**Module 2: Linear Regression Model**

ALY6020 – Predictive Analytics, Northeastern University, Boston

Professor Justin Grosz

CRN: 20387

03/05/2023

Submitted by: Abhinav Jain

**Introduction**

This report focus will be on a car dataset known for its large vehicles that was struggling due to a high mileage due to powerful engines. The company concerned that their vehicles wouldn't be sufficiently fuel-efficient to appeal to customers who were concerned about the environment. After using the linear regression model, have the information necessary to build a fuel-efficient automobile that will satisfy customer demand.

The initial stage was to compile information on the gas mileage of various automobile attributes. Looked at the data and found several characteristics, like mileage, weight, and horsepower, that may help vehicles get better gas efficiency. We created a linear regression model based on this data to forecast a car's gas mileage depending on these characteristics.

An accurate and reliable linear regression model was built in this project's initial stage, which involved cleansing the data. To look for any missing numbers, anomalies, or outliers, the dataset was thoroughly inspected, and those were then dealt with appropriately. When the data had been cleansed, the following phase was exploring the correlations between the variables utilizing new Python visualization tools and graphing. Hence, we were able to comprehend the data more fully and spot any patterns, trends, or correlations that would help our model.

**Data Cleaning**

Initially, we examine the summary statistics, data visualization, and other exploratory data analysis methods such as boxplot, bar graph, and correlation matrix. As through which to find data problems and discovered missing values in the dataset, denoted by a question mark (?). replaces with NaN missing values that might improve in future analysis for variable prediction. In the dataset imputer the missing value with median of the horsepower variable. There were only 6 values in the field of horsepower. As once missing values have been handled, the outliers in the dataset should be found and eliminated. Outliers are data points that are quite different from the other data points and have little or no impact on the performance of the model. Box plot approaches detected a few outliers that will not have an influence on the construction of the model.

To get the data ready for analysis or modeling, cleaning the data and eliminating outliers was essential. We can make sure that the data was clean and ready for additional analysis by resolving missing values, identifying outliers, and normalizing the data.

**Exploratory Data Analysis**

The dataset on automobiles contains 398 records and 8 variables that provide a full knowledge of the performance of the vehicles, of which 63% were produced in the US and 37% in other nations. The data reveals that the automobiles in the dataset had an average MPG of 23.5, showing that the cars were quite fuel-efficient. The data also reveals that most of the automobiles in the sample have eight cylinders, suggesting that they are strong and capable of producing a lot of torque. Also, the average displacement of the automobiles in the dataset is 262, indicating that they have sizable engines that can produce a big amount of power. The average weight of the automobiles in the dataset is 2970 pounds, which is rather low and indicates that the vehicles were made to be quick and agile. The average acceleration rate of 15 is also outstanding, showing that the vehicles can reach high speeds without losing control.

Chart, bar chart

Description automatically generatedThe automobiles in the dataset were made in 1973, indicating that they are older models. This does not imply that they are ineffective, though. In actuality, the year of the automobile might aid in our understanding of the utilization and effectiveness of the vehicle. For instance, a vehicle with a lower mileage and a history of good maintenance may be more efficient than one with a higher mileage.

A picture containing graphical user interface

Description automatically generatedAfter removing the variables with multicollinearity difficulties, the correlation matrix analysis shows that acceleration, MPG, and model year have the highest relationships with the other variables in the dataset. With a correlation value of 0.42, acceleration is positively correlated with other variables in a moderately strong way. With a value of 0.58, MPG's positive association with other variables is stronger, indicating that MPG should be considered when analyzing other dataset variables. Model year has a smaller positive association (0.29 coefficient) with other factors. It's worth noting that horsepower and MPG are negatively correlated, meaning that automobiles with more horsepower typically have lower MPG. This implies that performance and fuel efficiency in automobiles can be tradeoffs.

Chart

Description automatically generatedThe boxplot analysis demonstrates that the automobiles in the sample have an average acceleration that ranges between 13.8 and 17.2. The data does contain a few outliers, though, which could be worth looking into more. One vehicle in the sample had an exceptionally low acceleration of sub 8.5, which may be a sign of a problem. On the other hand, another vehicle in the dataset had an acceleration of 22, which is quite high. According to the quartile values, 25% of the vehicles in the sample had accelerations of 13.82 or less, and 50% had accelerations of 15.5 or less. The acceleration was at or below 17.17 for the 75th percentile. The dataset's minimum and highest acceleration values were 8, respectively.

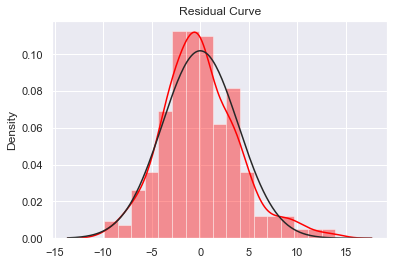
**Analysis**

While created a linear regression model to forecast a car's efficiency based on specific variables and assessed the model's performance using a train and test dataset.  The test score of 0.7778 indicates that the model generalized well to test data, while the train score of 0.7405 indicates that the model was able to explain around 74% of the variation in the training data. The model's accuracy of 0.7778 means that, on average, it was able to estimate the car's efficiency correctly 78% of the time. The model's projections were generally wrong by around 3.84 miles per gallon, according to the Mean Squared Error (MSE) of 14.7735 (assuming the goal).

The model's adjusted R-squared value, which considers the amount of features it contains, in this instance indicates that the model was still able to explain around 73.7% of the variation in the target variable.

The goodness of fit and model complexity was both considered when determining the model's quality, which is indicated by the AIC (Akaike Information Criterion) score of 1669. In this situation, the AIC number appears to be reasonably low, indicating that the model was a good match for the data. A lower AIC value suggests a better model fit.

This analysis gives a clear picture of the relationships between the various variables in the dataset and emphasizes the significance of MPG as a crucial variable to consider while examining other variables. Moreover, it implies that there can be trade-offs between performance and fuel efficiency in automobiles and that the location of a car's manufacturing may not be much influenced by the model year. Moreover, the study provides an intuitive understanding of the range of acceleration for the automobiles in the dataset and identifies several intriguing outliers that may be worth looking into further.

 The residual points' (-3 to 1) range indicates that the model has accurately and fairly captured all the relevant information in the data. Although the tendency of the positive data points to have greater values than the negative data points could sometimes hint that the model is somewhat overestimating the target variable, the pattern of the residuals generally shows that the model is well-fit to the data.

**Conclusion:**

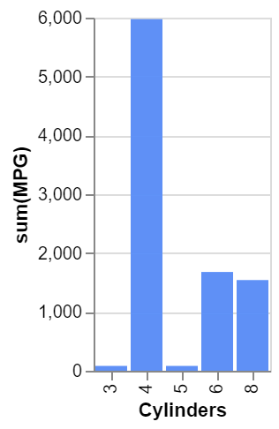
Consider gathering further information on especially early variables, such as engine type, transmission type, and car weight, that may have an impact on automotive economy if this model is used in a new environment. The model's accuracy and predictive ability may be enhanced by incorporating these more variables.

To check if the model could generalize successfully to new data, it may also be useful to assess its performance on a validation set or through cross-validation. To enhance the performance of the model even further, it could be advantageous to investigate other modeling strategies, such as non-linear models or ensemble models.

In terms of technology change or recommendation, we can implement the solar panels in car to provide the better efficiency with other hybrid technology. The incorporation of solar panels inside vehicles would be a creative and exciting idea that has the potential to significantly increase vehicle economy and lessen reliance on non-renewable energy sources. By supplementing the energy that the car's battery uses, solar panels can reduce the amount of energy that the engine needs to produce. Drivers may experience increased fuel economy, less emissions, and cheaper operating expenses consequently.

Appendix:

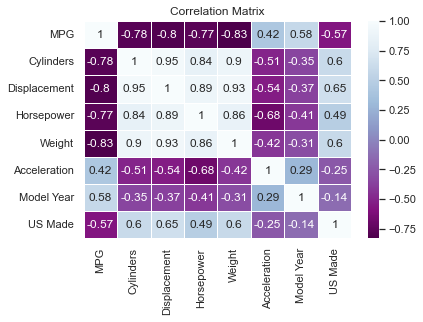
Insights for the data through the visualization tool in python:



Stack chart to analyze Year of Manufacture with cylinder and weights

Chart, bar chart

Description automatically generated

First, Correlation Matrix before dropping of the variable while implementing the model.