**PROJECT REPORT**

**Santander Customer Transaction Prediction**

*Submitted By:*

**Allwyn Joseph**

**Date:07/08/2019**

**Background​ -**

At ​Santander​, mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals. Our data science team is continually challenging our machine learning algorithms, working with the global data science community to make sure we can more accurately identify new ways to solve our most common challenge, binary classification problems such as: is a customer satisfied? Will a customer buy this product? Can a customer pay this loan

**Problem Statement:**

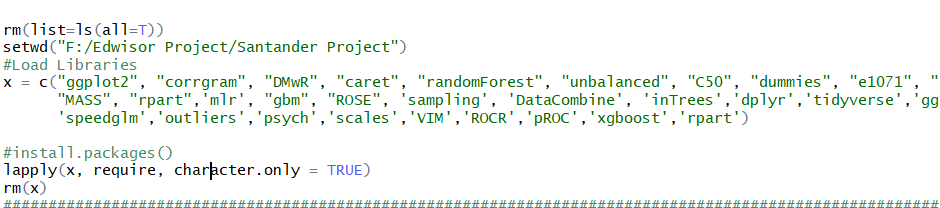
In this challenge, we need to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.

**PART A.**

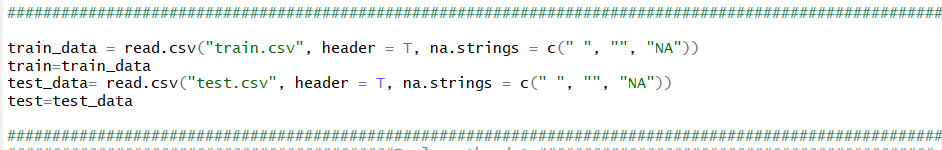
**-R**

**Loading the R Libraries:**

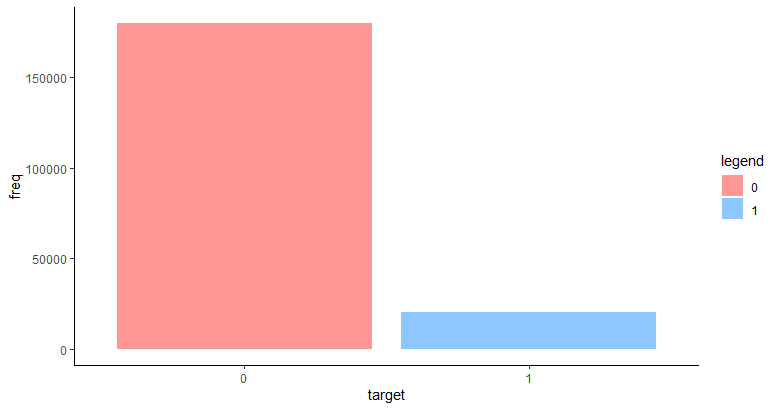
x = c("ggplot2", "corrgram", "DMwR", "caret", "randomForest", "unbalanced", "C50", "dummies", "e1071", "Information", "MASS", "rpart",'mlr', "gbm", "ROSE", 'sampling', 'DataCombine', 'inTrees' , 'dplyr' , 'tidyverse' , 'ggthemes' , 'data.table','speedglm','outliers','psych','scales','VIM','ROCR','pROC','xgboost', 'rpart')



**Importing the datasets:**

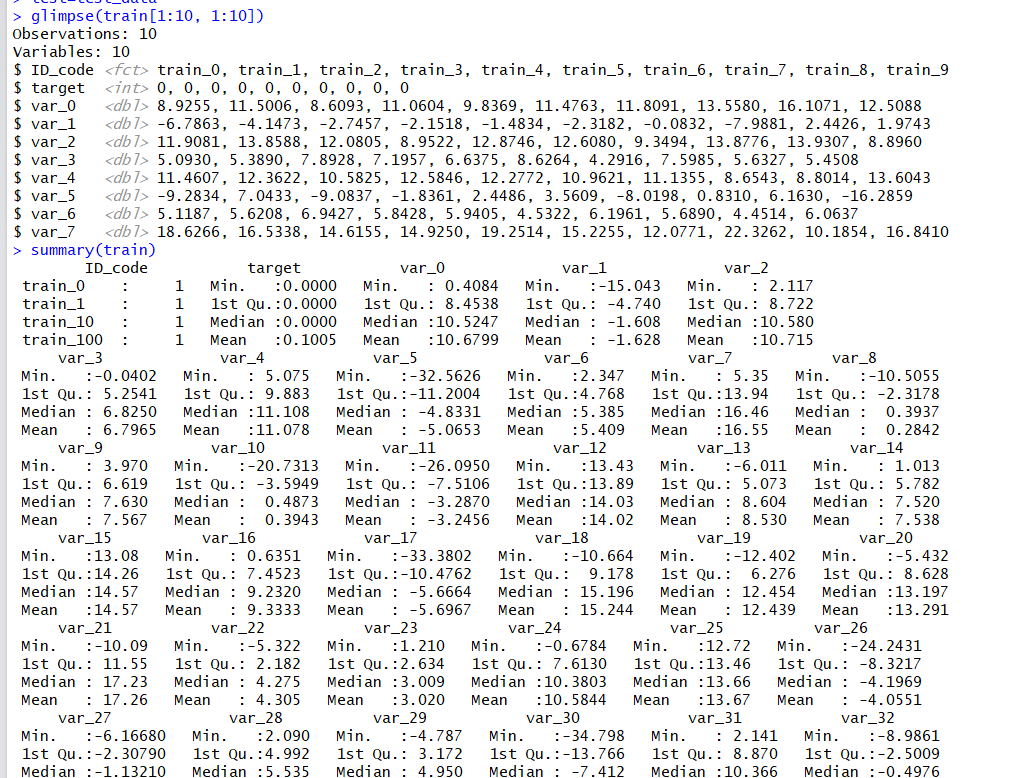


# Target Variable:

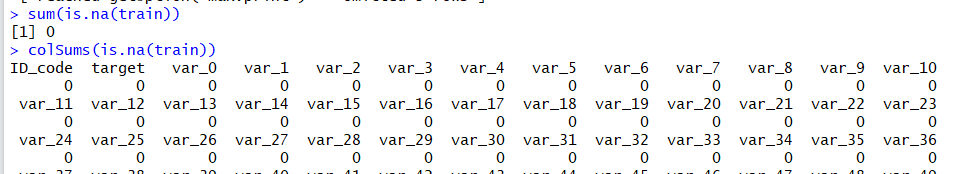
****10.04% is class 1 and 89.95% is 0.

It implies that the dataset is very highly unbalanced.

# Understanding the Data:



**Looking for Missing Values:**



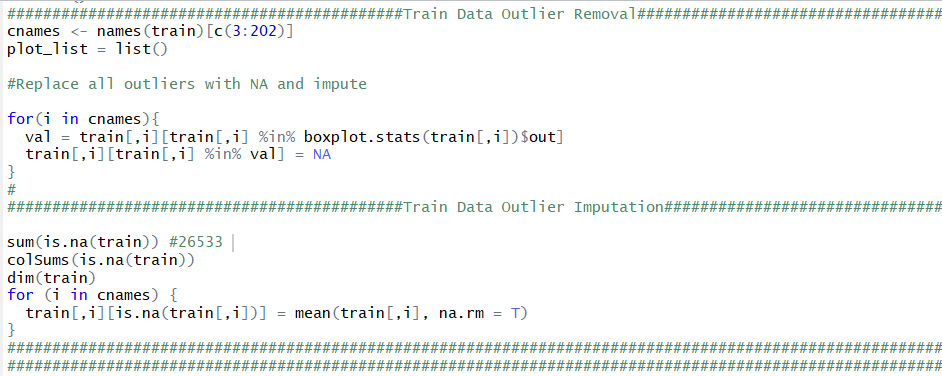
There are no missing values in the data.

1. **Pre-Processing: Outlier Analysis:**

Searching for outliers in the data.

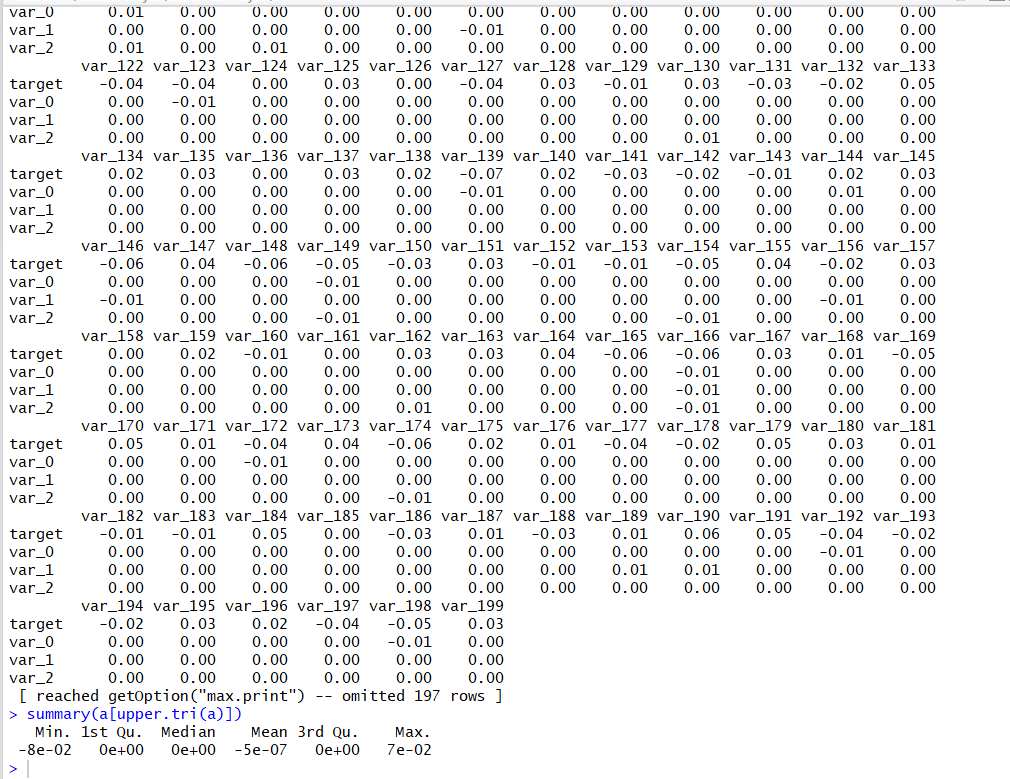
Replacing outliers with NA.

Imputing newly added NAs with mean.

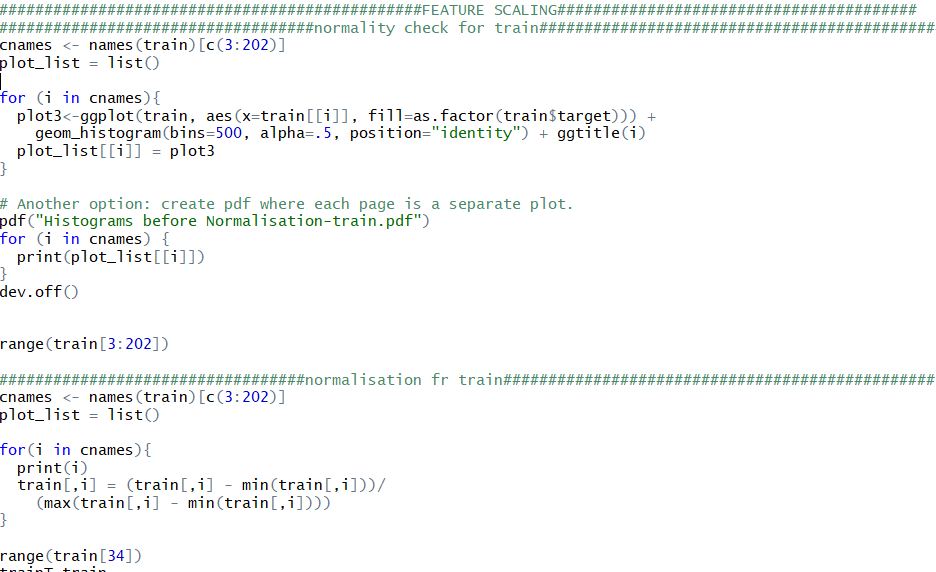


1. **Feature Selection: Correlation Matrix**

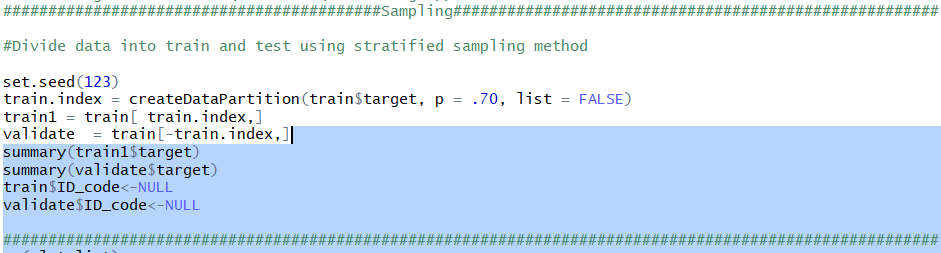
We can see that all the values are closer to 0, indicating that there is no auto correlation. Hence, we’ll have to consider all the features.



1. **Feature Scaling:**

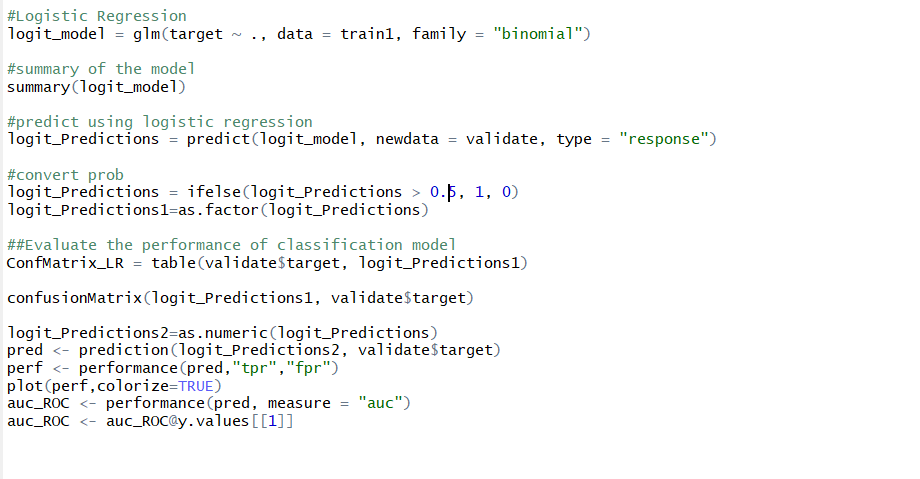


1. **Sampling:**

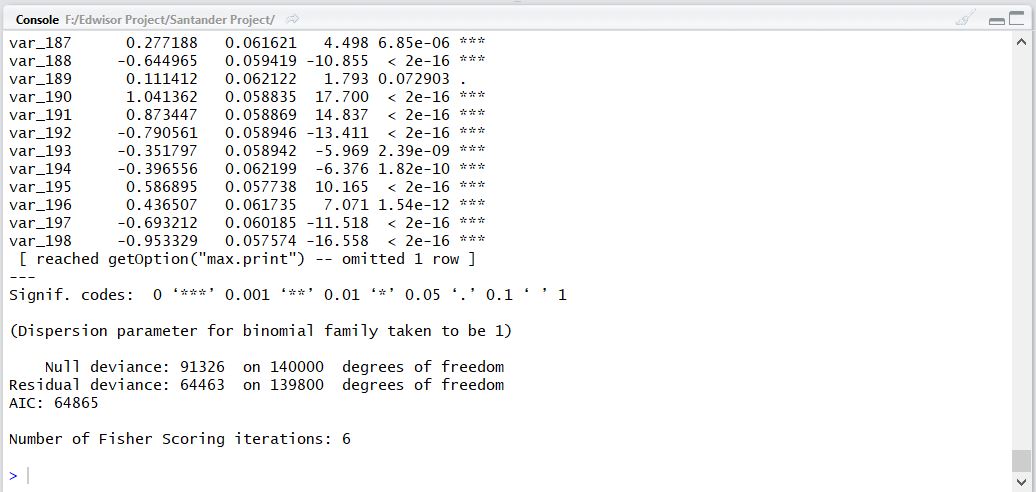


**MODELS**

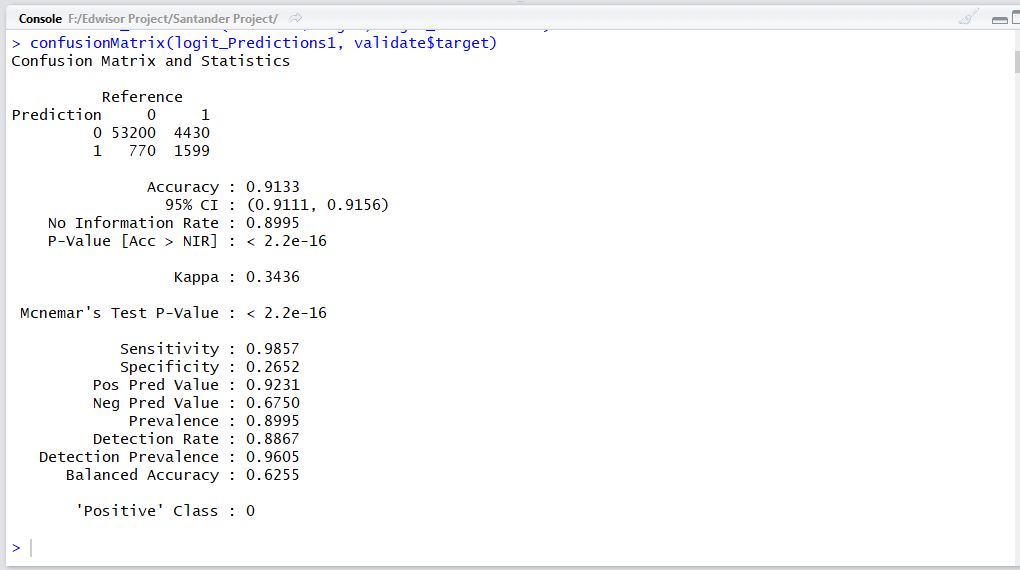
1. **Logistic Regression**
2. **code**



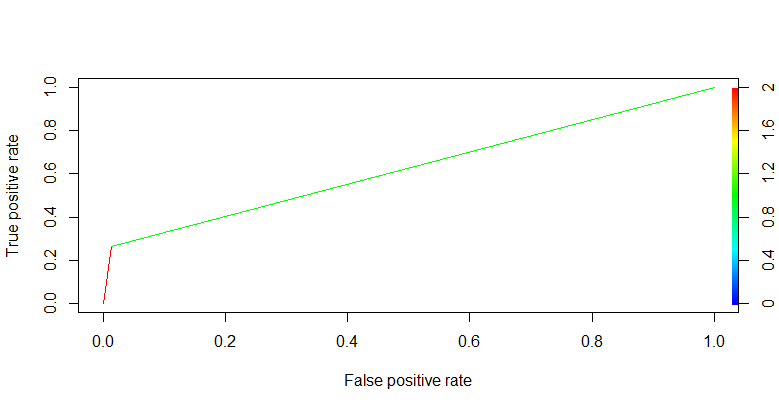
1. **Summary of the model**

****

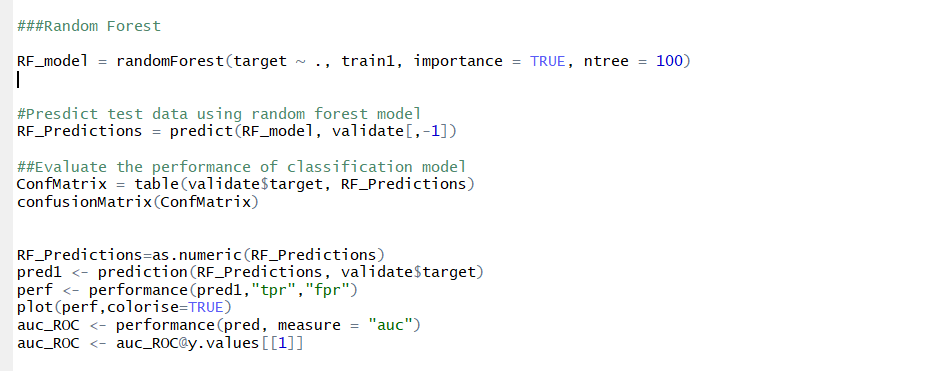
1. **Confusion Matrix**



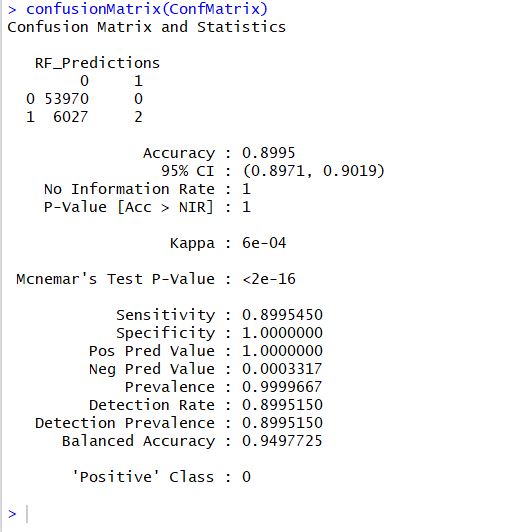
1. **AUC-ROC**



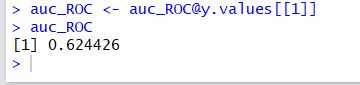
1. **Random Forests**
2. **Code**



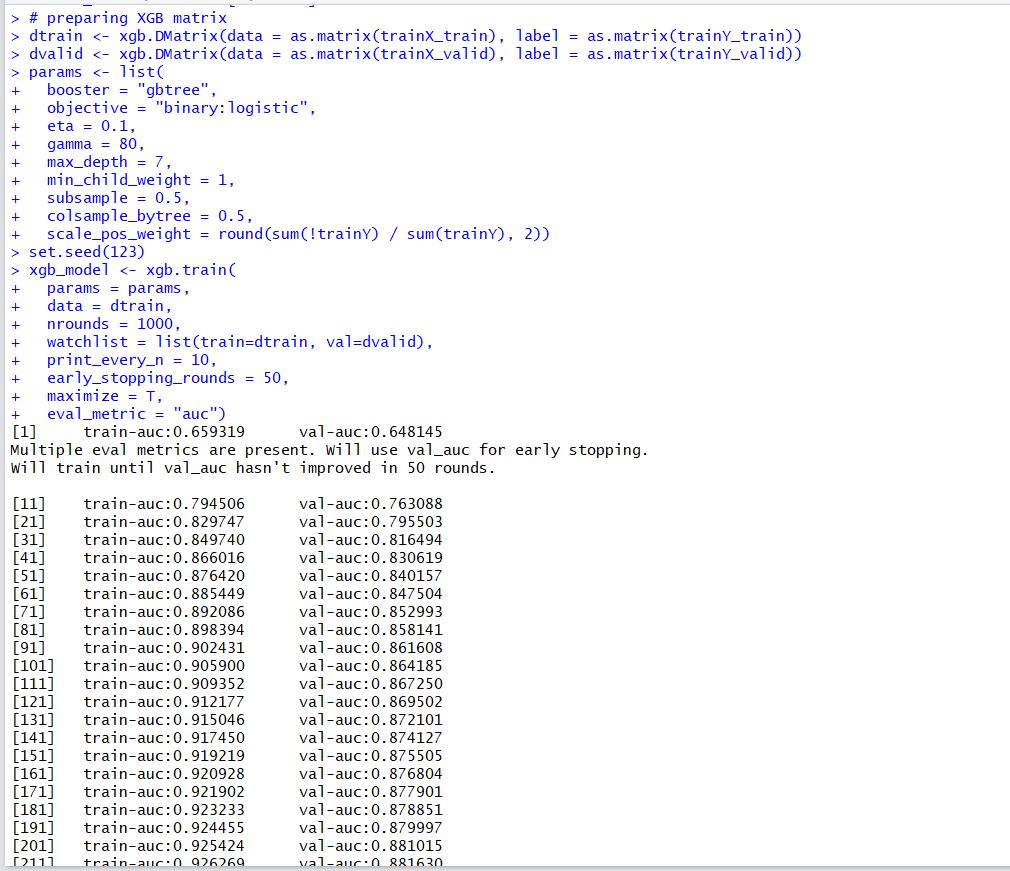
1. **Summary of the model:**

****

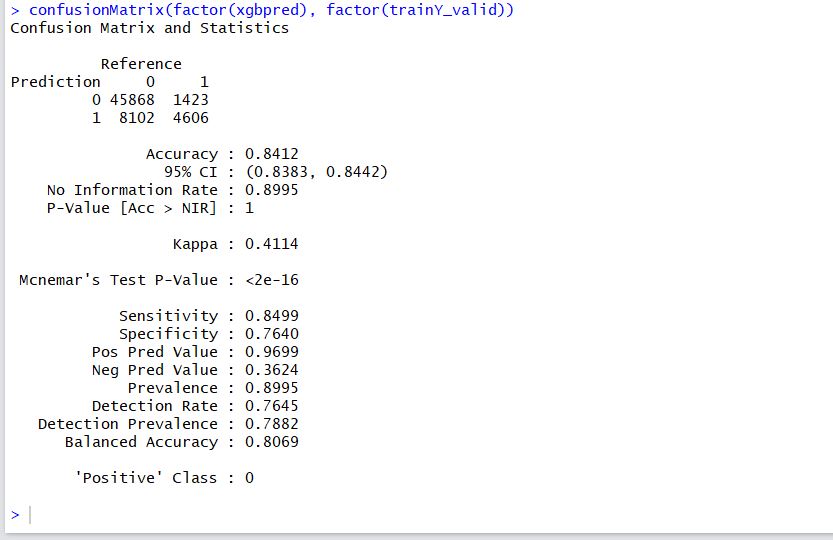
1. **AUC score**

****

1. **XG BOOST**
2. **Summary**

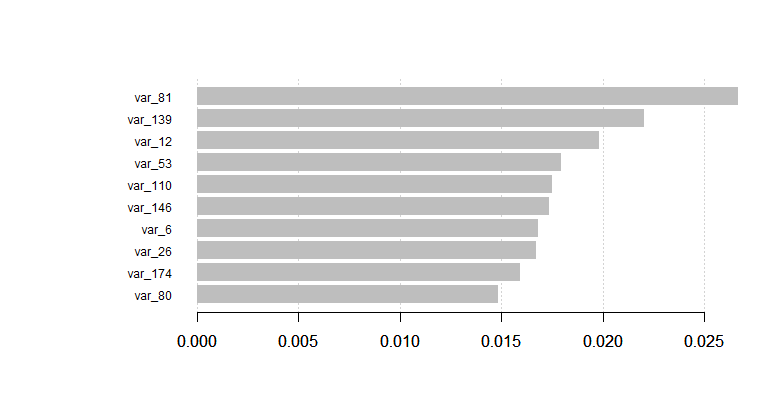
****

1. **Confusion Matrix**

****

1. **Var-IMP PLOT**

**The below variables are the 10 most important variables according to the model.**



1. **AUC score.**

**After hyper tuning the parameters we can observe that the model attains an AUC score of 0.8937on the predicted set.**

**CONCLUSION:**

**COMPARING THE 3 MODELS:**

|  |  |  |  |
| --- | --- | --- | --- |
| **MODEL** | **RECALL** | **PRECISION** | **AUC SCORE** |
| **Logistic Regression**  **Model** | **98.57** | **92.31** | **0.6254** |
| **Random Forests**  **Model** | **89.95** | **100** | **0.6257** |
| **XG BOOST Model** | **84.99** | **96.99** | **0.8937** |

The better AUC for XGB indicates that it is clearly good in separating the two classes.

Since the aim of this project is to identify which customers will make transactions in the future, we can consider XG Boost model as it has significantly good AUC score as well as a good precision value.