AUTOMOBILE GAS MILEAGE ANALYSIS

SEPTEMBER 20, 2021

KO TSZ NGA (VALERIE)
valerie.ktn@gmail.com

Introduction

The aim is to **predict whether a given car gets high or low mileage** with 5 classification models. They are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naive Bayes, Logistic Regression and K Nearest Neighbour (KNN).

Background of the dataset

The 'Auto MPG' dataset is gathered at UCI Machine Learning (ML) Repository. It originally contains 398 rows and 9 columns. The cleaned version that we will be using for this homework consists of **392 rows and 8 columns**. Here are the variables description:

1. mpg: continuous [**Dependent variable**]
2. cylinders: multi-valued discrete [Independent variable]
3. displacement: continuous [Independent variable]
4. horsepower: continuous [Independent variable]
5. weight: continuous [Independent variable]
6. acceleration: continuous [Independent variable]
7. model year: multi-valued discrete [Independent variable]
8. origin: multi-valued discrete [Independent variable]

The continuous dependent variable (ie. mpg) has to be transformed into binary. The 'mpg' column will be replaced by 'mpg01' which contains 1 if a value above its median, and a 0 if mpg contains a value below its median. Its distribution is:

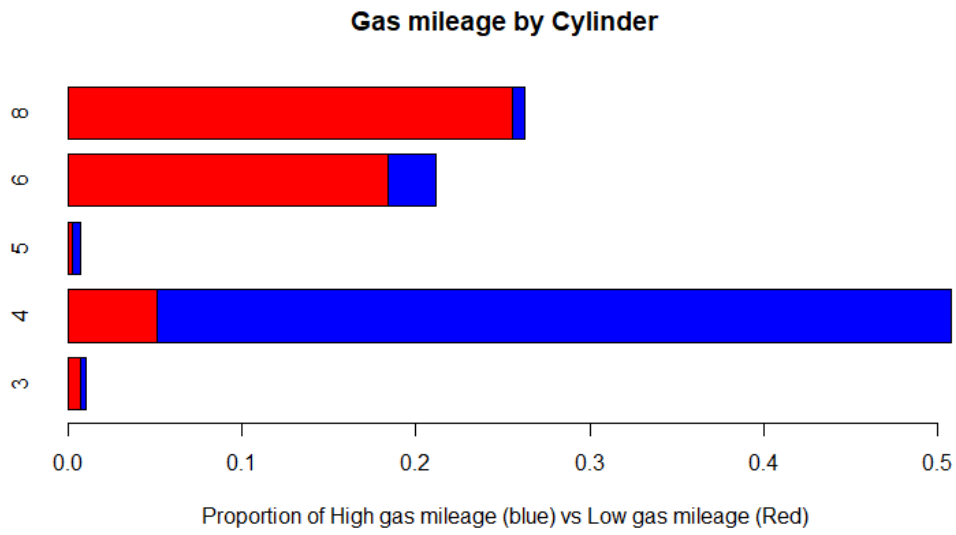
```
> table(Auto$mpg01)

  0   1
196 196
```

Exploratory data analysis

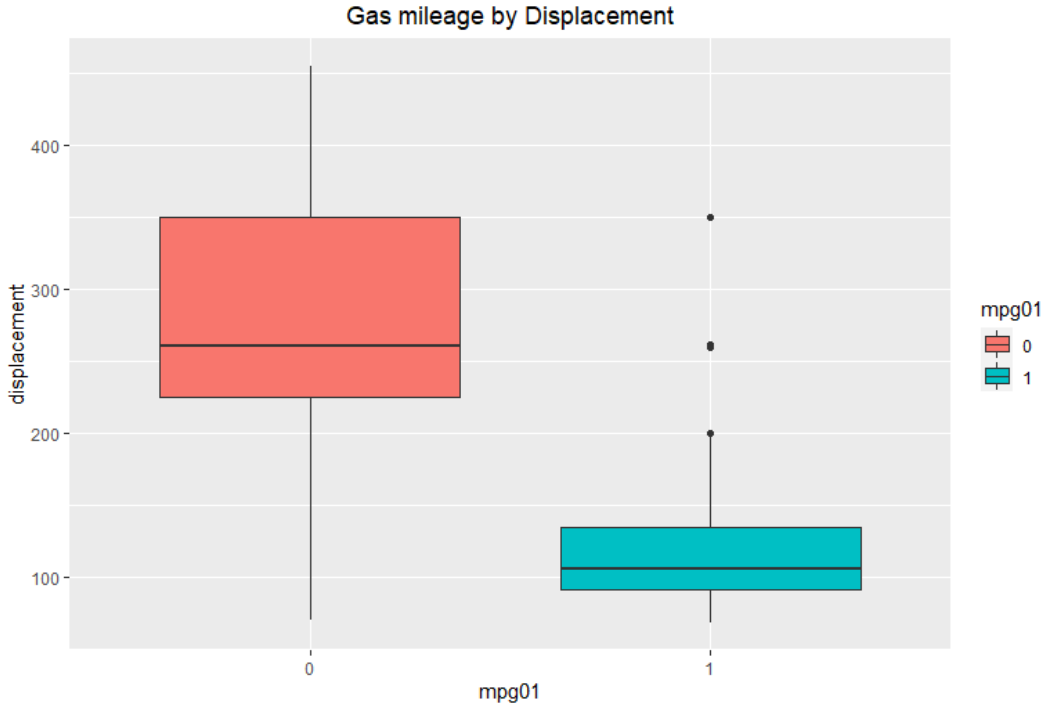
I aim to find the relationship between ‘mpg01’ and other independent variables.

a) ‘Cylinder’ vs ‘mpg01’



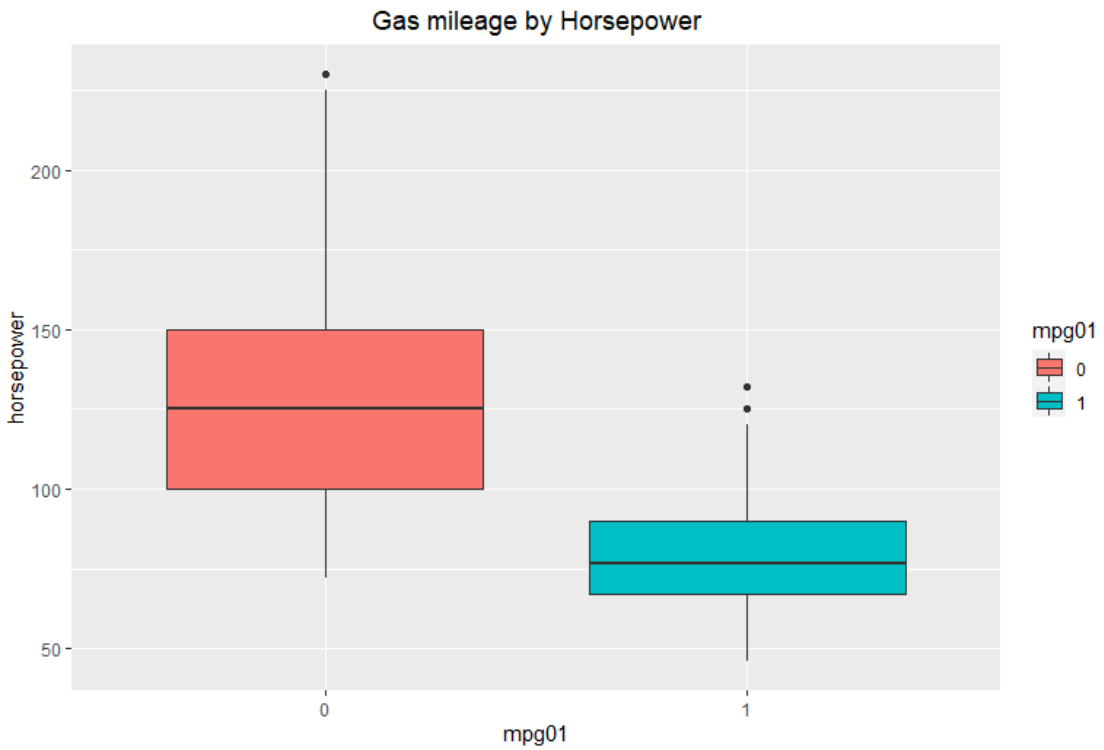
As shown, there is a **substantial amount of high gas mileage car for cylinder value =4.**

b) ‘Displacement vs ‘mpg01’



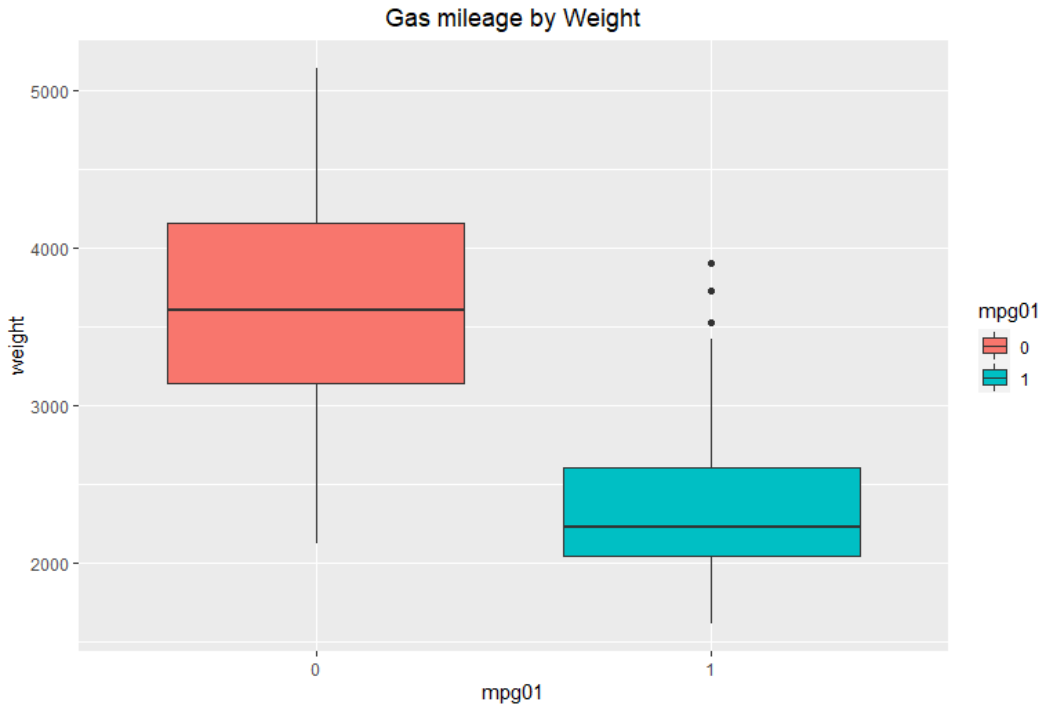
High gas mileage tends to have a smaller displacement.

c) 'Horsepower' vs ‘mpg01’



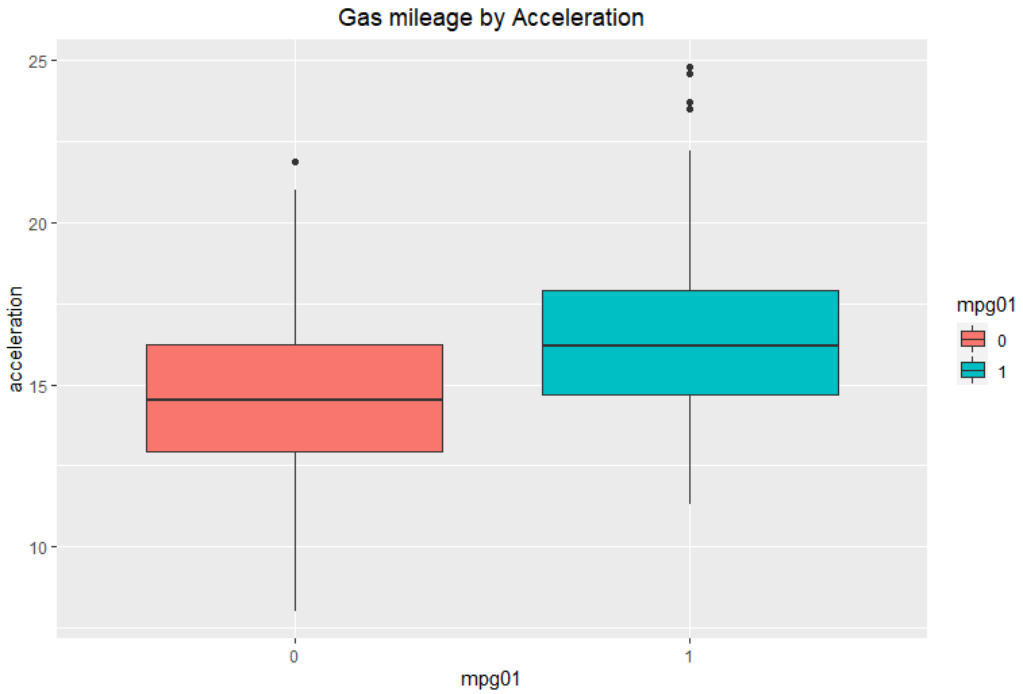
High gas mileage tends to have a smaller horsepower.

d) 'Weight' vs ‘mpg01’



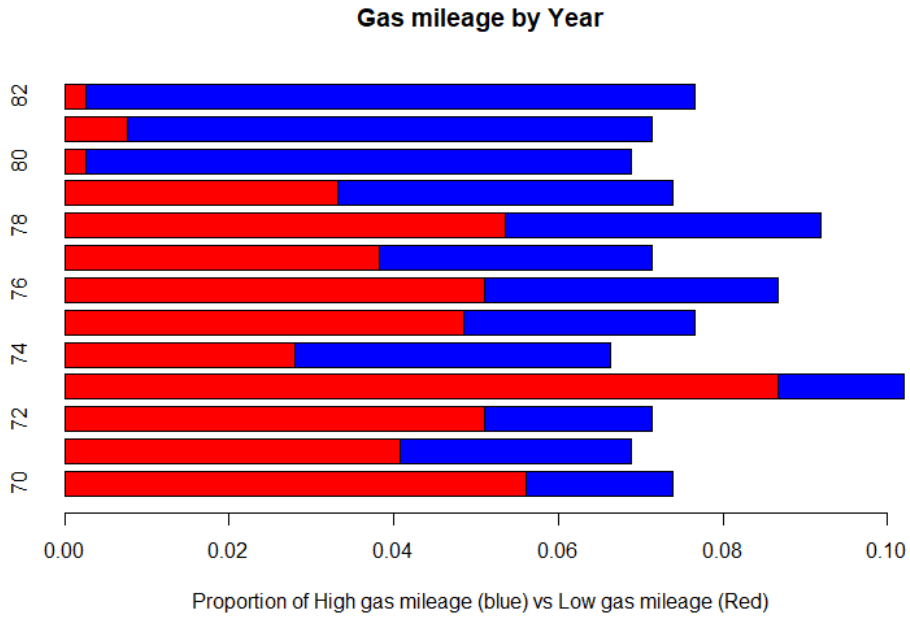
High gas mileage tends to have a lower weight.

e) 'Acceleration' vs 'mpg01'



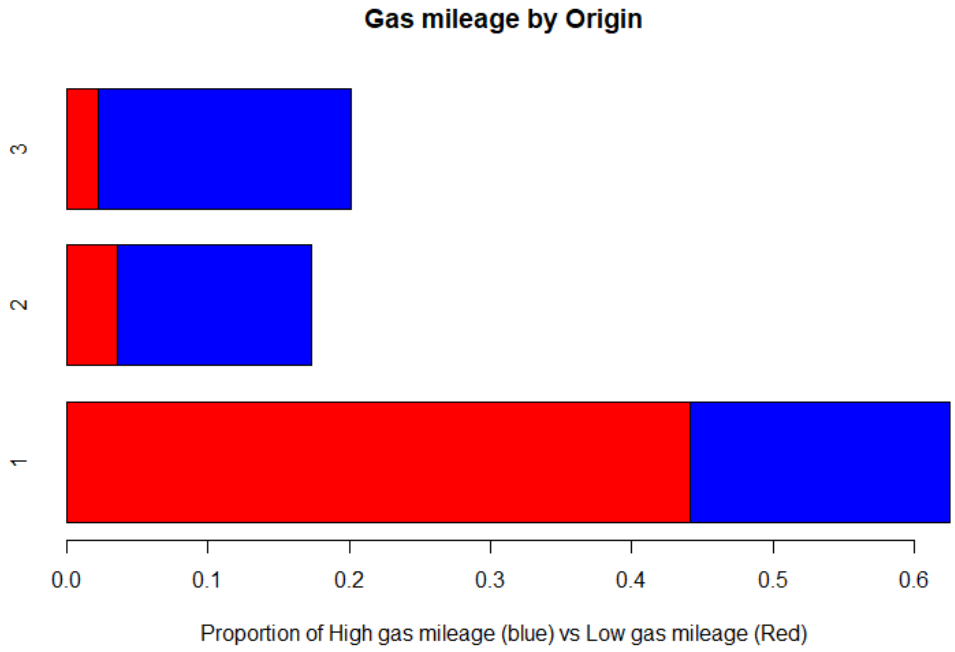
High gas mileage tends to have a higher acceleration.

f) 'Year' vs 'mpg01'



It is shown that there is a **substantial increase of high gas mileage cars when time progres**s. In short, the majority of car has changed from low gas mileage to high gas mileage across the 12 years' time.

g) ‘Origin’ vs ‘mpg01’



For **origin 2 and 3, their majority of cars are in high gas mileage**, while **the majority for origin 1 is low gas mileage.**

In my opinion, except for **‘acceleration’** in which the medians of both groups are relatively close together, all attributes seem to be useful in predicting ‘mpg01’.

Method

As found in our exploratory data analysis, **‘acceleration’ will be excluded** in our analysis. I have decided to use **80% of the data for training and the remaining 20% as testing** data. The split is reasonable by adding randomness which is to set a seed at 19. That results in **314 rows of data in training set and 78 rows in testing set.** Several classification methods (ie. (1) LDA (2) QDA (3) Naive Bayes (4) Logistic Regression (5) KNN with several values of K are performed on the training data and the corresponding test errors are recorded.

Results

(1) LDA

Confusion matrix:

```
          test_y
pred1test  0  1
        0 34  1
        1  6 37
```

Accuracy: 0.9102564

(2) QDA

Confusion matrix:

```
          test_y
pred2test  0  1
        0 36  1
        1  4 37
```

Accuracy: 0.9358974

(3) Naive Bayes

Confusion matrix:

```
          test_y
pred3test  0  1
        0 34  2
        1  6 36
```

Accuracy: 0.8974359

(4) Logistic Regression

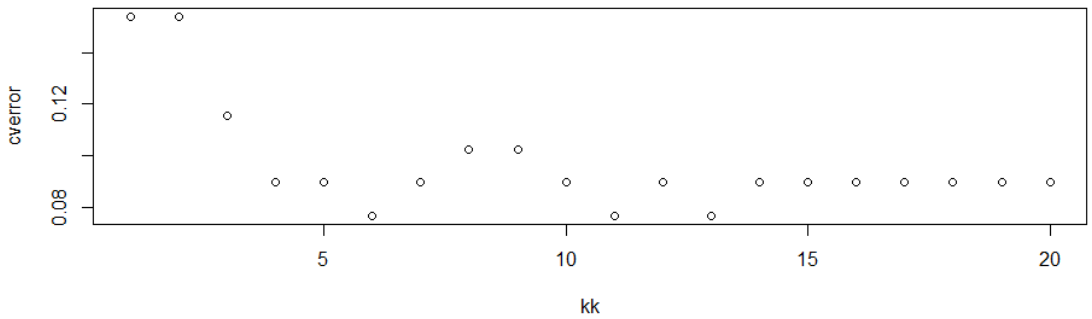
Confusion matrix:

```
          test_y
pred4test  0  1
        0 36  6
        1  4 32
```

Accuracy: 0.8717949

(5) KNN with several values of K.

I have tried to build the model with **K ranging from 1 to 20** and below shows the result.

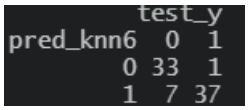


As shown, **K = 6, 11, 13 gives the smallest cross validation error.**

To find out which K to use, I have performed KNN with all 3 values and compare their respective accuracy.

(5.1) **K=6:**

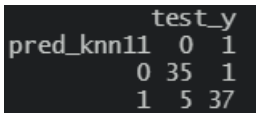
Confusion matrix:



Accuracy: 0.8974359

(5.2) **K=11:**

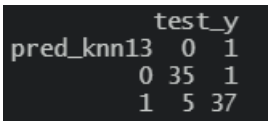
Confusion matrix:



Accuracy: 0.9230769

(5.3) **K=13:**

Confusion matrix:

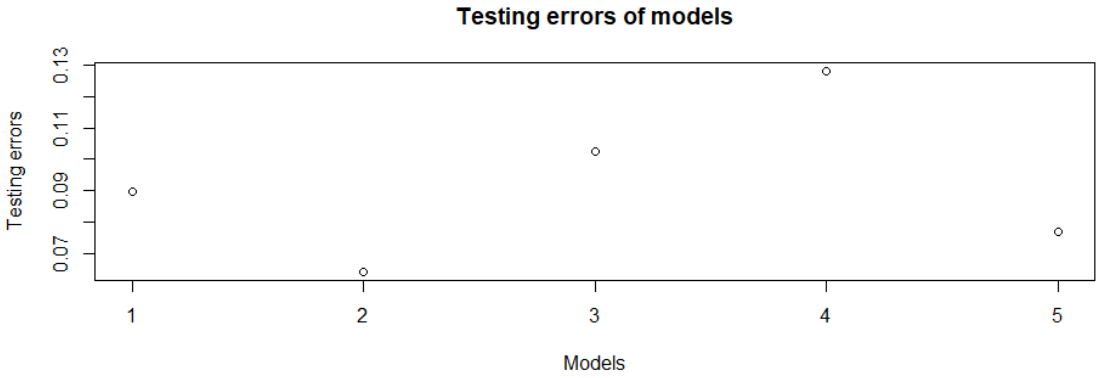


Accuracy: 0.9230769

As we can see, **K=11 and K=13 gives the best accuracy** and hence they should be used.

Findings

The testing errors of the 5 models are displayed as follow:



It is found that the **testing errors of Model 2(QDA) results in the lowest.** Therefore, it is the best model in this case.