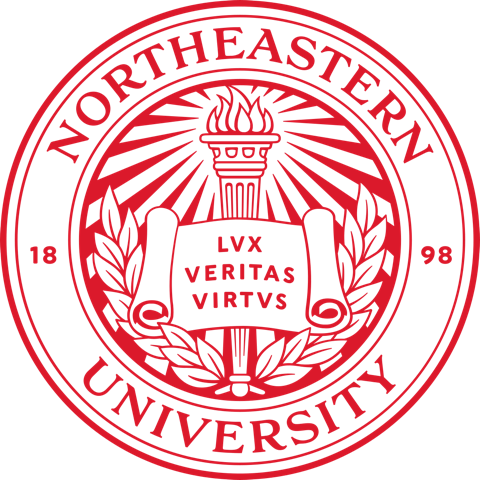
**Module 2 Project**

**Building the Car of the Future**

Sunil Raj Thota



College of Professional Studies, Northeastern University ALY 6020 – Predictive Analytics

Prof. Justin Grosz

November 12, 2021

**Introduction:**

A car manufacturer known for huge autos is having trouble selling them, and they've called for your aid in building a more energy-efficient vehicle. Determine which qualities may lead to higher gas mileage using the data acquired so that a more fuel-efficient automobile can be designed. Seven of the initial instances were eliminated from the prediction of the attribute "MPG" because the "MPG" attribute had uncertain values. The information is about city-cycle fuel usage in miles per gallon, which may be predicted using three multivalued discrete variables and five continuous attributes. Let's discuss it, which can assist vehicle manufacturers in advising clients on what motivates them to acquire additional cars from them. After evaluating the findings, we will advise the company on which factors contribute to an increase in car sales.

**Part 1:**

To begin, we must cleanse the data and identify any outliers (if any) to ensure that we have high-quality data for the model. We will develop a linear regression model to estimate Miles per gallon (MPG) based on the numerous features in the dataset using optimization approaches to see if we can improve the model's accuracy and compare both models, as stated in the question. After evaluating the findings, we will advise the company on which factors contribute to an increase in car sales. We've imported all of the packages and libraries we'll need for our initial data exploration. Using the scikit-learn to package, you can create, evaluate, and tune various regression models.

**Data Quality:**

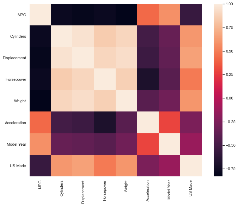
To extract certain information from the dataset, the '?' sign is substituted by "NaN" during data filtering. Since the symbol has been changed with NaN, which stands for a null value, these can now be easily identified and computed to verify the sum of the missing values in the collection. There are 398 records and 8 columns in the cars dataset. To get started, we've imported all of the packages that are required for undertaking model analysis. Pandas, NumPy, matplotlib, seaborn, train test split, Linear Regression, and preprocessing are all examples of Python libraries.

**Data Cleansing and Preprocessing:**

We tested for missing values in the dataset after putting it into a Python environment for further analysis. Missing values must be taken into account because they may have an impact on our analysis and AI models. But why is horsepower an object rather than a float, if the values we saw earlier were all numbers? Let's see if we can convert the column using astype (). The term "outlier" refers to a data point that differs dramatically from the rest of the dataset. Anomaly in the distance between the values, to be precise. This can happen as a result of experimental errors or measurement variability.

Let's look at the different aspects of horsepower to see if there are any differences. When we print out all of the unique horsepower values, we see that there is a '?' placeholder for missing values. Please don't delete these entries. As a result, all entries containing a '?' as a data placeholder have been erased. The horsepower data, on the other hand, is still an object type, not a float. Because of the '?' in pandas, the entire column was coerced as an object when we imported the data set, thus we didn't want to mess with it. The lowest figure is 9 and the highest is 46, but the average is 23.44, with a variation of 7.8. We can look at mpg: using our seaborn tool. Kurtosis of -0.51 and skewness of 0.45.

**Exploratory Data Analysis:**



Not surprisingly the feature describing the engine is strongly correlated to another. Many cylinders equal more displacement equals more horsepower which in turn is appropriate for heavier cars. So far, we've looked at the data to gain a sense of it, and we've seen the spread of the desired variable MPG over the many discrete variables, such as Origin, Year of Manufacturing, or Model, and Cylinders. Let's extract another discrete variable, this time the company name, and add it to the data. To create this new column, we'll use regular expressions and the pandas data-str.extract() frame's function. Now that we've looked at the data distribution along with discrete variables and seen some scatterplots using the seaborn pair plot, let's look at the data distribution along with continuous variables. Let's try to figure out what's causing the changes in mpg. This will aid us in deciphering the patterns in these relationships.

The dataset's property MPG is a parameter that will be forecasted using the GLM algorithm. It is a classification task, in which we must determine and describe which attributes contribute to higher MPG over others. MPG is negatively connected with a powerful engine and a larger vehicle. There is a link between newer cars and faster acceleration. We investigate the distribution of mpg more fully because several regression algorithms rely on a linear relationship between features and target.

**Part 2:**

**Data Modelling:**

A vast class of models is referred to as a generalized linear model. The response variable in this model is considered to have normal distributions, and the arbitrary function of the response variable is supposed to fluctuate linearly with the predictor variables. GLM is a model-fitting software that is commonly used. It was previously known as GLM but is now known as GLM Generalized linear models. Assumptions are This regression's data is dispersed randomly.

The dependent variable is chosen at random, but it is regularly distributed; the errors should be independent, but not distributed, and the values are estimated using MLE (Maximum likelihood estimation) rather than OLS (Ordinary least squares). We chose the GLM model to determine which values assist the manufacturer in increasing vehicle sales. We chose this GLM model to produce response variable predictions (MPG). We simply created the model fit between both variables by using the dependent variable Response and the independent variables.

In this step, we first built a linear regression model to accurately forecast miles per gallon (MPG) based on a vehicle's characteristics. The R-squared value is 0.824, and the Adjusted R-squared value is 0.821. 4 significant attributes were identified which are highly suitable for the MPG predictions. They are Model Year, US Made, Weight, and Displacement.

**Part 3:**

We can figure out which attributes are useful for the model by looking at the p-values and coefficients. The P-value aids in the selection of the most appropriate variables. Regardless of the other selection techniques, the findings Cylinders have a pvalue of 0.05, so they may be utilized for analysis. We have also taken some additional variables to know reliable results.

Backward stepwise selection (also known as backward elimination) is a variable selection approach that includes the following steps: Starts with a full model that includes all variables under consideration (called the Full Model), then removes the least significant variables one by one until the model reaches a pre-determined stopping rule or no variables are left in the model.

A variable that is the least significant is one that: When compared to other predictors, it has the greatest p-value in the model, or its removal from the model causes the smallest drop in R2, or its removal from the model causes the smallest increase in RSS (Residuals Sum of Squares). I have optimized the above model by removing the least significant feature i.e., Acceleration. In terms of achievement, a slight change can be found in the AIC score from 2057 to 2056 and BIC Score from 2089 to 2084. While the R-squared and Adj. R-squared values remain the same. In this Model Year, US Made, Weight, Displacement, and Horsepower are the key attributes in getting a higher MPG over others.

**Findings and Recommendations:**

In both, models, Model Year and the US Made are majorly contributing to higher gas mileage (MPG). So, I would recommend this car manufacturer to focus more on the Models years and manufacture the cars in the USA. We have a model that predicts fuel mileage for a variety of cars; our client can use this to plan for cars that achieve desired levels of fuel efficiency. With this detail, our client can plan future car production or purchase plans.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **R-squared** | **Adj. R-squared** | **AIC** | **BIC** | **F-statistic** |
| Linear Regression | 0.824 | 0.821 | 2057 | 2089 | 257.1 |
| Backward Stepwise Regression | 0.824 | 0.821 | 2056 | 2084 | 300.1 |

**Conclusion:**

When comparing the outcomes of the two models, GLM and optimized models, we can tell the difference between the two. We conducted this study to assist the manufacturer in determining what steps to take to increase sales. The optimized model has an accuracy of 82.4 percent, while GLM has an accuracy of 82.4 percent. As a result, we can conclude that the models are well balanced.

To be honest, the percentage of customers eager to acquire cars from this manufacturer is determined by their preferences and interests. However, based on the foregoing data outputs, we may conclude that MPG has an impact on car sales. Aside from that, several other factors can influence sales. But the major point is that people choose to acquire less expensive cars. As a result, the majority of consumers will purchase fuel-efficient vehicles with better MPG ratings. If he wants to enhance sales, he should look for more options that result in high MPG values, which will lead to more purchases.

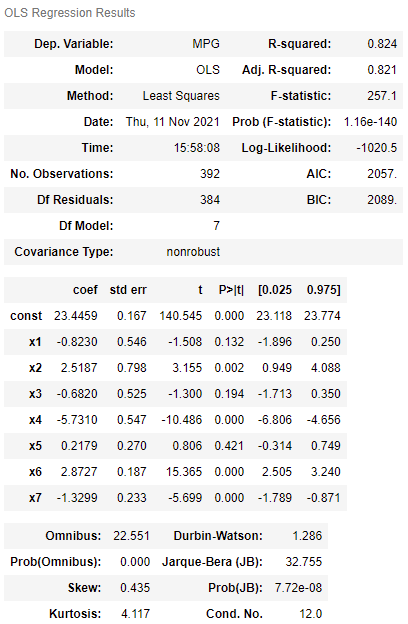
**References:**

Ashutosh, Tripathi. (Jun 10, 2019). Feature Selection Techniques in Regression Model. *towardsdatascience*.

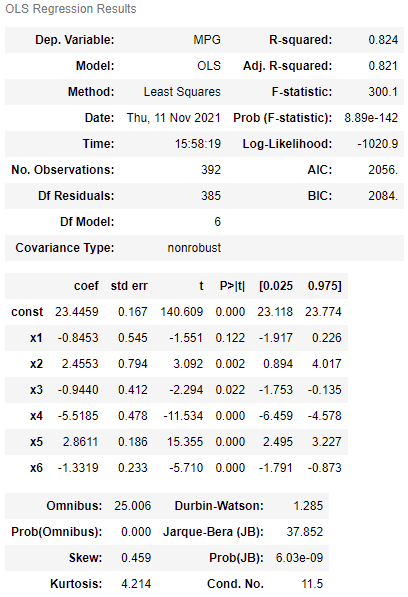
Retrieved from https://towardsdatascience.com/feature-selection-techniques-in-regression-model-26878fe0e24e

**Appendix:**

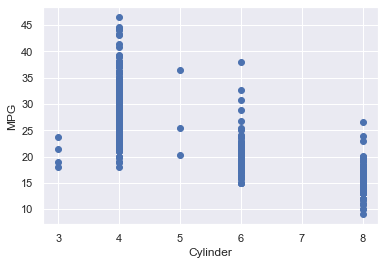
**Figure 1: Linear Regression Results**



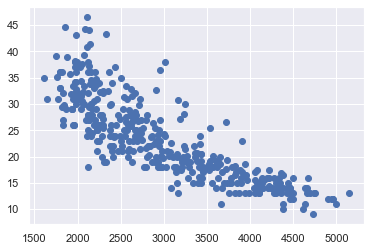
**Figure 2: After removing the 5th feature i.e, Acceleration**



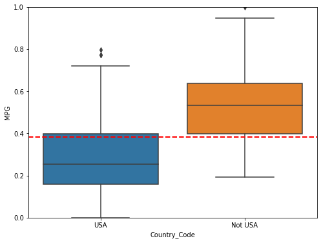
**Figure 3: Cylinder Count by MPG**

****

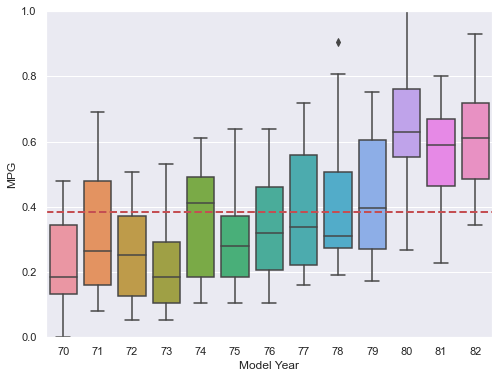
**Figure 4: Weight By MPG**

****

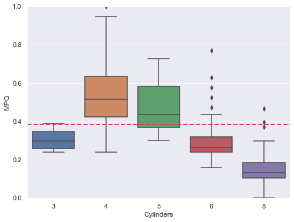
**Figure 5: Country Code vs MPG**

****

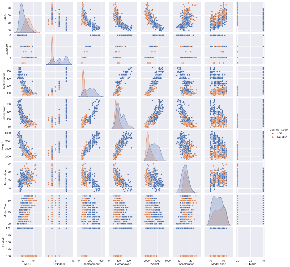
**Figure 6: Model Year vs MPG**

****

**Figure 7: Cylinders vs MPG**

****

**Figure 8: Pairplot**

****