Monash University Malaysia
BSc Computer Science
Assignment # 2

Bank Marketing Data

FIT3152 Data Analytics - Report

Prepared By:
Bazil Muzaffar Kotriwala

Date:
08/10/17

Table of Contents

Introduction

The purpose of this report is to analyse the Bank Marketing data set and predict whether a client will subscribe to a term deposit (yes or no) using various different classification models such as Decision Trees, Naive Bayes, Bagging, Boosting and Random Forest. We will be identifying the single best classifier using the accuracy and area under the curve (auc) calculated for each of the classifiers.

Exploring and Cleaning the Dataset

The Bank Marketing data set consists of a total of 40,000 observations with 17 attributes. 7 of these attributes are numerical (real-valued) and rest of the 10 are categorical. We create a subset from this data set by taking a random sample of 1000 observations with 17 attributes. We use this sample subset to do our analysis and to create our classification models.

The 'subscribed' attribute is the response variable (dependent variable) which will be crucial to our analysis. Therefore, to start off we find the proportion of successful to unsuccessful cases. The proportion turns out to be 0.1036 : 0.8964 which indicates that only 10.36% of total clients subscribe for a term deposit whilst 89.64% of total clients do not subscribe for a term deposit.

The sample subset of the data contains N/A value observations for both the real-valued and categorical attributes. The N/A values for some of the real-valued attributes have been completed such as 'age', 'pdays' and 'prev'. For the age column, the mean age is found for all 1000 clients in the subset, and a total of 10 observations in the 'age' column containing N/A are filled with the mean age.

Furthermore, there is a relation between 'pdays' and 'prev' since a value of -1 for 'pdays' indicates that the client has never been previously contacted which means that the 'prev' will be 0 as 'prev' tells us how many times the client has been contacted before. Therefore, all 'pdays' observations with -1 and 'prev' with N/A had the N/A replaced with 0 whilst all 'pdays' observations with N/A and 'prev' as 0 had the N/A replaced with -1. A total of 8 observations were completed for 'prev' and 'pdays' cumulatively.

All other observations of attributes containing N/A were omitted from the sample subset. A total of 149 observations were removed and 18 were completed. This left the subset with a total of 851 observations. The 'dplyr' package was used to create a subset of real-valued attributes and their descriptions were found such as mean, median, standard deviation etc.

```
> library(dplyr)
> PBD_rva = select(PBD, age, balance, day, duration, campaign, pdays, previous)
> summary(PBD_rva)
      age          balance            day          duration       campaign        pdays          previous      
 Min.   :20.0   Min.   :-1212   Min.   : 1.0   Min.   :   4   Min.   : 1.0   Min.   : -1   Min.   : 0.00  
 1st Qu.:33.0   1st Qu.:   68   1st Qu.: 8.0   1st Qu.:  98   1st Qu.: 1.0   1st Qu.: -1   1st Qu.: 0.00  
 Median :39.0   Median :  456   Median :16.0   Median : 169   Median : 2.0   Median : -1   Median : 0.00  
 Mean   :41.4   Mean   : 1213   Mean   :15.9   Mean   : 257   Mean   : 2.9   Mean   : 37   Mean   : 0.56  
 3rd Qu.:49.0   3rd Qu.: 1322   3rd Qu.:22.0   3rd Qu.: 314   3rd Qu.: 3.0   3rd Qu.: -1   3rd Qu.: 0.00  
 Max.   :86.0   Max.   :35589   Max.   :31.0   Max.   :2621   Max.   :29.0   Max.   :842   Max.   :51.00  
```

Fig 1.0: Summary of Real-Valued Attributes

```
PBD_rva %>%
  summarise_all(funs(sd(.)))
  age balance   day duration campaign pdays previous
10.87    2478 8.286      273    3.214 99.42    2.557
```

Fig 1.1: Standard Deviation of Real-Valued Attributes

From Fig 1.0 and Fig 1.1, we can make some noteworthy observations. The three attributes 'balance', 'duration' and 'pdays' have a large standard deviation which shows that spread of values from the mean is large for each attribute respectively. Therefore, this indicates that our data for these respective attributes is skewed. Thus, median would be a better measure of central tendency as opposed to the mean. Hence, the mean value for these three attributes may be misleading.

Classification Models

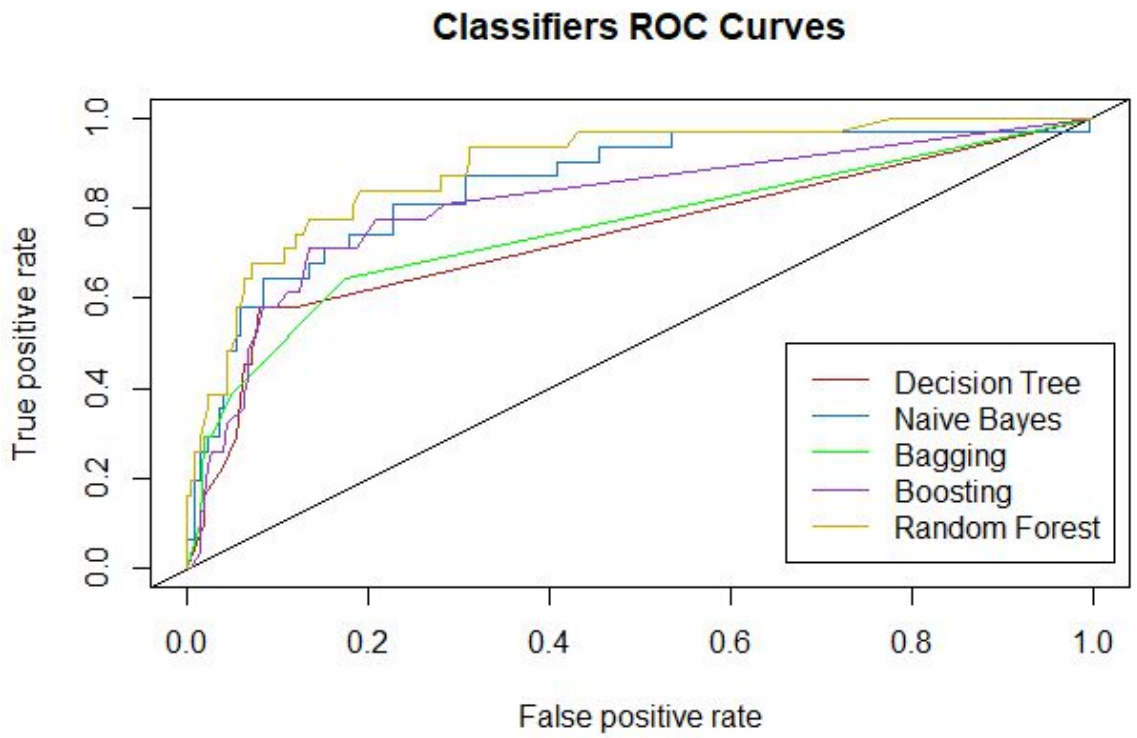
The sample subset was divided into a 70% training and 30% test set. The training set contains a total of 595 observations and the test set contains a total of 256 observations. The training set was used to build the classification models whilst the test set was used to make predictions and report the accuracy of each model. Five different classification models were created which were Decision Trees, Naive Bayes, Bagging, Boosting and Random Forest. For each of these classification models, 'subscribed' is the response variable and all the other input variables are the predictors.

For all the classification models, using the test data all the test cases were classified as subscribed either 'yes' or 'no'. A confusion matrix was created for each model and the accuracy was found. Fig 2.0 below shows that accuracy found for each classification model using the test data.

Classifiers	Accuracy (Test Data)	Area Under Curve (AUC)
Decision Tree	87.11%	0.7393
Naive Bayes	89.06%	0.8520
Bagging	89.06%	0.7558
Boosting	87.89%	0.8138
Random Forest	90.23%	0.8910

Fig 2.0: Classifier Results Table

Random Forest has the highest accuracy of 90.23% out of all the classification models constructed. Furthermore, using the test data, for each model the confidence level of prediction of subscribing 'yes' or 'no' was found. Using this, an ROC curve was plotted for each classifier which is shown by the graph below.



The area under the ROC curve for each classifier was calculated and tabulated into Fig 2.0 along with the classifiers respective accuracy. The area under the curve depicts the 'goodness' of each respective classifier. Random Forest has the highest 'True Positive Rate' at each 'False Positive Rate' value throughout the range from 0.0 - 1.0 amongst all classifiers. Thus, Random Forest has the highest area under the curve of 0.8910 out of all the classifiers constructed depicting 'excellent discrimination'. Since, Random Forest has the highest accuracy as previously noted, along with the highest area under the curve, we can conclude that it is the best classifier out of all the ones which we have constructed.

The most important variables found across all classification models constructed on whether or not a customer subscribes consisted of 'duration', 'month', 'balance', 'job' and 'age'.

Optimised Random Forest

Once we had all the classification models, we decided to experiment on an existing classifier to make it the best classifier out of all the ones we currently have. This would require the updated classifier to have an accuracy higher than any of the existing classifiers. The classifier we chose to improve is Random Forest which is currently our best classifier with the highest accuracy and area under curve.

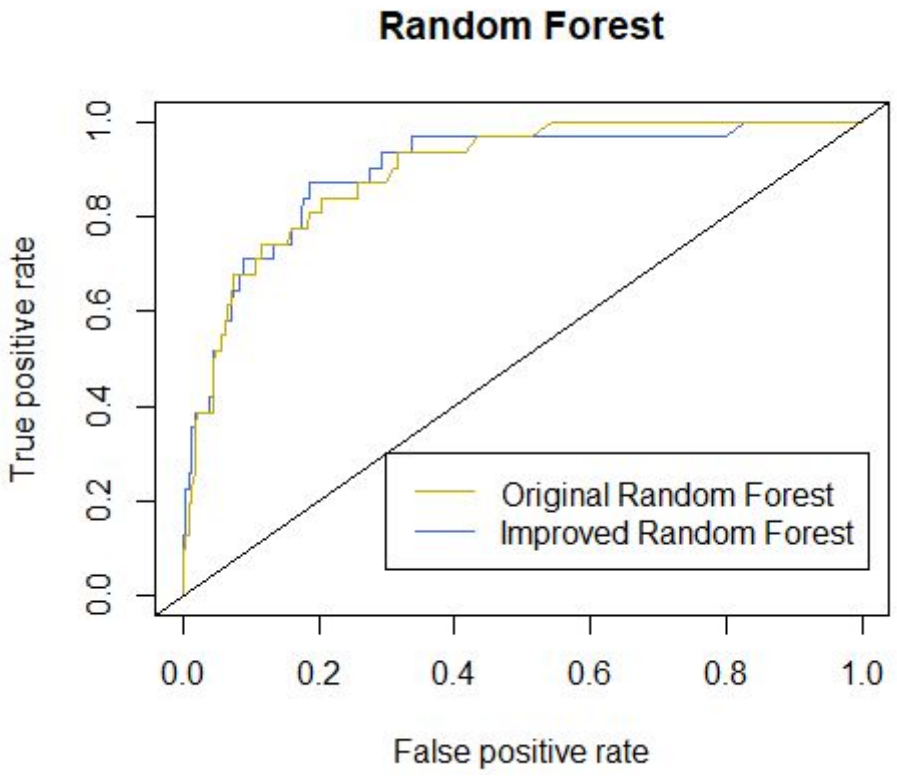
To optimise the Random Forest classification model we adjusted its parameters such as 'ntree' and 'mtry'. These two parameters are crucial to improving the accuracy of Random Forest. 'mtry' denotes the number of variables selected at each split whilst 'ntree' denotes the number of trees to grow. Using the tuneRF function we found the optimal 'mtry' value which was found to be 4 i.e. the value which minimised the Out-of-bag error. The 'ntree' value used was 301. Furthermore, increasing the number of trees increases the accuracy upto a certain point and is computationally expensive. Therefore, we need to find the optimal number of trees which were 301 in our case.

After adjusting these parameters using mtry as 4 and ntree as 301, we achieved an accuracy of 91.02% with an area under curve of 0.8871. This was a higher accuracy than what we had previously computed for any classification model. The figure below depicts the results.

Classifiers	Accuracy (Test Data)	Area Under Curve (AUC)
Original Random Forest	90.23%	0.8910
Improved Random Forest	91.02%	0.8871

Fig 2.1: Random Forest Comparison Table

The accuracy increased by 0.79% and the area under the curve decreased by 0.0039. We are going to visualise the difference between the two Random Forest models using ROC curves and area under the curve measures to demonstrate the improved accuracy and performance of the updated Random Forest classifier.



The Improved Random Forest has a higher ‘True Positive Rate’ at each ‘False Positive Rate’ than the Original Random Forest for most of the values between 0.0 - 1.0 apart from when the ‘False positive rate’ is between 0.5 - 0.8.

Conclusion

After our exhaustive analysis and construction of various classification models, we have found that an optimised Random Forest is the best classifier to use. This has been demonstrated by accuracy measures such as ROC and Area under curve (AUC). This classifier predicts with a 91.2% accuracy as to whether a client will subscribe (yes or no) to a term deposit.