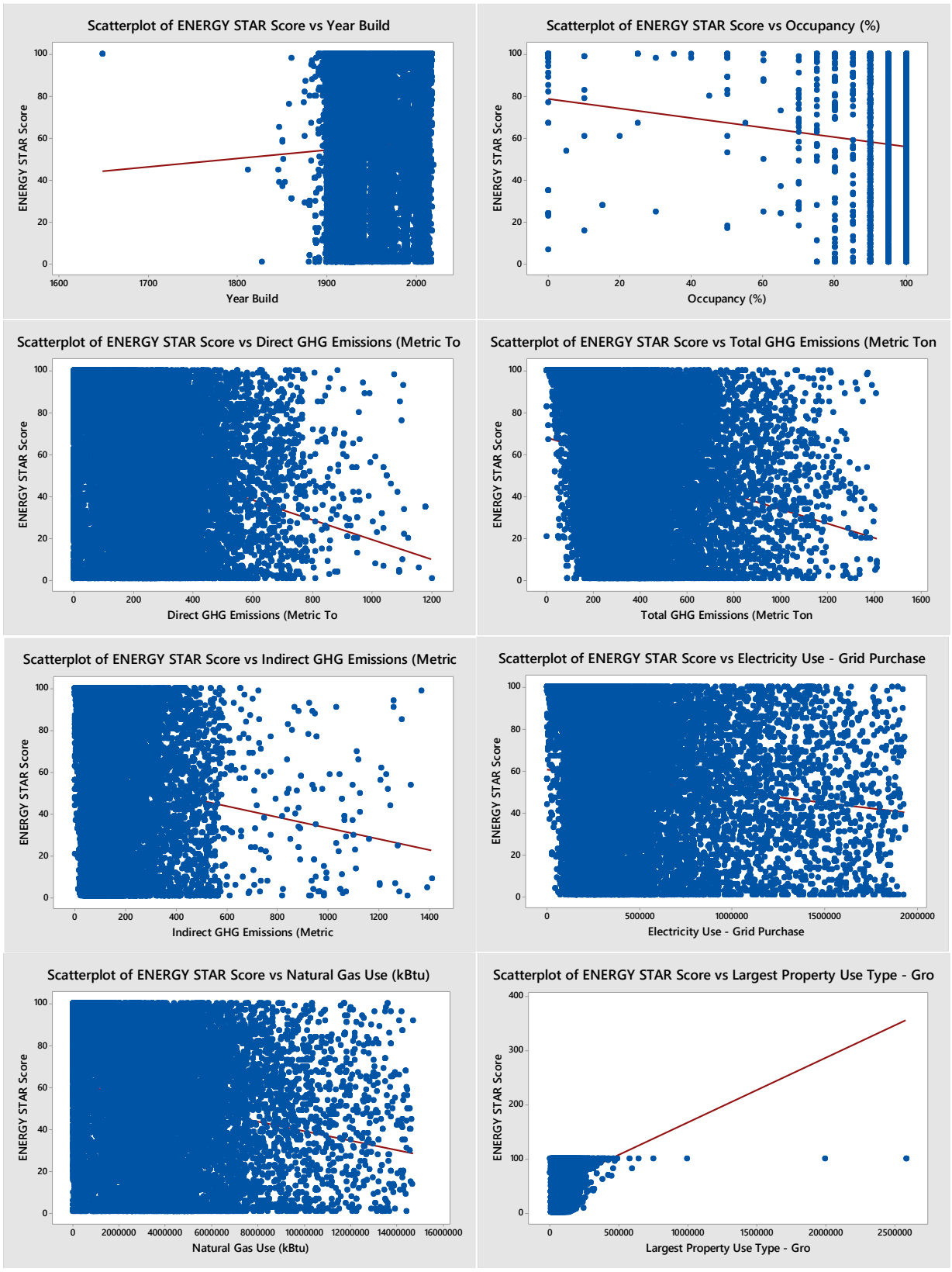


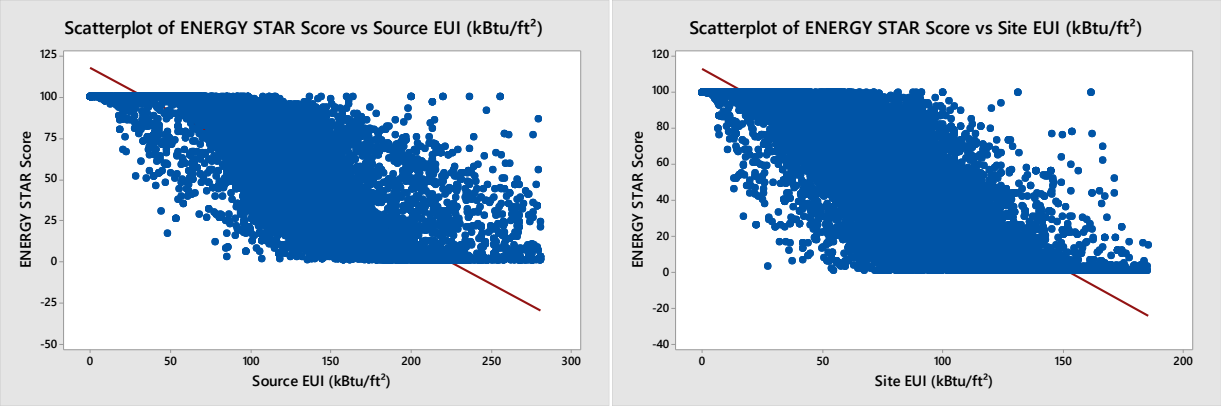
The Energy Star Score is highly concentrated around the highest rating of 100 and lowest rating of 1. This provides an unusual quantity of buildings having the minimum and maximum rating for the Energy Star Score. This response variable is based off self-reported energy usage response variables which is very indictive of self-reporting being biased as building managers and owners are likely to skew results in favor of themselves if asked to evaluate their building metrics. Further evaluation needs to be conducted either providing a more objective measure of a building's energy efficiency or improving energy benchmarking with self-reported data. The former and latter fall outside the scope of this report's analysis of focusing on predicting the Energy Star Score and not devising a methodology of screening building metrics.

3.2) Relationship Between Features and Energy Star Scores

The target or response variable is the Energy Star Score, which is determined by a variety of factors. To explore the relationships between the features in the data and target variable, a matrix plot containing subplots of the relationship among each feature and the target variable is conveyed in Figure 3. This matrix plots allows for the inspection of relationships and distributions between all the features and Energy Star Score.

Figure 3: Energy Star Score Matrix Plots





In inspecting both the Energy Star Score and Site EUI or Source EUI, there is an indication of an expected negative relationship present. The Pearson correlation determines a quantifiable expression of the relationship through a linear perspective conveying the strength and direction of the relationship. The Pearson correlation between Energy Star Score and Site EUI is -0.79, and Pearson correlation between Energy Star Score and Source EUI is -0.786. Figure 4 provides the Minitab operation to calculate the Pearson correlation.

Figure 4: Energy Star Score and EUI Pearson Correlation

Correlation: ENERGY STAR Score, Source EUI (kBtu/ft²)

Correlations

Pearson correlation	-0.786
P-value	0.000

Correlation: ENERGY STAR Score, Site EUI (kBtu/ft²)

Correlations

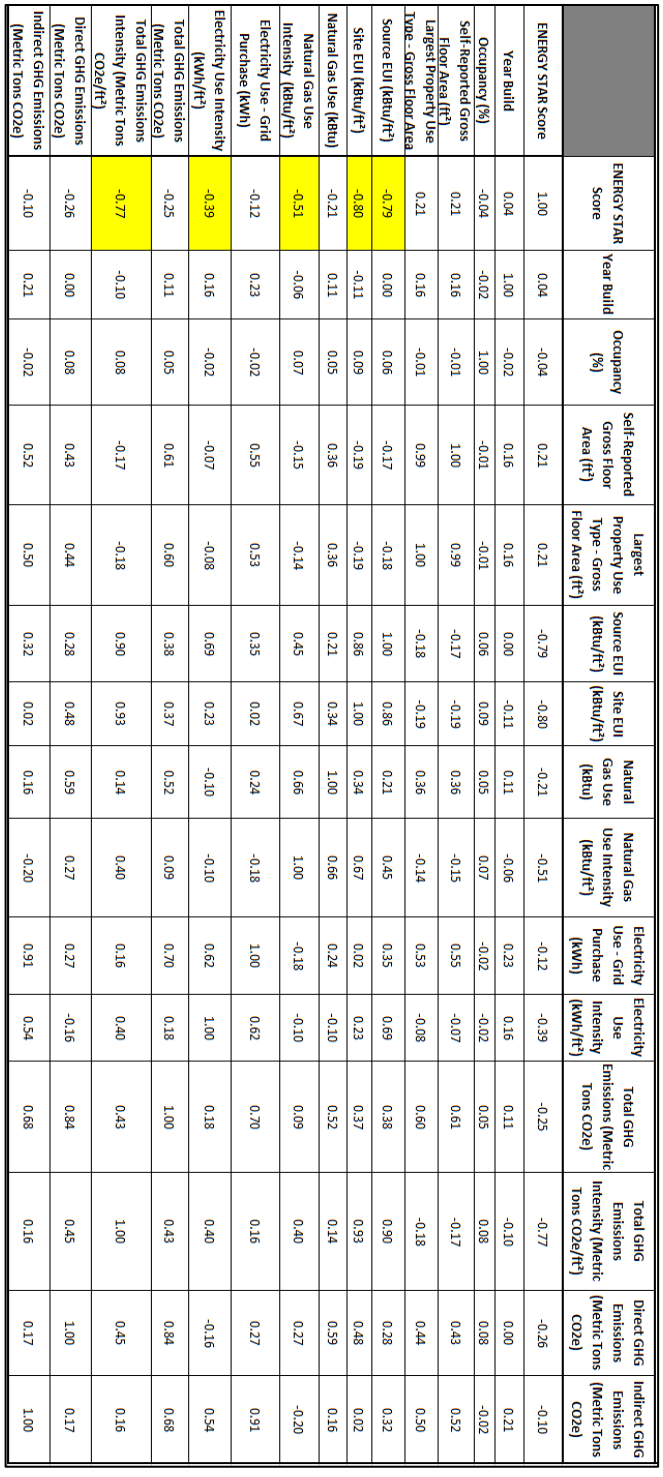
Pearson correlation	-0.795
P-value	0.000

Furthermore, a further analysis of the relationship between the features and the Energy Star Score can be done by constructing a Pearson Correlation Matrix as shown in Figure 5. The Pearson Correlation Matrix shows the linear relationship between all the features and target response. A value closer to 1 or -1 conveys a positive or negative relationship among the two variables.

Additionally, three new features were engineered and added to the dataset using existing features: Natural Gas Use Intensity (kBtu/ft²), Electricity Use Intensity (kWh/ft²), and Total GHG Emissions Intensity (Metric Tons CO2e/ft²). Feature engineering is the process of using existing or raw features and establish new features through transformation of variables in order to further understand the present data. These features were constructed to further demonstrate how the efficiency of a building's natural gas, electricity, and emissions were utilized in accordance to the building's total square footage – giving a better indication of how these three

features were used in proportion to the building parameters. These new features were determined by dividing the corresponding existing feature and dividing by the Self-Reported Gross Floor Area (ft^2).

Figure 5: Features and Response Pearson Correlation Matrix



	ENERGY STAR Score	Year Build	Occupancy (%)	Self-Reported Gross Floor Area (ft²)	Largest Property Use Type - Gross Floor Area (ft²)	Source EUI (kBtu/ft²)	Site EUI (kBtu/ft²)	Natural Gas Use (kBtu)	Natural Gas Use Intensity (kBtu/ft²)	Electricity Use - Grid Purchase (kWh)	Electricity Use Intensity (kWh/ft²)	Total GHG Emissions (Metric Tons CO2e)	Total GHG Emissions Intensity (Metric Tons CO2e/ft²)	Direct GHG Emissions (Metric Tons CO2e)	Indirect GHG Emissions (Metric Tons CO2e)
ENERGY STAR Score	1.00	0.04	-0.04	0.21	0.21	-0.79	-0.80	-0.21	-0.51	-0.12	-0.39	-0.25	-0.77	-0.26	-0.10
Year Build	0.04	1.00	-0.02	0.16	0.16	0.00	-0.11	0.11	-0.06	0.23	0.16	0.11	-0.10	0.00	0.21
Occupancy (%)	-0.04	-0.02	1.00	-0.01	-0.01	0.06	0.09	0.05	0.07	-0.02	-0.02	0.05	0.08	0.08	-0.02
Self-Reported Gross Floor Area (ft²)	0.21	0.16	-0.01	1.00	0.99	-0.17	-0.19	0.36	-0.15	0.55	-0.07	0.61	-0.17	0.43	0.52
Largest Property Use Type - Gross Floor Area (ft²)	0.21	0.16	-0.01	0.99	1.00	-0.18	-0.19	0.36	-0.14	0.53	-0.08	0.60	-0.18	0.44	0.50
Source EUI (kBtu/ft²)	-0.79	0.00	0.06	-0.17	-0.18	1.00	0.86	0.21	0.45	0.35	0.69	0.38	0.90	0.28	0.32
Site EUI (kBtu/ft²)	-0.80	-0.11	0.09	-0.19	-0.19	0.86	1.00	0.34	0.67	0.02	0.23	0.37	0.93	0.48	0.02
Natural Gas Use (kBtu)	-0.21	0.11	0.05	0.36	0.36	0.21	0.34	1.00	0.66	0.24	-0.10	0.52	0.14	0.59	0.16
Natural Gas Use Intensity (kBtu/ft²)	-0.51	-0.06	0.07	-0.15	-0.14	0.45	0.67	0.66	1.00	-0.18	-0.10	0.09	0.40	0.27	-0.20
Electricity Use - Grid Purchase (kWh)	-0.12	0.23	-0.02	0.55	0.53	0.35	0.02	0.24	-0.18	1.00	0.62	0.70	0.16	0.27	0.91
Electricity Use Intensity (kWh/ft²)	-0.39	0.16	-0.02	-0.07	-0.08	0.69	0.23	-0.10	-0.10	0.62	1.00	0.18	0.40	-0.16	0.54
Total GHG Emissions (Metric Tons CO2e)	-0.25	0.11	0.05	0.61	0.60	0.38	0.37	0.52	0.09	0.70	0.18	1.00	0.43	0.84	0.68
Total GHG Emissions Intensity (Metric Tons CO2e/ft²)	-0.77	-0.10	0.08	-0.17	-0.18	0.90	0.93	0.14	0.40	0.16	0.40	0.43	1.00	0.45	0.16
Direct GHG Emissions (Metric Tons CO2e)	-0.26	0.00	0.08	0.43	0.44	0.28	0.48	0.59	0.27	0.27	-0.16	0.84	0.45	1.00	0.17
Indirect GHG Emissions (Metric Tons CO2e)	-0.10	0.21	-0.02	0.52	0.50	0.32	0.02	0.16	-0.20	0.91	0.54	0.68	0.16	0.17	1.00

From examining Figure 5, there are five strong negative correlations between the features and the response of Energy Start Score. These features include: Source EUI (kBtu/ft²), Site EUI (kBtu/ft²), Natural Gas Use Intensity (kBtu/ft²), Electricity Use Intensity (kWh/ft²), and Total

GHG Emissions Intensity (Metric Tons CO2e/ft²). All the features are an indication of a building's Energy Use Intensity (EUI), which is the energy usage of a building per square footage. These features are a all measures of efficiency within a building; thus, a lesser value would indicate a higher Energy Star Score, which is intuitive.

3.3) Feature Selection

In order to develop a defensible model using regression techniques to predict the Energy Star Score, the most relevant features need to be selected for the model passes the so-called "regression checklist" (Allen, 2010). Selecting the features to be used by the model form need to be interpretable and are the most important factors of the response value, Energy Star Score. The features that will be selected from the previously identified strong negative correlated features with the response are:

- Site EUI (kBtu/ft²)
- Natural Gas Use Intensity (kBtu/ft²)
- Electricity Use Intensity (kWh/ft²)

Site EUI had a strong relationship with the response, along with both Natural Gas Use Intensity and Electricity Use Intensity. Additionally, these features were not highly collinear with one another. By having highly collinear features in the model, the variance inflation factors (VIFs) would be too large as VIFs detects multicollinearity in regression analysis, which can adversely affect your regression results in negative way; thus, nor meeting the regression checklist. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

4) REGRESSION MODEL APPROACH

4.1) First Order Model Investigation

As stated before, Allen (2010, pp. 392-397) bears a standard first order regression modeling process. This allows for a first order model form to be developed to predict the Energy Star Score corresponding to the selected features. For the model to be defensible, the regression checklist must be passed for on-hand data as this report is based off. This checklist is shown in Figure 6.

Figure 6: Regression Checklist

Issue	Measure	On-hand
Inputs: evidence is believable?	Randomization completed?	✕
Inputs: model is supported?	VIFs < 3	?
Outputs: outliers in the data?	Normal plot of residuals	?
Outputs: model is a good fit?	Summary statistics	?
Variation Explained	$R^2 > 0.3$	?
Model makes sense?	Subjective assessment	?

Below, the first order model in the three features: Site EUI (kBtu/ft²), Natural Gas Use Intensity (kBtu/ft²), Electricity Use Intensity (kWh/ft²). The corresponding coefficients were determined through a regression analysis. Figure 7 shows the results of the regression and model estimated for the first order regression fit of the features and response of Energy Star Score. Figure 8 shows a normal probability plot of the residuals from the first order model as described below. Un-coded inputs for each feature were used as the model is of first-order; hence; it does not affect the statistics including VIFs or hypothesis testing. Minitab software was utilized throughout these analyses. Below is the determined first-order model of the Energy Star Score grounded on the feature:

$$\boldsymbol{Y_{est} = 120.732 - 0.651x_1 - 0.057x_2 - 2.056x_3 + \varepsilon}$$

where …
Y_{est} *is the estimated average Energy Star Score,*
x_1 *is the Site EUI in* (kBt/ft^2),
x_2 *is the Natural Gas Intensity in* (kBt/ft^2),
x_3 *is the Electricity Use Intensity in* (kBt/ft^2)

Figure 7: Energy Star Model Estimates for First Order Regression Fit

Regression Analysis: ENERGY STAR Score

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	14215879	4738626	14977.09	0.000
Site EUI (kBtu/ft²)	1	4699875	4699875	14854.61	0.000
Natural Gas Use Intensity (kBt	1	42408	42408	134.04	0.000
Electricity Use Intensity (kWh/	1	1023465	1023465	3234.80	0.000
Error	20910	6615750	316		
Lack-of-Fit	15136	6593243	436	111.75	0.000
Pure Error	5774	22507	4		
Total	20913	20831629			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
17.7874	68.24%	68.24%	68.23%

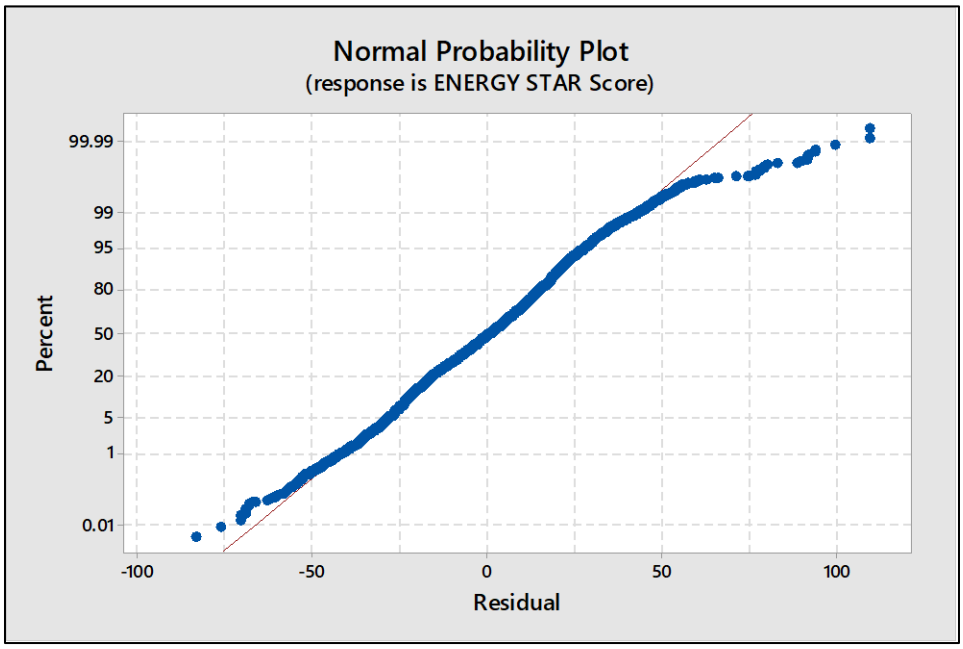
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	120.732	0.334	361.92	0.000	
Site EUI (kBtu/ft²)	-0.65132	0.00534	-121.88	0.000	2.15
Natural Gas Use Intensity (kBt	-0.05665	0.00489	-11.58	0.000	2.06
Electricity Use Intensity (kWh/	-2.0555	0.0361	-56.88	0.000	1.20

Regression Equation

ENERGY STAR Score = 120.732 - 0.65132 Site EUI (kBtu/ft²)
- 0.05665 Natural Gas Use Intensity (kBt
- 2.0555 Electricity Use Intensity (kWh/

Figure 8: Energy Star Model – Normal Plot of Residuals for First Order Regression Fit



The first order model passes the regression checklist through an assessment of approval of the subjective believability since the coefficients associated with higher response of Energy Star Score are negative. This would indicate buildings with higher Energy Star Score are associated lower Site EUI (kBtu/ft²), Natural Gas Use Intensity (kBtu/ft²), and Electricity Use Intensity (kWh/ft²). Furthermore, the variance inflation factors (VIFs) as explained earlier are all less than 3. Then, the summary statistics are very reasonable as p-values = 0. Thereafter, the moderately high R^2 or coefficient of determination (68.24%>>30%) means the variance in in the response variable, Energy Star Score, is adequately explained by the feature of the model.

Lastly, from analyzing the normal popularity plot of the residuals, it is acceptable to show normality for most of the data at hand. The major deviations from the normality line is explained the distribution of the Energy Star Score across the dataset of buildings in the NYC region as shown in Section 3.1. From the related histogram, the Energy Star Score is highly concentrated around the highest rating of 100 and lowest rating of 1. This provides an unusual quantity of buildings having the minimum and maximum rating for the Energy Star Score. Further inspection of the credibility of the dataset provided is required for a through explained of the outliers in the residual plot.

To further develop this model, in order to attempt to align the outliers in the normal popularity plot of the residuals, more terms can be added to the model as the VIFs are relatively low.

4.2) Second Order Model Investigation

As stated in the earlier section, the normal probability plot of the residuals is prone to outliers as indicated in Figure 8. To create even a more defensible model, a second order model is utilized. The model form will be with scaled units through subtracting the mean and diving the standard deviation creating the range of inputs between -1 to +1. Coded units aid in minimizing numerical errors with better ability for in model interpretation. The terms with the largest first order effects through the t-value (absolute value) in Figure 7 was Site EUI (kBtu/ft²). The second order term will be Site EUI* Site EUI through the Minitab Regression Model Dialog. Figure 9 shows the results of the regression and model estimated for the second order regression fit of the features and response of Energy Star Score. Figure 10 shows a normal probability plot of the residuals from the second order model as described below.

Figure 9: Energy Star Model Estimates for Second Order Regression Fit

Regression Analysis: ENERGY STAR Score

Method

Continuous predictor standardization
Subtract the mean, then divide by the standard deviation

Predictor	Mean	StDev
Site EUI (kBtu/ft²)	76.8972	33.7680
Natural Gas Use Intensity (kBt	45.1939	36.1044
Electricity Use Intensity (kWh/	5.8383	3.7233

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	14219480	3554870	11241.24	0.000
Site EUI (kBtu/ft²)	1	4688974	4688974	14827.52	0.000
Natural Gas Use Intensity (kBt	1	42734	42734	135.14	0.000
Electricity Use Intensity (kWh/	1	1025799	1025799	3243.79	0.000
Site EUI (kBtu/ft²)*Site EUI (kBtu/ft²)	1	3601	3601	11.39	0.001
Error	20909	6612149	316		
Lack-of-Fit	15135	6589642	435	111.70	0.000
Pure Error	5774	22507	4		
Total	20913	20831629			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
17.7830	68.26%	68.25%	68.24%

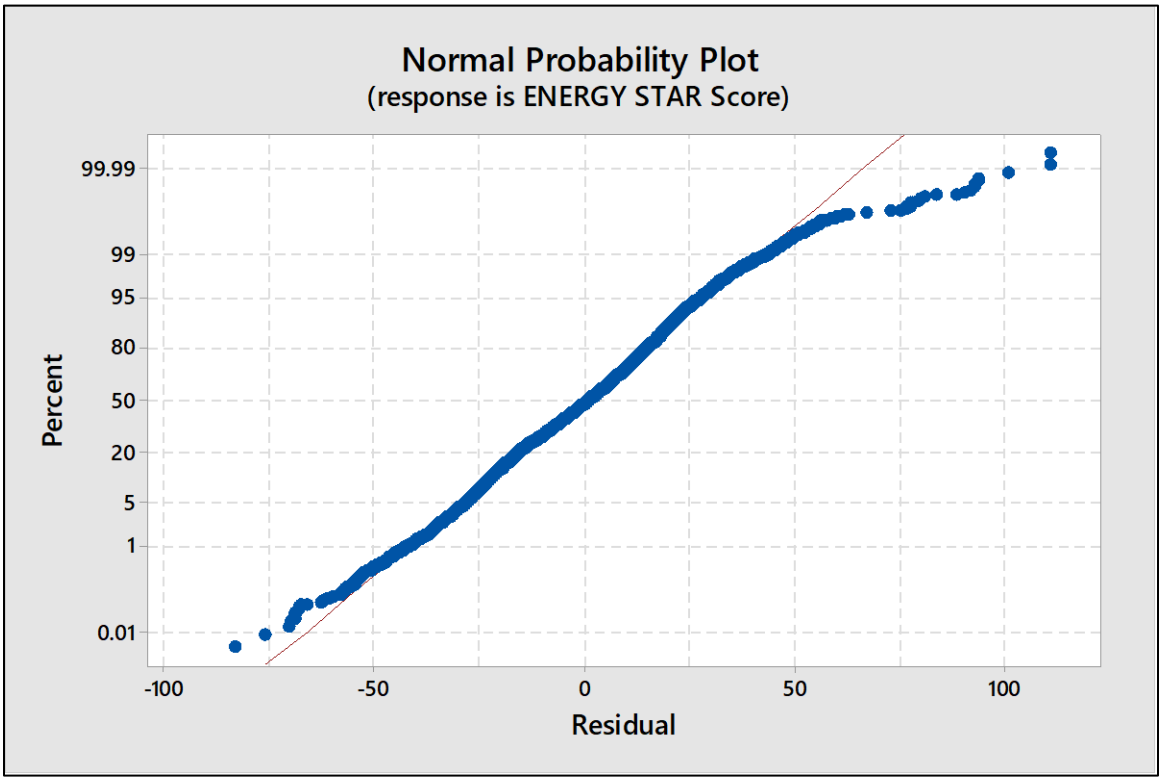
Coded Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	56.376	0.150	376.10	0.000	
Site EUI (kBtu/ft²)	-21.977	0.180	-121.77	0.000	2.15
Natural Gas Use Intensity (kBt	-2.053	0.177	-11.62	0.000	2.06
Electricity Use Intensity (kWh/	-7.664	0.135	-56.95	0.000	1.20
Site EUI (kBtu/ft²)*Site EUI (kBtu/ft²)	-0.2893	0.0857	-3.37	0.001	1.00

Regression Equation in Uncoded Units

ENERGY STAR Score = 119.509 - 0.6118 Site EUI (kBtu/ft²)
- 0.05687 Natural Gas Use Intensity (kBt
- 2.0584 Electricity Use Intensity (kWh/
- 0.000254 Site EUI (kBtu/ft²)*Site EUI (kBtu/ft²)

Figure 10: Energy Star Model – Normal Plot of Residuals for Second Order Regression Fit



The second order model also passes the regression checklist through an assessment believability since this model would indicate buildings with higher Energy Star Score are associated lower Site EUI (kBtu/ft²), Natural Gas Use Intensity (kBtu/ft²), and Electricity Use Intensity (kWh/ft²). Furthermore, the variance inflation factors (VIFs) as are all still less than 3. Then, the summary statistics are very reasonable as p-values for all terms/features are approximately 0. Thereafter, the moderately high R^2 or coefficient of determination (68.25%>>30%) means the variance in in the response variable, Energy Star Score, is adequately explained by the feature of the model.

Lastly, from analyzing the normal popularity plot of the residuals, it is acceptable as most of the data shows normality. The major deviations and outliers from the normality line is explained the distribution of the Energy Star Score across the dataset of buildings in the NYC region as shown in Section 3.1. From the related histogram, the Energy Star Score is highly concentrated around the highest rating of 100 and lowest rating of 1. This provides an unusual quantity of buildings having the minimum and maximum rating for the Energy Star Score. The data that corresponds to the response variable is indictive from self-reported metrics. Self-reported data is very biased as building managers and owners are likely to skew results in favor of themselves or not complete all information if asked to evaluate their building metrics. The high number of '100' and '0' rating is very unlikely and thus the normality plot of residual conveys this disinformation. Further inspection of the credibility of the dataset provided is required for a through explained of the outliers in the residual plot.

5) MODEL SELECTING AND FITTED MODEL

Overall, the second order model did not vary form the first-order model in term off normal probability plot of the residuals; hence it is advisable to continue to use the first order model for simplicity without attempting to overfit or underfit the training data. Again, this model is:

$$\boldsymbol{Y_{est} = 120.732 - 0.651x_1 - 0.057x_2 - 2.056x_3 + \varepsilon}$$

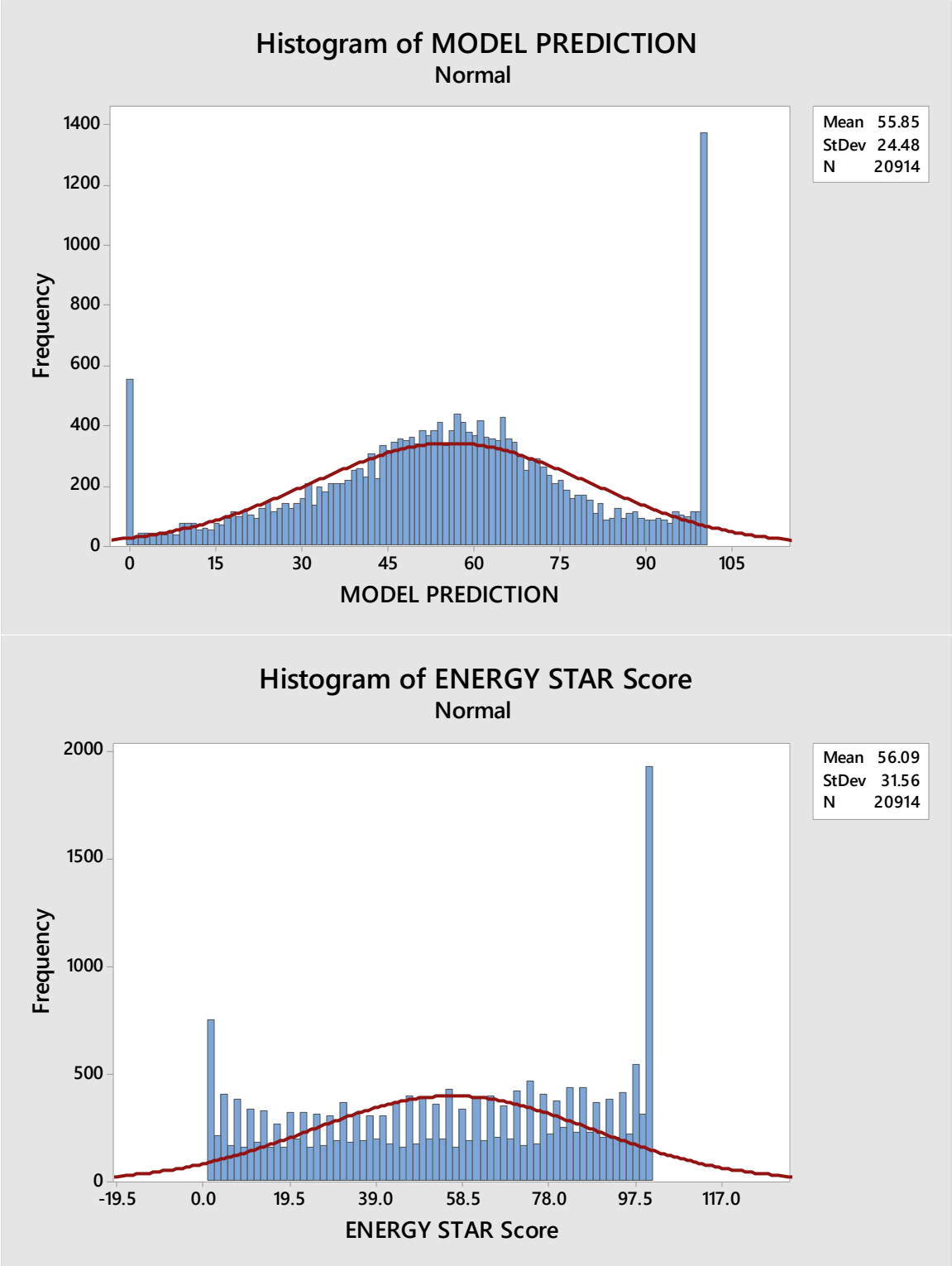
$where\ \ldots$
$Y_{est}\ is\ the\ estimated\ average\ Energy\ Star\ Score,$
$x_1\ is\ the\ Site\ EUI\ in\ (kBt/ft^2),$
$x_2\ is\ the\ Natural\ Gas\ Intensity\ in\ (kBt/ft^2),$
$x_3\ is\ the\ Electricity\ Use\ Intensity\ in\ (kBt/ft^2)$

The normality of the residuals is subjectively acceptable, and further tremendously acceptable id the reasoning explained in the previous section is considered. The model passes the regression checklist of VIF, R-squared, and satisfactory p-values. The model above is a defensible model to predict the Energy Score Rating of buildings, especially in the NYC region.

5.1) Regression Results with Testing Data

Using the first-order model that was developed and convey before, Figure 11 shows the distribution of the predicted Energy Rating Score compared to the true scores using the NYC Energy and Water Data Disclosure for 2018 (Data for Calendar Year 2018). This annual dataset is used a testing dataset that is displaying the data and metrics on water and energy consumption in buildings for FY2018, compared to the 2017 training dataset used for this report. This is considered the testing data it has all available features and pre-determined Energy Star Score to further analyze the model in application. Figure 11 displays the results of the model compared to testing run.

Figure 11: Energy Star Score Model Prediction vs Actual Distribution– Testing Dataset



The summary statistics show a very good analysis of the fit of the testing data, while the model is more normally distributed than what the true data suggest. Further analysis on the modeling technique can be done-possibly using regression techniques such as random forest regression. Random Forest regression technique yields the most important features in a quantifiable manner with greater ease of interpretation. Knowing the most import features variables is very helpful in making predations as random forest is known to be very accurate and can easily be applied on non-linear datasets and features through it feature section algorithm.

6) DISCUSSION

According to our first-order model, Site EUI or the energy use intensity as calculated by Portfolio Manager at the property site in kBtu per gross square foot (kBtu/ft²), for the reporting year is the most important feature, followed by Natural Gas Use Intensity (kBtu/ft²), and Electricity Use Intensity (kWh/ft²) are the most useful for determining the Energy Star Score. These features are inversely proportional to the Energy Star Score; thus, reducing these three intensity factors will have a profound impact on increasing a building's Energy Star Score. Mainly, reducing the Site EUI could be an effective strategy along with savings in natural gas and electricity use to increase energy efficiency of a building. This untimely provides a more favorable Energy Star Score-along with savings in money through expenditure and tax cuts.

The Energy Star score is a rating determined by the measure of benchmarking metrics of the total electricity, natural gas, district steam and heating fuel oil consumed in a building and for other factors. These factors allow governments, intuitions, or building/infrastructure owners to determine an Energy Star Score that aids in understanding how facilities are operating inefficiently. A primary use is model is allowing building owners and governments to predetermine the Energy Star score their building on average can get, prioritizing the Site EUI, Natural Gas Use Intensity, and Electricity Use Intensity. This can fundamentally lead to buildings being prepared for energy efficiency investments and to monitor building performance over time using the three most important features above.

In addition, having an energy efficient building is vital as energy becomes a critical issue of concern and discussion-intensely in sustainability and consumption areas. Fundamentally, this means having further insight into how well energy is being used in buildings is essential to save money as well as reduce greenhouse gas emissions. Buildings and their construction together account for 36 percent of global energy use and 39 percent of energy-related carbon dioxide emissions annually. This report establishes important features to consider for reducing building emissions; thus, thus increasing their efficiency will improve the reliance on non-renewable fuels for the future.

Finally, this report also identified areas for further evaluation in finding a more objective measurement of building energy performance and understanding the root causes of why the Energy Star Score distribution vary extensively.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Theodore Allen, an Associate Professor of Integrated Systems Engineering at The Ohio State University. for teaching about regression techniques and processes and how to work with real datasets in application for simple to complex solutions and analysis.

Furthermore, the author would like to thank The New York City Department of Citywide Administrative Services (DCAS) which annually submits benchmarking results for City buildings to the Department of Finance (DOF) for publication. Benchmarking measures the total electricity, natural gas, district steam and heating fuel oil consumed in a building and adjusts for other factors so that the City can understand which facilities are operating inefficiently.

Please the attached datasets to gain a further understanding of processing and use of the data submitted in the reports. These include the original raw dataset, the author cleaned dataset, and testing data for FY2018.

REFERENCES

"Introduction to Engineering Statistics and Lean Sigma: Statistical Quality Control and Design of Experiments and Systems." 2nd ed. London: Springer, Allen, T.T. 2010.

"Data Disclosure & Reports." GBEE - Greener, Greater Buildings Plan - LL84: Benchmarking - Benchmarking Scores & Reports, https://www1.nyc.gov/html/gbee/html/plan/ll84_scores.shtml.

"Energy and Water Data Disclosure for Local Law 84 2019 (Data for Calendar Year 2018)." Data.gov, Publisher Data.cityofnewyork.us, 20 Nov. 2019, https://catalog.data.gov/dataset/energy-and-water-data-disclosure-for-local-law-84-2019-data-for-calendar-year-2018.

AUTHOR BIOGRAPHY

HARSHIL PATEL is a third year Industrial and Systems Engineering student at The Ohio State University. He is currently taking a Statistical Quality and Reliability Engineering course, to which this publication is submitted to. His interests include machine learning, blockchain, and integrated robotics. He can be contacted at <patel.3001@osu.edu>.