**USE CASE ON CLASSIFICATION OF BANK MARKETING DATASET**



Team Members

Nikhil (001277045)

Manoj (001859255)

Data Mining in Engineering

(IE 7275)

Faculty Advisor

Prof. Sagar Kamarthi

Teaching Assistant

Ramin Mohamamdi

***Table of Contents***

|  |  |  |
| --- | --- | --- |
| **No.** | **Description** | **Page** |
| **1** | **Abstract** | **3** |
| **2** | **Methodology** | **4** |
| **3** | **Data Description and Preprocessing** | **5** |
| **4** | **Data Visualization** | **6** |
| **5** | **Model Implementation and Validation** | **13** |
| **6** | **Conclusion** | **24** |
| **7** | **Reference** | **25** |

**Abstract**

Banking Industry is prominent contributor for economy of any country. It was the main cause of the global economic crisis in 2008 because of the bad loan deposits. The project aims in finding the potential customers who are likely to take term deposit based on the marketing campaigns done by Portuguese Banking Industry. The marketing campaigns were done based on phone calls. More than one contacts to the same customer was required, in order to access if a customer would take term deposit. The dataset contains 41188 customer records with 20 predictor variable ordered by date from May 2008 to November 2010. The data was partioned to 60% of training 20% validation and 20% test data. The best model was decided by true positive and False negative critieria. Class of Interest was customer accepting the term deposit. Classifying an non-acceptor as potential customer wasn’t a problem but the converse would be big loss to the marketing. Five model were implemented to on training data and the best model was decided on validation data. The best model was applied on test data.

**Methodology**

Machine learning methodology was used to determine the final model. The basic outline and methods used are mentioned below.

**Training Phase**

Decision Trees

Logistic Regression

SVM

Neural Networks

Naïve Bayes

**Validation Phase**

Decision Trees

Logistic Regression

SVM

Neural Networks

Naïve Bayes

Test Data

Naïve Bayes

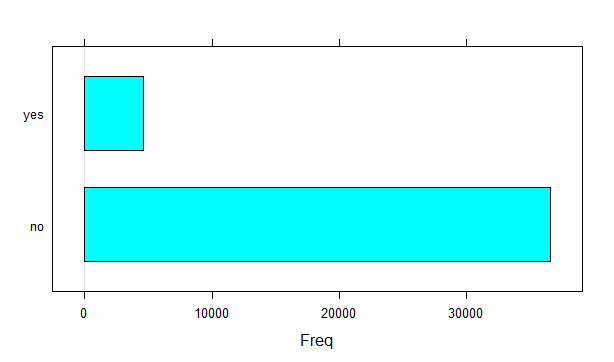
***Data Description and Preprocessing***

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Type of category** | **Transformed categories** |
| Age | Numeric | Not tranformed |
| 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***

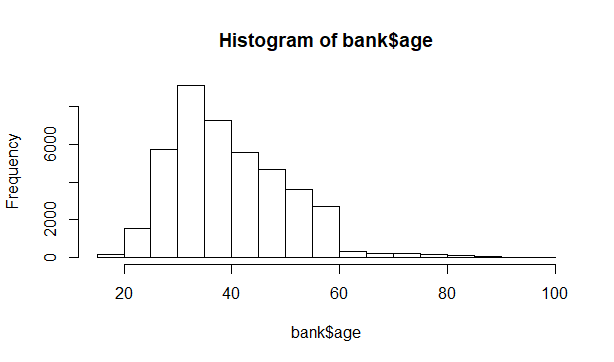
**Output variable (desired target):**

y - has the client subscribed a term deposit? (binary: "yes","no")

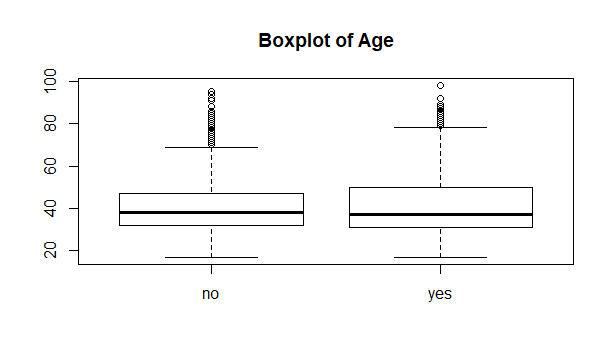
****

The Plot shows that most customers didn't prefer the term deposit.

**The Distribution of Age**

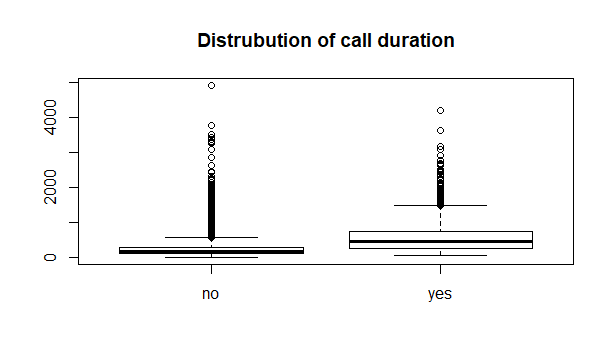


This below plot shows that the bank has contacted mostly in the age range of 20 to 60. Also we notice that maximum frequency age group is 30-35.



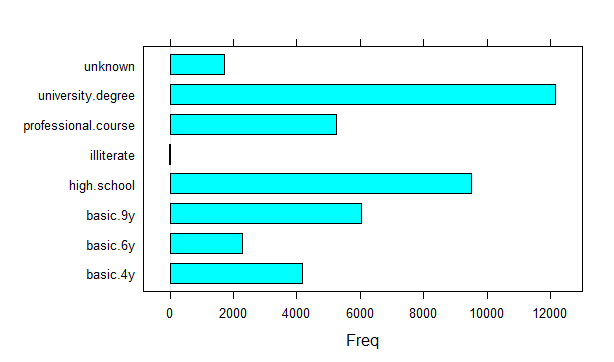
The box plot shows the distribution of age compared to term deposit. The outliers are more for customers not taking term deposit when compared to customers taking. The whiskers of NO are almost equal which mean customers not taking term deposit is equally distributed. In Yes group the upper whisker is longer. This shows that customers taking term deposit is more in the mid age group.

**The Distribution of Call Duration**



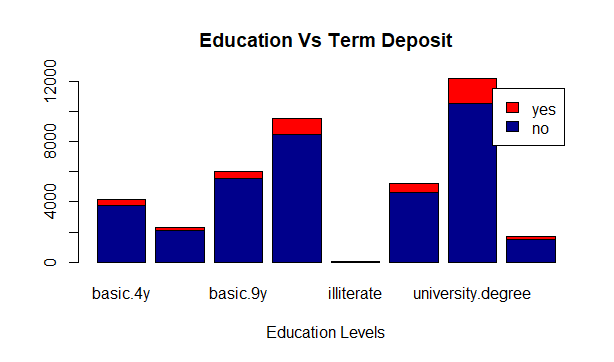
From the Above plot we infer that higher the call duration, the probability of customer accepting a term deposit is more.

**Barplot of Education Variable**

****

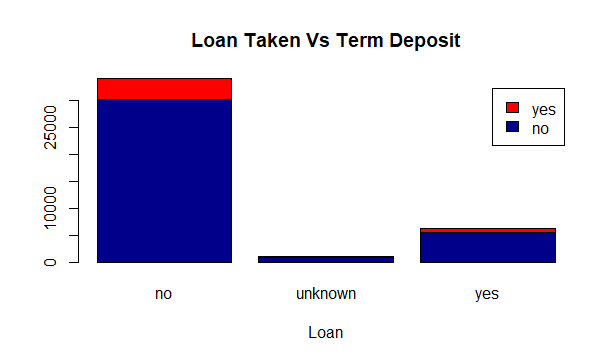
From this plot we notice that the bank contacted mostly to customers having higher education. University degree frequency is highest and the high school frequency is second highest.

**The Distribution of Education Variable**



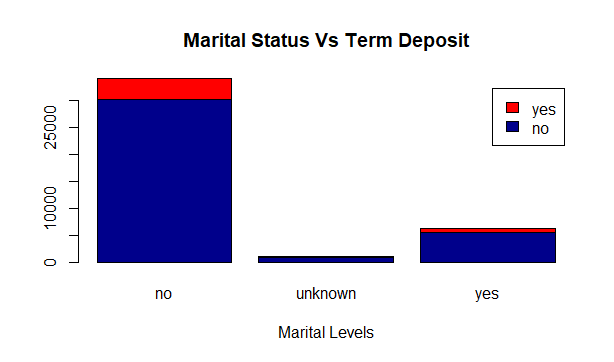
In all the education categories, the preference to term deposit is less. Compared to all education levels, customers having university degree opted more for term deposit.

**Customers Having Loan vs Term Deposit**



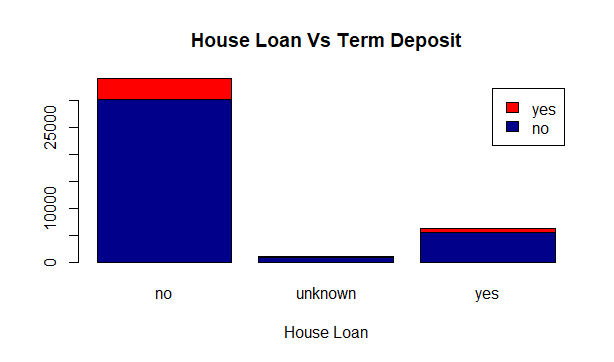
The customer having no loan prefered the term deposit when compared to customers having loan.

**Customers Marital Status vs Term Deposit**

****

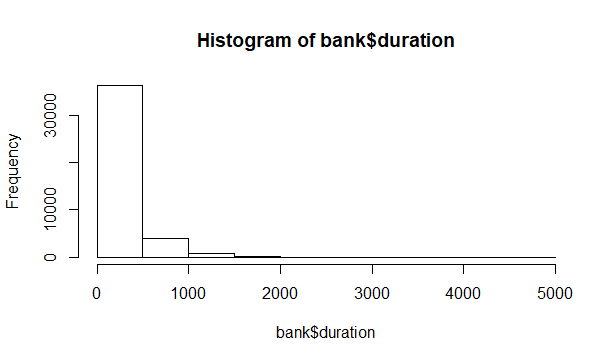
Here we see that married and single customers prefered term deposit more than divorced.

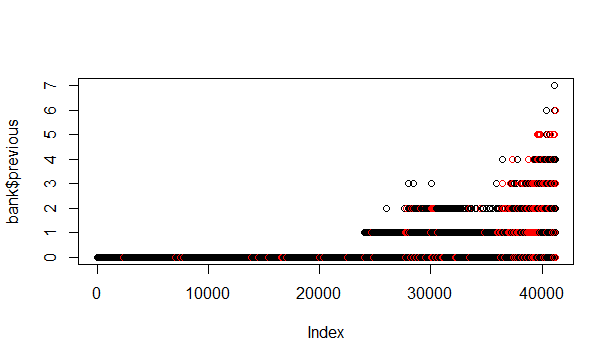
**Customer Having House Loan vs Term Deposit**



The bank contacted both the customers having house-house loans and non-house loans. Both categories preferred the term deposit almost equally.

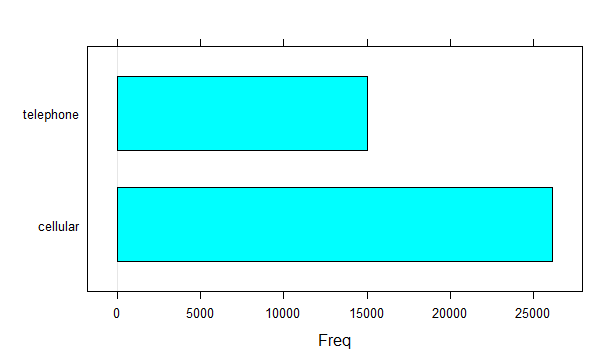
**Histogram of Customer Last Contacted duration**



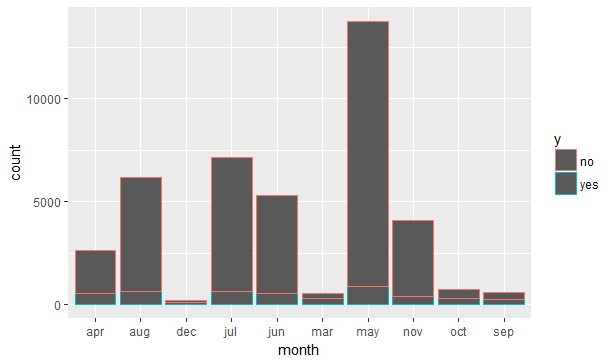
**No of Contacts Performed previously vs Term Deposit**

Here we notice that frequency for duration having 0 is more which indicates that bank mostly preferred new customer than existing. The scatter plot further classifies this with term deposit for new and existing customers.

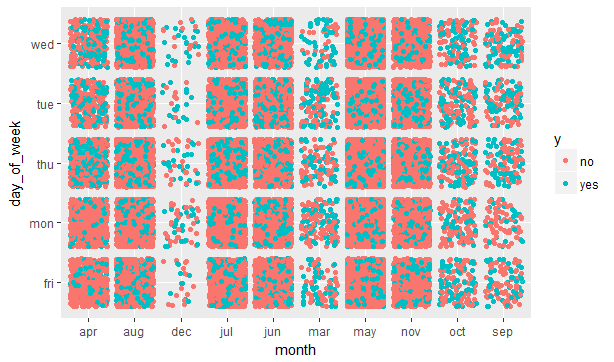
**Barchart of Contact variable**



Last Contact Month vs Term Deposit

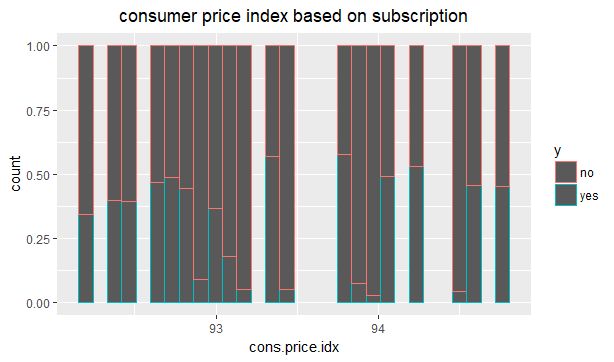
****

**Contact days of a week vs Term Deposit**

****

The plot of proportion table and scatterplot of month and day of the week on which people were contacted shows that the months december, march, october and september have a very high probability of people taking the plan compared to other months. But the histogram of month in accordance with y shows that very less number of people were actually contacted in those months.

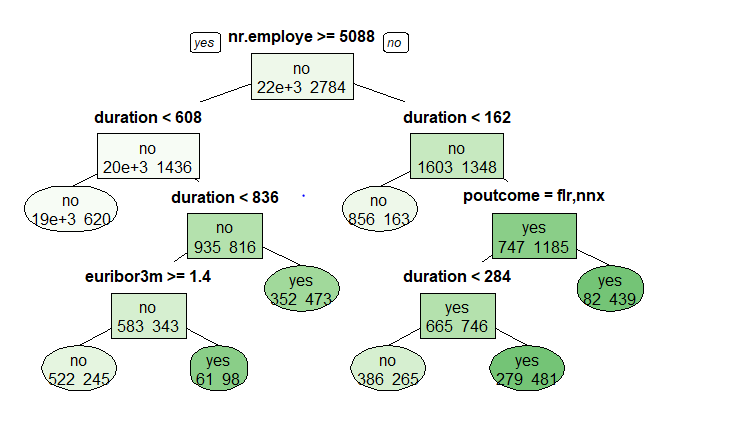
**Consumer Price Index vs Term Deposit**

****

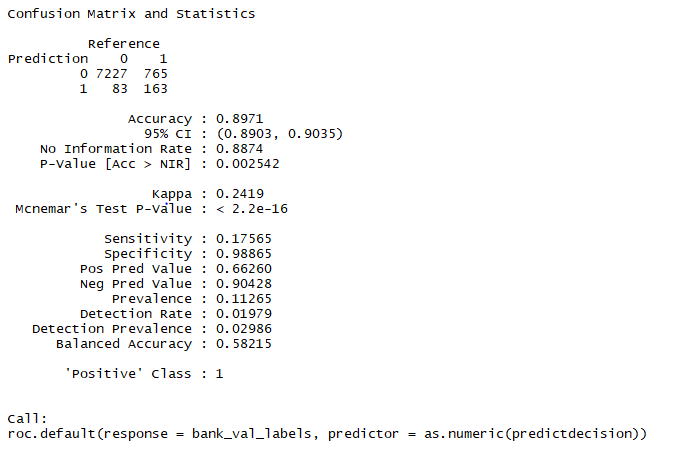
A histogram of proportions of the variable consumer price index shows that the concentration of people taking the plan is evenly spread on both higher and lower sides of CPI, so we can't really make any generalizations about a particular group being more favorable for saying "yes"

**Decision Trees**

Classification tree is constructed for full grown tree. Since Decision trees are not affected by the transformations we are proceeding without any normalization.



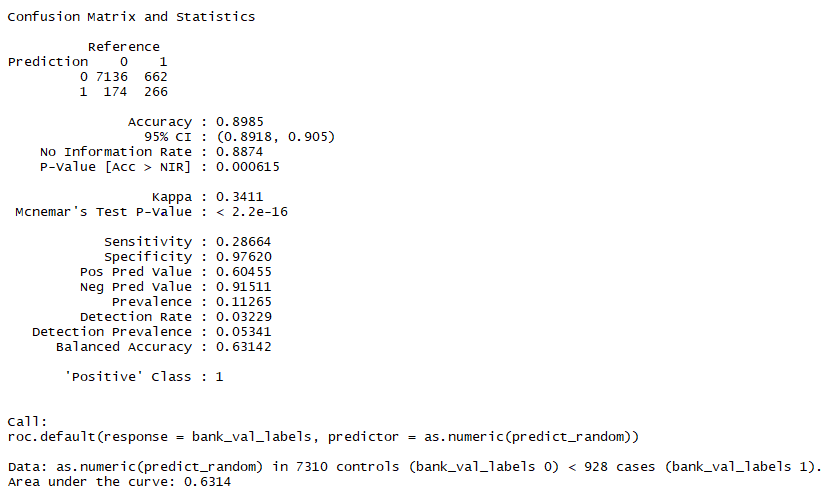
Confusion Matrix



#AUC is 58.21% we got 89 percent accuracy with decision trees. But let’s apply random forest to see if we can further increase our accuracy with random forest.

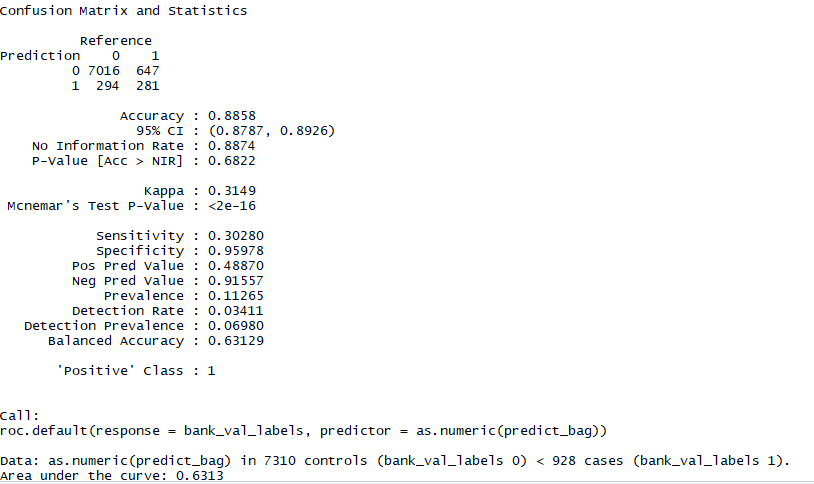
**Random Forest.**

Confusion Matrix



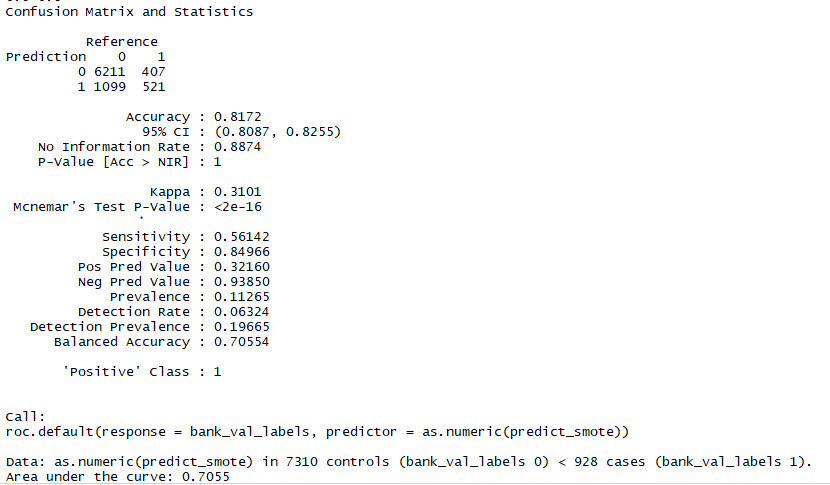
Accuracy for decision tree is 89%.AUC is 63.3 percent. AUC has increased from 58 to 63%. From the above confusion matrix, we can see that the classes were better classified than full decision tree. Also, we see the FN value to decrease and false positive value to increase. There is an increment in the sensitivity. However, the accuracy remains same. Losing a potential customer incurs more loss to organization than incorrectly classifying a non-potential customer. The random forest reduces the loss by decreasing the False negative value.

**Bagging**



AUC is 63% and remains the same. Accuracy is 88%. We see that bagging performed better than random forest with increment of sensitivity and decrement of false Negative.

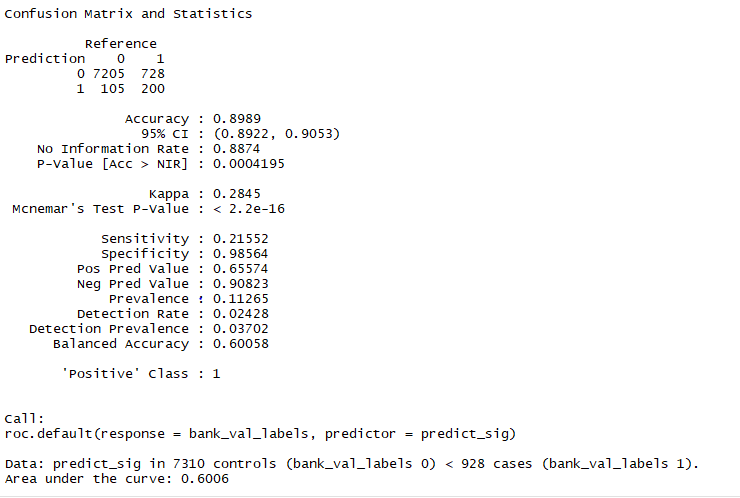
Since this is an oversampled data with 1:9 ratio of preferred variable, we have used oversampled concept with SMOTE package to see any difference.



AUC is 71%. Accuracy is 81%. But we see there is good improvement in sensitivity and false negative. Since our class of interest is and the oversampled concept gives us better results when compared to previous model, we finalize our decision tree as bank smote.

**Logistic Regression**

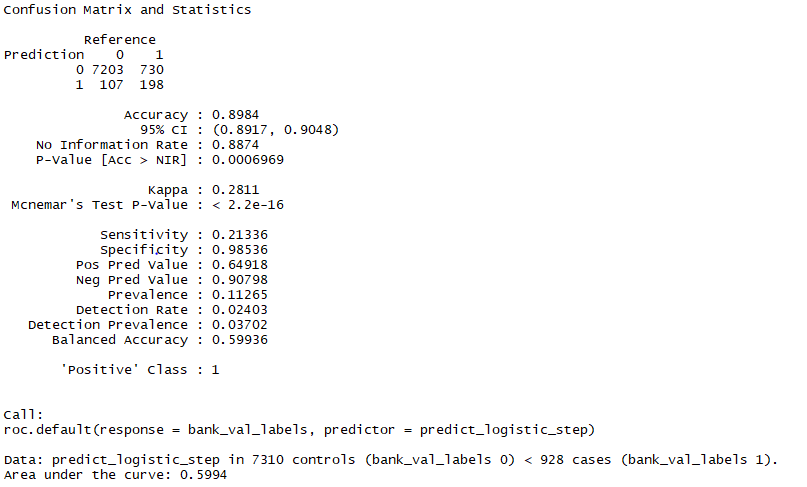
Logistic Regression is being applied to all the variables initially and the significant variables are taken from it and confusion matrix is created to see the accuracy.



From the summary, we see that the residual deviance has reduced to 13590 with the cost of degree of freedom. The confusion matrix above shows that sensitivity is just 21% and false negative value is 765 which is high. AUC is 60%

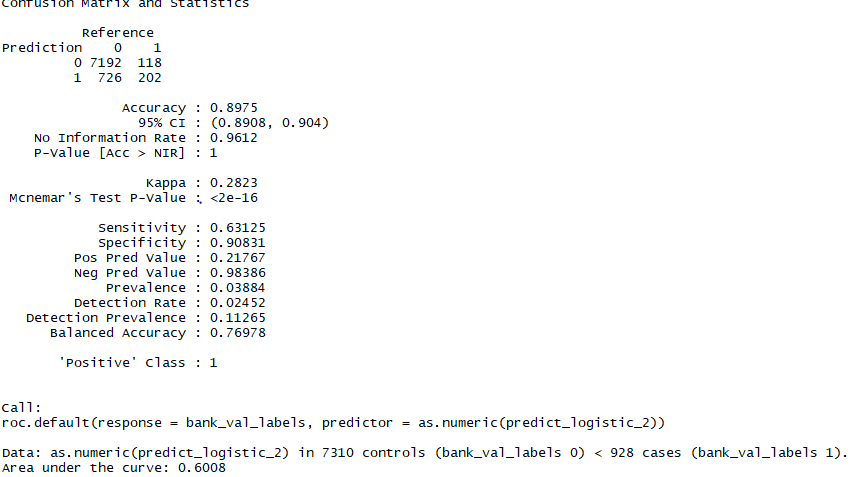
However, this deviance is also large. Applying backstep regression method to find the desired variables

Confusion matrix for the best model from Backstep Regression.



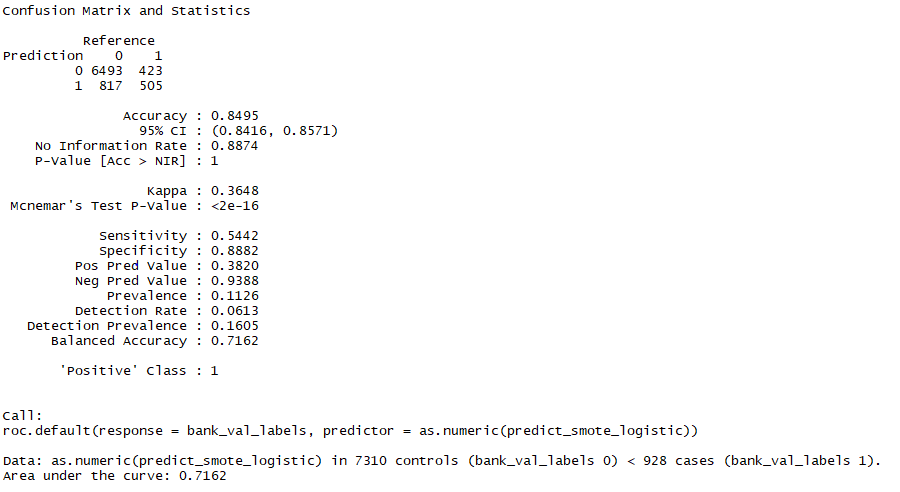
Accuracy 89% and AUC 60%. The above model too gives the same deviance as bank\_logistic. Also the sensitivity and false negative have not improved. Applying cross-validation to check any better model

**Using the cross validation in logistic regression.**



#Accuracy 89%. AUC 60% and no improvement in sensitivity and false negative.

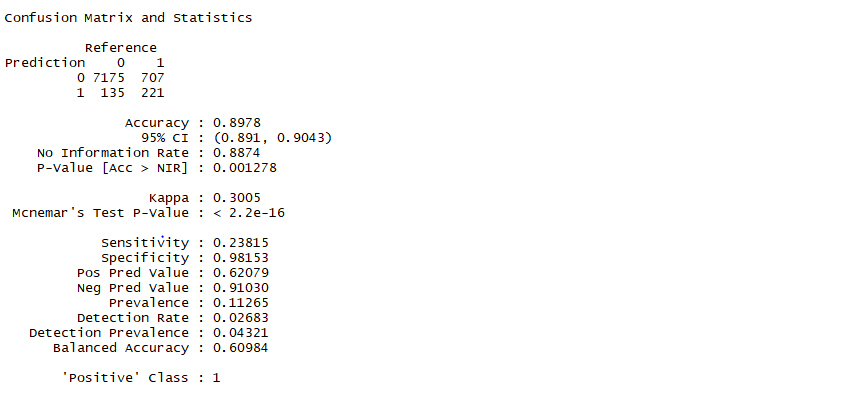
**Applying Smote function**



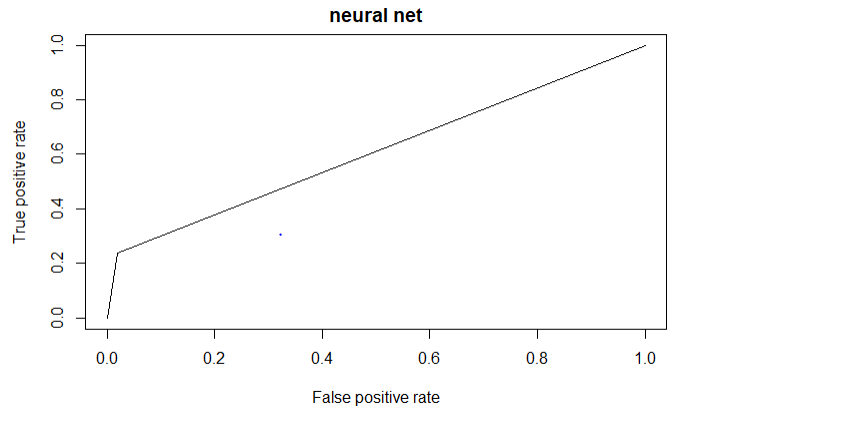
#Accuracy 85%. AUC 71%. The above confusion matrix says that sensitivity and true negative has improved has improved. But slightly lesser than decision tree smote.

**Neural Networks**

Normalization and created dummy variables for neural networks

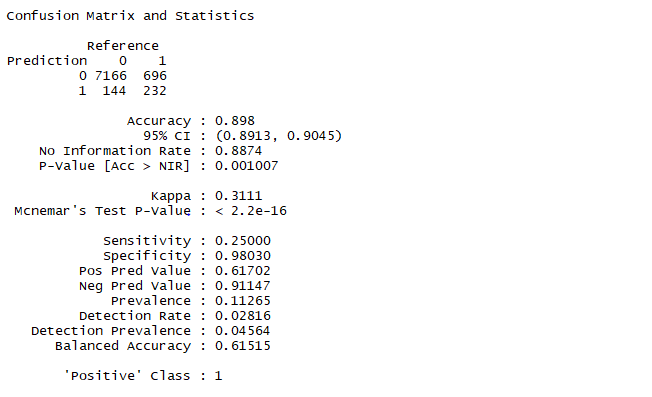


ROC Curve



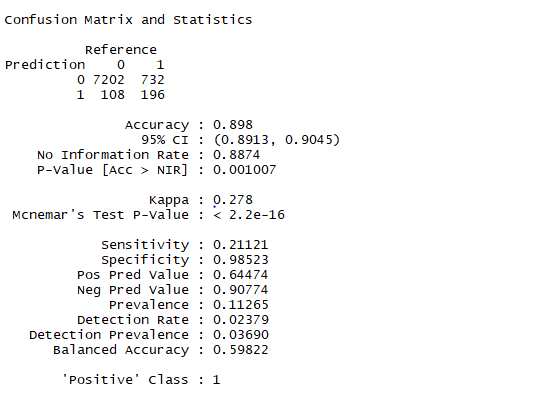
ROC curve of the predicted and true values indicating the relationship between true positive rate and false positive rate. The area under the curve for the plot is 0.7386739

Trying to improve the model performance by using the function pcaNNet which applies principal component analysis to the variables before building a neural network model. And also, size of the hidden layers were reduced to 2 for the model to generalize more on future data and to avoid overfitting

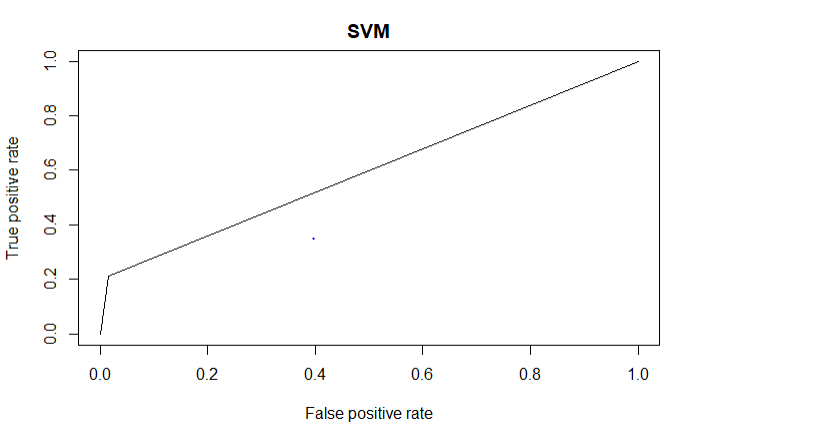


we can see an improvement in sensitivity and false negative.

**Support Vector Machine**

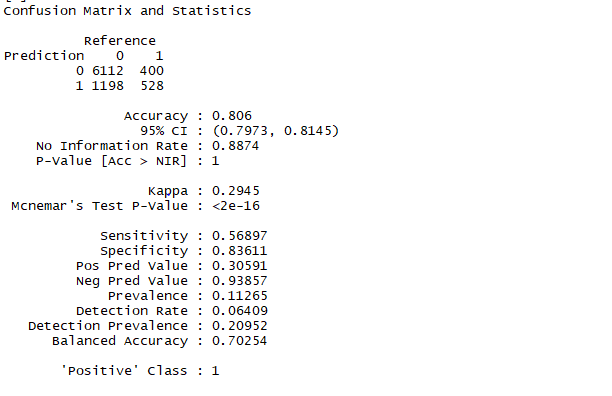


ROC Curve



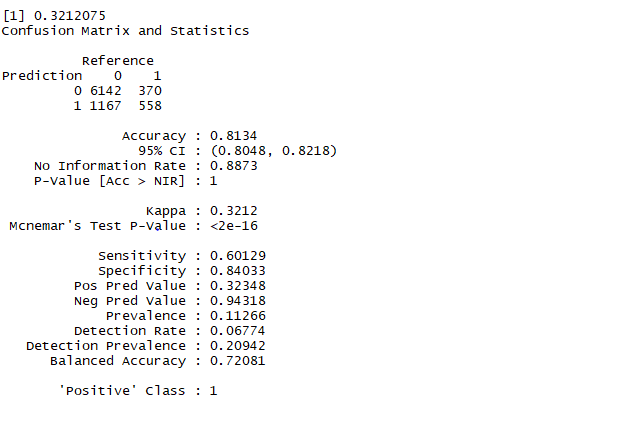
ROC curve of the predicted and true values indicating the relationship between true positive rate and false positive rate.

**Naive Bayes model**



Based on the sensitivity and false negative value, we choose our final model as Naive based model.

**Applying naive bayes on test data.**



**Conclusion**

From the confusion matrix above we notice that accuracy is 81% . Also the false positive value is 370 and true positive values are 558 we have used 20% for validation and 20% of test data. we got a better result for test data when compared to validation data in terms of true positive and false negative.

**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]