Logo

Description automatically generated

ALY6020 Module 2 Midweek Project

**Authored by: Prateek Singh**

Course: ALY6020 – Predictive Analytics

Submission Date: 04/26/2022

**Instructor’s name: Kristen Drobnis**

# Introduction

For this week project we have a car feature dataset which will be used to fit into a regression model. Objective of the exercise is to predict the features which affects the mileage of the car and hence can contribute to the higher mileage and manufacturing a more fuel cost-effective vehicle. We will be creating several graphs for fitting the regression model and find the distribution of several features data points. We will finally depict the results of mean squared error to predict the model accuracy and later present the best fit model with significant factors or features which enhances the miles per gallon of the automobile.

In this model we will be using multiple or multivariate linear regression technique. Our target or the independent variable is “MPG”, and we will determine our most significant independent variables during the modeling stage. This will help us interpreting the current dataset but also enable us to predict the unknown datapoints of future.

# Analysis

The dataset has 398 observations and total of 8 columns, out of which 7 are of numeric type and 1 is object type. However, the object variable is Horsepower, and it is because it contains ‘?’ values inside it. This also means we have some missing data and hence data cleansing is required.

Text

Description automatically generated

Fig 1: Dataset Summary

There are total of 6 ‘?’ values present only in Horsepower variable.

Text

Description automatically generated

Fig 2: Number of ‘?’ values

We then replaced it with the ‘NaN’ values before moving forward to replace it with the median values.

Text

Description automatically generated

Fig 3: Number of Null values

Before we can decide which imputation technique can be used to replace the ‘NaN’ values from the Horsepower variables, we first need to check the data distribution of it.

We are going to plot a boxplot first to see how data is distributed in terms of minimum, maximum, quantiles, and median values. We cannot see any outliers which can affect our mean values of the variable.

Chart, box and whisker chart

Description automatically generated

Fig 4: Boxplot of Horsepower to understand datapoint distribution

Further we want to see the density distribution of the variable to understand if there is any skewness. We found that the variable distribution is right-skewed.

Chart, histogram

Description automatically generated

Fig 5: Density Distribution of Horsepower

Hence, we can take the decision to replace the ‘NaN’ values in the variable with either with median or mode. In this case I decided to replace it with median because there were not much ‘NaN’ values present and I wanted to replace it with mid value, so it doesn’t affect the population much. Hereafter I checked if there are any more missing values present in the dataset to deal with or not.

Text

Description automatically generated

Fig 6: No Missing values post replacing

However, the variable is still having object datatype and hence we went ahead to change the datatype to numeric and got the description of the final dataset to be used for the analysis and creation of Data and Target set. We can see the average miles per gallon value is 24 for 5 cylinder and 104 horsepower cars. On an average a car displaces for 193 units to achieve an average MPG and must be 2970 pounds. Maximum cars are US make and generally from the year 76.

Table

Description automatically generated

Fig 7: Statistical Analysis of the dataset

It is now time for understanding the correlation between the variables, and because we know that MPG is our target variable, we will see all the variables against it. We found that Model Year is strongly positively related to MPG and weight is strongly negatively related to MPG. Least related variable is acceleration. The second most significant variable is displacement however I believe it will have dependency on weight which is also our independent and most significant variable. Hence, we might have multicollinearity which we will check in the later stage of our analysis.

Table

Description automatically generated

Fig 8: Correlation Table between variables

To visually understand the correlation between the target and independent variables, we plotted a pair-plot. We can see that only Acceleration and Model Year are positively related and all other variables are negatively impacting the MPG. Displacement and Weight as discussed earlier have the most negative influence on MPG with Cylinders and Horsepower not so behind. Below graph presents how each variable is dependent on each other however we will focus on the first row which shows MPG on y-axis as a target and rest variables on x-axis as independent variables which affects the dependent variable.

A picture containing shoji

Description automatically generated

Fig 9: Pair plot or correlation plot

We now want to check the multicollinearity in the dataset and found that VIF values are very high Cylinders, Displacement, Horsepower and Weight. We induced earlier that displacement and weight are correlated to each other, however theoretically when the number of cylinders increases the weight of the car will increase and it will somewhere affect the horsepower too. To find that we generated the VIF score and got the below figures.

Text

Description automatically generated

Fig 10: VIF Score

However, we cannot just drop the variables mere on the fact of high VIF score. We hence will fit the linear regression model to see the p-value and coefficient constant to understand the most significant variables. Before fitting the model, we divided the dataset into Independent and Target dataset and created train and test set for both.

We can see that only Weight, Model Year and US Made are the most significant variables and others not because they have high p-value of more than 0.05. We can now deduce based on VIF and regression model results, variables which can be included in our regression model can be Weight, Acceleration, Model Year and US Made. For now the accuracy of our model is coming out to be 82.4% with AIC and BIC value as 1653 and 1683 respectively.

Table

Description automatically generated

Fig 11: Regression Model with all the Independent Variables

However, when I used the Forward Selection Technique to see the most significant variables for our model I got only Weight, Model Year and US Made variables in the result. Hence, I decided to keep these variables only for my final best fit model.



Fig 12: Forward Selection Technique Result

After dropping all the insignificant variables, my model accuracy comes out to be 82.3% which is almost like previous model however with no insignificant variables and improved AIC and BIC values as 1650 and 1665 respectively.

Table

Description automatically generated

Fig 13: Regression Model with the significant Independent Variables

# Conclusion

1. The best fit model has three independent variables as Weight, Model Year and US Made.
2. The best fit model accuracy comes out to be 82.3%.
3. AIC and BIC values are improved and comes out to be 1650 and 1665 respectively.
4. Weight has a negative influence on MPG. With every 1 unit change in weight the MPG values goes down by 0.005 units.
5. US Made cars have lower MPG when compared with non-US brands. This should be investigated further as to what is causing that and how cars manufactured out of US different from the cars made in US. Material used, Weather-Conditions, Mechanical Technology, etc. can be few of the variables factoring such result. Cars manufactured outside of US are different on MPG by almost 2.3 times than the US manufactured automobiles.
6. Model Year has the positive influence on MPG which makes perfect sense and the technology have improved over the years and many mechanical techniques have changed over the time. For every consecutive year the MPG is getting improved by 0.77 times than the previous year.

# Reference

1. Grosz, J. (n.d.). Lesson 2-3— How and When to Use Linear Regression. Canvas. Retrieved April 22, 2022, from <https://canvas.northeastern.edu/>
2. Moffitt, C. (2017, February 6). Guide to encoding categorical values in python. Practical Business Python Atom. Retrieved April 22, 2022, from <https://pbpython.com/categorical-encoding.html>
3. Team, D. C. (2020, September 3). Python Pandas Select Columns tutorial. DataCamp Community. Retrieved April 22, 2022, from <https://www.datacamp.com/community/tutorials/python-select-columns>
4. Zach. (2021, November 11). How to fix: Pandas data cast to Numpy dtype of object. check input data with np.asarray(data). Statology. Retrieved April 22, 2022, from <https://www.statology.org/pandas-data-cast-to-numpy-dtype-of-object-check-input-data-with-np-asarraydata/>
5. Convert character column to numeric in pandas python (string to integer). DataScience Made Simple. (2020, July 26). Retrieved April 26, 2022, from <https://www.datasciencemadesimple.com/convert-character-to-numeric-pandas-python-string-to-integer-2/>
6. Cosine. (2020, August 29). Detecting multicollinearity with VIF - python. GeeksforGeeks. Retrieved April 26, 2022, from <https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/>
7. Das, P. (2019, May 7). Linear regression in python with large dataset example. CodeSpeedy. Retrieved April 26, 2022, from <https://www.codespeedy.com/fitting-dataset-into-linear-regression-python-model/>
8. Gandiban, K. (2020, January 7). MODULENOTFOUNDERROR: No module named 'plotly’. GitHub. Retrieved April 26, 2022, from <https://github.com/plotly/plotly.py/issues/2039>
9. Kumar, A. (2021, October 3). Python - replace missing values with mean, median & mode. Data Analytics. Retrieved April 26, 2022, from <https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/>
10. kumar, A. V. (2019, February 11). Stepwise-regression/step\_reg.py at master · AakkashVijayakumar/stepwise-regression. GitHub. Retrieved April 26, 2022, from <https://github.com/AakkashVijayakumar/stepwise-regression/blob/master/stepwise_regression/step_reg.py>
11. McCullum, N. (n.d.). Linear regression in python - a step-by-step guide. Nick McCullum Headshot. Retrieved April 26, 2022, from <https://nickmccullum.com/python-machine-learning/linear-regression-python/>
12. Pedamkar, P. (2021, July 10). Introduction to Pandas Dataframe.iloc[]. EDUCBA. Retrieved April 26, 2022, from <https://www.educba.com/pandas-dataframe-iloc/>
13. Seaborn Distplot. Seaborn Distplot - Python Tutorial. (n.d.). Retrieved April 26, 2022, from <https://pythonbasics.org/seaborn-distplot/>
14. Toth, G. (2020, November 23). Feature selection methods with python. DataSklr. Retrieved April 26, 2022, from <https://www.datasklr.com/ols-least-squares-regression/variable-selection>

# Appendix

Note: Code is attached separately.