Project 1: Fuel Efficiency Prediction Using Linear Regression

This project aims to use a linear regression machine learning model trained on a dataset that contains information about different cars to predict expected fuel efficiency given some information about the car.

Before the machine learning model could be implemented, the dataset needed to be preprocessed. The original dataset consisted of 398 cars, containing data on fuel efficiency (in miles per gallon), number of cylinders in the engine, engine air displacement, engine horsepower, weight, acceleration, the year the model was produced, the "origin" of the car, and the car's name. While most of these columns held datatypes that made sense for their respective names, the horsepower was listed as "object" data instead of numeric values. Upon closer inspection, this was because it contained six rows with "?" as the value. These values were temporarily replaced with NaNs to allow the horsepower to be converted to floats. To obtain a better approximation for the NaN values than the average of the entire dataset, the vehicle data was split by the number of cylinders in each car and separate horsepower averages were calculated for each group. This was performed under the assumption that horsepower and number of cylinders would be correlated. These separate averages were then used to fill in the NaN values according to the cylinder group the vehicle belonged to. The "origin" data of the car was also a bit strange, being an integer valued column that only held 1, 2, or 3. This format seemed to better lend itself to a categorical data type, so the origin data was converted to this via one-hot encoding.

Upon analysis of the preprocessed dataset, I could see a few trends. The univariate plots of the data showed that most of the numeric columns held either a uniform (model year), skewed (displacement, horsepower, and weight), or normal distribution (fuel efficiency and acceleration).

The number of cylinders data was discrete and more proportionally separated compared to the smoother continuity of the other data, with very few cars holding 3 or 5 cylinders compared to 4, 6, or 8. Additionally, some of the car names were held by more than one car, but each of these data rows represented different years/designs, so there was no duplicated data. The bivariate correlation heat map showed that fuel efficiency was strongly negatively correlated with number of cylinders, engine displacement, horsepower, and weight, while only being moderately correlated with acceleration and model year. The plot also shows that horsepower is strongly correlated with the number of cylinders, helping to validate the reasoning behind the horsepower data filling during preprocessing.

The linear regression model for predicting fuel efficiency was trained using the chosen features of number of cylinders, engine displacement, horsepower, weight, acceleration, and model year. The car names were dropped because they could not be assigned appropriate numeric values and attempting to classify the column using one-hot encoding would result in a very high-dimensional model, which would encourage overfitting. The origin column was also dropped because its meaning was vague and undocumented; even though including it increased the model's predictive accuracy for this dataset, it would likely perform worse on new/external car data that wouldn't be able to fill this column. The data was shuffled and split into training and test datasets in a reproducible manner, with the training dataset consisting of 70% of the original dataset and the test dataset making up the other 30%. Over multiple shuffles of the dataset, the model obtained an average R2 score of around 0.8 on both the training data and the test data. Because of these consistent, similar high scores, I have high confidence in this model's ability to predict fuel efficiency for new cars.