Customer prediction for term deposit based on marketing campaign

Data Mining Applications

Final Project Report (ALY6040)

(March 27, 2018)

Team Members

Jainik Rajiv Majmudar

Joydeep Singh

Vikas Warudkar

Professor:

Professor Sergiy Shevchenko

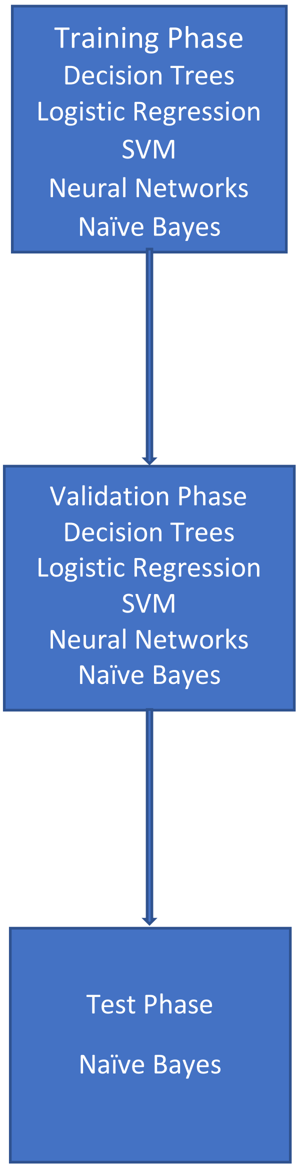
Table of Content

Introduction:

Banking Industry is an important factor for the economy of any nation. It was the fundamental cause of the worldwide economic crisis in 2008 due to the bad loan deposit. The project points to find the potential clients(customers) who are probably going to take term deposit based on the marketing campaigns done by Portuguese Banking Industry. The marketing campaigns efforts were performed via telephone calls. More than one contacts to a similar customer were required, with a specific end goal to get to if a customer would take term deposit. The dataset contains 41188 customer records with 20 predictor variables ordered by date from May 2008 to November 2010. The data was partitioned to 60% of training 20% validation and 20% test data.

Data Model:

Machine learning Techniques and the process given below where used to determine the final model of the marketing campaign data set of Portuguese Banking Industry. The overview and the methods used are described below.



Data Description and Processing:

Here in the dataset, there were lots of variable with their different type of categories and we used that categories and transformed the categories to visualize the dataset. In variable age, there was only one category which is numeric so there was no transformation in that variable (not transformed). Where as in other variable called Education there where multiple of categories. They were categorical with Basic 4y, basic 6y, high school, illiterate, professional course, university degree, unknown and their transformed categories where Basic Education in that following categories Basic 4y, basic 6y, illiterate, whereas in High School basic 9y, high school and unknown and lastly Uni & pro (Professional Course and university Degree). With this and many more variables, types of category and transformed category are described below in the table which were there in the marketing campaign data set.

Variable Name	**Type of category**	**Transformed categories**
Age	Numeric	Not transformed
Job	Catgegorical with admin, Blue collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown	**High-Pay-Job**(Admin, Blue-collar, management, services) **self-pay-job**(self-employed,technician, enterprenuer**)** **No-pay-Job**(student, technician, enterprenuer)
Maritial	Catgegorical with 'divorced','married','single','unknown'	Not transformed
Education	Catgegorical with Basic 4y, basic 6y, high school, illiterate, professional.course, university degree, unknown	**Basic Education**(Basic 4y, basic 6y, illiterate) **High School**(basic 9y, high school and unknown) **Univ&pro** (Professional Course and university Degree)
Default	Categorical-Yes, No and Unknown	Not tranformed
Housing	Categorical-Yes, No and Unknown	Not tranformed
Laon	Categorical-Yes, No and Unknown	Not tranformed
Contact	Categorical – Cellular, telephone	Cellular as 1 and telephone as 0
Month	Categorical- Jan-Nov	Q1, Q2, Q3,Q4
Day_of_Week	Categorical- Mon-Sun	Mon-Sun

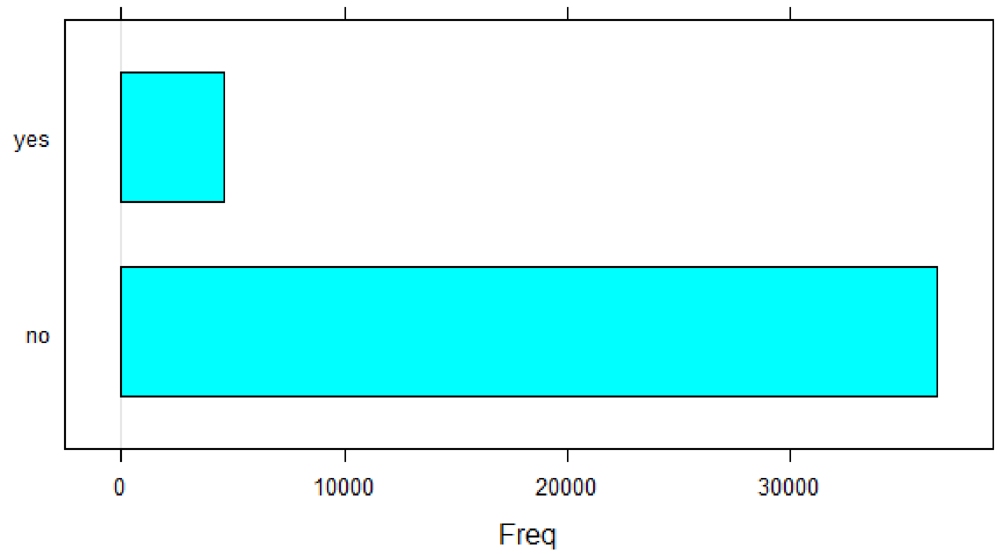
Duration	Numerical variable Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.	Removed from the dataset as it us not used for predicting task.
Campaign	Numeric	Not tranformed
Pdays	Numeric	Transformed to new and previous customer. New Customers are 0 and Old customers as 1
Previous	Numeric	Not tranformed
Poutcome	Categorical Failure, Non-Existent, Success	Not tranformed
Emp.Var.rate	Numeric- Employment Variation Index	Not tranformed
Cons.price.Idx	Numeric-Consumer Price Index	Not Transformed
Cons.Conf.indx	Numeric-Consumer Confidence Indx	Not Transformed
EuriBor3M	Numeric- Eurobor 3month Rate	Not transformed
Nr.employed	Numeric-Number of employees	Not transformed.
Y (Response Variable)	Categorical- Yes(Accepted term Deposit) No(Didn't Accept)	Transformed to 0 and 1, 1 took deposit 0 didn't take

Data Visualization:

Visualization of the dataset in which we visualized the data frequency where the customer is going to buy the term deposit on the phone called made.

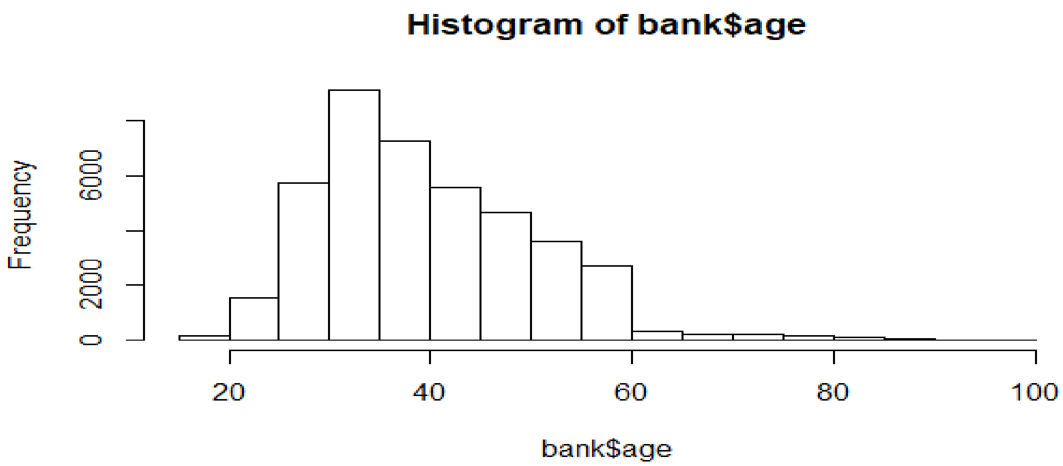
Output variable (Desired Target):

The customer subscribed a term deposit? (Binary: “yes” or “no”)



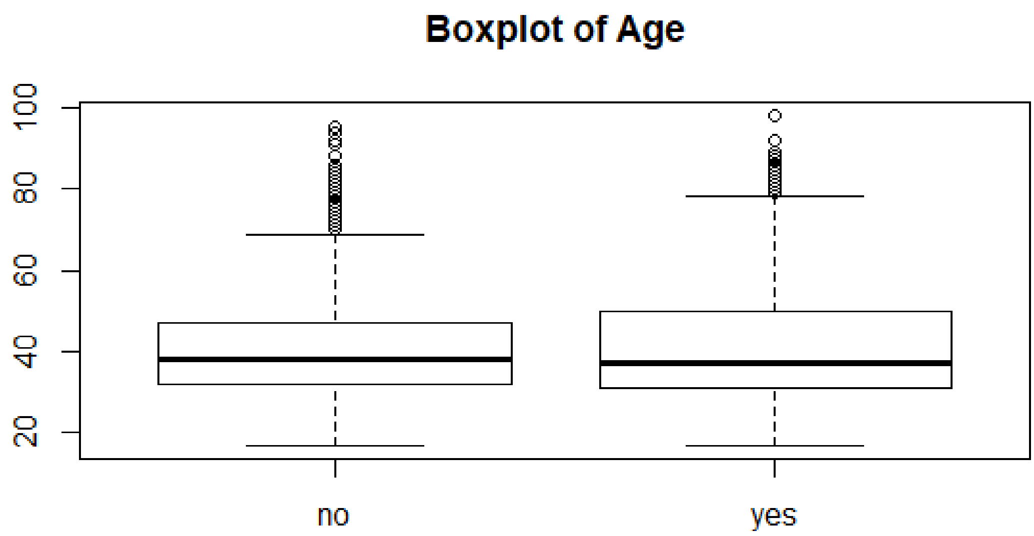
The plot of the data shows that most of the customer didn’t subscribed to the term deposit.

The Distribution of Age:



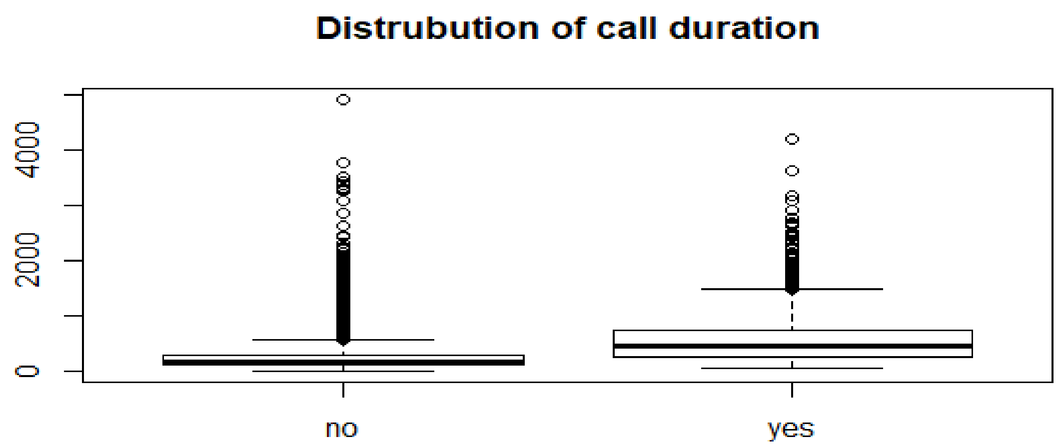
This frequency plot graph shows that mostly bank has contacted the customer with the age between 20 to 60. And from the plot we can also observe that age group between 30 to 35 where contacted the most.

The Box Plot of Age:



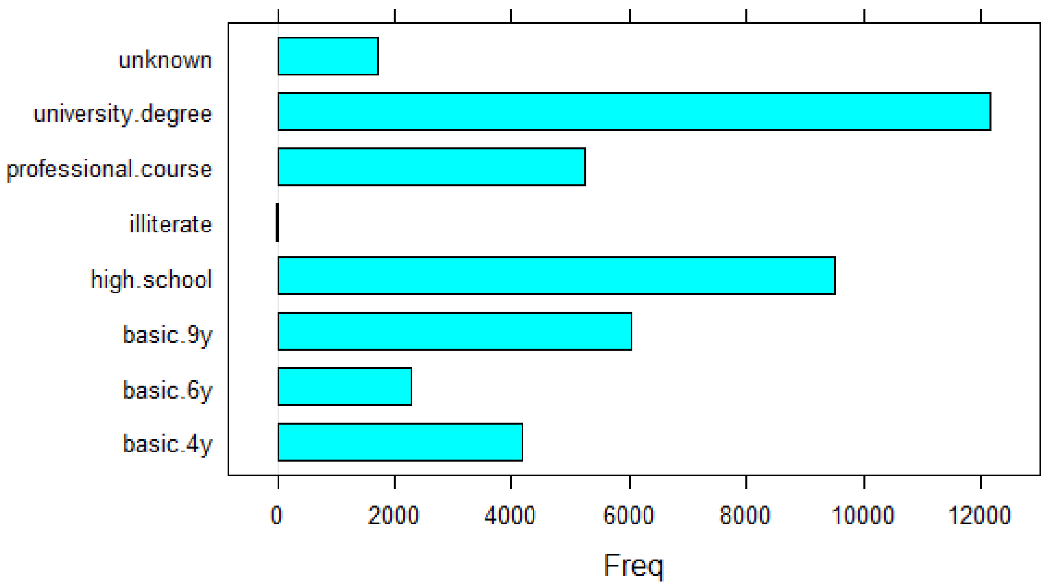
The Box Plot of the dataset of age shows the distribution of age compared to term deposit. The outliers in box plot are more for the customers which are not taking the term deposit from the compared customer which are taking the term deposit. The whisker of No is almost equal that means customer not taking the term deposit are equally distributed in all the age group. In yes, the whisker is more in the upper hand, so we can say that the customer between that range prefer to take term deposit compared to the other. Means that age group is mid-age which are taking the term deposit.

The Distribution of Call Duration:



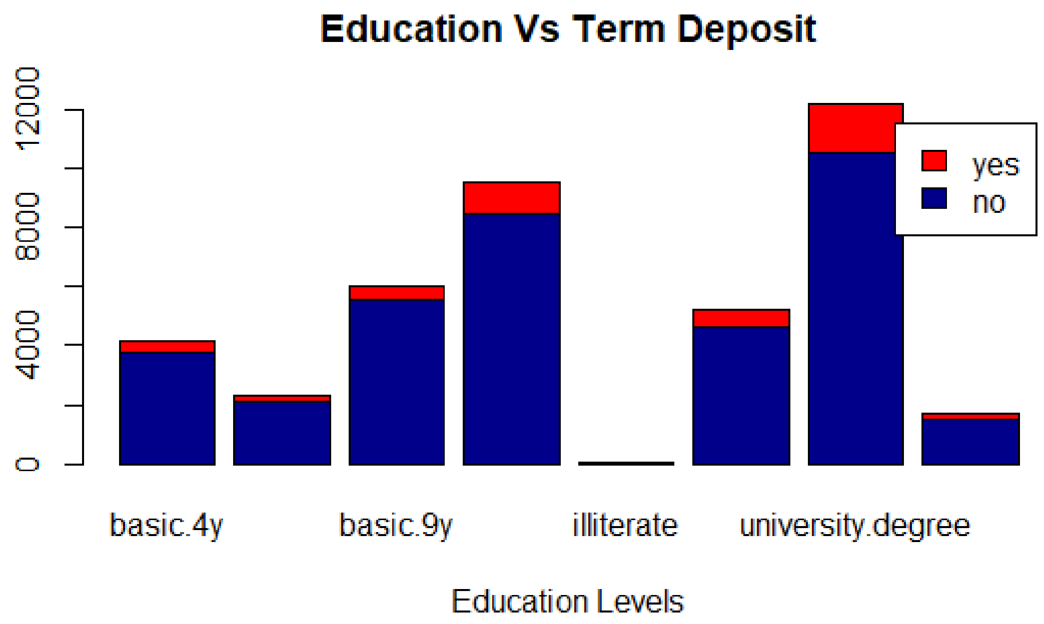
For the above distribution we can say that longer the call lasts the probability of taking the term deposit by that customer increases.

Bar plot of Education Variable



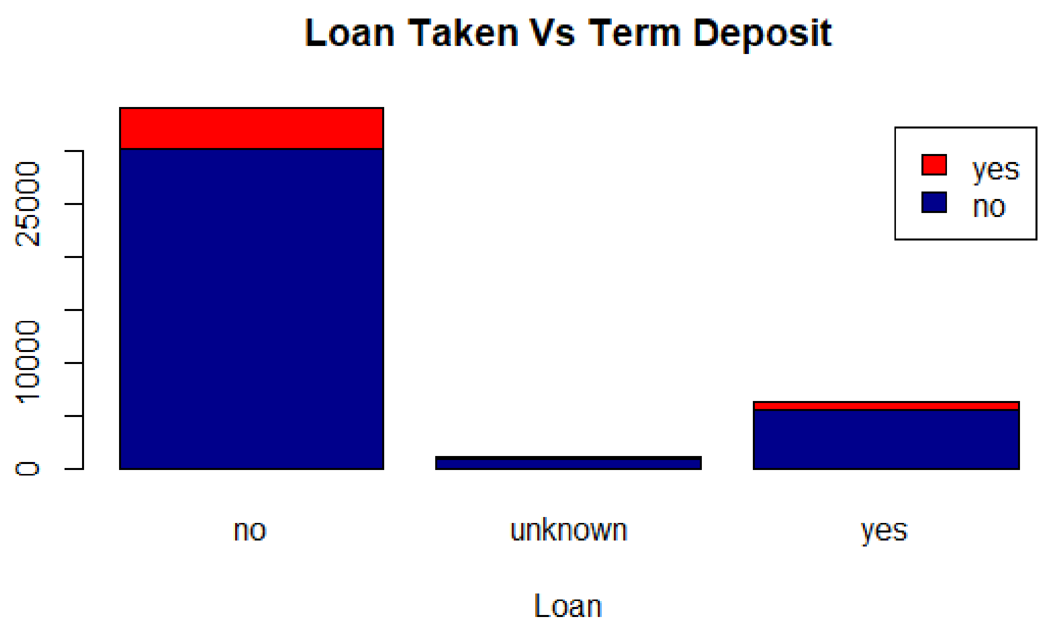
For the bar plot above, we can say that bank contacted the person or customer having the higher education. University degree is the highest whereas the second highest is high school.

The Distribution of Education Variable



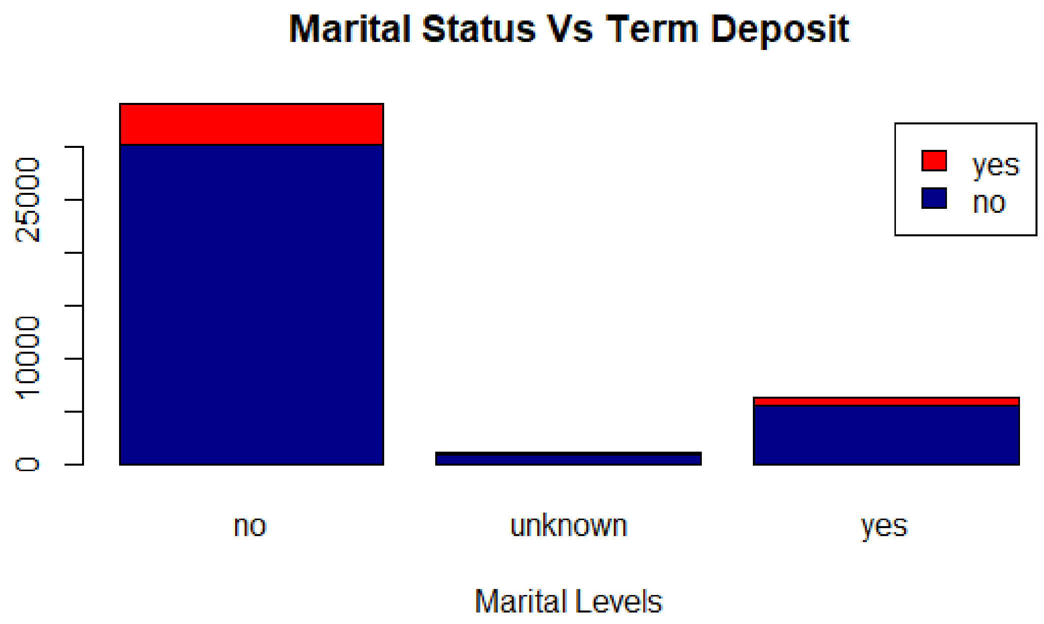
Here, we can see that the university degree is the highest where customer take term deposit.

Customers Having Loan vs Term Deposit



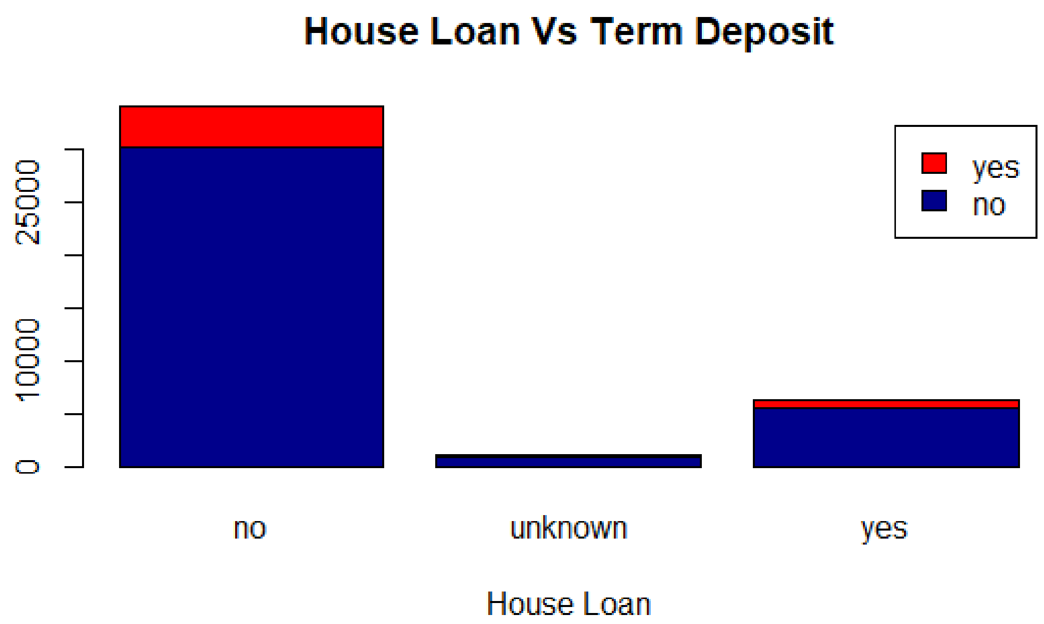
The customer having loan not preferred to take the term deposit, where as the customer which are not having the loan are taking the term deposit.

Customers Marital Status vs Term Deposit



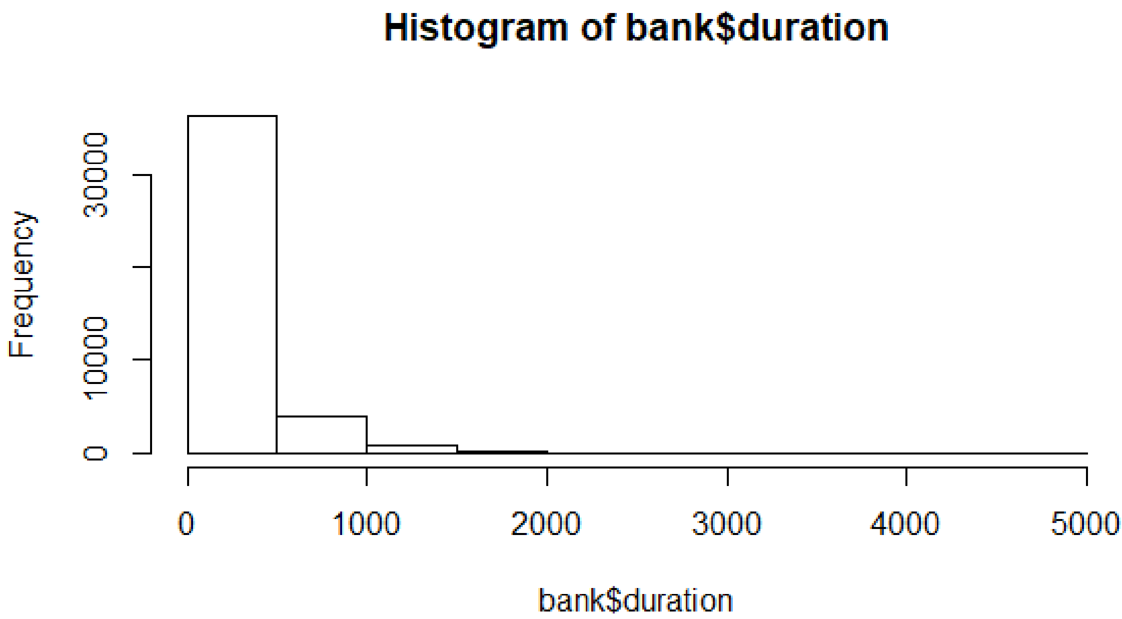
Based on the marriage status, we can observe that the person or the customer which are married and the customer which are single prefer term deposit more than the customer which are divorced.

Customer Having House Loan vs Term Deposit

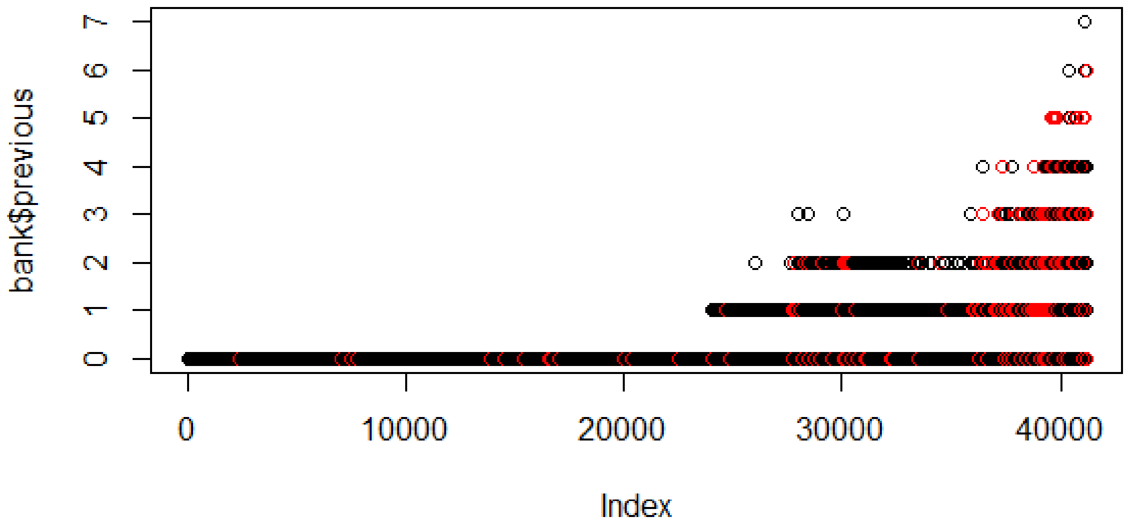


The bank connected the customer with both having house loan and not having house loan. Both the categories preferred the term deposit almost equally.

Histogram of Customer Last Contacted duration

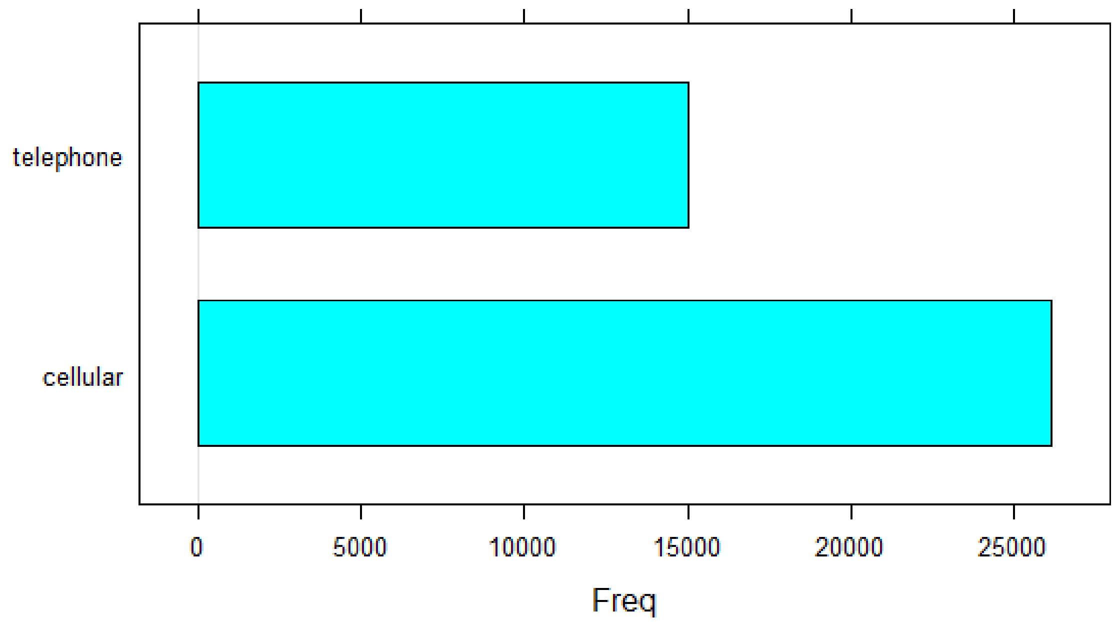


No of Contacts Performed previously vs Term Deposit

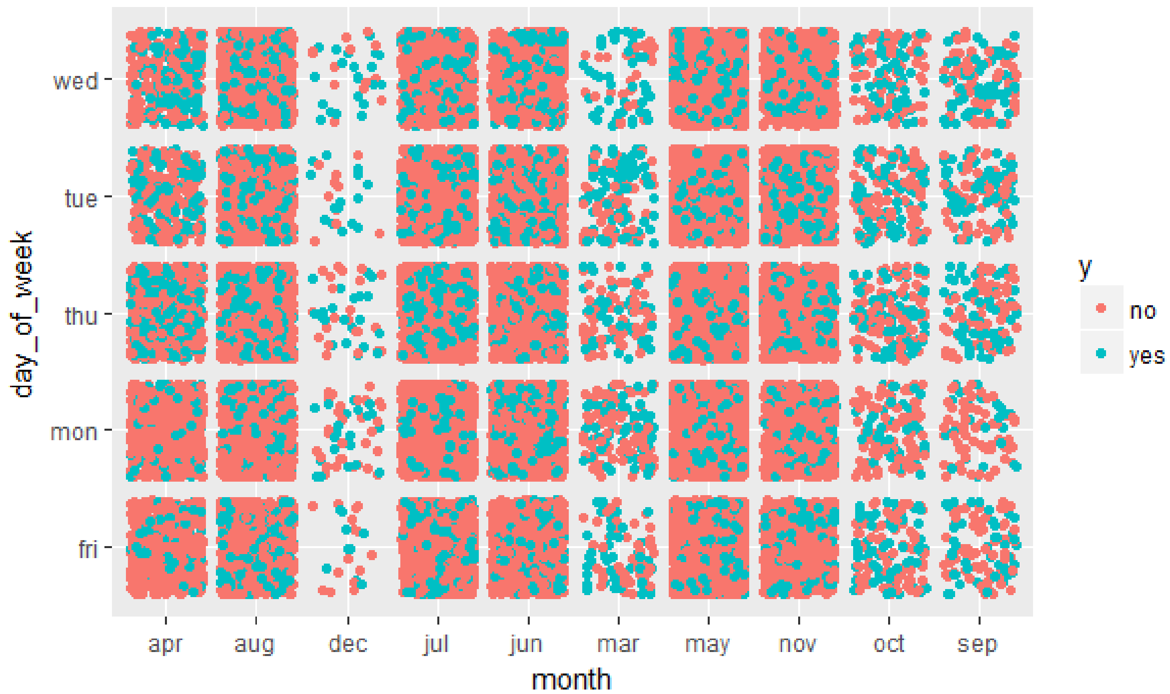


Here we can see that the frequency with duration having 0 is higher so that we can say that the bank prefers to contact new customer than the existing one. The scatter plots than future classifies with the term deposit with new and exciting customer.

Bar chart of Contact variable



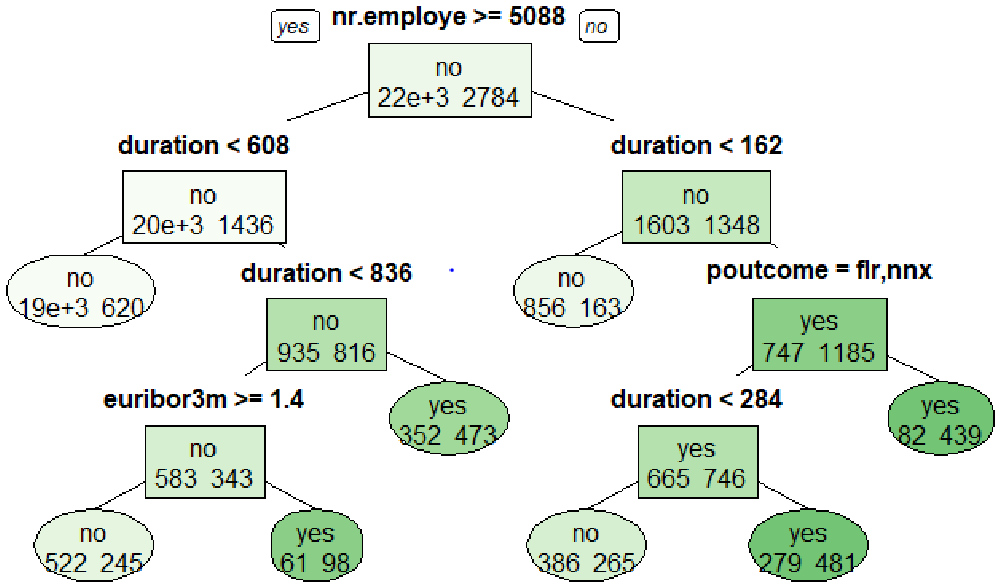
Contact days of a week vs Term Deposit



The plot of proportion table and scatterplot of month and day of the week on which the bank contacted the customer where seen to be the months in which it is high probability of people taking the term deposit are December, march, October and September compared to other months.

Decision Tree

Classification tree is generated for full grown tree. Since decision tree are not affected by the transformation so we are procced by not performing normalization.



Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7227  765
         1   83  163

               Accuracy : 0.8971
                 95% CI : (0.8903, 0.9035)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.002542

                  Kappa : 0.2419
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.17565
            Specificity : 0.98865
         Pos Pred Value : 0.66260
         Neg Pred Value : 0.90428
             Prevalence : 0.11265
         Detection Rate : 0.01979
   Detection Prevalence : 0.02986
      Balanced Accuracy : 0.58215

       'Positive' Class : 1


Call:
roc.default(response = bank_val_labels, predictor = as.numeric(predictdecision))
```

Balanced Accuracy is 58.21%, we got 89.71% accuracy by decision tree. But we can not decide by only one, so now let’s try the other method.

Random Forest

Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7136  662
         1  174  266

               Accuracy : 0.8985
                 95% CI : (0.8918, 0.905)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.000615

                  Kappa : 0.3411
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.28664
            Specificity : 0.97620
         Pos Pred Value : 0.60455
         Neg Pred Value : 0.91511
             Prevalence : 0.11265
         Detection Rate : 0.03229
   Detection Prevalence : 0.05341
      Balanced Accuracy : 0.63142

       'Positive' Class : 1


Call:
roc.default(response = bank_val_labels, predictor = as.numeric(predict_random))

Data: as.numeric(predict_random) in 7310 controls (bank_val_labels 0) < 928 cases (bank_val_labels 1).
Area under the curve: 0.6314
```

Here we can see that there is not much difference in accuracy which is 89.85%. But Balance Accuracy is increased from 58 to 63.14%. And here the classes where well classified compared to the decision tree. There is an increment in the sensitivity. Here we can see that False Negative is decrease so we can say that random forest reduces the loss by decreasing the False negative value.

Logistic Regression

Logistic Regression is firstly applied to all the variable from that the significant variable is taken out and confusion matrix is created to see the accuracy of the data by this method.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7205  728
         1  105  200

               Accuracy : 0.8989
                 95% CI : (0.8922, 0.9053)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.0004195

                  Kappa : 0.2845
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.21552
            Specificity : 0.98564
         Pos Pred Value : 0.65574
         Neg Pred Value : 0.90823
             Prevalence : 0.11265
         Detection Rate : 0.02428
   Detection Prevalence : 0.03702
      Balanced Accuracy : 0.60058

       'Positive' Class : 1


Call:
roc.default(response = bank_val_labels, predictor = predict_sig)

Data: predict_sig in 7310 controls (bank_val_labels 0) < 928 cases (bank_val_labels 1).
Area under the curve: 0.6006
```

From the summary of the method, we can say that residual deviance has reduced to 13590 with the cost of degree of freedom. The confusion matrix in the logistic regression shows that sensitivity is just 21% and the false negative value is 728 which is high, and the balanced accuracy is 60%.

This deviance is also high so applying the back-step regression to find the desired variables from it. Confusion matrix is as follows:

Back-Step Regression

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7203  730
         1  107  198

               Accuracy : 0.8984
                 95% CI : (0.8917, 0.9048)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.0006969

                  Kappa : 0.2811
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.21336
            Specificity : 0.98536
         Pos Pred Value : 0.64918
         Neg Pred Value : 0.90798
             Prevalence : 0.11265
         Detection Rate : 0.02403
   Detection Prevalence : 0.03702
      Balanced Accuracy : 0.59936

       'Positive' Class : 1


Call:
roc.default(response = bank_val_labels, predictor = predict_logistic_step)

Data: predict_logistic_step in 7310 controls (bank_val_labels 0) < 928 cases (bank_val_labels 1).
Area under the curve: 0.5994
```

This model also gives the same result as the logistic regression with accuracy of 89% and balanced accuracy of 60%. We can also see that sensitivity and false negative both the things are not improved in this method. So now we applied cross validation to see some better and changed results.

Cross validation in logistic regression

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7192  118
         1  726  202

               Accuracy : 0.8975
                 95% CI : (0.8908, 0.904)
    No Information Rate : 0.9612
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2823
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.63125
            Specificity : 0.90831
         Pos Pred Value : 0.21767
         Neg Pred Value : 0.98386
             Prevalence : 0.03884
         Detection Rate : 0.02452
   Detection Prevalence : 0.11265
      Balanced Accuracy : 0.76978

       'Positive' Class : 1


Call:
roc.default(response = bank_val_labels, predictor = as.numeric(predict_logistic_2))

Data: as.numeric(predict_logistic_2) in 7310 controls (bank_val_labels 0) < 928 cases (bank_val_labels 1).
Area under the curve: 0.6008
```

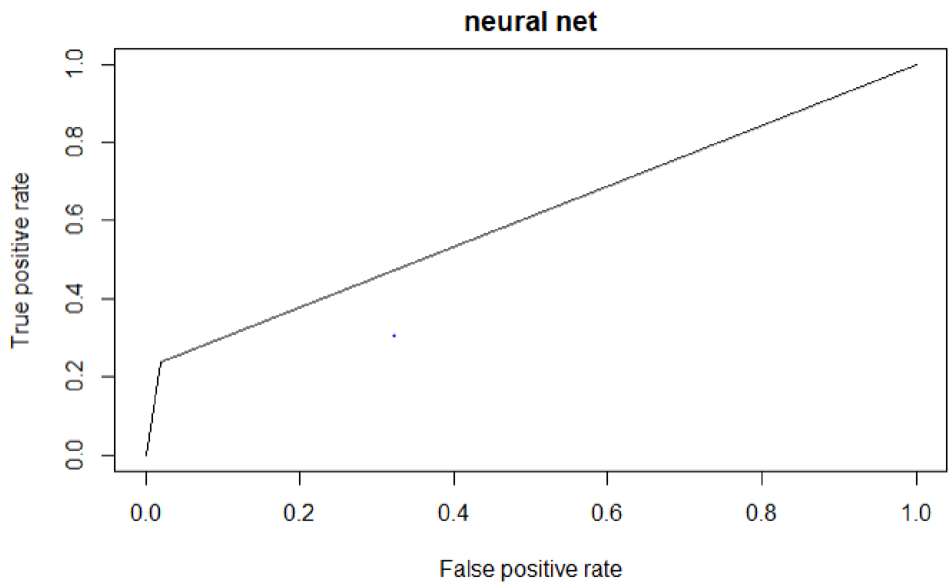
Here the results are almost simpler as above of accuracy of 89% and there is also no improvement in false negative and sensitivity.

Neural Network

Normalization and dummy variables are required for the neural network, so they where been created.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7175  707
         1  135  221

               Accuracy : 0.8978
                 95% CI : (0.891, 0.9043)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.001278

                  Kappa : 0.3005
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.23815
            Specificity : 0.98153
         Pos Pred Value : 0.62079
         Neg Pred Value : 0.91030
             Prevalence : 0.11265
         Detection Rate : 0.02683
   Detection Prevalence : 0.04321
      Balanced Accuracy : 0.60984

       'Positive' Class : 1
```

ROC Curve



ROC Curve of the predicted and true values indicates the relation between true positive rate and false negative rate. The area which is under the curve for the plot is 0.7386739.

To improve the performance of the model we use the function pcaNNet which applies the principal component analysis to the variables before building a neural network model. And the hidden layer size was reduced to 2 so that model can generalize it for future data prediction.

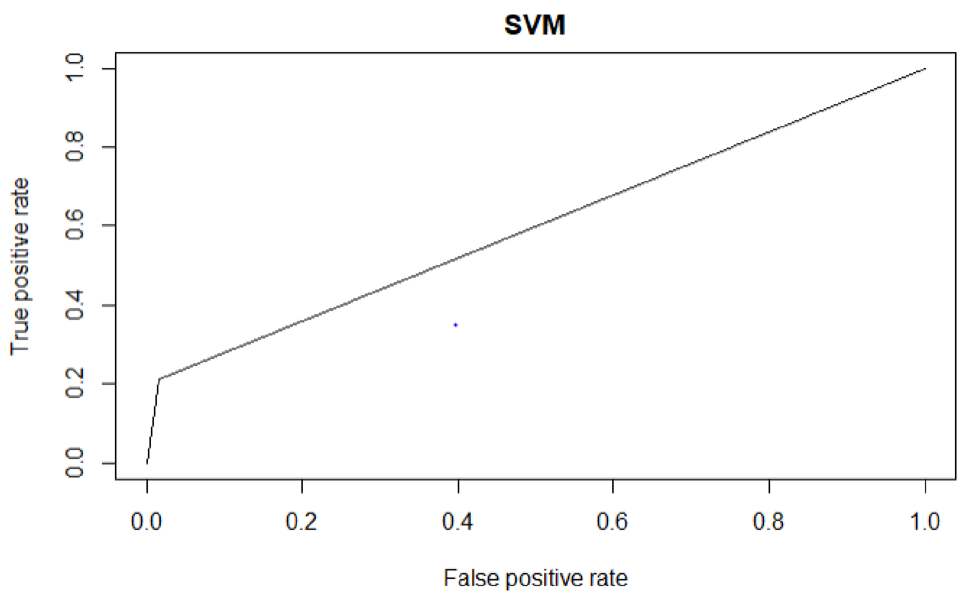
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7166  696
         1  144  232

               Accuracy : 0.898
                 95% CI : (0.8913, 0.9045)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.001007

                  Kappa : 0.3111
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.25000
            Specificity : 0.98030
         Pos Pred Value : 0.61702
         Neg Pred Value : 0.91147
             Prevalence : 0.11265
         Detection Rate : 0.02816
   Detection Prevalence : 0.04564
      Balanced Accuracy : 0.61515

       'Positive' Class : 1
```

We can see in this that there is improvement in sensitivity and false negative.

Support Vector Machine (SVM)

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7202  732
         1  108  196

               Accuracy : 0.898
                 95% CI : (0.8913, 0.9045)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.001007

                  Kappa : 0.278
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.21121
            Specificity : 0.98523
         Pos Pred Value : 0.64474
         Neg Pred Value : 0.90774
             Prevalence : 0.11265
         Detection Rate : 0.02379
   Detection Prevalence : 0.03690
      Balanced Accuracy : 0.59822

       'Positive' Class : 1
```

ROC Curve



ROC curve of the predicted and true values indicating the relation between the true positive and false positive rate.

Naive Bayes model

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6112  400
         1 1198  528

               Accuracy : 0.806
                 95% CI : (0.7973, 0.8145)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2945
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.56897
            Specificity : 0.83611
         Pos Pred Value : 0.30591
         Neg Pred Value : 0.93857
             Prevalence : 0.11265
         Detection Rate : 0.06409
   Detection Prevalence : 0.20952
      Balanced Accuracy : 0.70254

       'Positive' Class : 1
```

Based on the values of sensitivity and false negative, we choose that our final model is Naïve based model as it has the best values compared to the upper models.

Applying Naïve Bayes on test data

```
[1] 0.3212075
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6142  370
         1 1167  558

               Accuracy : 0.8134
                 95% CI : (0.8048, 0.8218)
    No Information Rate : 0.8873
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3212
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.60129
            Specificity : 0.84033
         Pos Pred Value : 0.32348
         Neg Pred Value : 0.94318
             Prevalence : 0.11266
         Detection Rate : 0.06774
   Detection Prevalence : 0.20942
      Balanced Accuracy : 0.72081

       'Positive' Class : 1
```

Conclusion:

From the confusion matrix above we can see that accuracy is 81%. The false positive value is 370 and true positive value is 558. We have used 20% validation and 20%. And from validation and test data we got the better results in test data when compared to the validation data in terms of false negative and true positive.

References:

https://archive.ics.uci.edu/ml/datasets/bank+marketing

[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

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

S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]

https://cran.r-project.org/doc/contrib/Short-refcard.pdf