Project 1

Nikhil Sharma

Introduction

The goal of Project 1 was to introduce the class to exploratory data analysis and pre-processing, as well as utilizing various Python libraries. The data was uploaded to a dataframe in pandas which allowed for visual and statistical analysis. All relationships between variables were visualized using matplotlib and seaborn. The dataset was then fitted to a linear regression model using sklearn, which yielded a model equation incorporating each of the eight features from the dataset. Lastly, the R^2 values of the training and test data were calculated to determine the goodness of fit in the linear regression model.

Data Preparation

As mentioned previously, the data was uploaded to a pandas dataframe, where it could be parsed into columns and analyzed. The data contained 398 rows, or cars, and 9 columns—miles per gallon (or fuel efficiency), the number of cylinders, the displacement, the horsepower, the weight, the acceleration, the model year, the origin, and the car name. Initially, the data contained 3 columns of type float64, 4 columns of type int64, and 2 columns of type object.

The only column that should have been type object was the name of the car, but the data type horsepower was also. Upon further inspection, there were 6 null entries holding a value of ?, which were eventually substituted for the statistical mean of the horsepower values. While those cars may not necessarily have had a horsepower value equal to the statistical mean, replacing them with such a value allowed for the mean of the data to remain constant.

When looking at the car_name variable, it was the only column whose values were not quantities. In order for a categorical variable to have significant impact in a linear regression model, there cannot be many unique values, and car_name had 305 unique values. Therefore, the column was dropped from the dataset.

The dataset was also transformed using one-hot encoding on the origin column, whose values for all 398 rows were either a 1, 2, or 3. The original column, origin, was split into two new columns, origin_2 and origin_3, which each had type boolean. This allows the data to be used later in the linear regression model (values of 0 or 1).

Some statistical analysis was performed on the dataset, including univariate and bivariate plots. A histogram of fuel efficiency was created to see the distribution of

values; this plot exhibited a slight right skew and a mode at 20. A box plot of horsepower was also generated to see if there were any outliers, which there were. Each quantitative variable was plotted against fuel efficiency to see the trend between the two, and it showed that there was a strong, indirect correlation between miles per gallon and displacement, horsepower, and weight.

Training the Linear Regression Model

Once the exploratory data analysis and pre-processing were complete, the data was split into training and test datasets using an sklearn method. 30% of the data was reclassified to test data, while the other 70% remained as data to train the model. A linear regression model was then created, and the feature and label training data were fit to the model. Because the model had eight features, it was not able to be visualized in a two- or three-dimensional space, but would need nine dimensions. To test the model, an equation was created using weights, or coefficients, on each of the features, as well as an intercept value. The equation derived from the model is as follows:

```
mpg = -0.396*cylinders + 0.029*displacement + -0.021*horsepower +
      -0.007*weight + 0.066*acceleration + 0.838*model_year +
      2.991*origin_2 + 2.378*origin_3 + -21.380
```

If the features from new data, such as the test data, were inputted into this model equation, a predicted fuel efficiency value would be calculated.

Model Performance

The accuracy of the model was tested using both the data that was used to train the model, as well as the test data that was separated and set aside earlier in the process. An R^2 test was performed on both datasets to determine a quantitative value that represents the fit of the data to the model. The test compares the predicted label value, in this case, miles per gallon, calculated in the equation to the value given in the data. Higher R^2 values means the model was more accurate, and vice versa. The training data outputted an R^2 value of 0.814, while the test data outputted a value of 0.843.

The R^2 value of the test set is higher than that of the training set, which is interesting as the model was made to fit the training data. The model may be a bit overfit for the training data, yielding some inaccurate results, while accurately predicting the values of the new test data that it has not seen before.

With R^2 scores of above 0.80, I am confident in the model's ability to predict a car's fuel efficiency given the number of cylinders, the displacement, the horsepower, the weight, the acceleration, the model year, and the two origin booleans.

References

Class repository: https://coe-379l-sp24.readthedocs.io/en/latest/index.html
Dataset: https://raw.githubusercontent.com/joestubbs/coe379L-sp24/master/datasets/unit01/project1.data