Project 1: Exploration & Linear Regression

Pranjal Adhikari

Introduction

In this project, the main objective was to explore an automobiles data set and utilize various different Python libraries, such as *pandas*, *seaborn*, and *scikit-learn* to run data analysis. Furthermore, the data was prepared to fit a linear regression model to predict the fuel efficiency of vehicles. The data set itself contains various different attributes of vehicles, with most being utilized in this project. The *pandas* library was first used to process the data and run exploratory data analysis. Next, *scikit-learn* was used to train a linear regression model to predict the fuel efficiency given the different attributes of the vehicles. Lastly, this model's accuracy and performance was determined in predicting the fuel efficiency of vehicles.

Data Preparation

The data set used in this project (found here) contains 9 attributes of 398 different vehicles, including:

1. name - name of the car
2. origin - origin of the car (1: America, 2: Europe, 3: Japan)
3. model_year - year of the model
4. acceleration - time (in seconds) to accelerate from 0-60 mph
5. cylinders - number of cylinders in the engine
6. horsepower - engine horsepower
7. displacement - engine displacement (in inches)
8. weight - vehicle weight (in pounds)
9. mpg - fuel efficiency measured in miles per gallon (mpg)

Before the model could be trained on the data set, the data set had to be prepared for optimization. First, the data type of each column was determined to verify the values within each column were set to the appropriate data type. In this data set, the horsepower attribute was set to data type object. Using *pandas*, it was converted to data type float. Next, vehicles missing values for any attributes were dropped from the data set, as it could not be utilized in creating the model. For the horsepower attribute, 6 vehicles were missing values, thus were dropped from the data set. Additionally, the name attribute was dropped from the data set as it would not be utilized in training the model for fuel efficiency. The origin attribute, a categorical data type, had three total unique values,which could be transferred to create two new columns utilizing the one-hot encoding technique. After running the preparations all attributes, except name, and two new attributes titled origin_2 and origin_3 (from one-hot encoding) remained for further use.

After preparing the data, visualization and analysis was performed using *seaborn* to determine any pre-existing relationships between the attributes. It could be seen that the cylinders, displacement, horsepower, and weight attributes all had relatively close negative correlations with fuel efficiency in the data set. This gives the insight that these attributes have a considerable impact on the fuel efficiency of vehicles.

Training the Model

First, two new variables were created to split the attributes. The attribute mpg was saved as the y variable (dependent variable) and other attributes saved as the X variable (independent variable). This was done so the model could be trained to predict the fuel efficiency using the vehicle attributes. Next, the data contained within the two variables were split into training and testing sets using *scikit-learn*, with 30% allocated to the testing set and 70% allocated to the training set. After the allocation, the training data was fit to a linear regression model.

Conclusion

To evaluate the model's performance in predicting the fuel efficiency, *scikit-learn* was utilized to calculate the coefficient of determination, R^2, value for the training and testing data sets. This coefficient value is the measure of the variation in the dependent variable which is predicted by utilizing the independent variable. The R^2 value ranges from 0 to 1, with closer to 1 meaning the model is better at fitting the data set. In this case, the R^2 value for the training set was 0.8205, and the value for the testing data set was 0.824. With these values, we can conclude that the model performed well. Furthermore, it can be concluded that I have high confidence in this model predicting the fuel efficiency of vehicles overall.

References

Class website and materials: https://coe-379l-sp24.readthedocs.io/en/latest/index.html