Auto-MPG Report

Dataset

The dataset examined in this report was obtained from UCI Machine Learning Repository. It contains 398 instances and 8 attributes.

- Input Variables:
  - 2 - Cylinders
  - 3 - Displacement
  - 4 - Horsepower
  - 5 - Weight
  - 6 - Acceleration
  - 7 - Model Year
  - 8 - Origin
  - 9 - Car Name (unique for each instance)
- Output Variable:
  - 1 - MPG

The input variable of "Car Name" was eliminated due to its object type being a unique string to each instance.

Exploratory Data Analysis

Table 1. Summary Statistics of auto-mpg.csv.

	mpg	**cylinders**	**displacement**	**weight**	**acceleration**	**model year**	**origin**
count	392.0000	392.0000	392.000000	392.00000	392.00000	392.0000	392.0000
mean	23.44591	5.471939	194.411990	2977.5841	15.541327	75.97959	1.576531
std	7.805007	1.705783	104.644004	849.40256	2.758864	3.683737	0.805518
min	9.000000	3.000000	68.000000	1613.0000	8.000000	70.00000	1.000000
25%	17.00000	4.000000	105.000000	2225.2500	13.775000	73.00000	1.000000
50%	22.75000	4.000000	151.000000	2803.5000	15.500000	76.00000	1.000000
75%	29.00000	8.000000	275.750000	3614.7500	17.025000	79.00000	2.000000
max	46.60000	8.000000	455.000000	5140.0000	24.800000	82.00000	3.000000

Table 1 displays basic statistics over the dataset. A note of point is the reduction of the files original 398 instances to 392 as 6 of the instances contained one or more missing attributes.

As the amount of instances with missing data points were low, a complete removal of them was found to be appropriate.

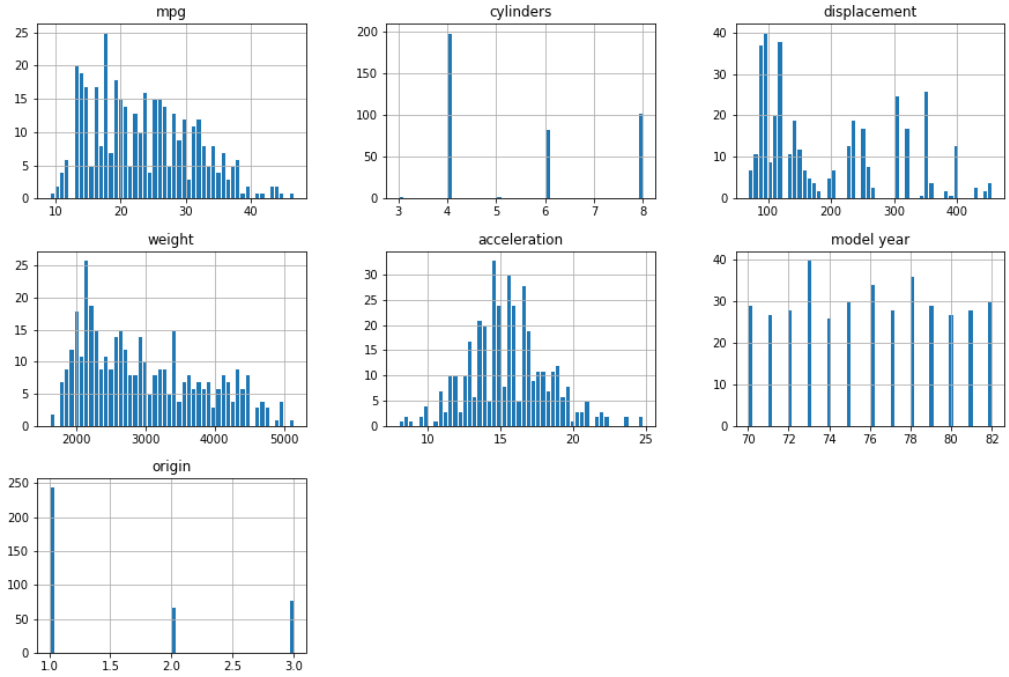

Figure 1: Histogram plots of the distribution of each parameter against the response variable.

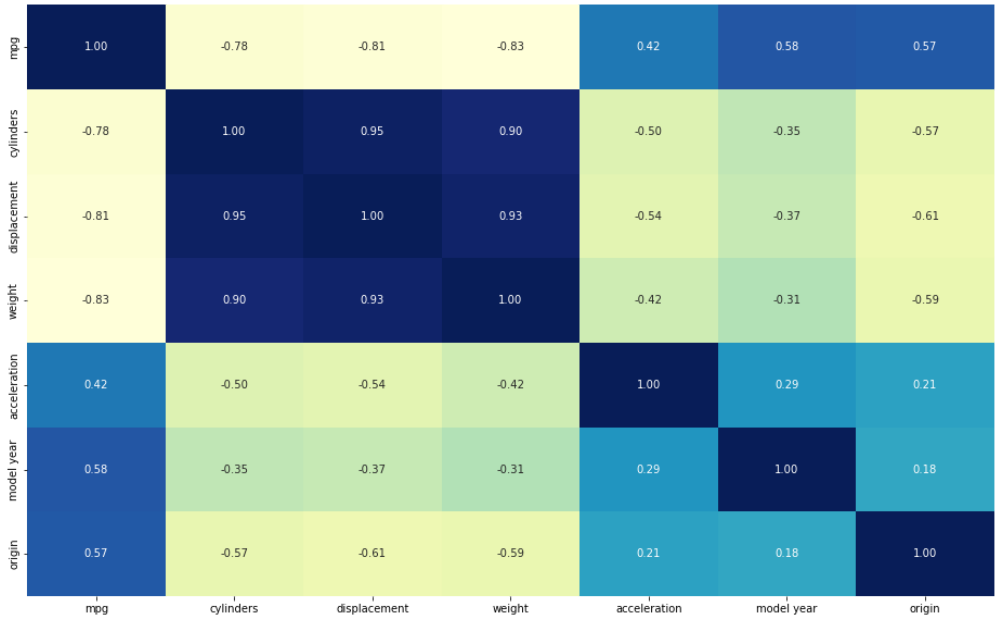

Figure 2: Correlation ‘heatmap’ of the auto-mpg.csv data file.

The histograms in Figure 1 allowed for preliminary insight on what possible variables would impact the response variable the most. Both “acceleration” and “model year” show a uniform distribution while the rest of the variables have different degrees of being skewed right.

Moreover, the correlation 'heatmap' displayed in Figure 2 gave insights into collinearity in the data set. For instance, there is a weak positive correlation between acceleration and model year which is intuitively sound as machining of parts and improvements in engines have taken place over time. Multicollinearity is also present in the dataset, with both displacement and cylinders having strong positive correlation with weight as expected. These observations, as well as the moderately strong positive correlation of origin, model year and acceleration with the response variable, give insight on how feature selection algorithms used on the dataset will perform.

Feature Selection

In the Python file, three feature selection methods from sklearn were selected to be used on the models: forward, backward, and recursive selection. The forward and backward selection methods resulted in "weight", "model year" and "origin" being chosen to predict "mpg". The recursive selection method instead used "cylinders", "model year", and "origin" as the optimal predictor variables.

In the Scala file, the used feature selection methods from Scalation were forward selection, backwards elimination and stepwise elimination. In comparison to how sklearn's feature selection works in Python, Scalation's feature selection was dependent on the model used and thus every model resulted in various used features. The forward selection models often opted to use all the features to predict "mpg". Likewise, the stepwise elimination models used all the features except "acceleration". In comparison, the backward elimination models commonly used "cylinders" and "model year" as predictor variables.

Results

The full results, including features used in a model, R^2, adjusted R^2 and R^2 cross-validated , can be found in the accompanied summary file.

In the Python component of the analysis, both the forward and backward selection models performed the same as both feature selection methods resulted in the same set of features. The symbolic models for both of these selection methods had the highest R^2 cross-validated value of 0.86365. In comparison, the quadratic models for both methods had the highest adjusted R^2 of 0.8466. Thus if performing on new sets of automobiles to predict miles per gallon, the symbolic model would be more ideal. The recursive selection method models R^2 results for this data set were on average 25% lower than the other two selection method models.

In the Scala component, all models performed better than their counterparts in Python. But likewise with the Python models, the symbolic model (with intercept and first cross term) performed the best in both the forward and backward selection models with R^2 cross-validated values of 0.87394 and 0.873536 respectively. As well both these selection methods saw the highest R^2 adjusted value of 0.917528 in Symbolic Ridge Regression with both cross terms present. Comparatively, the stepwise selection method slightly outperformed the other two selection models with both Symbolic (no intercept, with first cross term) and Symbolic Ridge (neither cross terms included) regression resulting in a R^2 cross-validated value of 0.880934.

Likewise with forward and backward selection, the Symbolic Ridge regression with both cross terms present resulted in the highest R^2 adjusted value (0.897624) out of all the models tested for the stepwise selection method.

Overall, in both languages, the Symbolic models appear to result in best R^2 cross validated values with either forward or backward feature selection.

Discussion

The auto-mpg data set resulted in several strong R^2 values given from the various models and feature selection methods. Despite this success, more systematic approaches to the data set could yield with better models. Such approaches could include a more in depth exploration into the right-tailed skewing of most of the variables in comparison with the response variable (one such solution could be to mean center the data). Furthermore, at least in regards to the Python component, a better understanding of how sklearn's feature selection methods could be taken place along with testing of several of their other feature selection methods as well.