****

 CIS 4321 Data Mining, Section 1, Fall 2021

December 1, 2021

Professor Fadi Batarseh

Jacob Yim

Table of Contents

[Introduction 2](#_Toc89228652)

[Problem Statement 3](#_Toc89228653)

[Understanding the Dataset 4](#_Toc89228654)

[Data Analytics & Results 7](#_Toc89228655)

[Conclusion 11](#_Toc89228656)

[References 12](#_Toc89228657)

# **Introduction**

As an individual who is graduating this December, it has been my interest to work within a sector that deals with energy. I found an interesting data set on Kaggle that included a weather and energy CSV file. I decided to only work with the energy dataset because it had the most significance to my curiosity. Working with this dataset, I got to explore the vast amounts of information that revolve around the sector. Although these utility companies have built lots of infrastructure around the world, there are other domains that I can work within that market. Since my knowledge within this field is very limited, I would be making various predictions based upon the column descriptions.

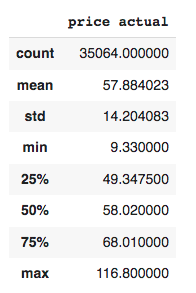
Machine learning is very useful for stakeholders because it allows them to understand specific importance within the domain. Although each training algorithm per model depicts various accuracy measurements, these insights can allow stakeholders to imagine an approximated value for a certain object. These individuals can utilize this knowledge to better prepare them for future decision-making. The professional stakeholders within this field of domain include the government, utilities, financial institutions, contractors, retailers, and manufacturers. The government would contact utility operators when certain emissions are at an unsustainable level, and these restrictions would affect the cost of operations. Financial institutions would directly manage the operations and the cost of goods required for built infrastructure and incurred overhead. Contractors, retailers, and manufacturers help to influence the use and product of energy-efficient products in the market which would help laboring and construction of operations.

# **Problem Statement**

In this dataset, I would be performing an analysis on the target column "actual price" to answer my questions on the dataset. Furthermore, the machine learning techniques used to predict the data would be primarily linear regression due to the data being continuous. In addition, I will be diversifying the model into three different groups: the linear regression model with no adjustments, linear regression with grid search implementation, and linear regression model with recursive feature exploration (RFE). I hypothesize that grid search and RFE will be vital for gaining accuracy in linear regression. Since grid search helps to optimize the features for framing the model, I believe this setup would help to gain the highest accuracy in all my models. I also hypothesize that the column "price day ahead" will have the greatest correlation and the generation of hard fossil coal, hard fossil gas, and other renewables will have the highest correlation to the price. The specific column "price day ahead" appears to be a predictive algorithm from an unknown organization. I also wondered whether or not the correlations in the dataset would have any direct implications for the RFE analysis. These questions would directly impact the decision-making of the stakeholders for features in reference to price.

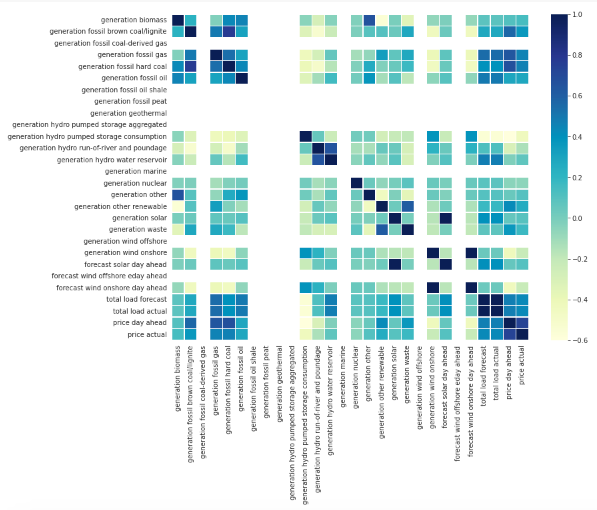
# **Understanding the Dataset**

Taking a glance at the first five rows of data after inserting it into the notebook, I noticed that there was going to be a lot of cleaning required. The first thing I did was to check the shape of the dataset to see how many rows and columns I was working with; the shape displayed 35046 rows and 29 columns. I also went ahead to check for duplicates within the rows to make sure I wasn’t working with any information that may be repeated. As there were none, I went ahead to describe the data to get an idea of the values that are within each column. I was able to see that there are some columns with lots of null values, and some of the features may need to be removed because their values were a lot of 0s. An important feature of the describe method on the target “price actual” column is the utilization of the min and max descriptions.



Pricing in our case should never be negative, and the data portrayed from the output showed that there were no such negative values. All of the values in this column were also listed as there was the same number of counts as the shape showed us. I went ahead to check out the column by using the info method to make sure that all the objects displayed were accurate. There was one object listed as “time,” but I was going to remove it due to it being unrelated to the features. Since all the column objects looked accurate, I went ahead to start looking at the null counts. I found it interesting that some columns had equal amounts of null values as the number of rows. I even went further and placed the dataset into a correlation heatmap to see which columns were missing, and I found that there were around 8 columns that needed to be removed from the list.

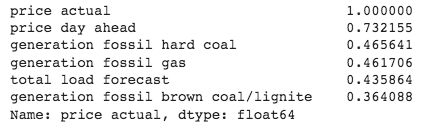
**Correlation Matrix HeatMap**



The biggest challenge in this cleaning the data was figuring out the excess of nan values. By extracting the nan index, I noticed there were some rows with 14 nan values. After inspecting the remaining values, I noticed there were still 5 extra rows remaining. Although there would be data loss from removing the items, I concluded that the items were not relevant to the machine learning that I would be conducting. As time is not inclusive to the original hypothesis, the total removal of these values would only impact 1% of the entire dataset. The result of the dataset that I would be working with has a shape of 35041 rows and 19 columns.

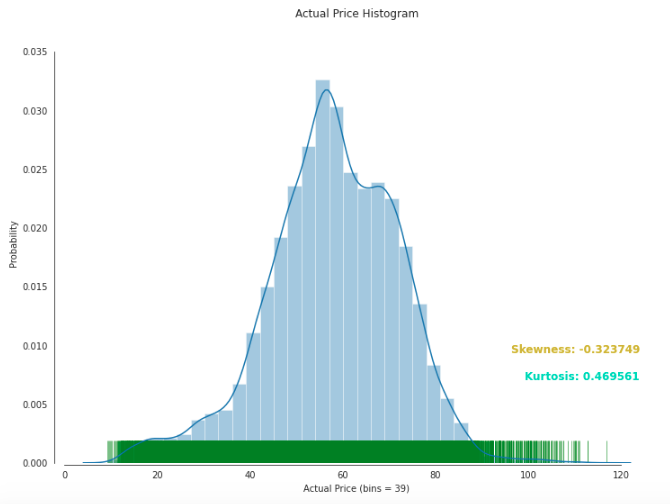
# **Data Analytics & Results**

The initial exploratory analysis that was performed on the cleaning of the dataset was used to understand the "actual price" column. Starting with the correlation matrix, I wanted to see if there was any relation to the feature selection in reference to the target variable price. The correlation matrix would also help to see if the prices fluctuate upon any specific column as this is important to decision-making for management. The results showed that the column, "price day ahead," had relatively the highest positive correlation to price with a p-value of 0.732. As this column is related to the TSO predictive algorithm, I ignored this value because it is possibly directly related to the price column. The next top 5 positive correlations are as indicated:

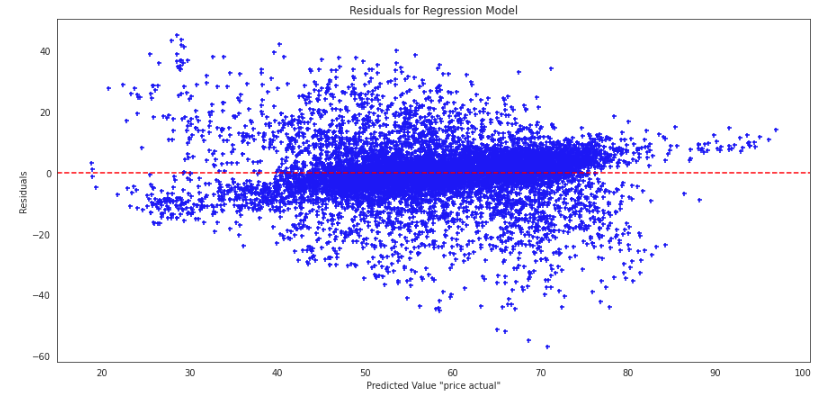


These p – values outputted for each column show that they have a relatively low positive correlation to the price. This means that these features have relation to the target value, but their association with each other is low.

The research also shows that the dispersion of values within the target column is relatively normal. For this assessment, I used a histogram and rug plot overlay, and I calculated the number of bins by using Sturge’s Rule. The histogram showed a skewness of -0.32 and a kurtosis of 0.47. From the skewness, the histogram is fairly symmetrical with a relatively normal distribution. The kurtosis shows that the prices are platykurtic, which is a good indicator of fewer outliers. This is backed up from the rug plot, as the values show that the data are mostly centralized.

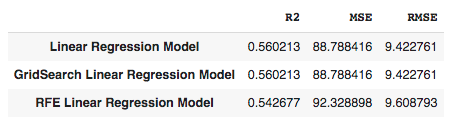


With a dataset of this caliber, it was important to normalize the dataset before modeling to prevent overfitting. I ran the features into a standard scaler and split the data into testing and training categories. After splitting the dataset, I made a linear regression model and decided to show the residuals.



The patterns in the visual graph indicated that there isn't any form of trending pattern in the algorithm. Linear regression was probably not a good fit to predict the target value. This residual plot also shows the relationship with the correlation matrix, for they both describe the relationship of feature columns to the target value; it is indicated that there is positively little to no correlation to the target value. We can also deduct that the value of the residuals lies between 40 and 80 of the predicted target values; this is similar to the distribution of the histogram.

To enhance the accuracy of the model I decided to make two additional models. GridSearch would help to find the perfect parameters for the most amount of accuracy, and RFE feature selection would best find the most relevant columns to select for the predictor. I noticed that the 6 features that RFE had selected were different from the correlation matrix. The only similar columns were the “price day ahead.” The accuracy reports were measured with R-Squared, MSE, and RMSE.



The figure above shows that each model had very low accuracy on all metrics. GridSearch also showed the same accuracy on all metrics in comparison to the first model. With Recursive Feature Elimination (RFE), there is a significant amount of data removal, which should stem with higher accuracy. Since this metric shows lower accuracy than without feature importance, there must be other columns that are important within the dataset that justifies its accuracy.

# **Conclusion**

Although the value of a certain target price point is strongly determined by its features, my research shows that there were no such implications. My original hypothesis about the feature importance being relevant to the correlation proved to be untrue. There could have been some other interesting factors that are relevant due to the size of the dataset column. The grid search model with additional hyperparameter had the same accuracy as the linear regression. I presume that the linear regression model without hyperparameters was already the best result. Although my main theories on the correlation of hard fossil coal and hard fossil gas in reference to price were accurate, I was incorrect about the importance of renewable gas. This came as a surprise to me because I thought that our generation was in the age of finding new and sustainable energy. I had thought that the companies within this realm viewed this ideology as being more important than fossil fuels.

This project was a very fun experience for me because I rarely get to use my knowledge of machine learning. With the collaboration of data preprocessing and explorative data analysis, the development of the end product helped to understand how machine learning can be used in real-life scenarios. In the past, I had done many projects revolving around data visualization and machine learning, but I never worked on a dataset of this caliber. Looking back on the project, I think it would have been interesting to use get dummies and a binary target value to evaluate a classification dataset. In a future analysis, I would bring together the weather dataset to bring forth more possibilities and help better my understanding of computer vision techniques on different models.

# **References**

*Energy Efficiency Program stakeholders*. U.S. Agency for International Development. (2018, October 1). Retrieved December 1, 2021, from https://www.usaid.gov/energy/efficiency/basics/stakeholders.

Jhana, N. (2019, October 10). *Hourly Energy Demand Generation and weather*. Kaggle. Retrieved December 1, 2021, from https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather.