FIT3152 Data science

Assignment 2

By: Matthew Cullen

Objective: The objective of this assignment is to gain familiarity with classification models using R. We will be exploring bank marketing data provided, and using learnt classification models to predict whether or not a client will subscribe.

**Question 1**

By splitting data into both subscribed and not subscribed, and dividing by the total number of examples found it was easy to obtain the proportion of both successful and unsuccessful cases in the data.

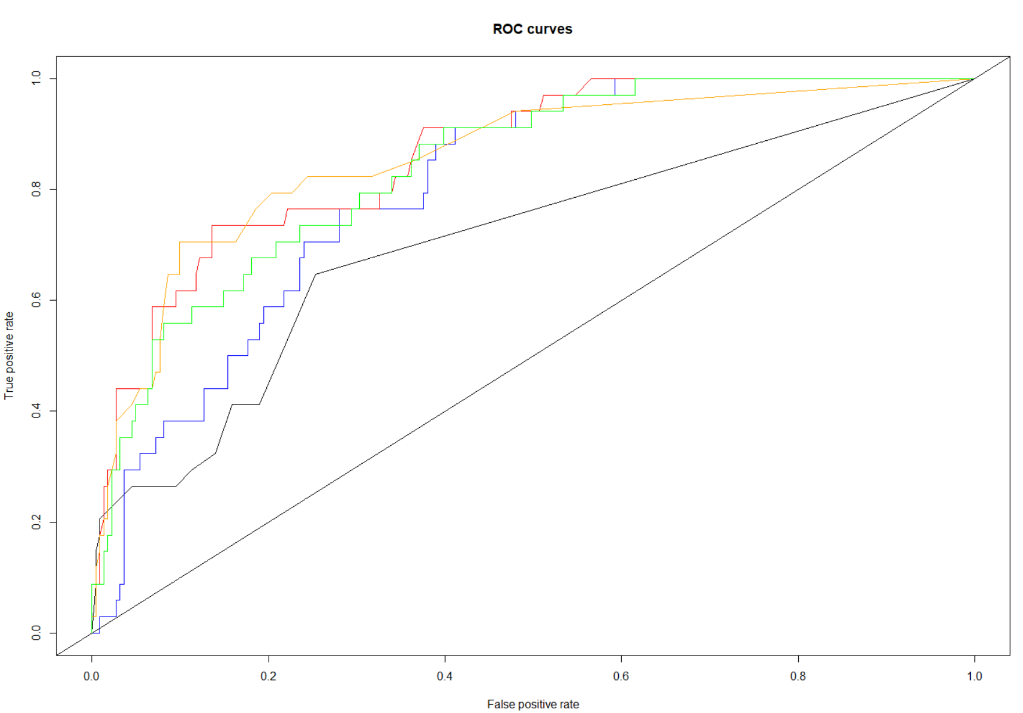
|  |  |  |
| --- | --- | --- |
| **Case type** | **Proportion (%)** | **Number of rows** |
| Successful | 12.35 | 105 |
| Unsuccessful | 87.65 | 745 |

In the code provided, descriptors are explored in both case types, as well as the overall dataset. The data also needed to be cleaned to remove NA values before beginning.

**Question 4**

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Accuracy (value)** | **Accuracy(%)** |
| Tree Model | 0.855 | 85.5% |
| Naïve Bayes Model | 0.796 | 79.6% |
| Bagging Model | 0.878 | 87.8% |
| Boosting Model | 0.886 | 88.6% |
| Random Forest Model | 0.886 | 88.6% |

**Question 5**

ROC curve diagram

Legend:

Black: Tree  
Blue: Naïve Bayes  
Orange: Bagging  
Green: Boosting  
Red: Random Forest

|  |  |
| --- | --- |
| **Classifier** | **AUC** |
| Tree Model | 0.704 |
| Naïve Bayes Model | 0.803 |
| Bagging Model | 0.856 |
| Boosting Model | 0.841 |
| Random Forest Model | 0.861 |

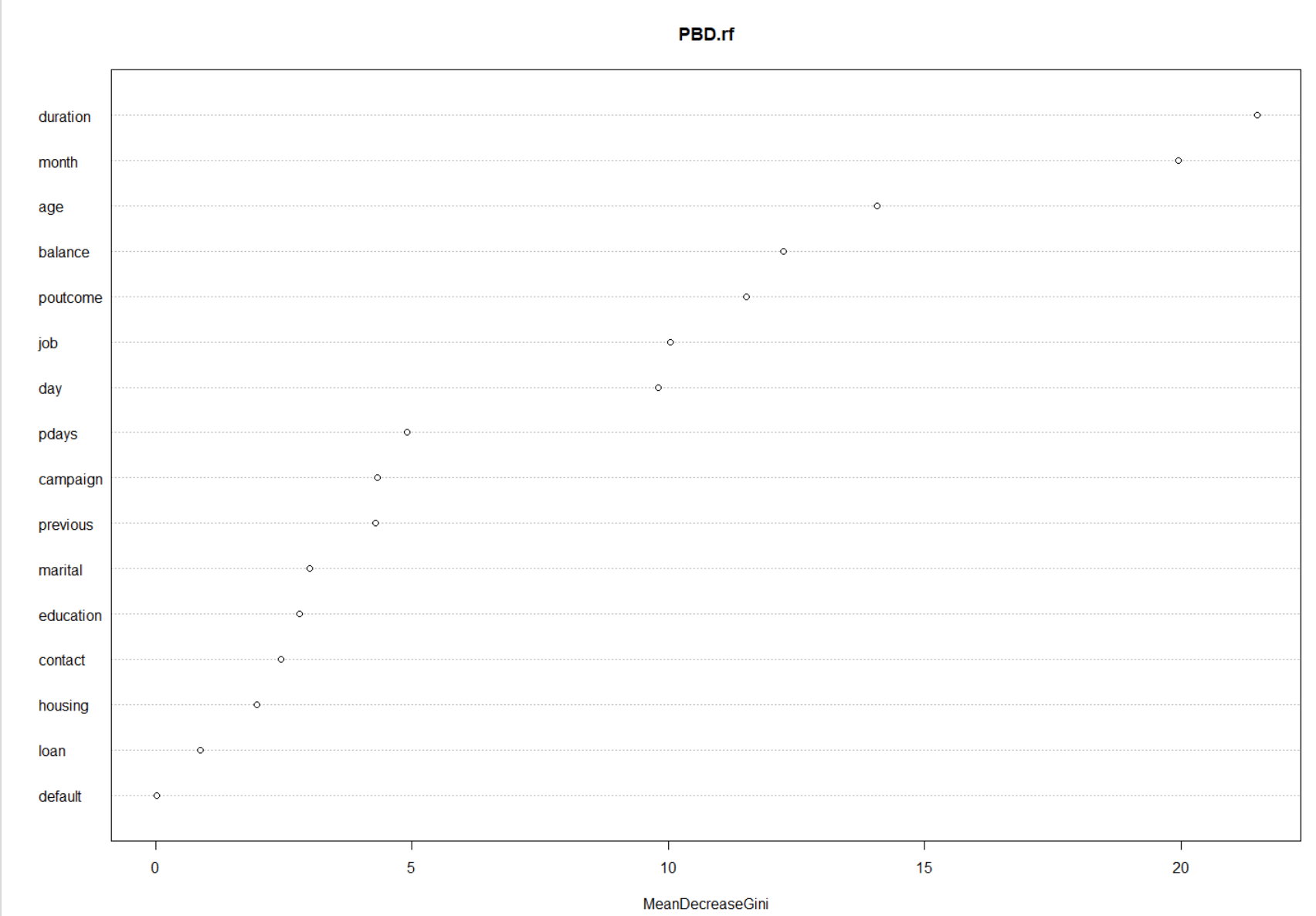
**Question 6**

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Accuracy (value)** | **AUC** |
| Tree Model | 0.855 | 0.704 |
| Naïve Bayes Model | 0.796 | 0.803 |
| Bagging Model | 0.878 | 0.856 |
| Boosting Model | 0.886 | 0.841 |
| Random Forest Model | 0.886 | 0.862 |

From the table, the clear winner out of the classifiers for this data set is the random forest model. Not only does it have the highest accuracy, it also has the highest ‘area under the curve’ value, meaning it has the highest probability to rank a randomly chosen positive example higher than a randomly chosen negative example.

**Question 7**

Using random forest function “varImpPlot”, we can physically see the most important variable tp predict subscription is clearly duration.

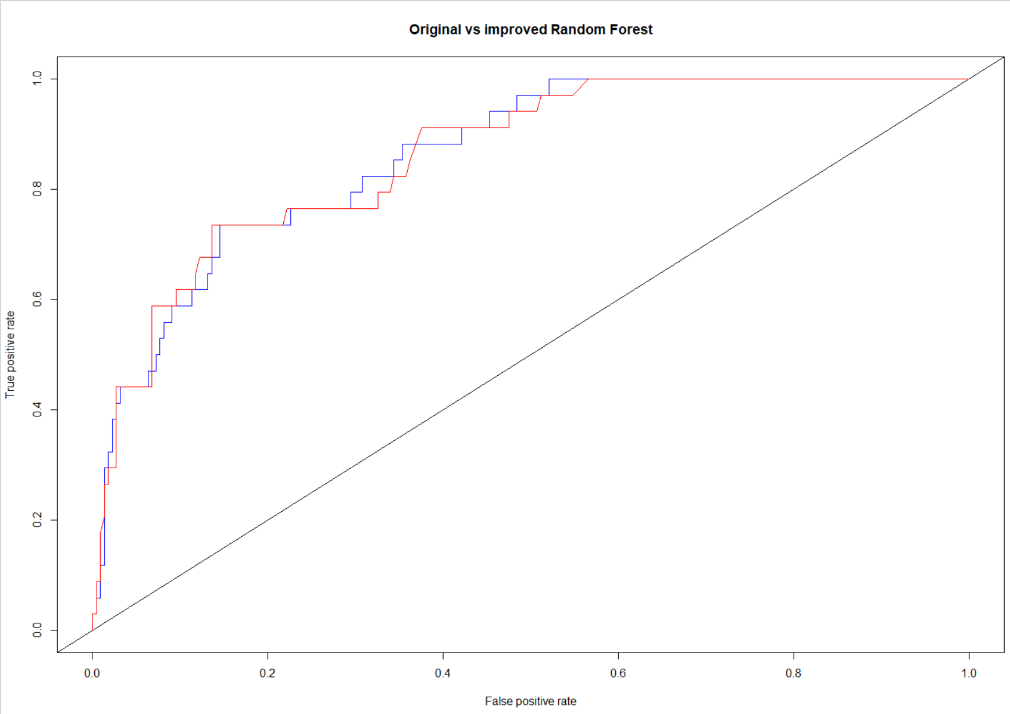


**Question 8**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy (original)** | **Accuracy (improved)** | **Difference (improvement)** | **AUC (original)** | **AUC (improved)** | **Difference (Improvement** |
| Tree Model | 0.855 | 0.875 | 0.02 | 0.704 |  |  |
| Random Forest Model | 0.886 | 0.894 | 0.08 | 0.862 | 0.862 | 0.01 |

Tree model: pruned using misclass with best value of 2  
Random forest model: Ntree value set to 10000, mtry set to 4.

For random forest, there was a slight improvement if 0.08 in accuracy, and a slight improvement of AUC by 0.01. Here is a ROC graph comparing the two.



Legend:  
Blue: Improved  
Red: Original