Module 1 Assignment: Building the Car of the Future

Dhairyav Jatin Shah

College of Professional Studies, Northeastern University

Professor: Justin Grosz

January 22, 2023

**Introduction**

This assignment makes use of a dataset with automotive specifications provided by a car manufacturer. The auto maker is having trouble with sales and is trying to figure out how to develop the car with the highest MPG thereby maximizing sales. We will use a linear regression model to predict the dependent variable, MPG, utilizing predictor factors like Model Year, Cylinders, Displacement, Acceleration, etc. The manufacturer can then use these discoveries to create cars that are more fuel-efficient.

**Data cleaning**

There are 398 records in the dataset, along with 8 variables. The data set has 8 variables total, including 1 date type variable that contains the car's make and model year and 7 numerical variables. We discovered anomalies in a column during analysis i.e., **“Horsepower”**. These values are swapped out for null values. In addition, we handle these null values by substituting the mean of the given column. Since the column with the aberrant values is a continuous numerical variable, we have selected mean in this stage because mean is the best match replacement in such a situation.

The "Model Year" column is examined in the following phase, after which "19" is concatenated to it and the YY format is changed to a YYYY format. We would also treat this column as a categorical variable and create dummy variables while fitting the model since we only have the year of the car in which it is manufactured. This would make it clearer how much a car made in a certain year contributes to greater the fuel efficiency.

**EDA**

During the exploratory data analysis process, we learn about our target variable, **“MPG”** and its relationships with the other independent variables. First, we generate a table with descriptive statistical information about the numerical variables in the data frame. From table no 1 (In appendix 1), we can see that the average milage a car provides is 23.5 MPG with a minimum of 9 MPG and a maximum of 46 MPG. The 2nd variable is the number of cylinders present in the engine, an essential component of the engine where the fuel is burned, and power is produced is called a cylinder. A car can have 3 to 8 cylinders in its engine to produce power. The 3rd variable is the displacement, it is the combined swept volume of the pistons inside the cylinders. The average displacement generated in the engine with an average of 5 cylinders is 193.42 m, with a maximum displacement of 455 m. The 4th variable is the horsepower, it is the power produced by the engine. In our data set the engine is producing an average of 104.46 horsepower’s, with a minimum of 46 and maximum of 230 horsepower. Looking at these observations we can understand that if the car manufacturer has more number of cylinders in the engine then eventually a greater displacement would be generated which would increase the burning of fuel generating higher horsepower and decreasing the milage of the car. In the next step we have generated a histogram plot to understand the distribution of the variables in the data set. From fig no 1( in appendix 2), we can see that our target variable is slightly skewed towards the right. Further we can see that there are more than 200 cars having only 4 cylinders which is almost 50% of the total data. We can see that acceleration column is normally distributed, where as weight is skewed slightly towards the right as well as the horsepower and displacement. Since 50% of the cars have only 4 cylinders the displacement in the engine and the horsepower generated for most of the cars would be a less hence the skewness is towards the right. Further we plot a pair plot to study the correlation of the variables on each other. From fig 2 (in appendix 2) we can see that our target variable has a negative slope with features like “Cylinder”, “Displacement”, “Horsepower”, and “Weight” this shows that an increase in any of these variables would cause a decrease in the target variable “MPG” as we discussed earlier. Whereas columns like “Acceleration” and “Model Year” have a positive slope with the target variable “MPG”. Further to get a better understanding of correlation and multicollinearity we plot a heatmap. From fig 3 (in appendix 2), we can confirm that columns “Cylinder”, “Displacement”, “Horsepower”, and “Weight” have a negative correlation with our target variable “MPG”. Also, we can see a very strong multicollinearity among the columns “Cylinder”, “Displacement”, “Horsepower”, and “Weight”. To confirm this we calculate VIF, where if the VIF value is more than 5 then it states that the column is correlated to a independent variable and it needs to be dropped before modelling. From table no 2.1 (in appendix 1) we can see that, columns “Cylinder”, “Displacement”, “Horsepower” has the VIF values greater than 10 and they need to be dropped. However, the column “Horsepower” has a VIF value greater than 5 but after dropping the other 3 columns “Horsepower” doesn’t have a VIF value greater than 5 as shown in table no 2.2 ( in appendix 1) hence we would no longer need to drop this column.

**Analysis**

**Modeling**

Once the initial analysis is complete, we start preparing the data set for modeling. In the first step we will generate dummy variables for the column “Model Year” and drop the original column from the data set. Further we will drop the columns having multicollinearity and create a copy data frame X1. Next, we will split the data set into train and test data set along with the X and Y variables. Our split ratio would be 70% and 30% since our data set consists of just 398 records. Now we fit the first model with the selected features. From table no 3 (in appendix 1), we can see that we obtained a R2 of 80.8 and AIC = 1494. Now we start with backwards feature selection and start eliminating features with a p-value higher than 0.05% as our confidence interval is 95%. From the 1st Model we eliminate “MY\_1972”, “MY\_1973”, “MY\_1974”, “MY\_1975”, “MY\_1976” since these columns are not so significant and p value is very high. We now fit the 2nd model with the remaining variables. From table no 4 (in appendix 1) , we can see that “MY\_1971” needs to be eliminated. We now fir the 3rd model, from table no 5 (in appendix 1) we can see that “MY\_1970” also is insignificant hence we will have to eliminate the column. We now fit the 4th model, with the final selection of variables. From table no 6 (in appendix 1) we can see that our R2 has decreased to 80.3 and AIC has decreased 1489. The final equation we obtained is given below.

Further we will use this fitted model on our test data frame. From fig 4 (in appendix 2) we have plotted the actual values of MPG in our test data set and the predicted values of MPG we obtained after running the model against the test data set. Here we can see that 80% of the predicted values are plotted close to actual values of MPG.

**Conclusion**

Based on our analysis and modeling, we have created a linear regression model with a R2 value of 80.3% which is not bad but due to multicollinearity we had to drop several features which would play an important role in predicting the milage of the car. From the model that we have created we can say that if the horsepower and acceleration of the car increases that would bring a drop in the milage. Similarly, we can also see that cars which are US made have lower milage compared to the cars manufactured out of the United States. Also we can see that as the manufacturing year of the car increases the coefficient of that variable also increases, hence the cars manufactured after the year 1980 would give the best milage compared to earlier years. We can say that the overall model is good and is able to predict the milage, but we could use other regularization methods to check and see if it improves the accuracy of the model.

**Recommendation**

Using our research and modeling, we can give the following recommendations to the manufacturer to increase the fuel efficiency of the car and maximize the sales.

1. Usage of more number of cylinders could lead to a increase in the displacement generated and increased in horsepower, which would decrease the efficiency.
2. 4 cylinders would be an ideal number to generate decent displacement in the engine and produce power keeping efficiency in mind.
3. While designing the car, manufacturers can refer the models manufactured 1980’s to get better fuel efficiency.
4. Designers can refer to cars made outside the United States as the cars made in the United States have a negative coefficient in the regression model.

**Reference**

1. Module 2 Canvas (n.d.) *Lesson 2-5 — Improving Regression Model Results* <https://northeastern.instructure.com/courses/131156/pages/lesson-2-5-improving-regression-model-results?module_item_id=8228128>
2. Data to fish (July 2022) *Linear Regression in Python using Statsmodels* <https://datatofish.com/statsmodels-linear-regression/>

**Appendix 1**

**Table no 1 Descriptive Statistics of the Data set**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **count** | **mean** | **min** | **max** |
| **MPG** | 398.0 | 23.51 | 9.0 | 46.6 |
| **Cylinders** | 398.0 | 5.45 | 3.0 | 8.0 |
| **Displacement** | 398.0 | 193.42 | 68.0 | 455.0 |
| **Horsepower** | 398.0 | 104.46 | 46.0 | 230.0 |
| **Weight** | 398.0 | 2970.42 | 1613.0 | 5140.0 |
| **Acceleration** | 398.0 | 15.56 | 8.0 | 24.8 |
| **Model Year** | 398.0 | 1976.01 | 1970.0 | 1982.0 |
| **US Made** | 398.0 | 0.62 | 0.0 | 1.0 |

**Table no 2 VIF Values to check Multicollinearity.**

|  |  |  |
| --- | --- | --- |
|  | **Table 2.1** | **Table 2.2** |
|  | **VIF** | **VIF** |
| Const | 152.78 | 116.95 |
| **Cylinders** | **10.67** | **Dropped** |
| **Displacement** | **22.61** | **Dropped** |
| **Weight** | **10.15** | **Dropped** |
| Acceleration | 2.49 | 1.91 |
| US Made | 1.90 | 1.33 |
| Horsepower | 8.71 | 2.34 |

**Table no 3 OLS Regression Results Model 1**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Table no 4 OLS Regression Results Model 2**

Table

Description automatically generated

**Table no 5 OLS Regression Results Model 3**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Table no 6 OLS Regression Results Model 4**

**A screenshot of a computer

Description automatically generated with medium confidence**

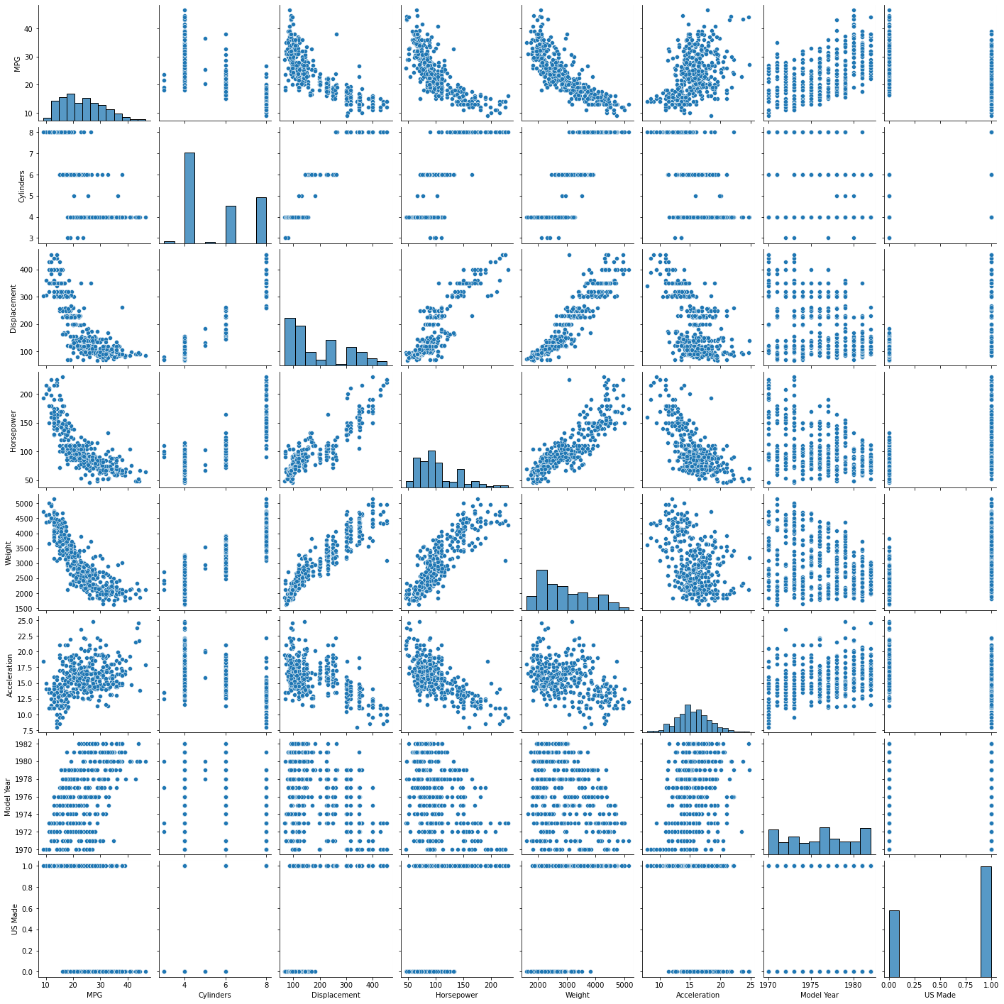
**Appendix 2**

**Fig No 1 Histogram**

**A picture containing shoji, crossword puzzle

Description automatically generated**

**Fig no 2 pairplot**



**Fig no 3 Heatmap**

**Graphical user interface

Description automatically generated with medium confidence**

**Fig no 4 Actual MPG vs Predicted MPG Scatter Plot**

**Chart, scatter chart

Description automatically generated**