Report for:

Delinquency Telecom Model

Done by:

Siva Krishna

Contents

Part 1 Project Background

Many donors, experts, and microfinance institutions (MFI) have become convinced that using mobile financial services (MFS) is more convenient and efficient, and less costly, than the traditional high-touch model for delivering microfinance services. Nowadays, telecom companies are also collaborating with an MFI to provide micro-credit on mobile balances.

Main Objectives: -

Delinquency is a condition that arises when an activity or situation does not occur at its scheduled (or expected) date i.e., it occurs later than expected. The Consumer is believed to be delinquent if he deviates from the path of paying back the loaned amount within 5 days.

Problem Statement: -

Create a delinquency model which can predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan (Label '1' & '0').

Part 2 About the Data

There are 2,09,593 observations in the dataset, with no missing values. Each represents an existing customer. For each observation, the dataset records 36 input variables that stand for both qualitative and quantitative attributes of the customers.

There is a single binary output variable that denotes "yes"(1) or "no(0)" revealing whether the customer will be paying back the loaned amount within 5 days of insurance of loan.

Part 3 Data Cleaning

3.1 Outlier Removal:

Several changes were made to the dataset to prepare it for analysis. As there are no null values in the data set there is no need to perform any null value imputation for the data set. There are outliers for many variables in the data set.

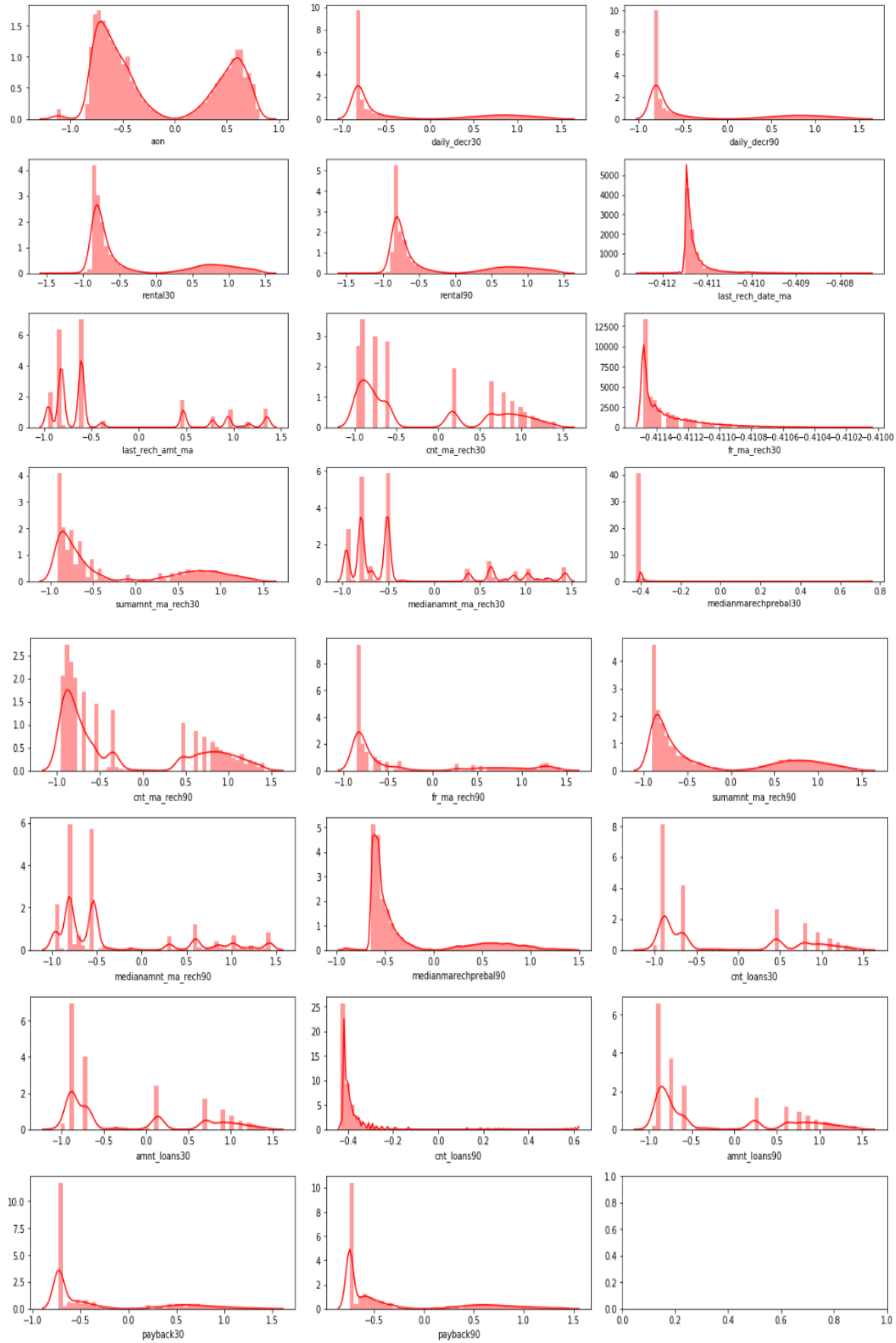
By observing these features, I found way of doing an outlier's imputation technique for the data of the features whose z-score >3. There are many ways to deal with outliers such as imputing outlier's with mean, median, mode (categorical), k-NN imputation, mice imputation or simply removing and others.

For this data set I simply choose mean for imputing the outliers with the respective features. After performing mean, I also applied cube root for the data to bring data closer as to make the distribution normal.

After performing the mean imputation and also applying cube root to the data become so what normally distributed compared to the data which haven't undergone any type of imputation or outlier transformation.

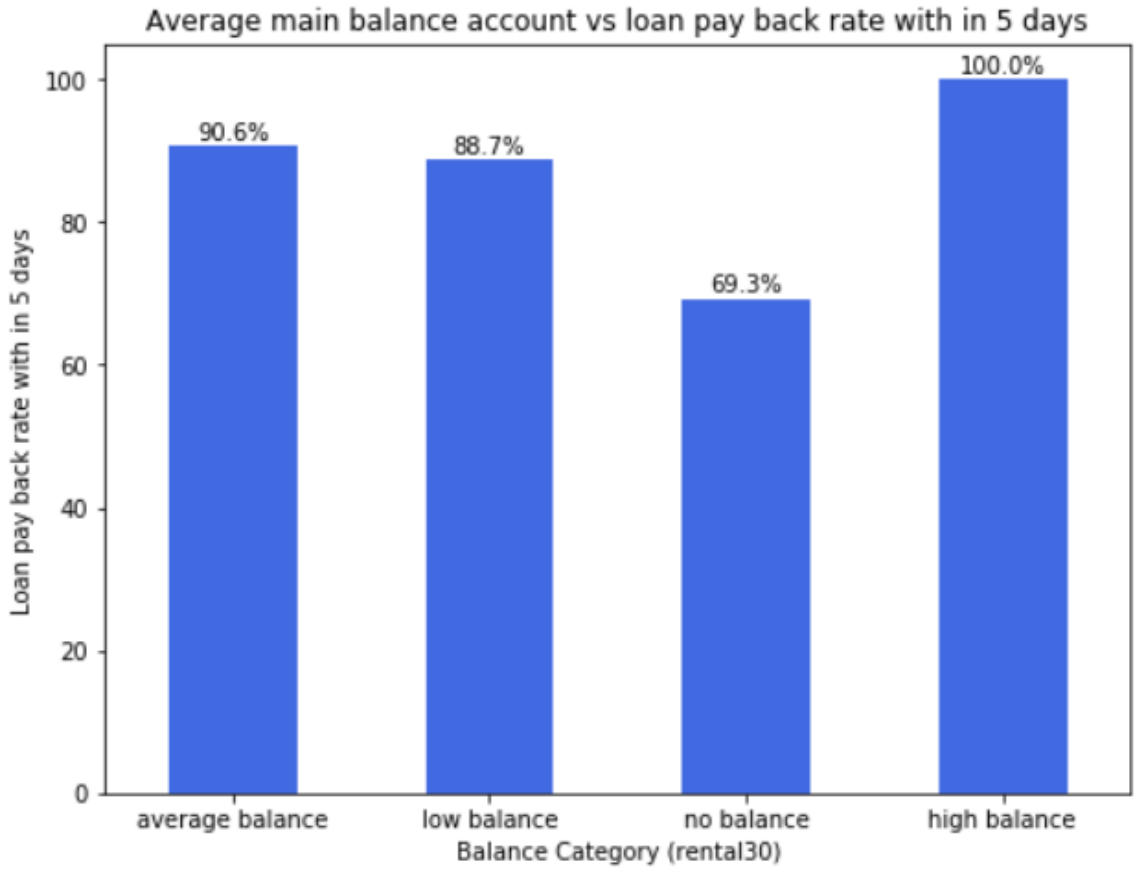
So, outlier imputation is far better than simply removing the outliers from the data. As the data set belongs to the loan defaulters or not the outliers are also important for us to get the unbiased results after performing machine learning algorithms.

Normality of the features after outlier's treatment:



Part 4 Exploratory Data Analysis:

4.1 Average main balance account vs loan pay back rate within 5 days

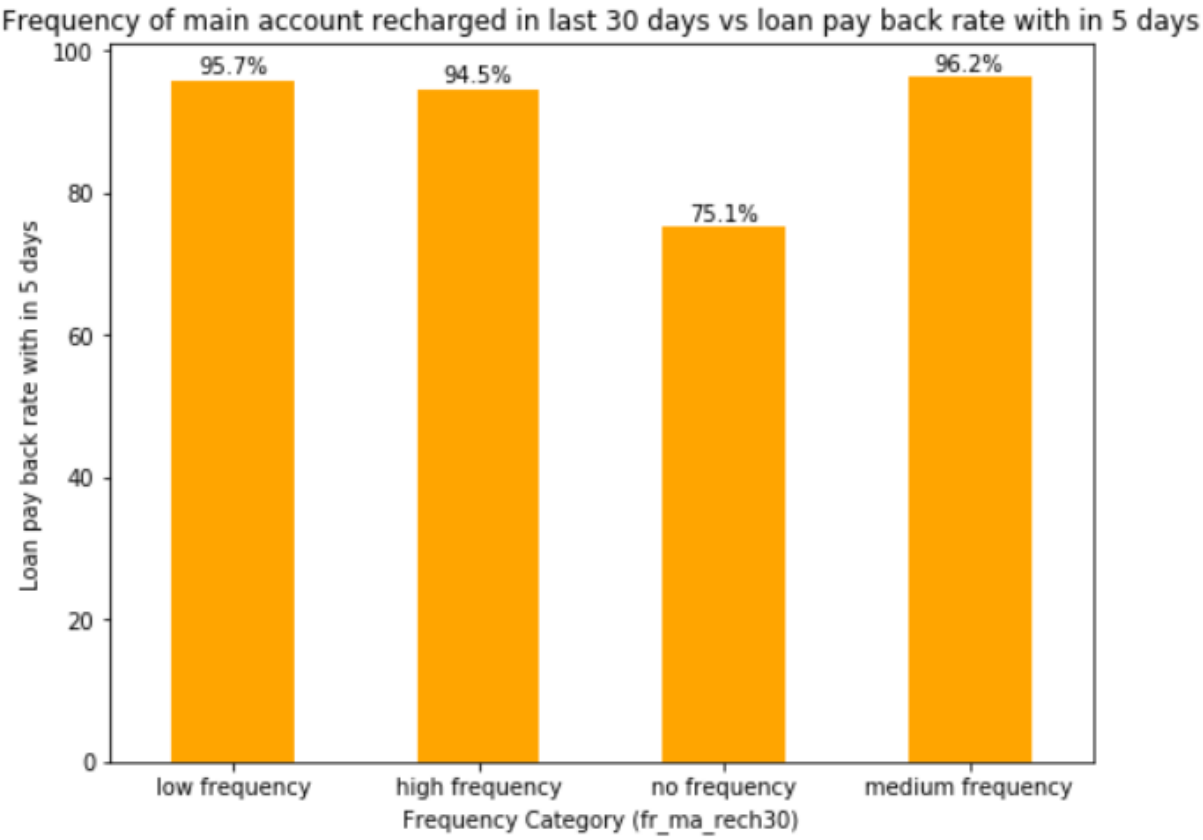


The above bar plot infers us how customers with different main balance levels are paying back the loan with in five days. The high balance level people are with 100% rate i.e they are paying loan within 5 days. Coming to the average and low balance people it is observed that around 10%-12% of people are not paying the loan within 5 days.

Coming to low balance level people, it is observed that around 30% of people are not paying back the loan with in stipulated 5 days of time. The 30% of people with no balance or negative balance people are creating a major loss to the company without paying back the loan within five days of time.

In order to decrease loss to the company, the company should start some marketing strategies like sms alerting and notifications and others on the people with no balance, average and high balance level people notifying them to pay the loan back within five days of time.

4.2 Frequency of main account recharged in last 30 days' vs loan pay back rate within 5 days

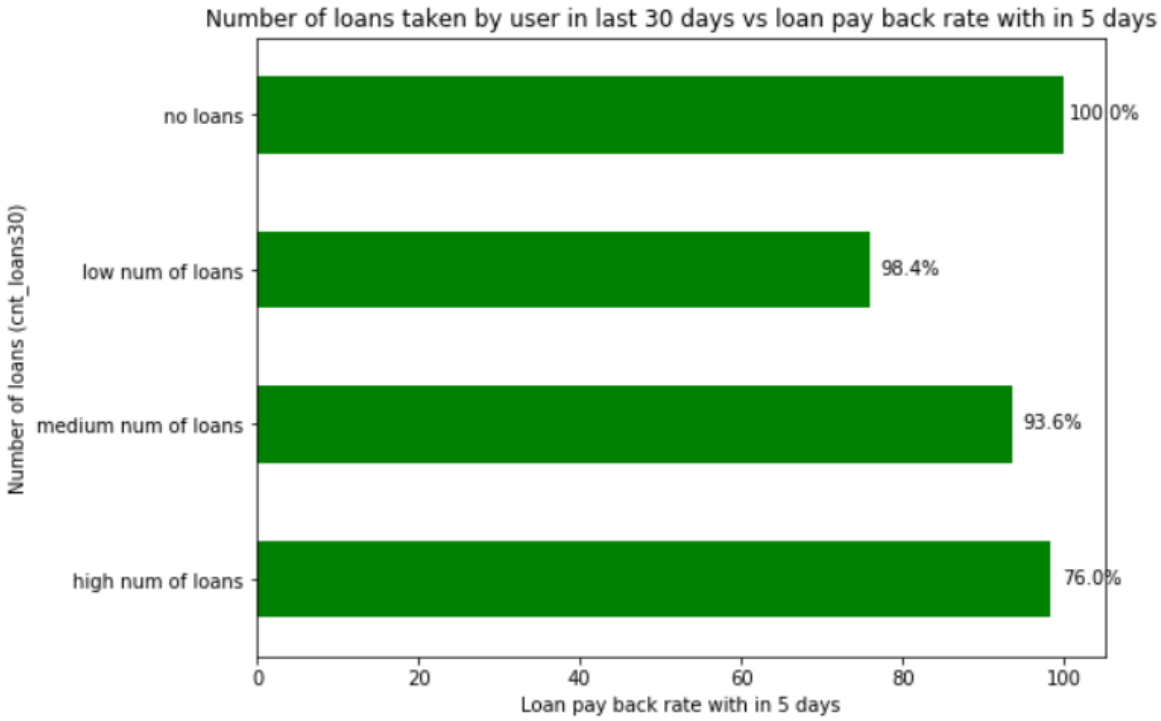


The above bar plot infers us how customers with different frequency levels (main account recharge) are paying back the loan within five days. The is no 100% rate in any of the frequency levels to pay back the loan within 5 days. Coming to the average and low & medium frequency people it is observed that around 5%-6% of people are not paying the loan within 5 days.

Coming to low frequency level people, it is observed that around 25% of people are not paying back the loan with in stipulated 5 days of time. The 25% people who are not getting their main account recharge for 30 days creating a major loss to the company without paying back the loan within five days of time.

In order to decrease loss to the company, the company should start some marketing strategies like sms alerting and notifications and others on the people with all frequency levels and especially on no frequency level people notifying them to pay the loan back within five days of time.

4.3 Number of loans taken by user in last 30 days' vs loan pay back rate within 5 days



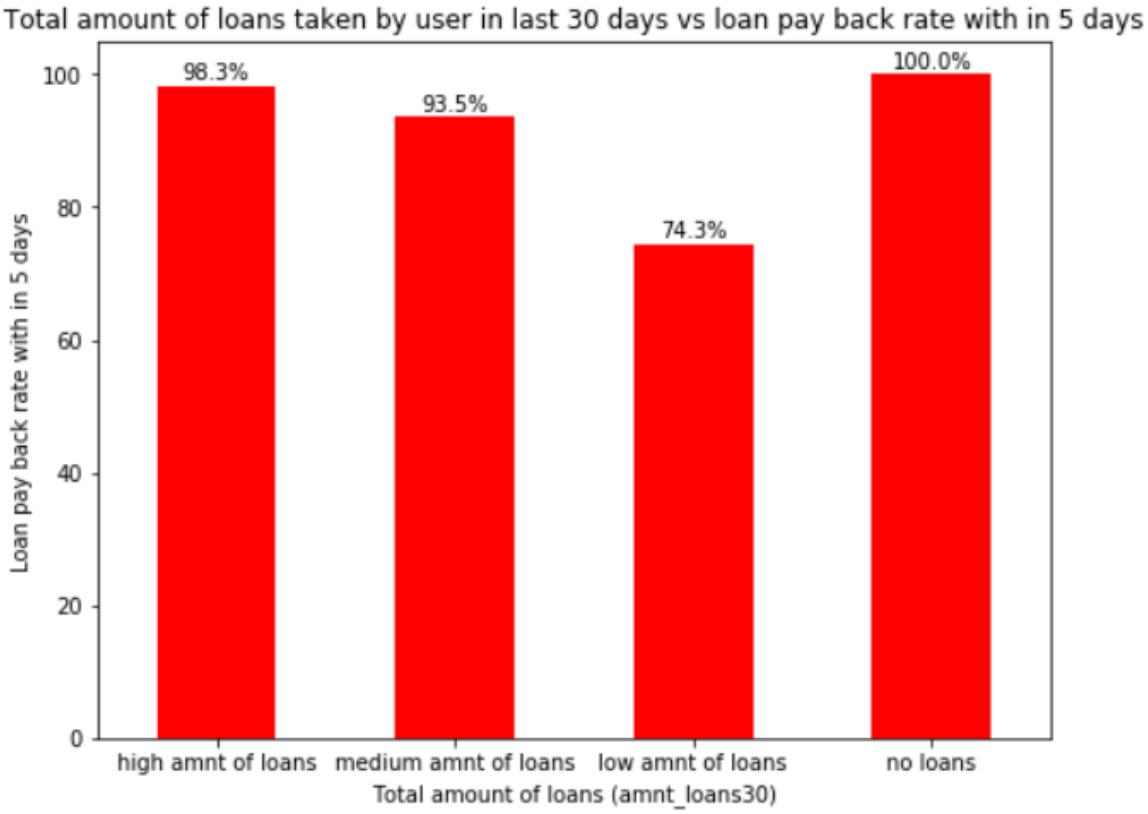
The above bar plot infers us how customers with different loans levels taken are paying back the loan within five days. In the data set people not taken loans are labelled as '1'. So we should not consider the people with no loans labelled in the above graph.

Considering the remaining levels, there is no 100% rate in any of the loan levels to pay back the loan within 5 days. Coming to the high number of loan level people it is observed that around 25% of people are not paying the loan within 5 days.

Only 2% of the people from low number of loans category are not paying the loan within 5 days. This is followed by the people with medium number of loans having defaulters of 7% approximately.

In order to decrease loss to the company, the company should start some marketing strategies like sms alerting and notifications and others on the people with all loan levels and especially on low & high level people notifying them to pay the loan back within five days of time.

4.4 Total amount of loans taken by user in last 30 days' vs loan pay back rate within 5 days



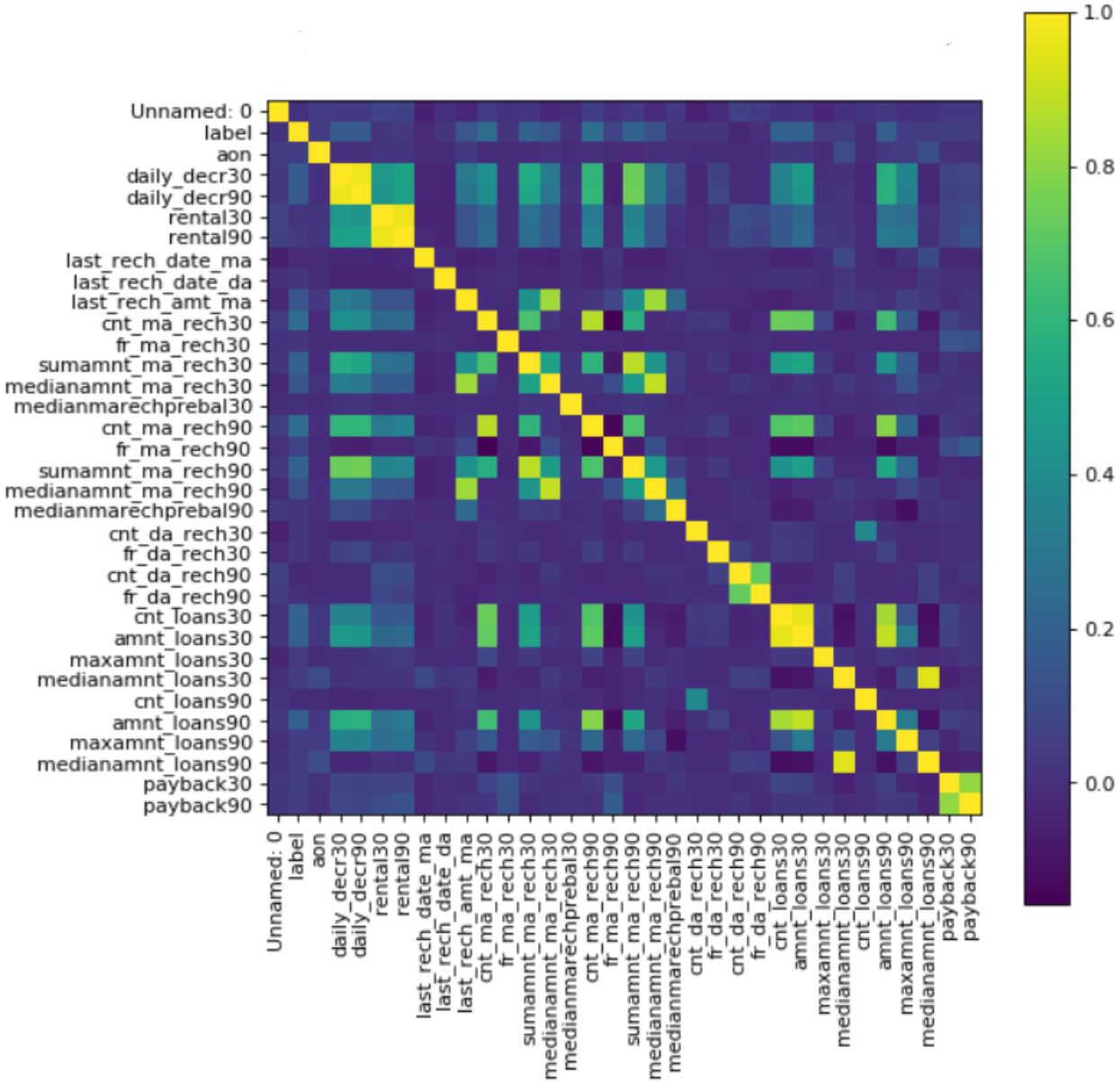
The above bar plot infers us how customers with different loans levels taken are paying back the loan within five days. In the data set people not taken loans are labelled as '1'. So we should not consider the people with no loans labelled in the above graph.

Considering the remaining levels, there is no 100% rate in any of the loan levels to pay back the loan within 5 days. Coming to the low amount level people it is observed that around 25% of people are not paying the loan within 5 days.

Only 2% of the people taken high amount of loans are not paying the loan within 5 days. This is followed by the people with medium number of loans having defaulters of 7% approximately.

In order to decrease loss to the company, the company should start some marketing strategies like sms alerting and notifications and others on the people with all loan levels and especially on low & high level people notifying them to pay the loan back within five days of time.

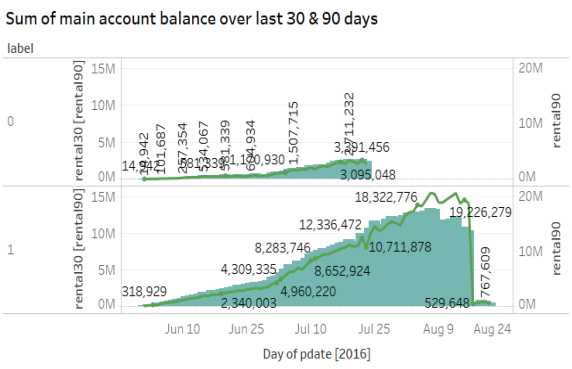
4.5 Correlation Matrix

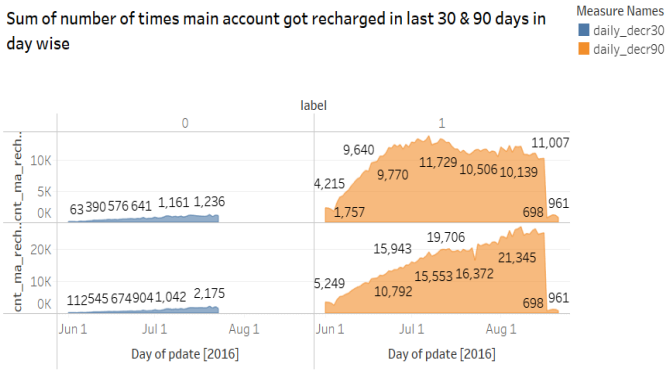


To investigate more about correlation, a correlation matrix was plotted with all qualitative variables. This correlation matrix infers that there is strong correlation (multi collinearity) among the features and it resembles that there is much noise in the data.

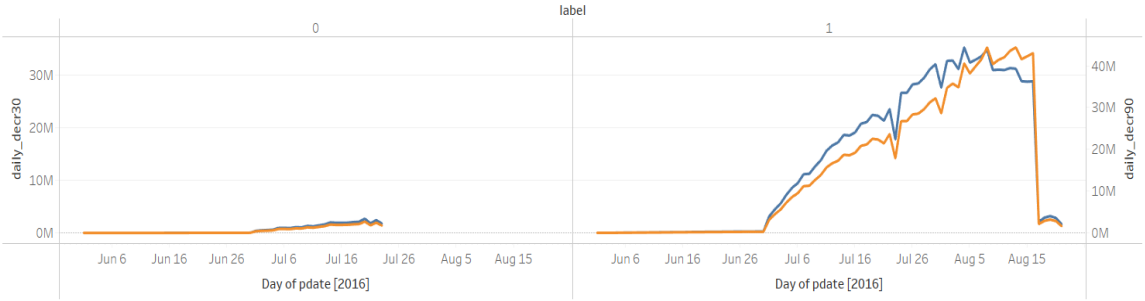
In order to get clear idea about noise or multi collinearity in the data it is advice to perform **VIF (variance inflation factor)**, where the vif values gives the clear about the noise in the data. The vif for this data set is performed below and the multicollinearity effect is also reduced by performing **PCA (Principal component analysis)** on the data set before mode building.

4.6. Plots regarding sum of main account balance, sum of number of times main account got recharged, daily sum of amount spent from main account.

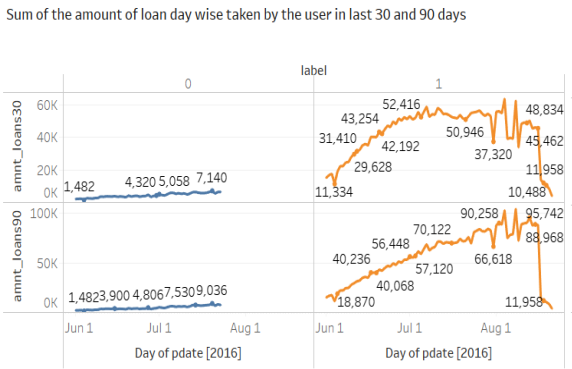


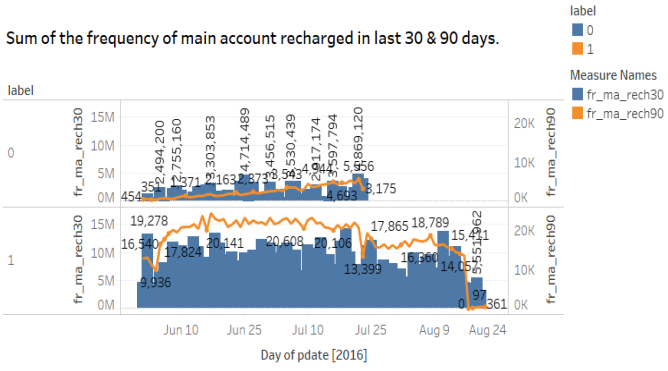


Daily sum of amount spent from main account, averaged over last 30 & 90 days (in Indonesian Rupiah)



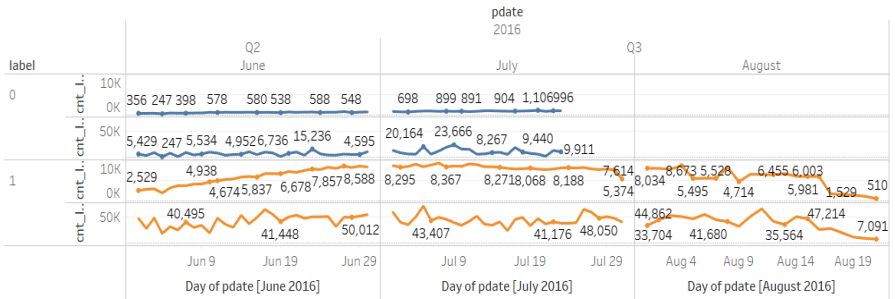
4.7 Plots regarding sum of amount of loans, sum of frequency of main account recharge, share of the label feature, sum of number of loans taken.





Number of 0's and 1's in the data set

Sum of number of loans taken by user in last 30 and 90 days month wise



Part 5 Statistical Analysis

Statistical tests were performed to see the whether the independent variables have a significant relationship with the dependent variable, 'LABEL'

5.1 Two-sample t test

For all the numeric variables, a two-sample unpaired t tests were performed between values of the variable for two classes of target variables to compare their means.

Null Hypothesis H0: The means of the two samples are EQUAL

Alternate Hypothesis Ha: The means of the two samples are NOT EQUAL

If the means of the two samples are significantly different form each other, then we can conclude that the variable does have a significant relationship with the target variable.

Variable	P value	Decision
aon	2.47e-297	**Reject H0**
daily_decr30	0.000	**Reject H0**
daily_decr90	0.000	**Reject H0**
rental30	1.61e-201	**Reject H0**
rental90	1.23e-282	**Reject H0**
last_rech_date_da	7.53e-30	**Reject H0**
last_rech_amt_ma	0.000	**Reject H0**
cnt_ma_rech30	0.000	**Reject H0**
fr_ma_rech30	0.000	**Reject H0**
sumamnt_ma_rech 30	0.000	**Reject H0**
medianamnt_ma_r ech30	0.000	**Reject H0**
medianmarechpre bal30	1.843e-43	**Fail to reject H0**
cnt_ma_rech90	0.000	**Reject H0**

Variable	P value	Decision
fr_ma_rech90	0.000	**Reject H0**
sumamnt_ma_rech9 0	0.000	**Reject H0**
medianamnt_ma_rec h90	0.000	**Reject H0**
medianmarechprebal 90	0.000	**Reject H0**
cnt_da_rech30	0.765	**Fail to reject H0**
fr_da_rech30	0.0108	**Reject H0**
cnt_da_rech90	5.2165e-28	**Reject H0**
fr_da_rech90	0.308	**Fail to reject H0**
cnt_loans30	0.00	**Reject H0**
amnt_loans30	0.000	**Reject H0**
maxamnt_loans30	0.029	**Reject H0**
medianamnt_loans30	2.16e-46	**Reject H0**
cnt_loans90	0.000	**Reject H0**

Variable	P value	Decision
amnt_loans90	0.000	**Reject H0**
maxamnt_loans90	0.000	**Reject H0**
medianamnt_loans 90	0.000	**Reject H0**
payback30	0.000	**Reject H0**
payback90	0.000	**Reject H0**
Month_jun	0.001	**Reject H0**
Month_mar	0.000	**Reject H0**
Month_may	0.000	**Reject H0**
Month_nov	0.000	**Reject H0**
Month_oct	0.000	**Reject H0**
Month_sept	0.000	**Reject H0**
Poutcome_other	0.583	**Fail to reject H0**
Poutcome Success	0.000	**Reject H0**
Poutcome Unknown	0.000	**Reject H0**

Based on the above results it is observed that most of the features p-values is less than 0.05 which indicates that most of the features are statistically significant with the target variable and helps in the prediction of the term depositor's.

The features whose p-value >0.05 are removed are the machine learning models were built on the features whose p value <0.05 only.

As categorical variables such as 'msisdn', 'pcircle', 'pdate' are removed from the data set as they are not having much importance in the model prediction. So as there are no categorical columns in the data set there is no need to perform any categorical statistical test's such as **chi-square, goodness of fit test** for the variables.

5.2 check for multicollinearity (VIF)

Variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

Variable	VIF	Decision
aon	615.529713	Vif>4
daily_decr30	16.105033	Vif>4
daily_decr90	16.871304	Vif>4
rental30	7.059434	Vif>4
rental90	7.748310	Vif>4
last_rech_date_da	1562.711869	Vif>4
last_rech_amt_ma	1585.516831	Vif>4
cnt_ma_rech30	4.974675	Vif>4
fr_ma_rech30	1557.637247	Vif>4
sumamnt_ma_rech30	4.828908	Vif>4
medianamnt_ma_rech30	5.340574	Vif>4
medianmarechprebal30	251.139277	Vif>4
cnt_ma_rech90	5.580615	Vif>4

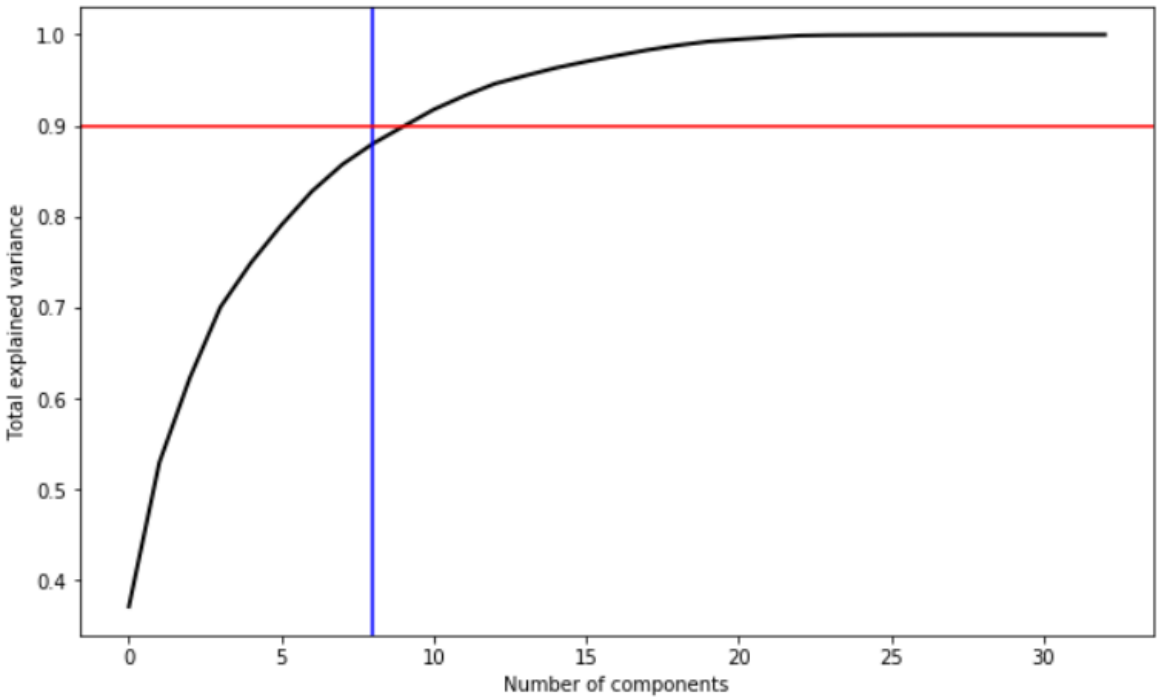
Variable	VIF	Decision
sumamnt_ma_rech90	5.418015	Vif>4
medianamnt_ma_rech90	5.359924	Vif>4
cnt_da_rech30	121.676642	Vif>4
fr_da_rech30	1565.521928	Vif>4
fr_da_rech90	87.795789	Vif>4
cnt_loans30	13.009286	Vif>4
amnt_loans30	16.175951	Vif>4
maxamnt_loans30	131.313993	Vif>4
medianamnt_loans30	9.761667	Vif>4
cnt_loans90	30.233758	Vif>4
amnt_loans90	5.934311	Vif>4
medianamnt_loans90	10.424172	Vif>4
payback30	4.084829	Vif>4

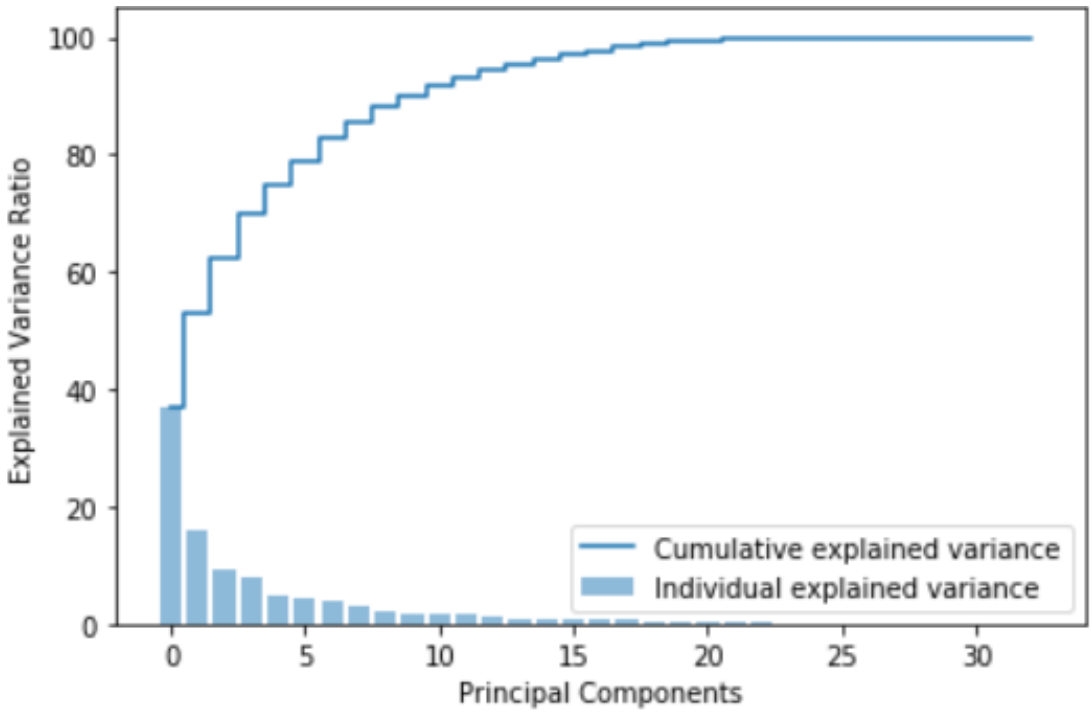
From the above results we can infer that the many features are having strong multicollinearity in the data set. This resembles that there is need to go for PCA (Principal Component Analysis).

If we won't perform PCA the noise or correlation between the independent variables will affect the model prediction and model results.

More than 50% of the features are having vif >4 so it is mandatory to perform PCA in order to reduce the multicollinearity effect among the independent variables.

5.3 PCA Results of the data set





From the above results it is observed that 90% of the data is covered at the pca components (n=13). So total number of pca components were taken as 13.

Part 6 Machine Learning: Classification

The main objective of this project is to identify whether the customer is paying back the loan within 5 days of time period or not. To achieve this objective, classification algorithms will be employed. By analysing customer statistics, a classification model will be built to classify all clients into two groups: "Yes" and "No".

Prepare Data for Classification

Select the most relevant customer information i.e feature whose p-value <0.05 and the data on which PCA was performed with n=13. There are no categorical variables in the data set as they were removed them before because of having less importance for prediction. So there is no need to perform encoding techniques such as dummy encoding one hot encoding. 70% of the data was used to build the classification model and 30% was reserved for testing the model.

Part 7 Evaluation Metrics

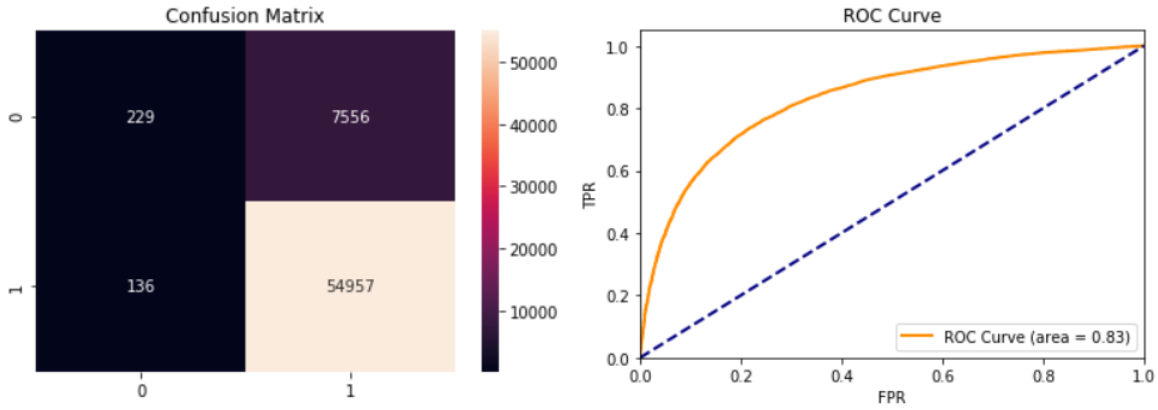
The Evaluation Metrics that can be used for a Binary Classification problem are:

- ✓ Accuracy - Proportion of correctly identified instances
- ✓ Precision - proportion of positive predictions that are correct
- ✓ Recall - Proportion of Actual positives predicted correctly
- ✓ F1 Score - Harmonic mean of Precision and Recall
- ✓ ROC AUC - Area Under Receiver's Operating Characteristics Curve (tradeoff between sensitivity and specificity for different thresholds)

Part 8 Base Model

Base classification algorithms (Logistic Regression) was run on the dataset and identified the classification metrics such as Accuracy, Precision, Recall, F1-score.

Accuracy	0.877667864753968
Precision	0.8791291411386432
Recall	0.9975314468262756
F1-Score	0.9345951737156268

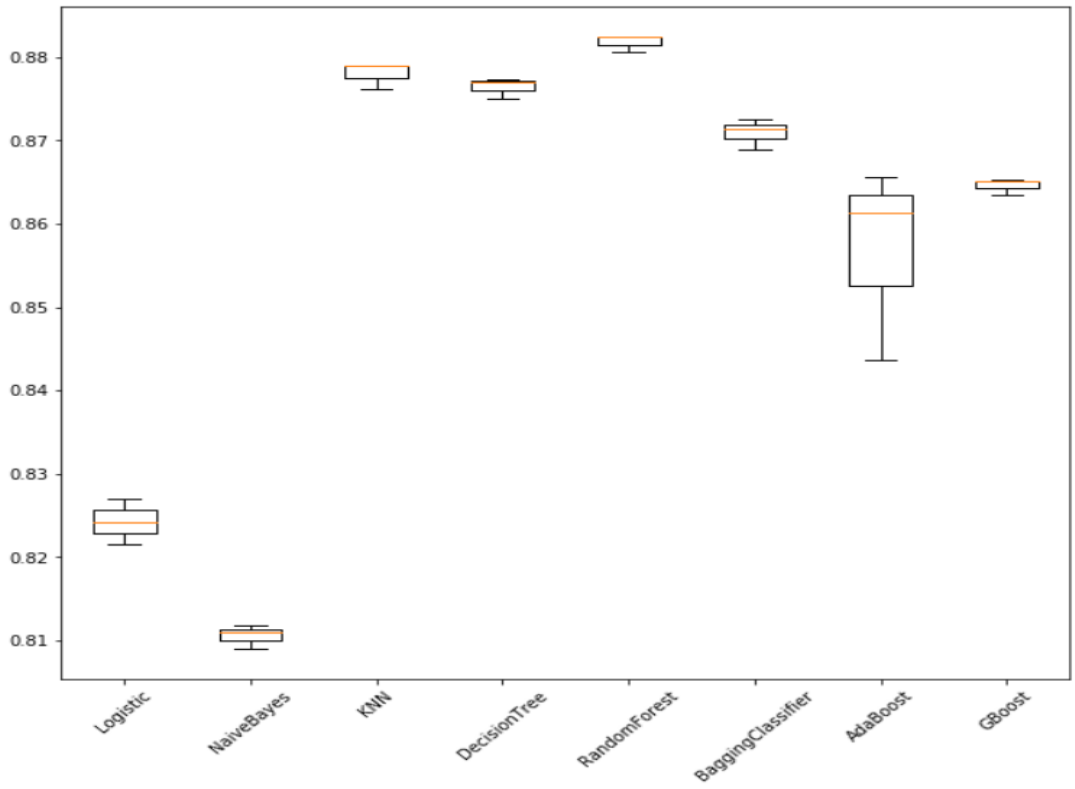


From the above results it is observed that all the metrics accuracy, precision, recall and f1-score are good. In order to improve the score very much high I also tried with various other models such as decision tree, random forest, naïve bays, knn, and ensemble models also.

Four different classification algorithms (Logistic Regression, K-Neighbours Classifier, Decision Tree Classifier, and Gaussian NB, Random Forest, Ada boost, Gradient Boosting) were run on the dataset through K-fold cross validation and the best-performing one was (identified by observing bias and variance errors) and used to build the classification model.

Classification Algorithm	Bias Error	Variance Error
Logistic Regression	0.824303	0.000007
NaiveBayes	0.810621	0.000002
KNN	0.87999	0.000003
DecisionTree	0.876401	0.000001
RandomForest	0.881770	0.000001
BaggingClassifier	0.870911	0.000003
AdaBoost	0.856886	0.000136
GBoost	0.864559	0.000001

Algorithms Comparison through boxplots:

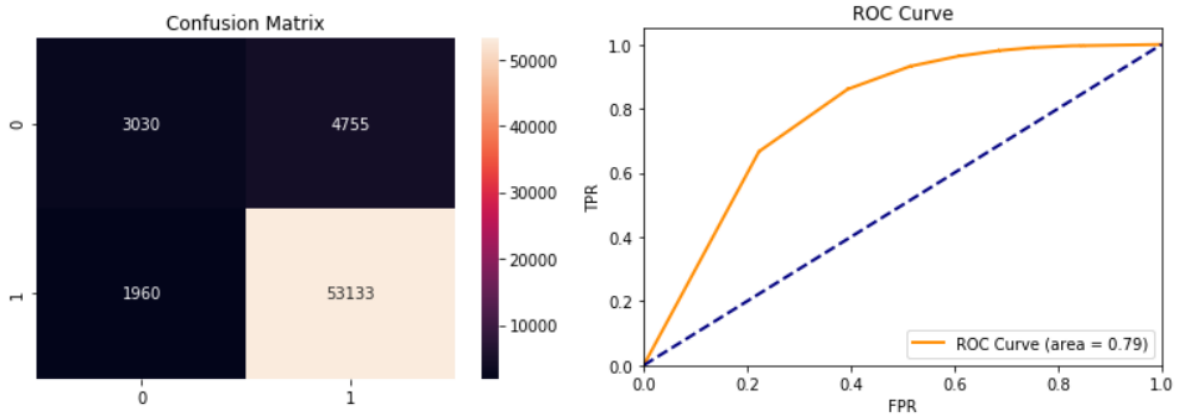


From the above results it is observed that Random Forest is the best performing model.

By comparing all algorithms bias error and variance error, random forest is observed to be the best so it would be used to predict loan defaulters. The test of random forest with base estimator (Decision Tree (which is default for random forest), n_estimators=7) model successfully achieved a weighted F1_score of 98%, suggesting high level of strength of this model to classify loan defaulter's.

Best performing classification algorithms (Random Forest) was run on the dataset which was under sampled and identified the classification metrics such as Accuracy, Precision, Recall, F1-score.

Accuracy	0.8932058907726073
Precision	0.917858623548922
Recall	0.9644237924963244
F1-Score	0.9405652277816624



Classification Report:

Total no of records in X_test and y_test is 62,878. The results of our random forest classifier are better than the before base line model.
TP=53133
TN=3030
FP=1960 (type-1 error)
FN=4755 (type-2 error)

From the above results we can infer that only 10% of the data are under type-1 & type- 2 errors. Our model is unable to classify well only 10% records whether the customer is loan defaulter or not.

Part 9 Conclusion

The main objective of this project is to classify the customer's whether they are loan defaulter's or not, which was successfully met through data analysis, visualization and analytical model building. A target customer profile was established while classification models were built to predict customers' will pay back loan within five days or not.

By applying Random Forest classifier algorithm, classification and estimation model were successfully built. With this model, the telecom company will be able to predict a customer's behaviour on the loan payment.