BANK MARKETING

Tamanna Baig (C13768206)

1. Introduction

The Bank marketing dataset can be used to predict and forecast the revenue model and subscribers for various campaigning strategies. This prediction model will assist banks to better understand what type of marketing strategy to use and what customer groups to target. The main classification goal is to predict which customer will subscribe a term deposit (response variable - y).

In this project, I have performed two important and famous classification model, Logistic Regression and Random forest classifier. Using the classification performed on the dataset, a confusion matrix is drawn to evaluate the accuracy, precision and recall.

2. Dataset

The dataset used here, is the Bank Marketing dataset from S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014. This dataset consists of **45211** observations with 17 predictors. The response variable (y) denotes if the client subscribes a term deposit (yes) or not (no).

The following are the 17 predictors that are included in the dataset:

age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, y

- *Predictors:* Age (age of the client), Job (Profession of the client), Marital (Marital status), Default (Does customer default on credit?), Balance (Overall balance in the account), Housing (Has housing loan?), Loan (Does the client have a loan?), Contact (How to contact the customer?), Day (Last contact day of the week), Month (Which month was the last contact made in?), Duration (Duration of last call, in seconds), Campaign (Number of clients contacted during this campaign), pdays (Number of days elapsed since last contact), Previous (Number of campaigns performed prior to this campaign), poutcome (What was the outcome of the previous campaign)
- *Response:* y (Has the client subscribed to the term deposit?)

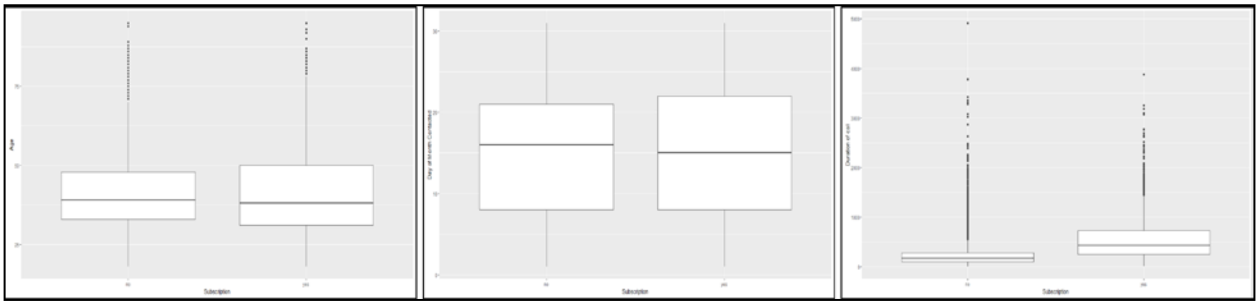
This dataset is made available on UCI Machine Learning Repository:

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Citation: [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

3. Exploratory Data Analysis (EDA)

Before fitting the model on the dataset, I split the data into a training dataset and a test dataset with a 70:30 split. To better understand the data and the relationship between the predictors and the response, I have plotted boxplots to depict trends and they are as follows:

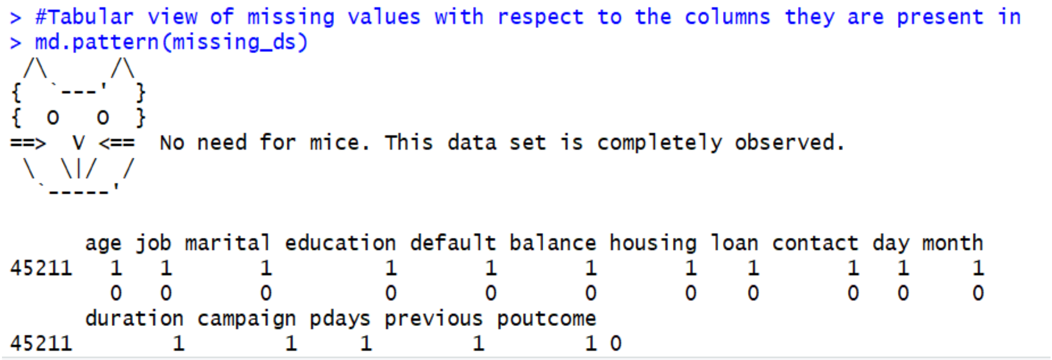


Age vs Subscription Day vs Subscription Duration vs Subscription

- *Age versus Subscription* (y): From the above boxplot, it can be observed that 50% of clients who declined lie in the age range of about 32-48 years whereas 50% of those who subscribe fall in the range of 31-50 years
- *Day of month versus Subscription* (y): 50% of subscribers lie in the range of 8th day-22nd day and 50% of declined lie in the range of 8th day-21st day. Considering that these two boxplots overlap almost exactly, we cannot draw any conclusions.
- *Duration of call versus Subscription* (y): 50% of those who subscribe have a call duration of 300-700 seconds with the bank representative whereas 50% of those who do not subscribe have a call duration of 100-300 seconds.

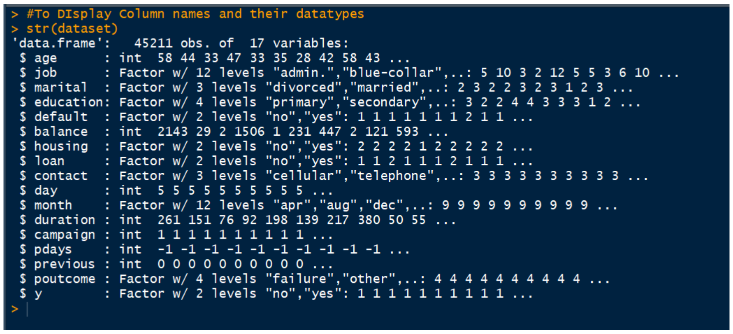
The response variable i.e. y (Subscription of a deposit) comprises of binary results i.e. Yes (1) or No (0) which can be inferred as: Client subscribes a deposit or Client does not subscribe a deposit.

I also used the library MICE, to find and impute missing data, if any. Upon execution, the mice function shows that the dataset does not contain any NA or Missing data.



```
> #Tabular view of missing values with respect to the columns they are present in
> md.pattern(missing_ds)
 /\     /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'

      age job marital education default balance housing loan contact day month
45211   1   1       1         1       1       1       1    1       1   1     1
        0   0       0         0       0       0       0    0       0   0     0
      duration campaign pdays previous poutcome
45211        1        1     1        1        1 0
```

During the EDA, I also found that the dataset contains 10 categorical variables and 7 numerical variables. I used the str function to do this, the result is shown below:

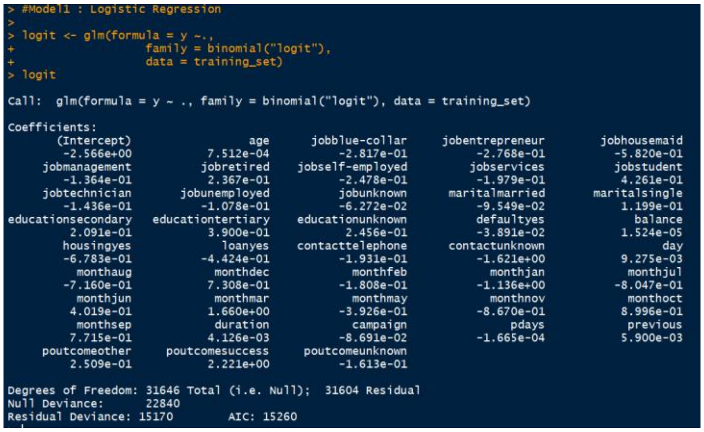


```
> #To DIsplay Column names and their datatypes
> str(dataset)
'data.frame':	45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
 $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
 $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
 $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
 $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
> 
```

4. Model selection and Validation

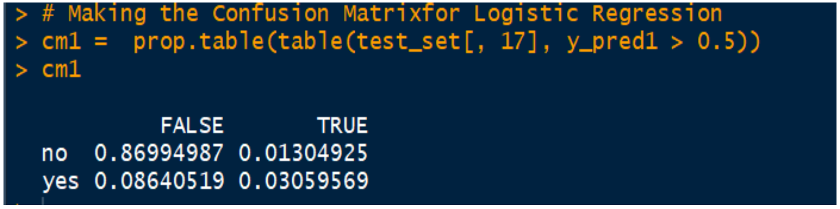
The Bank Marketing dataset primarily fits best with a Classification model as the response variable has a binary outcome. I have fit two models for this dataset: one is a logistic regression classifier and the other is the Random forest algorithm.

- **Logistic Regression**: Logistic regression is a method for fitting a regression curve, y = f(x), when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical or a mix of both. The dataset used is the best example for mix of both predictors. Also, here the response variable y, is binomial, hence the model can also be called as binary logistic regression model.

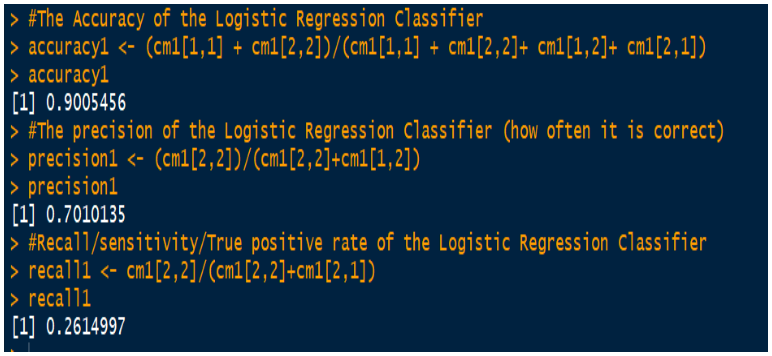


```
> #Model1 : Logistic Regression
> 
> logit <- glm(formula = y ~.,
+                  family = binomial("logit"),
+                  data = training_set)
> logit

Call:  glm(formula = y ~ ., family = binomial("logit"), data = training_set)

Coefficients:
      (Intercept)                age      jobblue-collar     jobentrepreneur        jobhousemaid
       -2.566e+00          7.512e-04          -2.817e-01          -2.768e-01          -5.820e-01
    jobmanagement         jobretired   jobself-employed         jobservices          jobstudent
       -1.364e-01          2.367e-01          -2.478e-01          -1.979e-01           4.261e-01
    jobtechnician      jobunemployed          jobunknown      maritalmarried       maritalsingle
       -1.436e-01         -1.078e-01          -6.272e-02          -9.549e-02           1.199e-01
educationsecondary  educationtertiary    educationunknown          defaultyes             balance
        2.091e-01          3.900e-01           2.456e-01          -3.891e-02           1.524e-05
       housingyes            loanyes    contacttelephone      contactunknown                 day
       -6.783e-01         -4.424e-01          -1.931e-01          -1.621e+00           9.275e-03
         monthaug           monthdec            monthfeb            monthjan            monthjul
       -7.160e-01          7.308e-01          -1.808e-01          -1.136e+00          -8.047e-01
         monthjun           monthmar            monthmay            monthnov            monthoct
        4.019e-01          1.660e+00          -3.926e-01          -8.670e-01           8.996e-01
         monthsep           duration            campaign               pdays            previous
        7.715e-01          4.126e-03          -8.691e-02          -1.665e-04           5.900e-03
    poutcomeother    poutcomesuccess     poutcomeunknown
        2.509e-01          2.221e+00          -1.613e-01

Degrees of Freedom: 31646 Total (i.e. Null);  31604 Residual
Null Deviance:	    22840
Residual Deviance: 15170 	AIC: 15260
```

This logit model gave an accuracy of 0.90.

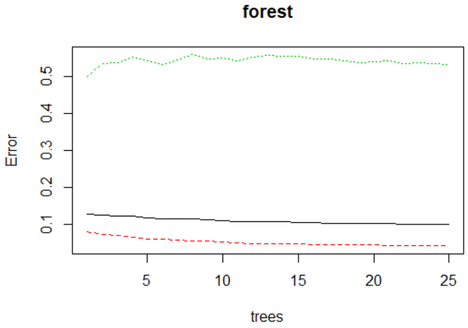


```
> # Making the Confusion Matrixfor Logistic Regression
> cm1 =  prop.table(table(test_set[, 17], y_pred1 > 0.5))
> cm1
     
           FALSE       TRUE
  no  0.86994987 0.01304925
  yes 0.08640519 0.03059569
```

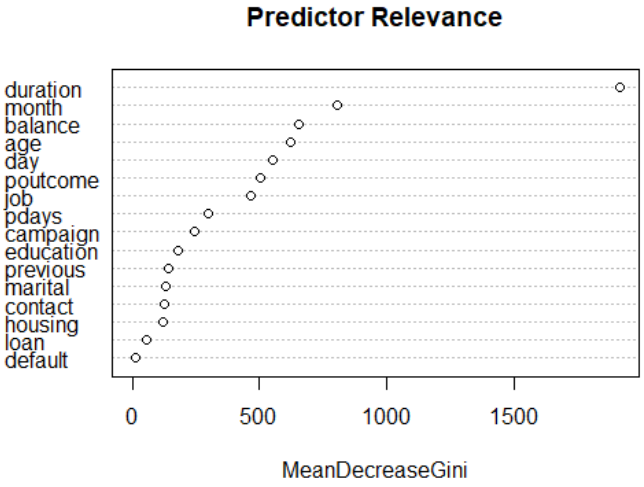


```
> #The Accuracy of the Logistic Regression Classifier
> accuracy1 <- (cm1[1,1] + cm1[2,2])/(cm1[1,1] + cm1[2,2]+ cm1[1,2]+ cm1[2,1])
> accuracy1
[1] 0.9005456
> #The precision of the Logistic Regression Classifier (how often it is correct)
> precision1 <- (cm1[2,2])/(cm1[2,2]+cm1[1,2])
> precision1
[1] 0.7010135
> #Recall/sensitivity/True positive rate of the Logistic Regression Classifier
> recall1 <- cm1[2,2]/(cm1[2,2]+cm1[2,1])
> recall1
[1] 0.2614997
```

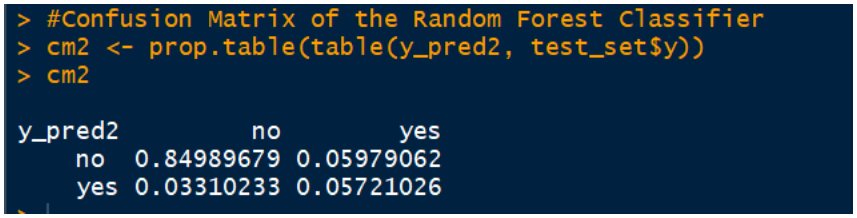
- **Random Forest:** The second model I fit to the dataset is, random forest model using the randomForest function in R. Considering that the data consists of binary data, random forest model is another perfect fit. A random forest basically constructs a set of decision trees and links them together to form a forest. The majority vote of the result of each tree is considered. Here, I used value of ntree = 25



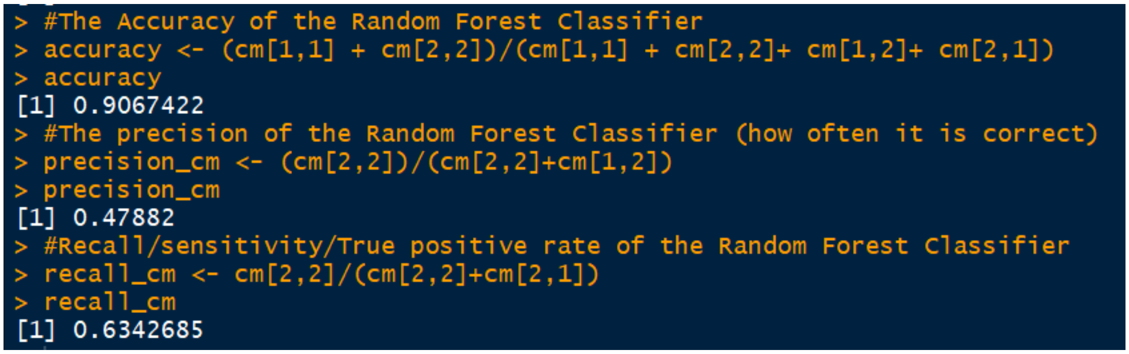
The model is fit using all the predictors present in the dataset. I used the varImpPlot() function to plot a graph that denotes the most relevant predictors for the model and it is visualized as follows:



Random Forest when predicted using the test data produced an accuracy of approximately 0.91



```
> #Confusion Matrix of the Random Forest Classifier
> cm2 <- prop.table(table(y_pred2, test_set$y))
> cm2

y_pred2          no        yes
    no   0.84989679 0.05979062
    yes  0.03310233 0.05721026
```



```
> #The Accuracy of the Random Forest Classifier
> accuracy <- (cm[1,1] + cm[2,2])/(cm[1,1] + cm[2,2]+ cm[1,2]+ cm[2,1])
> accuracy
[1] 0.9067422
> #The precision of the Random Forest Classifier (how often it is correct)
> precision_cm <- (cm[2,2])/(cm[2,2]+cm[1,2])
> precision_cm
[1] 0.47882
> #Recall/sensitivity/True positive rate of the Random Forest Classifier
> recall_cm <- cm[2,2]/(cm[2,2]+cm[2,1])
> recall_cm
[1] 0.6342685
```

5. Conclusion

In conclusion, the logistic regression produced a satisfactory model which assists in predicting whether a client subscribes to the deposit or not with an accuracy of 90%. The second model I fit to the data is the Random Forest which assists in providing information regarding the predictors with the highest relevance. The random forest prediction produced an accuracy of 91% with iteration of 25. The model identified that the duration of call had the highest influence on the prediction. I also increased the iterations with 50, 100, 500, the accuracy only increased by 0.01 in each of the cases. Hence it would be safe to use 25 trees instead of exhausting the model with greater number of tree iterations only for such insignificant progress. Based on these results the Portuguese retail bank can channel their resources into relevant aspects to increase the number of clients who subscribe a deposit.

Further study and analysis can include the using XGBoost library and the XGBoost classifier to see how much accuracy can be increased by this model.