Project 1

COE 379L

Joe Stubbs

Braulio Lopez

In project one, we were tasked to load a data set and work around it by using data analysis skills to understand the behavior of the data and the information it's giving. It was also important to perform machine learning techniques to predict the behavior of said dataset. For this project, it was necessary to work on the automobile dataset available in the class repository called “project_1”. This data set provides useful information regarding automobiles, the columns included are: miles per gallon (mpg), cylinders, displacement (in cubic inches), horsepower, weight (lbs), acceleration (s), model year, origin (1: American, 2: European, 3: Japanese) and car name.

The first step was to load the data using the pandas library in Python. To prepare the data it was necessary to identify the shape and size of the raw data to check the number of columns and rows in the data set. It was also important to get information about the data using the build function “info()” which gets information about the type of each column in the data frame along with the instances they appear. When looking at the information provided by the function it was clear that all the columns were in the correct type of data except for the horsepower column. This column was an object and it was changed to a float, this was accomplished by using the pandas' “to_numeric” function that changes objects and strings to a numeric values, it is important to mention that Chatgpt helped me to debug this implementation as it added the “error=coerse”

which ensures that it does not fail if there are invalid characters. After successfully converting the type to float, I checked the information in the data set and it was evident that there were missing values in the data set, as a result, I replaced these values with the mean to ensure we make accurate machine learning outcomes. Some insights I gained from the data preparation process is that it was important to detect these types of inconsistencies in as it will make inaccurate predictions on the model you are trying to achieve.

To train the model, I split the data into training and testing sets by using the train_test_split() function which gives the X_train, X_test, y_train, and y_test and the value you want to use as your independent variable and the dependent variable. It is important to note that I use a test size of 0.3 meaning that the training data is utilizing 70 % of the data to train and it's going to test it using 30%. I used a linear regression model to fit the training data set and we predict the linear regression using the testing dependent variable.

Based on the Training and testing accuracies it is safe to say that the model performs wells as it shows that the Training Accuracy is 81.4% and the Test Accuracy is 84.3%. This means that the model generalizes well to unseen data. In terms of confidence, I am confident in terms of what the data provided but it is important to acknowledge that other factors can affect the accuracy of the model, for example, if the data was not taken care of correctly at the beginning, meaning that if the data set is clean and unbiased then we can say that it is accurate from my project I am confident that the data was taken care correctly to have an accurate model.