ISYE 7406 - Homework 3

Matthew Bernstein

February 13, 2023

Introduction

In this experiment, different methods of fitting and comparing classification models were tested. The dataset included 7 characteristics of cars with fuel efficiency as the response. The goal was to determine a model using the characteristics to determine if a car is expected to have fuel efficiency that is better (or worse) than the median.

Five models were fit after selecting relevant features for 100 Monte Carlo cross validation trials. Accuracy for each model was assessed using classification error of the test set.

Exploratory Data Analysis

The dataset contains 7 characteristics. Of these, 2 are categorical variables: *cylinders* and *origin*. The other 5 are continuous variables. While *year* is listed as a catergorical feature, due to the number of levels and context, it is treated as a continous variable for this analysis. Also, given the low amount of 3 and 5 cylinder vehicles in the dataset, those levels were combined with 4 cylinders. The features are a combination of mechnical information such as *weight* and *acceleration* as well as "biographical" such as *year* and *origin*.

A pairwise analysis of the categorical variables is show in Figure 1. It appears that both variables have correlation with fuel efficiency. Lower cylinder cars are much more likely to have better than median fuel efficiency, as well as cars from region 2 and 3.

Figure 2 shows a pairwise analysis of the continuous variables while Figure 3 shows the correlation heatmap. *Displacement*, *weight*, and *horsepower* all are very strong predictors of fuel efficiency. The correlation heatmap shows extremely high colinearity between those variables as well. It is likely that only one of those variables is required for a strong classifier. *Year* appears to show a weak correlation to fuel efficiency with later years having better odds of being above the median. *Acceleration* appears to be the least correlated and might not be needed for a classifier.

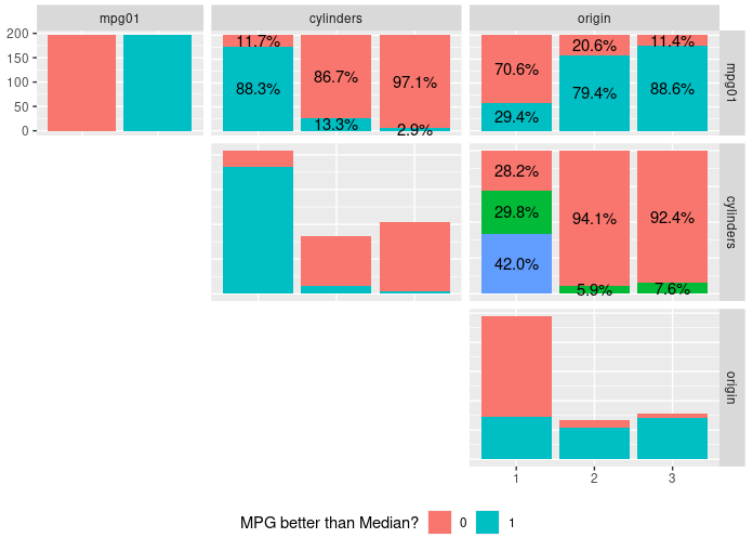

Figure 1: Pairwise analysis of categorical variables

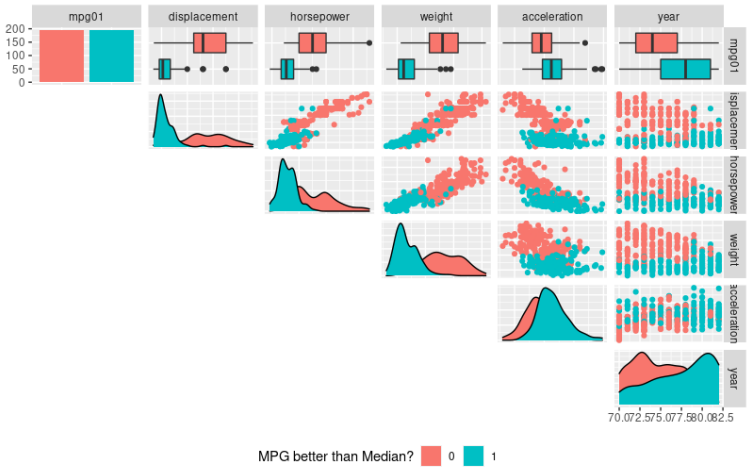

Figure 2: Pairwise analysis of continuous variables

Methodology

Variable Selection

In order to simplify the model, two rounds of variable selection were performed.

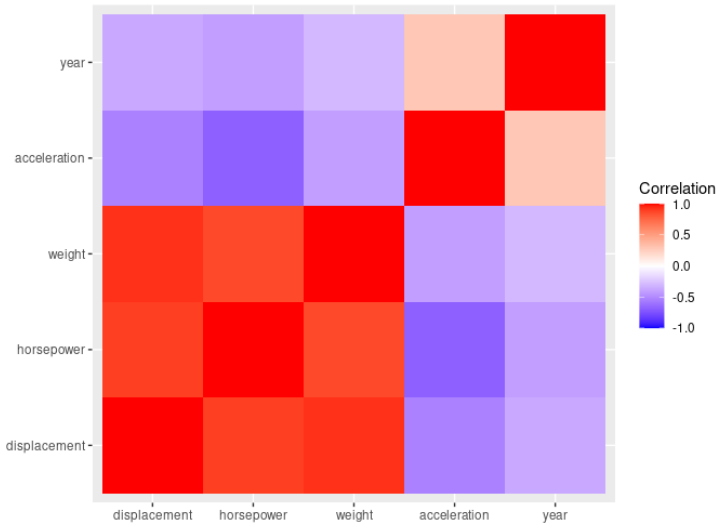

Figure 3: Correlation heatmap

Acceleration

Of all the continuous variables, *acceleration* seems the least likely to impact the gas milage. Two logisitic regression models were made, one with all variables and one with all variables except for *aceeleration*. Each model was trained on all data with 5-fold cross validation, repeated 3 times. The cross validation accuracy and AIC are shown in Table 1.

Table 1: Acceleration variable selection results

Model	Accuracy	AIC
Full	0.9090	159.4781
Without Acceleration	0.9115	157.8898

Moving forward, *acceleration* was omitted from the feature set.

Displacement, Weight, Horsepower

There was high colinearity between *displacement*, *weight*, *horsepower*. It is likely that only one of those three variables is sufficient for the model. Four models were created, one with all variables and one model each excluding two of the three variables in this round of selection. Each model was trained on all data with 5-fold cross validataion, repeated 3 times. The cross validation accuracy and AIC are shown in Table 2

All future models were created with *displacement* and *horsepower* omitted.

Table 2: Displacement, weight, horsepower variable selection results

Model	Accuracy	AIC
Full	0.9064	157.8898
Only Displacement	0.9098	195.0084
Only Weight	0.9183	162.3589
Only HP	0.8928	187.9729

Model Comparisons

Five models were created for comparison: Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naive Bayes Classifier, and K-Nearest Neighbors classifier. For the KNN model, k was chosen using a 3 times repeated, 5-fold cross validation. All models were subject to 100 trials of Monte Carlo cross validation. For the KNN model, the optimal k was determined each trial.

The test error was reported for each trial and aggregated.

Results

After 100 Monte Carlo cross validation trials, the aggregated test errors are shows in Table 3.

Table 3: Aggregated model test classification error

	Mean	Median	Variance
Logistic Regression	0.9158	0.9103	0.0008
Naive Bayes	0.9040	0.9103	0.0008
LDA	0.9035	0.9103	0.0008
QDA	0.8981	0.8974	0.0008
KNN	0.8719	0.8718	0.0011

As expected, the Logistic Regression model performed the strongest. Logisitic Regression is the most robust of the model choices since it makes no assumptions about the data. Naive Bayes and LDA performed similarly. Although, they also make assumptions on the data, it appears the common covariance assumption that LDA assigns is more accurate than the per-class normality assumption that QDA makes. Given that the model variables were downsized and sources of colinearity were removed, it follows that the independence assumption that Naive Bayes classifiers make would be more accurate. However, I suspect that removing those variables hurt the accuracy of the KNN classifier since those variables would help close distance between applicable neighbors. It is also possible that transforming the continuous response to a discrete variable could cause misclassification on points near the boundary.

Conclusion

When creating a classification model, there are many options to evaluate. LDA, QDA, and Naive Bayes place assumptions upon the dataset which can cause them to fall in prediction accuracy. KNN models are also an option however they are susceptible to data quality issues and unimportant features. Logisitic regressions are usually the most robust models and the analysis on this dataset supports this. In conclusion, I would recommend using logisitic regression models for classification.