COE 379L: Project 1

Ashton Cole
AVC687

February 19, 2024

1 Introduction

In this project, we analyze a data set of cars from the 70's and early 80's. Our core question is *What impacts the fuel economy of automobiles?* We also seek to model it with machine learning. First the data is cleaned, before being analyzed and used to train a simple linear regression machine learning model. The model is found to represent the data fairly well, but imperfectly.

This assignment demonstrates several important machine learning principles. The data is carefully inspected, cleaned, and analyzed thoroughly before a model is even built. The model is then evaluated on independent data to determine its worthiness. The model and prior analysis combined inform the conclusions.

2 Methodology

Analysis is conducted using a Python 3 kernel in two Jupyter Notebook files. The first file cleans and inspects the data. The second trains and tests the machine learning model.

First, the data set is loaded into a Pandas data frame. Then, its contents are inspected and cleaned through Pandas functions. Horsepower is converted from an object to an integer. A few records with unfilled entries are removed, since these 6 out of 392 will. The fields of car name and origin are removed: the name, because there are too many categories to encode with the one-hot method, and the origin, because its meaning is unclear. Year is a factor that could also be removed, because strictly speaking, the time when a given automobile is made does not determine its efficiency. However, the field is ultimately kept, as it could represent improvements in engine technology. A final important note is that none of the values are provided with units, except for mileage, inferred fromt the field name mpg.

Next, the data is analyzed without machine learning. Pandas has built-in functions to calcualte statistics, and Matplotlib is used to generate plots. The distributions of each variable are found to be reasonable, with few statistical outliers. Then, the correlation matrix in Tabel 1 is generated, indicating that mileage is most impacted by displacement, cylinders, weight, and horsepower. These correlate in a negative fashion. All of the correlations are considered to be reasonable, although interestingly, acceleration correlates somewhat positively with mileage. None of them are particularly strong, with the highest-magnitude correlation only being 0.83. This could reflect that a complex combination of variables determines mileage, or hints that there may be other underlying variables. Some scatter plots are generated to confirm the correlatons. Some, especially the highly correlated ones, are noted as having curved, nonlinear trends, like weight in Figure 1. This suggests that a liner model may not be the best fit to the data.

After this, the clean data set is used to train a machine learning model. It is split into two subsets: 70% of records are used to train the model, and 30% are used as test cases. This allows the model to be validated against some cases that were not used to inform it. The split is pseudo-random, using a fixed seed to ensure reproducibility. The training data are fed into a Scikitlearn linear regression model, part of the `scikitlearn.linear_model` submodule.

Table 1: Correlations between each of the variables of analysis.

	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year
mpg	1.000000	-0.777618	-0.805127	-0.778427	-0.832244	0.423329	0.580541
cylinders	-0.777618	1.000000	0.950823	0.842983	0.897527	-0.504683	-0.345647
displacement	-0.805127	0.950823	1.000000	0.897257	0.932994	-0.543800	-0.369855
horsepower	-0.778427	0.842983	0.897257	1.000000	0.864538	-0.689196	-0.416361
weight	-0.832244	0.897527	0.932994	0.864538	1.000000	-0.416839	-0.309120
acceleration	0.423329	-0.504683	-0.543800	-0.689196	-0.416839	1.000000	0.290316
model_year	0.580541	-0.345647	-0.369855	-0.416361	-0.309120	0.290316	1.000000

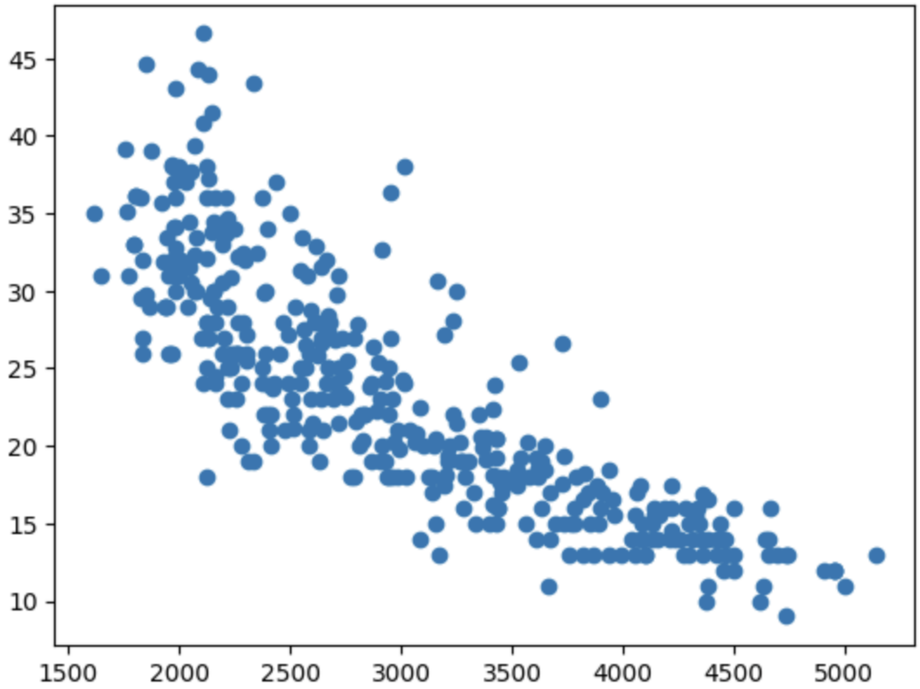

Figure 1: A scatterplot of mileage as a function of weight.

3 Results

After training, the model is tested on both the training and testing data. When tested on the testing data, the model shows a Mean Absolute Error (MAE) of about 2.71. This is the average magnitude of error for a given point. For the original training data, the MAE is about 2.61. These are roughly a third of the standard deviation of the mileages.

A scatter plot of weight, this time using the testing data, is shown again in Figure 2. It can be seen far more strikingly that many of the predicted points do not align with the truth at all. This applies especially for points with more outlying mileages, like the high and low points in the 2000-3000 acceleration region. One prediction, for the point around (2000, 45), is off by at least 10 mpg!

4 Conclusions

From the variables included in the analysis, it seems that weight, displacement, and horsepower are the best predictors of a car's mileage, all being negatively correlated.

Looking at the model results, it is clear that the linear model is a moderately capable predictor of efficiency. It generally captures whether cars should have a higher or lower mileage, but the predictions

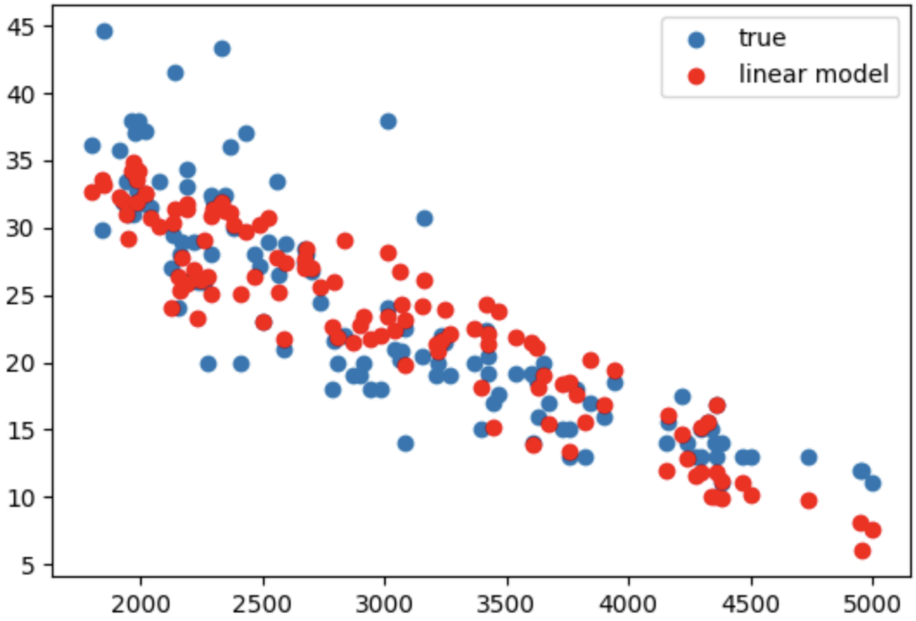

Figure 2: A scatterplot of mileage as a function of weight. The linear model is tested on data which was not used to train it.

are not very accurate, at times having severe discrepancies. Our confidence in the model is partial. The similarity of the MAE between the test and training data at least shows that the model is not overly biased towards the training data, and an average error smaller than the standard deviation lends some credibility. However, the model seems to be more a good guesser of “efficient” or “not” than a predictor of mileage.

Using a nonlinear model and including additional variables in the analysis would be two good ways to improve the model. Several of the scatterplots, like displacement, horsepower, and weight, show curved, nonlinear correlations. The amount of spread in the plots hints that there may be other underlying factors determining the mileage of a car. Since mileage is not a random phenomenon, there are likely some other determining factors not included. For example, surface area, front cross-sectional profile area, and the drag force they help dictate are notable determinants which do not strictly correlate with the variables of analysis.