OWEN LEE RUI JIE

Student ID: 10219528

ST3189

Machine Learning Course Work

Word Count: 1500

**Contents**

[**Part 1** 2](#_Toc98765533)

[1.1 Data Visualisation 2](#_Toc98765534)

[1.2 Principal Component Analysis 3](#_Toc98765535)

[1.3 K-Means Clustering 4](#_Toc98765536)

[**Part 2** 5](#_Toc98765537)

[2.1 Data Visualisation 5](#_Toc98765538)

[2.1.1 MAT & POR Dataset 5](#_Toc98765539)

[2.2 Math Regression Model 5](#_Toc98765540)

[2.2.1 Forward Selection 5](#_Toc98765541)

[2.2.2 Backward Selection 6](#_Toc98765542)

[2.2.3 K-Fold Cross Validation 6](#_Toc98765543)

[Conclusion 7](#_Toc98765544)

[2.3 Portuguese Regression 7](#_Toc98765545)

[2.3.1 Forward 7](#_Toc98765546)

[2.3.2 Backward 8](#_Toc98765547)

[2.3.3 K-Fold 9](#_Toc98765548)

[Conclusion 9](#_Toc98765549)

[**Part 3** 10](#_Toc98765550)

[3.1 Data Visualisation 10](#_Toc98765551)

[3.2 Regression 10](#_Toc98765552)

[3.3 Receiver Operator Curve 11](#_Toc98765553)

[3.4 Random Forest 11](#_Toc98765554)

# **Part 1**

## Data Visualisation

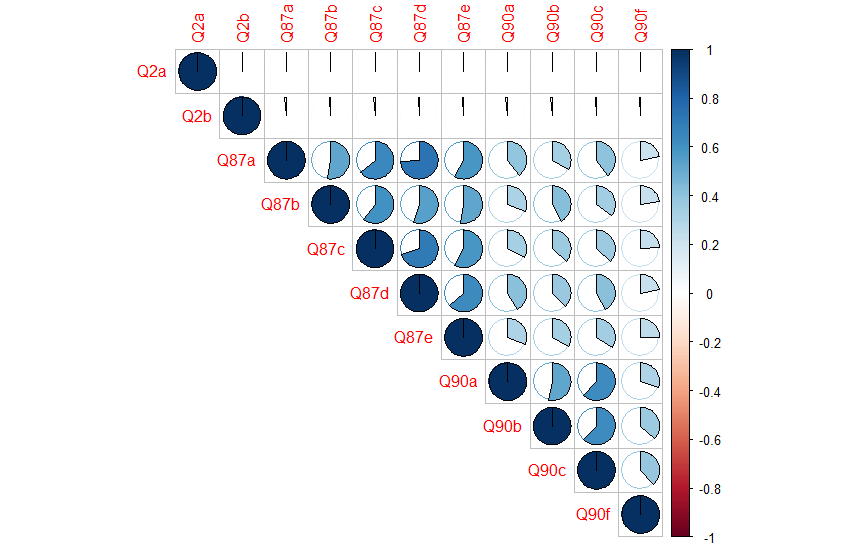


Figure 1.1

Figure 1.1 shows the correlation between the different variables of European Working Conditions Survey (EWCS). From the correlation plot, it can be inferred that Q2a and Q2b have little to no correlation with variables Q87a to Q90f. This means that Age and Gender do not affect an individual’s opinion on the working conditions in Europe.

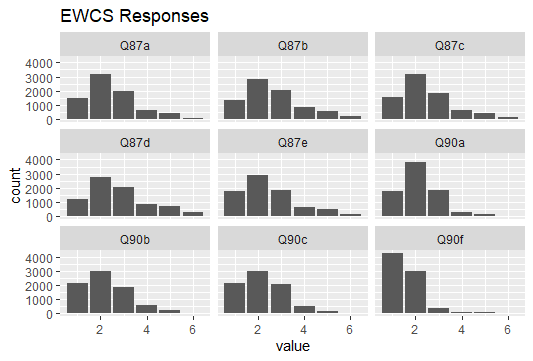


Figure 1.2

From figure 1.2, most respondents chose response 2 for all the questions except for Q90f. This indicates that the response, “most of the time”, is the most popular answer. Whereas for Q90f, most respondents are confident in their ability to perform their job as the most popular response is 1, which is “always”.

## 1.2 Principal Component Analysis

Summary of Principal Component Analysis

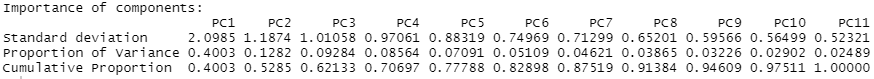


Figure 1.3

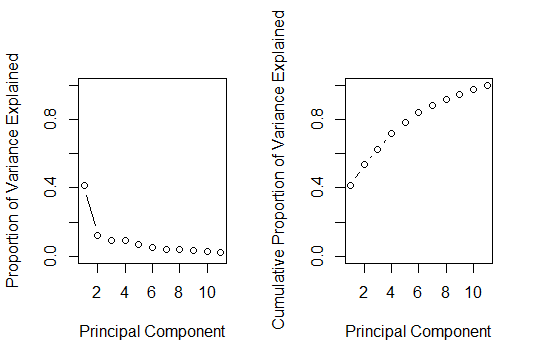


Figure 1.4

Figure 1.3 and 1.4 shows a summary of the Principal Component Analysis on the dataset. Based on the cumulative proportion, PC1 explains the most proportion of variance as compared to the other PCs, followed by PC2. PC1 and PC2 explains 52.85% of the total variance.

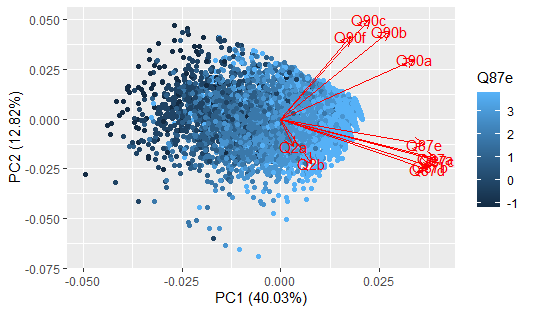
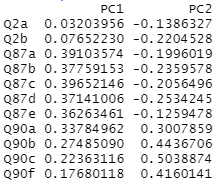


Figure 1.5 & 1.6

Generally, Figure 1.5 shows that the variables in Q87 and Q90a are better represented by PC1 while the remaining variables in Q90 are better represented by PC2. The variables from Q87a to Q90a showed higher loading on PC1 compared to PC2. Figure 1.6 also shows that PC1 explains more of the variables in Q87 as the variables are scaled more by the x-axis while PC2 explains more of the Q90 variables as the variables are scaled more by the y-axis.

## 1.3 K-Means Clustering

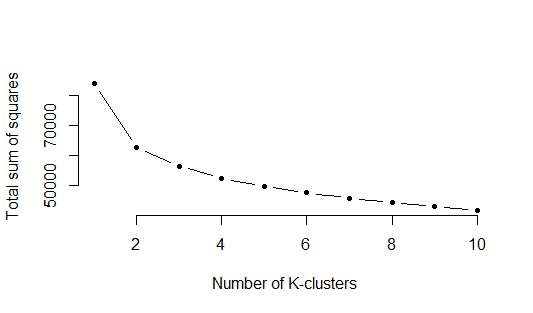


Figure 1.7

Figure 1.7 shows that that the optimal number of K-cluster. From the figure, we can see that the value added for a better model does not increase after the second cluster. Therefore, we can conclude that the optimal number of K-cluster is two.



Figure 1.8

Figure 1.7 shows the mean values of the data from different variables in the two clusters. With a mean score of 1.52 and 1.47, it shows that there is similar number for both genders in both clusters. There is a score of 45.5 and 41.9 for age, which shows that both clusters consist of data from respondents with similar age, with cluster two having more respondents that are slightly younger.

Generally, the scores are higher for cluster 1 as compared to cluster 2 for the variables Q87a to Q90f. Since the responses are ranked from ‘1’ to ‘5’, with ‘1’ being a more positive response and ‘5’ being a more negative response. This means that cluster 2 consists of more respondents who have a more positive mindset towards the European working conditions, as compared to cluster 1.

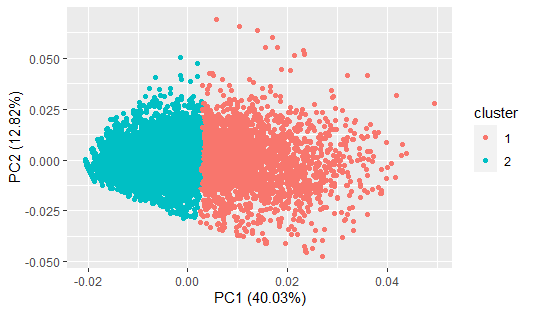


Figure 1.9

Figure 1.9 further explains the two clusters with regards to PC1 and PC2. It shows that PC1 affects the clustering more as the two clusters are separated evenly along the x-axis.

# **Part 2**

## 2.1 Data Visualisation

To provide a general data visualisation, this report will produce a correlation plot after removing the variables G1 and G2.

## 2.1.1 MAT & POR Dataset

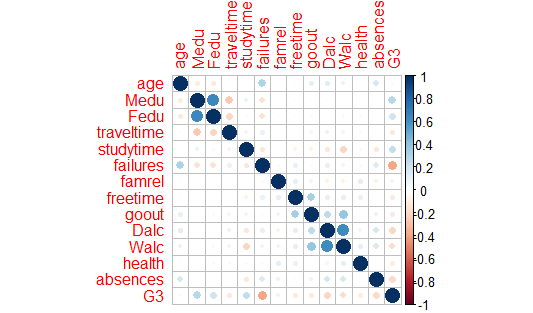
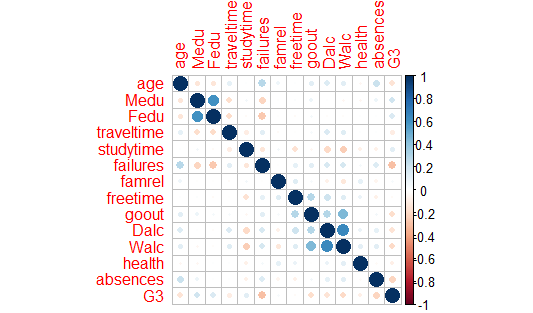


Figure 2.1 & 2.2

Figure 2.1 shows the correlation plot between variables of the Math course dataset. From the figure, we can see that there is little correlation for all the variables except for Mother and Father’s education, as well as workday and weekend alcohol consumption. Figure 2.2 shows the correlation plot between variables of the Portuguese language course. Similarly, it shows that most of the variables are not highly correlated except for Mother and Father’s education, as well as Workday and Weekend alcohol consumption.

## 2.2 Math Regression Model

Given that the data set is not huge, the report will use a 70:30 split ratio for training and testing data respectively.

### 2.2.1 Forward Selection

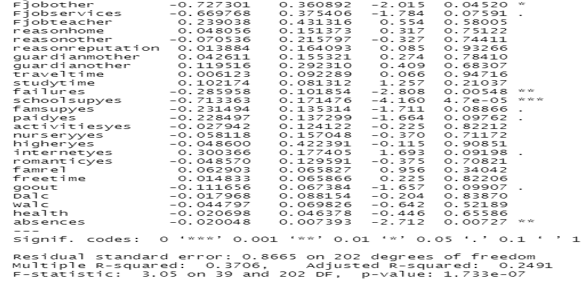


Figure 2.3

From the figure above, we can see that conducting linear regression produced a linear model with an adjusted R-square value of 0.2491. This means that the model explained 24.9% of the response variables. Apart from that, through prediction of the training and test datasets, a Root Mean Square Error of 11.84 and 11.85 was obtained respectively. With a 5% significance level, we can conclude from the model that variables such as Father’s Job (Other), Failures, School Supplementary (Yes) and Absences affect the scores of Math course significantly. This is due to their p-values of less than 0.05.

### 2.2.2 Backward Selection

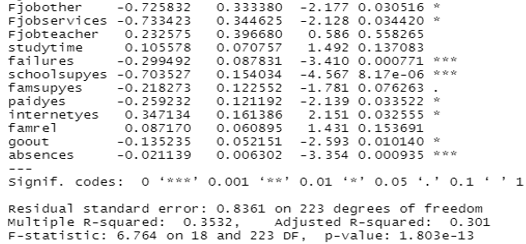


Figure 2.4

By conducting the linear regression with backward selection method, the model produced an adjusted R-square of 0.301. This means that it was able to explain 30.1% of the response variable, which is higher than that of the forward selection method. This model also produced a Root Mean Square Error of 11.84 and 11.86 on both training and test datasets respectively. With a significance level of 5%, we can conclude that variables such as Father’s Job (Other and Services), Failures, School Supplementary (Yes), Extra Paid Classes (Yes), Internet Access (Yes), Going Out and Absences contribute significantly to the scores of Math course.

### 2.2.3 K-Fold Cross Validation

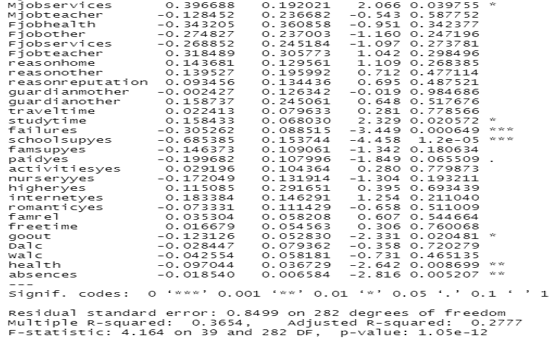


Figure 2.5

An adjusted R-squared value of 0.2777 was obtained from the 10-fold cross validation method, accompanied with a Root mean square error of 11.91 on the test set. This means that the model was able to explain 27.77% of the response variable, which is lower than that of the backward selection method. With a 5% significance level, we can see that Mother Job (Services), Study Time, Failures, School Supplementary (Yes), Going Out, Health and Absences contribute significantly to the students’ Math scores.

### 2.2.4 Conclusion

This report concludes that the backward selection method provides the best model due to the highest adjusted R-squared value. The variables, Failures, School Supplementary (Yes) and Absences contribute significantly to Math grades as they recurred in all three models.

## 2.3 Portuguese Regression

Similarly for the Portuguese language course dataset, the report will use a 70:30 split ratio for training and testing data respectively.

### 2.3.1 Forward

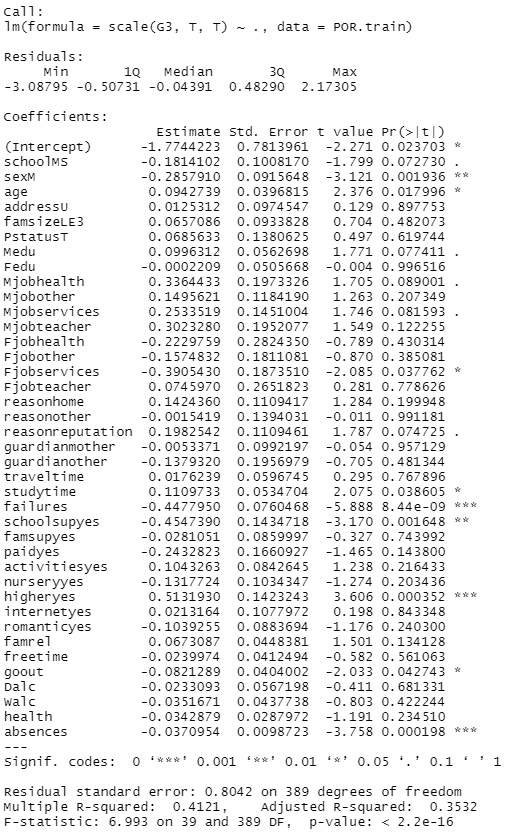


Figure 2.6

Conducting a forward selection method on the linear model produced an adjusted R-squared value of 0.3532 and a Root Mean Square Error of 12.48 and 12.27 from training and testing set respectively. This means that the model was able to explain 35.32% of the total response variable. With a 5% significance level, it shows that the variables, Sex (Male), Age, Father Job (Services), Study Time, Failures, School Supplementary (Yes), Higher Education (Yes), Going Out and Absences contribute significantly to Portuguese language course scores.

### 2.3.2 Backward

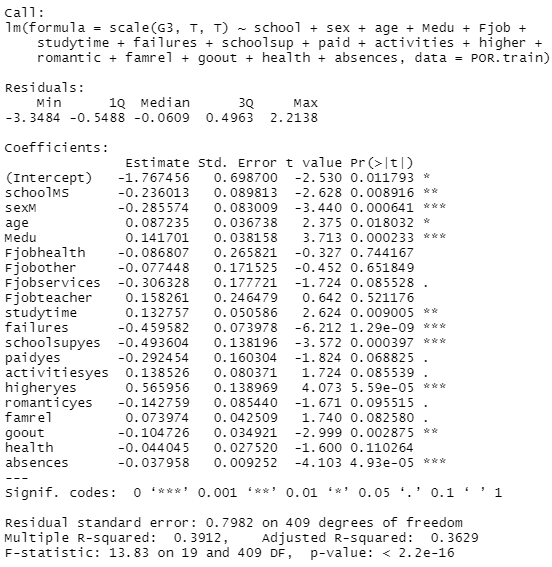


Figure 2.7

The backward selection method produced an adjusted R-squared value of 0.3629, and Root Mean Squared Error of 12.48 and 12.26 for training and test set respectively. This means the model was able to explain 36.29% of the response variable, which is higher than that of the forward selection method. With a 5% significance level, the variables, School (MS), Sex(M), Age, Mother’s Education, Study Time, Failures, School Supplementary (Yes), Higher Education (Yes), Going Out and Absences contribute significantly to Portuguese Language Course scores.

### 2.3.3 K-Fold

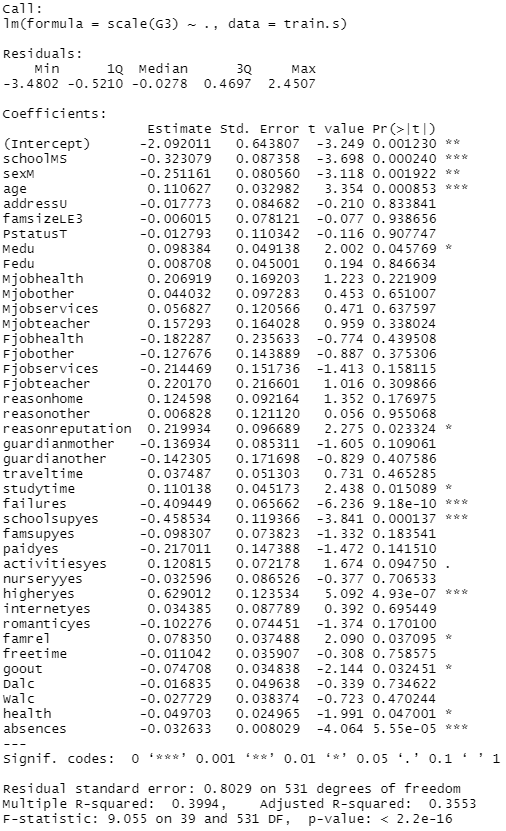


Figure 2.8

The 10-fold cross validation method produced an Adjusted R-squared value of 0.3553, and Root Mean Squared Error of 12.42. This means the model was able to explain 35.53% of the response variable, which is lower than that of the backward selection method. With a 5% significance level, the variables School (MS), Sex(M), Age, Mother’s Education, Reason (Reputation), Study Time, Failures, School Supplementary (Yes), Higher Education (Yes), Family Relationships, Going Out, Health and Absences contributed significantly to the scores of Portuguese Language Course.

### 2.3.4 Conclusion

This report concludes that the backward selection method provides the best model as it has the highest adjusted R-squared value. The variables, Sex(M), Age, Study Time, Failures, School Supplementary (Yes), Higher Education (Yes), Going Out and Absences are the most significant to Portuguese Language grades as they recurred in all three models.

# **Part 3**

## 3.1 Data Visualisation

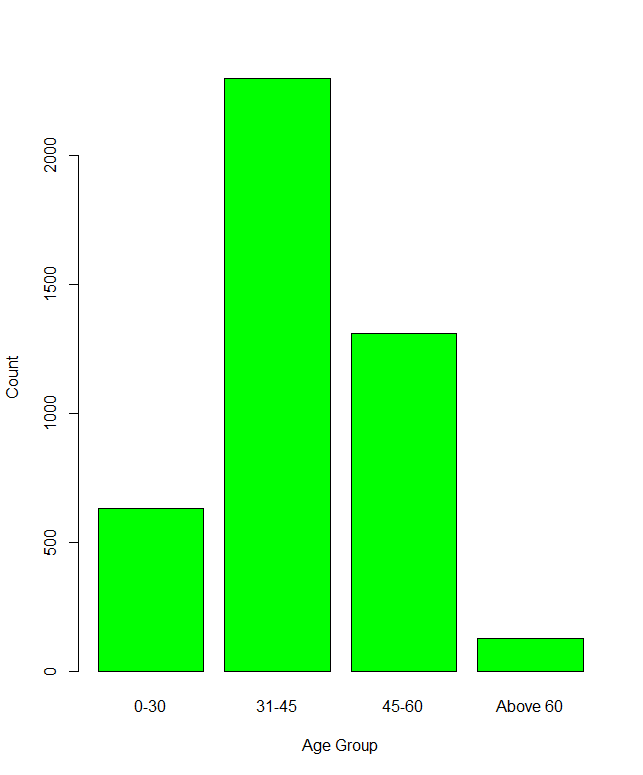


Figure 4.1

Figure 4.1 shows the demographics of the bank client data. The number of clients is highest between the age group of 31 – 45, followed by 45-60, 0-30 and finally aged above 60.

## 3.2 Regression

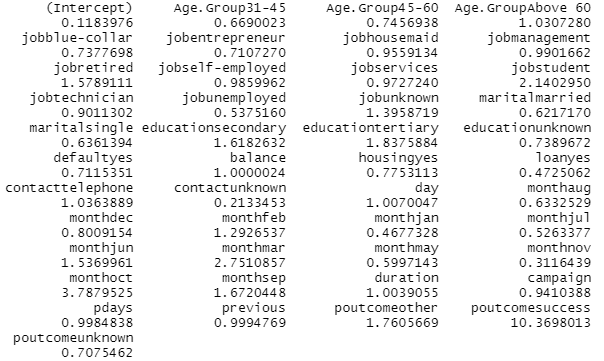


Figure 4.2

Figure 4.2 represents the odds ratio for this model. With the highest value of 10.37, we can infer that the outcome of the previous marketing campaign affects the model most significantly. The value of 10.37 explains that the odds for the outcome of current marketing campaign is increased by 10.37 depending on the outcome of previous campaign.



Figure 4.3

Figure 4.3 shows that the model was able to accurately predict 1173 out of 1303 cases. This means that it has an accuracy of 90.02%

## 3.3 Receiver Operator Curve

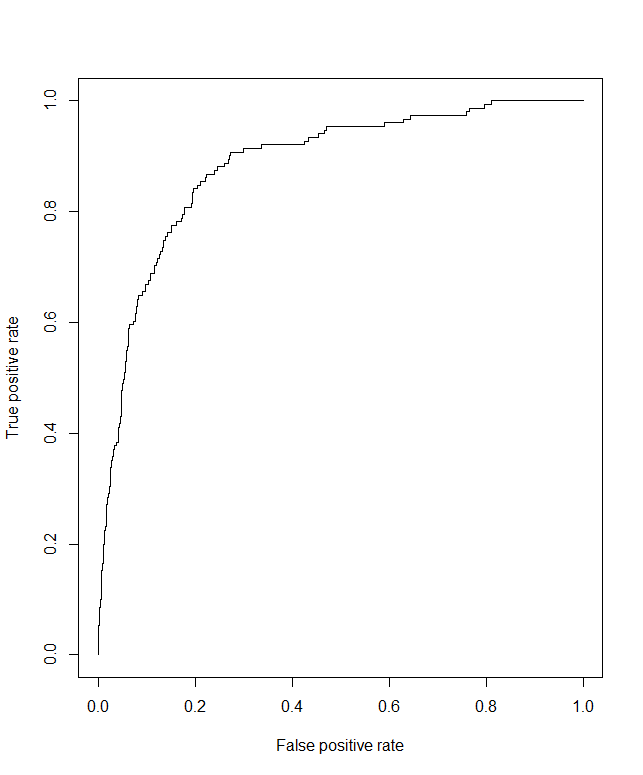


Figure 4.4

Figure 4.4 shows a Receiver Operator Curve. The area under the curve represents the probability that the model is predicting accurately. Therefore, the figure indicates that the model is accurate as the curve is close to the top left corner of the graph.

## 3.4 Random Forest

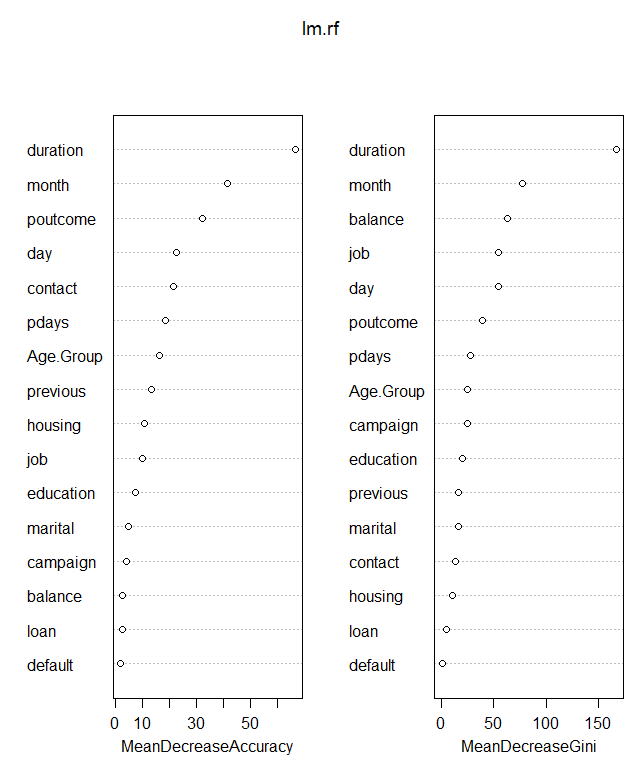


Figure 4.5

Figure 4.5 shows that duration is the most significant factor that affects the prediction of the market campaign outcome, while loan is the least significant factor.