Logo

Description automatically generated

ALY6020 Module 2 Midweek Project

**Authored by: Prateek Singh**

Course: ALY6020 – Predictive Analytics

Submission Date: 04/22/2022

**Instructor’s name: Kristen Drobnis**

# Introduction

For this midweek assignment we were supposed to analyze the Car Price dataset to model the multi-linear regression algorithm and predict the prices of the car. We were given the dataset which had many car features listed in it. Initially there were total of 205 observations and 26 variables out of which 8 were numeric, 18 were categorical (10 object and 8 float). There are no missing values in the dataset to be treated, hence the data can be considered as clean and ready to be used for modelling. However, later in our analysis we determined that some variables were not contributing much towards the objective of the model hence were dropped. Also, we converted the categorical variables level with the continuous label values.

# Analysis

We understood that with all the cars in our dataset, the average safety rating comes out to be 0.8 with maximum as 3 and minimum as -2. Higher the rating means less safe the car is. The average horsepower and peak rpm of the cars listed are 104 and 5125 respectively. The average car price of our dataset is 13277 going as low as 5118 to all the way up to 45400. The average engine size is 127. Maximum of our cars are of gas fuel type and standard aspiration. The most common car in our dataset is sedan majorly having forward drive wheel, front engine location, OHC engine type and four-cylinder cylinder types.

We need to fit the linear regression model; however, we have many categorical variables in our dataset. We need to replace them with either the dummy variables or label all the levels. I chose the latter because I do not want to populate the dataset with multiple columns, which we need to eventually drop as they might not be influential for our analysis and model. I provided continuous numerical values to every level of the categorical variables starting with zero. (*Refer “Fig. 1: Labelling for the Categorical Variables” from appendix*)

We then split our dataset into test set data and train set data for both of our Target (dependent) and Independent variables. Our Target variable is “Price” as stated in the problem statement and we want to determine what are the major factors which influence the price of the cars. With the help of the multi-linear regression model, we will determine the best model and the important factors.

I got the accuracy of 88.8% when I ran the model with all the factor variables, however there are many such variables whose p-value is way higher than 0.05 value and hence they are rendered ineffective factors. (*Refer “Fig 2: Regression Model with all the Factor Variables” from appendix*)

We then tried deleting those variables and ran the model again. Not as per our expectation, the model accuracy went down almost by 5% and came out to be 84.0% this time. Interestingly we only got two inefficient variables in our model whose value was more than p-value of 0.05. (*Refer “Fig 3: Final Regression Model with best fit Factor Variables” from appendix*)

Hence, I believe we can keep this model with lower accuracy but maximum effective factor variables.

# Conclusion

In the final model of accuracy 84% we had variables like carbody, enginelocation, cylindernumber, enginesize, stroke and peakrpm. Out of these variables we found out that carbody and enginelocation are no longer serving the purposes of our model and analysis. Apart from that although rest of the variables are rendered effective, the most significant variables for our model were cylindernumber, enginesize and stroke.

The effectiveness of cylindernumber is 1066, that of enginesize is 172 and for stroke it comes negative -2602. This means as the cylinder number increases the price increases by 1066, with increase in engine size the price increase by 172 and at last the price takes a dip of 2602 with an increase in stroke. Clearly number of cylinders has the greatest positive effect over the process of car.

# 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&nbsp;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/>

# Appendix

A screenshot of a computer

Description automatically generated with medium confidence

Fig. 1: Labelling for the Categorical Variables

Table

Description automatically generated

Fig 2: Regression Model with all the Factor Variables

Table

Description automatically generated

Fig 3: Final Regression Model with best fit Factor Variables

Note: Code is attached separately.