Predictive Classification of Automobile Fuel Efficiency

ADS526Z2 - Machine Learning II

James Sears

Applied Data Science Program, Graduate School, Longmeadow, MA

Submitted To: Dr. Xiaoxia Liu

Summer 2021

Jim Sears
ADS526Z2 - Machine Learning II
Final Project

Abstract

The objective of this project is to determine the best – or most plausible – method of predicting the classification of an automobile's fuel efficiency in terms of a high or low measurement of miles per gallon (mpg). The data analyzed included the weight, year built, country of origin, engine specifications (cylinders, displacement horsepower) and outputs (acceleration, mpg) for 392 different automobiles built between 1970 and 1982. The median mpg was calculated to classify each vehicle as one with a high or low mpg, and therefore bad or good fuel efficiency. Feature selection based on univariate and bivariate analysis was performed and the labeled data set was used to train and evaluate four different classification approaches including Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and K-Nearest Neighbors (KNN). The LDA model performed the best in terms of error rate and variance, supporting the noted separability of classes amongst the variables. With an accuracy rating of almost 90%, the LDA model is considered a very plausible predictor for classifying an automobile's fuel efficiency.

Introduction

The objective of this project is to determine a method of predicting the classification of an automobile's fuel efficiency. Fuel efficiency is measured as the average number of miles the vehicle will travel for every gallon of fuel consumed by its engine and is reported as *miles per gallon* (mpg). Vehicles with lower, or worse, fuel efficiency have lower measures of mpg; conversely, vehicles with higher, or better, fuel efficiency have higher measures of mpg. Fuel efficiency is a function of the attributes of the vehicle and its engine. The interest of this project is determining if those attributes and a classification modeling technique can enable the prediction of high or low fuel efficiency.

The `Auto` dataset selected for this project is from the ISLR package associated with *An Introduction to Statistical Learning with Applications in R (ISLR)* and was originally from the StatLib library which is maintained at Carnegie Mellon University. It was selected as it includes 392 different automobiles built between 1970 and 1982 and their observations of the following variables:

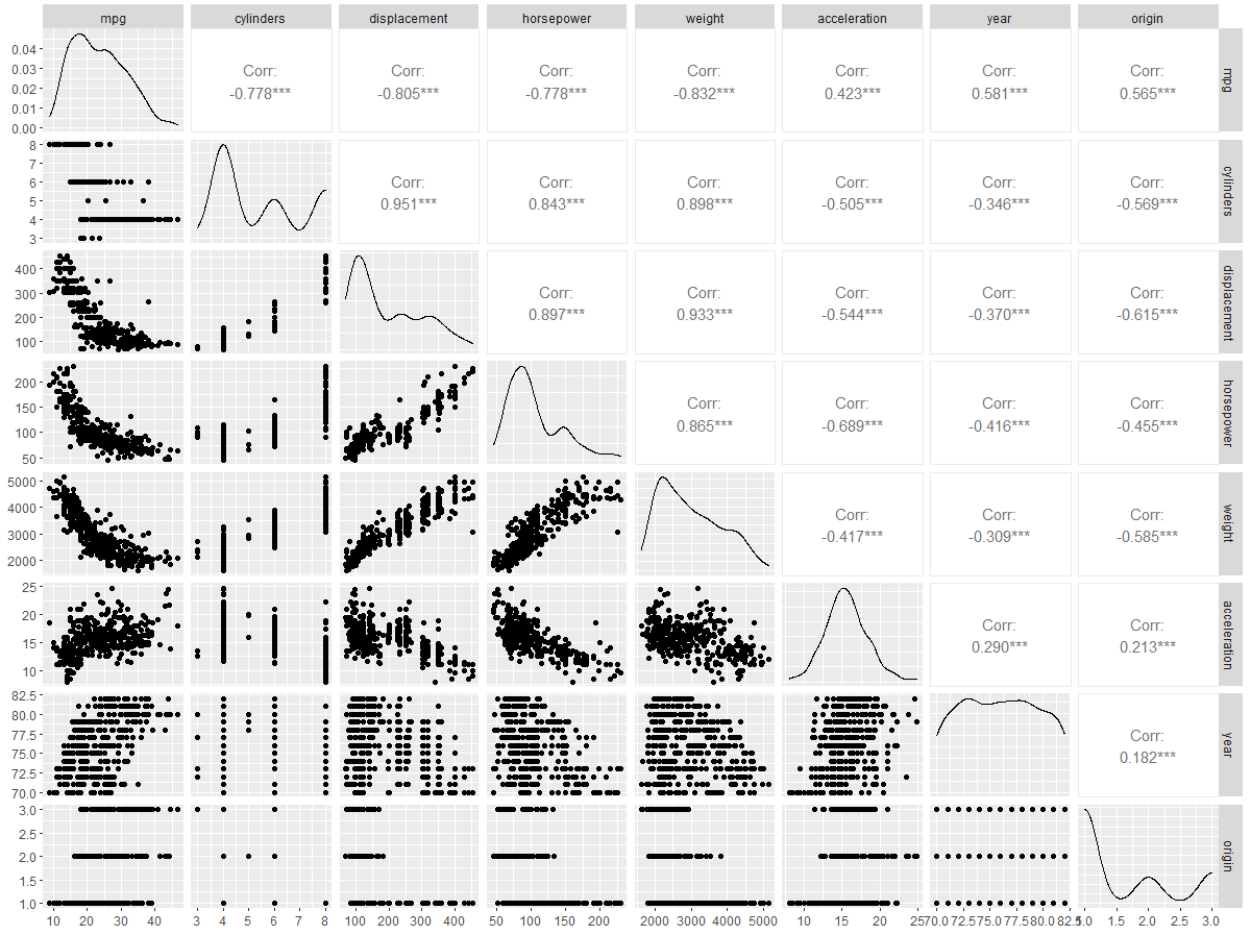
Table 1. `Auto` dataset variables and descriptions.

Variable	Description
mpg	miles per gallon
cylinders	Number of cylinders, between 4 and 8
displacement	Engine displacement (cu. inches)
horsepower	Engine horsepower
weight	Vehicle weight (lbs.)
acceleration	Time to accelerate from 0 to 60 mph (sec.)
year	Model year
origin	Origin of car (1. American, 2. European, 3. Japanese)
name	Vehicle name

The continuous mpg variable represents each vehicles' fuel efficiency and is selected as the dependent, or response, variable. The remaining independent, or explanatory, variables are either continuous (displacement, horsepower, weight, acceleration), discrete (cylinders, year, origin) or nominal (name). Removing the nominal variable and generating a scatter plot matrix (Figure 1) reveals more about the independent variables themselves, along with their relationship with mpg.

The variables that have strong relationships with mpg are cylinders (r = -0.778), displacement (r = -0.805), horsepower (r = -0.778) are weight (r = -0.832). These strong negative relationships are reasonable as an engine that is larger (more cylinders) will consume more fuel (higher displacement) and is likely going to produce more power (horsepower) but therefore be less fuel efficient; this is

Figure 1. Scatterplot matrix of variables.



compounded as a heavier (weight) vehicle needs more power than a lighter one to move and therefore consumes more fuel. Year (*r* = 0.581) has a moderately strong positive relationship with mpg and is reasonable as oil and gas prices were high in the 1970s and the demand for more fuel-efficient cars increased over time. Origin (*r* = 0.565) also has a moderately strong positive relationship with mpg; however, it is a discrete, but not ordinal, variable and therefore does not lend itself as a reliable predictor. Acceleration has a weak (*r* = 0.423) relationship with mpg as expected as it is likely a function, or result, of the combination of horsepower and weight and not an actual engine specification.

Therefore, the variables cylinders, displacement, horsepower, weight and year are potential predictors of mpg.

The values of mpg range from 9.00 to 46.60 with a mean of 23.45 and median that is slightly lower at 22.75. Therefore, it's distribution is slightly right skewed as observed in Figure 2. Dividing the

Figure 2. Histogram of mpg.

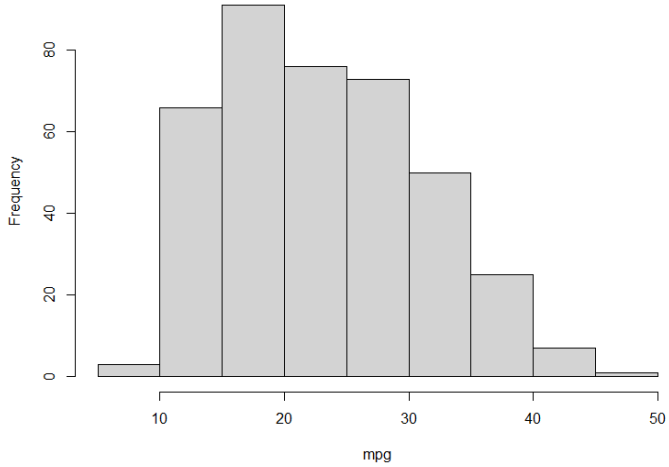
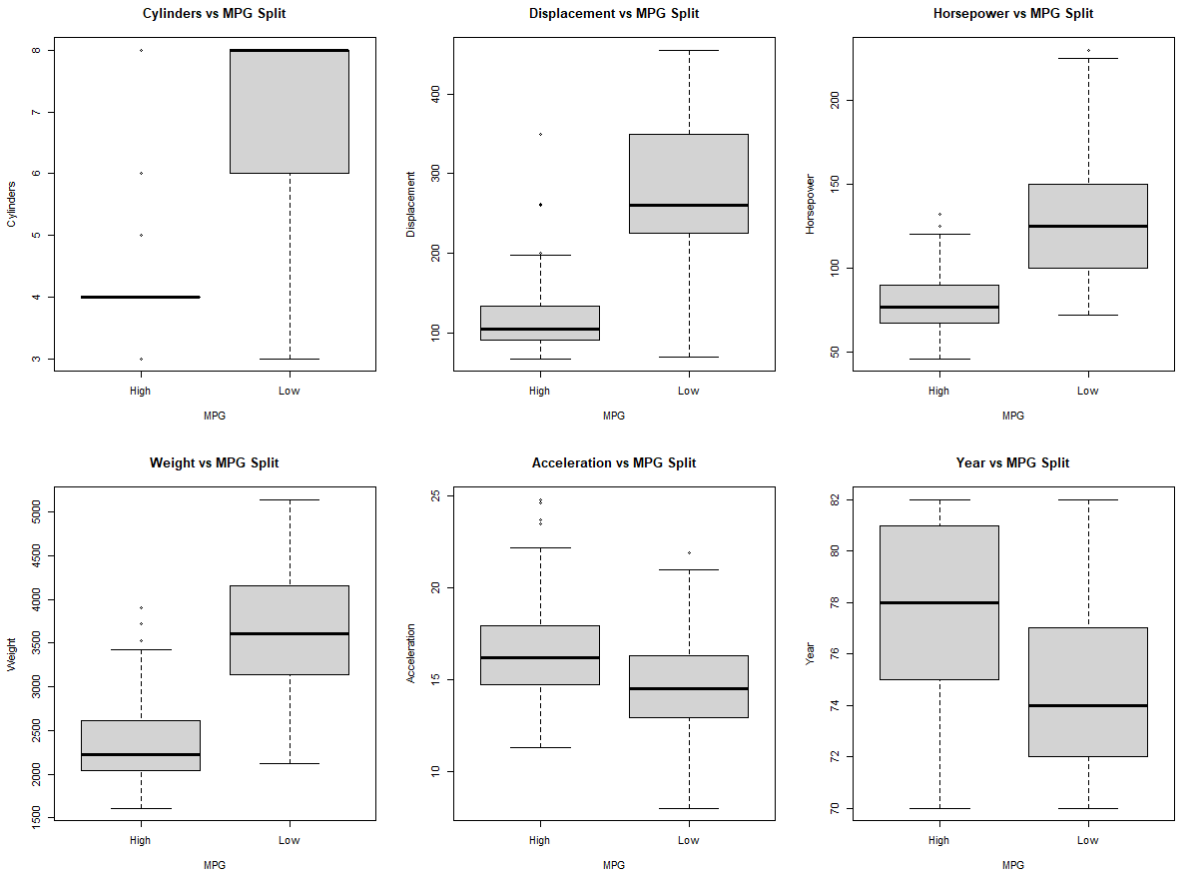


Figure 3. Scatterplot matrix of variables classified by mpg: low (red) and high (green).

observations in half based on the median results in categorizing 192 vehicles as “low” (below 22.75 mpg, poor fuel efficiency), and 192 as “high” (above 22.75 mpg, good fuel efficiency). Generating a scatterplot matrix that incorporates these classifications (Figure 3) reveals that there is separability within some of the variables. This is noted by the clustering of low and high mpg in the scatterplots involving displacement, horsepower and weight.

The separation is also seen in their boxplots (Figure 4) when the interquartile range of one class does not overlap with the other. The variables cylinders, displacement, horsepower and weight display this characteristic; acceleration and year do not.

Figure 4. Boxplots of variables classified by mpg: low and high.



Methods

As stated, the variables cylinders, displacement, horsepower and weight are correlated with mpg, and they display notable separability in terms of mpg classification. They are therefore potential predictors for a classification model to predict the classes of fuel efficiency. Several model types were selected for evaluation, including Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and K-Nearest Neighbors (KNN). The data was split into a training set (262, or two-thirds of the observations) and test set (130, or one-third of the observations). Each model was trained on the training set and evaluated on the test set in terms of their rates of predictive accuracy and error.

Logistic Regression

This classification method works especially well when the number of classes (*k*) equals two and dimensionality (the number of *p* predictors) is relatively low. In this case, there are two classes ("high" and "low") and only four predictors have been selected for the model. It is also noted that the number of observations (*n*) is 392, which is of moderate size, but certainly greater than $p = 4$. Therefore, in this low-dimension setting, classic logistic regression is appropriate and penalized logistic regression is not.

Logistic Regression is like Linear Regression, except that it uses the regression (multiple in this case) equation based on *X* not to predict *Y* quantitatively, but rather to calculate a probability that *Y* belongs to a given category, or class. As $p(X) = Pr(Y = 1|X)$, or the probability of *X* equals the probability of *Y* equals one given *X*, the probability will be between 0 and 1. And in the case where $k = 2$ and the distribution of mpg is almost normal, an observation is assigned to one of the two classes when $p(X)$ is less than or greater than 0.5.

It is noted that Logistic Regression's stability is susceptible to higher class separability.

Linear Discriminant Analysis (LDA)

As stated, the variables cylinders, displacement, horsepower and weight display notable separability in terms of mpg classification. As LDA maximizes the separability of the classes, it is a reasonable model for classification in this case. Separability can be evaluated as a ratio of the difference in means and the sum of the variances of the observations of each of the two classes within a variable.

LDA creates a new axis for each variable such that the separability between two classes is optimized. It then generates new dimensions (linear combinations of features in the dataset) that prioritize those variables whose separability is greatest. It optimizes the predictive power of the collective variables by prioritizing those which classify better, thus performing dimension reduction as well. New data is fed through the LDA model to generate estimates for the parameters that are plugged into Bayes' theorem to calculate the probability of belonging to one class or the other. (LDA Explained)

It is noted that LDA assumes that the variance of each class within each variable is normal and share the same covariance matrix. And while the selected variables of cylinders, displacement, horsepower and weight exhibit separability in terms of mpg classification, they do not exhibit normal distribution amongst the classes as seen by the skewed boxplots in Figure 4. Linear Discriminant Analysis, as its name implies, also assumes that a linear decision boundary between classes exists.

Quadratic Discriminant Analysis (QDA)

Like Linear Discriminant Analysis, a QDA classifier results from assuming that the observations from each class are drawn from a normal distribution and plugging estimates for the parameters into Bayes' theorem to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix (ISLR, page 149). As a result, QDA needs to estimate *k* times more parameters than LDA and is therefore a more flexible model that has less bias at the potential cost of more variance.

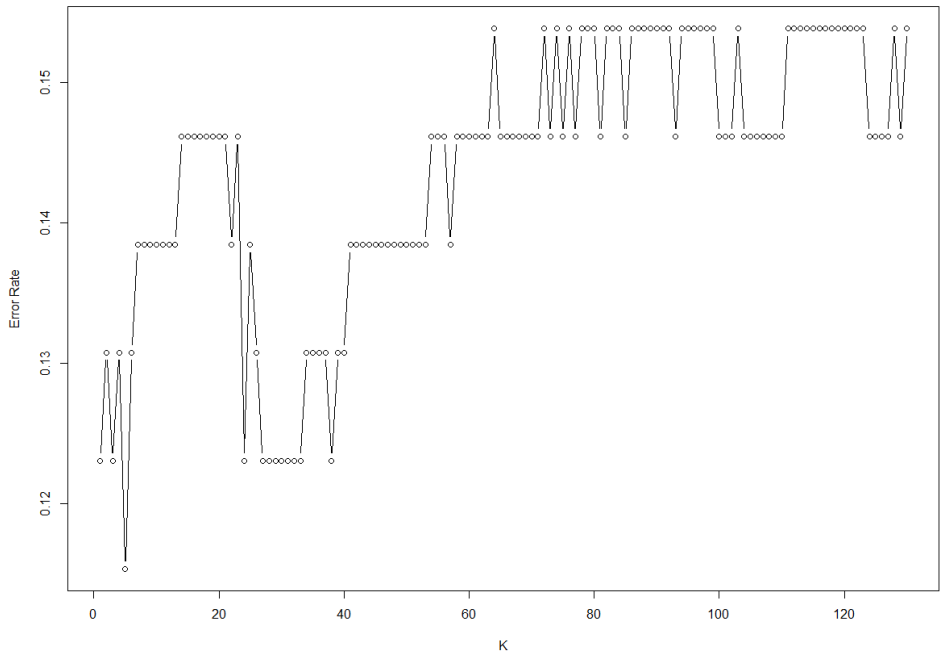
However, if the decision boundary between classes is non-linear (quadratic), QDA will likely perform better.

K-Nearest Neighbors (KNN)

This non-parametric technique for classification was also evaluated as an alternative to the models above that have greater bias. KNN assumes that similar data points are found close to each other. In other words, and in terms of classification, data points associated with one class are very likely to be grouped together. As these groupings are not bound by a linear or quadratic equation, they enable unbiased models to be created to predict classification. The probability of a new data point with similar characteristics is very likely to be of that same class such that Bayes' theorem can be applied to predict classification.

The selection of the number (any positive integer) of groups, *k*, can be set manually or optimized by evaluating the error rates of every possible value of *k* (1 to *n*-1). In the case of this project, the optimal value of *k* was determined to be five (Figure 5).

Figure 5. Evaluation of KNN model; lowest error rate when k = 5.



Results & Discussion

Each classification model was simulated 10,000 times: 100 simulations were run with new training and test data sets created each time with 100 different random number generator seeds. The mean error rate and variance for each model is listed in Table 2, and the variance is demonstrated in Figure 6.

The KNN (*k*=5) model (12.19% error rate) was outperformed by the Logistic Regression (10.84%), LDA (10.48%) and QDA (10.64%) models. It was noted that the separability of the classes among the

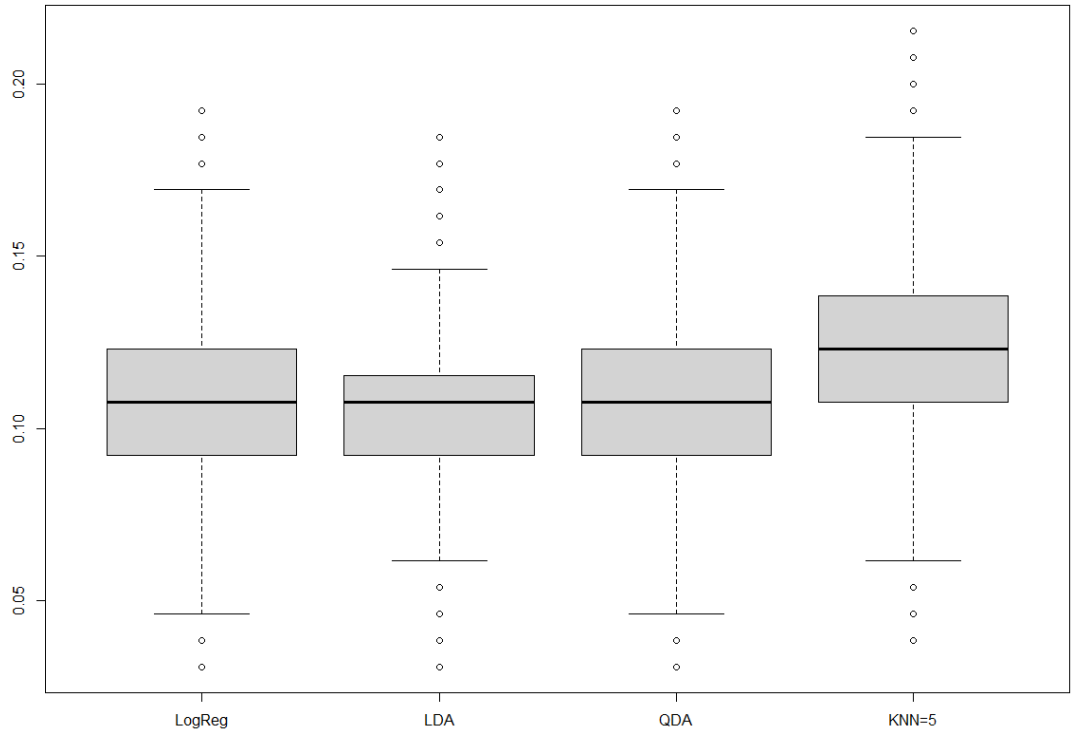
variables is good. That the evaluated optimal value of *k* is more than two (low, high) suggests that the model generated did not pick up on the separability and suffered as a result.

Table 2. Evaluation from 10,000 simulations of each classification model.

	Logistic Regression	LDA	QDA	KNN (k=5)
Accuracy Rate	89.16%	89.52%	89.36%	87.81%
Error Rate	10.84%	10.48%	10.64%	12.19%
Error Rate Variance	0.0508%	0.0435%	0.0462%	0.0580%

The performances of the Logistic Regression (10.84%), LDA (10.48%) and QDA (10.64%) models are very similar. The two Discriminant Analysis approaches performed slightly better than Logistic Regression, supporting the notion that the latter does not work well when the classes are well separated.

Figure 6. Boxplots of model simulation error rates.



That that the Linear Discriminant Analysis model outperformed the Quadratic version suggests that the separation of classes amongst the variables is linear, and that QDA suffered performance (and higher variance) in trying to fit a quadratic decision boundary. In addition to having the lowest error rate, the LDA model has the least variance, suggesting that its results are more reliable than the other models. Linear models are biased, and that the variance of the test errors did not suffer as a result again supports a linear decision boundary is more appropriate for the data set and variables.

The LDA model's 10.48% error rate and relatively low variance means that it is very reliable classifier of high and low mpg (good or bad fuel efficiency) but could be improved. To do so, more intensive univariate and bivariate analysis should be performed to better understand the relationships between the variables, such as potential quadratic relationships noted in the scatterplot matrix (Figure 1). A deeper and better understanding of automobile engines would also benefit in identifying any potential multicollinearity of variables.

References

ISLR. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R*, www.StatLearning.com, Springer-Verlag, New York

LDA Explained. Ye, Andrew. (2020, June). *Linear Discriminant Analysis, Explained in Under 4 Minutes*. Medium. https://medium.com/analytics-vidhya/linear-discriminant-analysis-explained-in-under-4-minutes-e558e962c877