Project 1 write up

Mihiro Suzuki

What did you do to prepare the data?

First I took out rows that were missing data, specifically the ones with ? in the horsepower column. Then, I made sure the data types were consistent with what it actually was. For example, horsepower was an object so I switched it to a numeric value. I also used one-hot encoding to change the origin column to be 3 columns of true false. This way I could run analysis with it. Finally, I dropped the string column (car names) in order to run numeric analysis on them.

What insights did you get from your data preparation?

The pairplot and correlation matrix gave a lot of insights. Notably

MPG and weight: There is a very strong negative correlation of -0.831741, suggesting that as the weight of a vehicle increases, its fuel efficiency (MPG) tends to decrease significantly.

MPG and displacement: Similarly, there is a strong negative correlation of -0.804203 between MPG and displacement, indicating that vehicles with larger engines tend to have lower fuel efficiency.

MPG and cylinders: The negative correlation of -0.775396 shows that cars with more cylinders tend to have lower MPG, which is consistent with the understanding that more cylinders generally mean larger engines and hence lower fuel efficiency.

There were also a lot of other insights. This highlighted that the given variables could be good predictors for the model.

What procedure did you use to train the model?

I cleaned up the data to be numeric values and took out the mpg as the independent variable (y) and the stored the rest as the dependent variable (X as a vector). Then I split the data into training and testing datasets using the built in function. Using the training dataset, I trained it using the model fit built in function.

How does the model perform to predict the fuel efficiency?

The model has an accuracy r^2 value of 0.8238159433262657 on the testing data and is around the same as the training r^2 of 0.8205282655302875. Which means that the model is performing at its expected accuracy on data outside of its training set as well.

How confident are you in the model?

I am pretty confident in the model since r^2 of 0.83 indicates that the predictors are pretty accurate but is not an absolute indicator of its fuel efficiency. It's definitely better than anything an average human can predict.

Use of ChatGPT

I used chatgpt in the section where I graphed insights for the data. It was useful to use the GPT's intuition to figure out what variables were good to graph against eachother. I also used it to give me an idea on how to one-hot encode since I was stuck on how to approach it. One error I couldn't figure out even with it was "/usr/local/lib/python3.11/site-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead. with pd.option_context('mode.use_inf_as_na', True):" in the graphing section. Although it didn't seem to have an effect on the data exploration.

[ ]: