Micro-Credit Defaulter Model

**INTRODUCTION**

* **Business Problem Framing**

This project includes the real time problem for Microfinance Institution (MFI) offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income, MFI provides micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).

* **Conceptual Background of the Domain Problem**

Generally, Credit Scores plays a vital role for loan approvals, and is very important in today’s financial analysis for an individual, Most of the loan lending vendors rely heavily on it, so in our case users has 5 days’ time to pay back the loan or else they are listed as defaulters which will impact the loan the credit score heavily, so there are few thing to lookout in this dataset as users who are taking extensive loans, user who have most frequent recharges in their main account have a good chance of 100% payback rate, and user who never recharged their main account for them loan should have never been approved as there is high chance for single user or default user taking multiple connections in name or documents of the family members.

* **Review of Literature**

Random Forest are evaluated for their ability to classify defaulters using several cross-validation approaches and the latter model performed best with low log loss and high f1 score. When the default rate is below 2%, it is better to offer everyone a loan. For higher default rates, the model Substantially enhances profitability. The model quadruples the tolerable level of default rate for breaking even from 8% to 32%. Nonlinear classification models offer considerable potential for credit scoring, coping with higher levels of default and therefore allowing for larger volumes of customers.

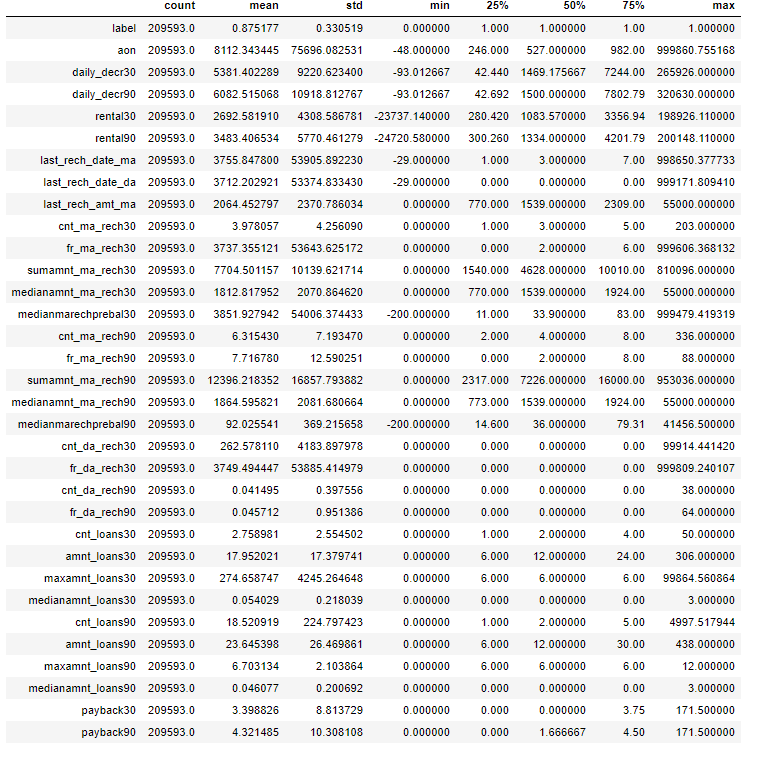
* **Motivation for the Problem Undertaken**

This project was highly motivated project as it includes the real time problem for Microfinance Institution (MFI), and to the poor families in remote areas with low income, and it is related to financial sectors, as I believe that with growing technologies and Idea can make a difference, there are so much in the financial market to explore and analyse and with Data Science the financial world becomes more interesting.

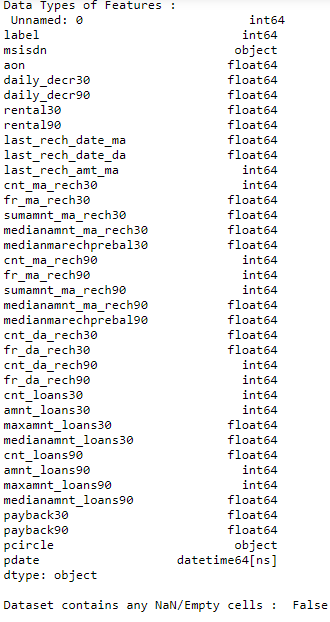
**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

This problem is a classification problem,the target variable is itself a stastistical parameter.we have to 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. In this case, Label ‘1’ indicates that the loan has been payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed .for a loan amount of 5 payback amount should be 6,and for loan amount of 10 payback amount is 12.

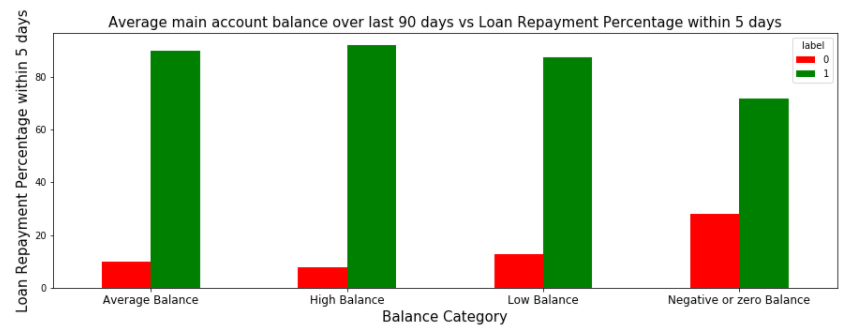
****

* From an initial statistical overview of the dataset, we infer that some data features are binary or ordinal, whereas other features are continuous. Further, the minimum is negative which is not even possible for most of the features notably daily recharge , main account balance, aon, and last recharge which can't be negative and maximum values for some features, notably for aon ,maxamnt\_loans30,medianmarechprebal90,medianmarechprebal30 are unrealistic. Most the features has mode is greater than median this suggests the presence of outliers in the data and All Features are not Normally Distributed( Theortically if feature is normally distributed, Mean = Median = Mode ) like weight and height are right and left skewed.
* **The Dataset we are having, consists of some features giving information about the user for the time span of 30 days and 90 days. According to me if we have data of large number of days for a particular user then we could interpret User's behavior more precisely because many users have the tendancy of repeating the same things. Thus the features having the data with a time span of 90 days gives more information about the user as compared to the features with a time span of 30 days.**
* **All the categories that is being made to make the visualizations easy are solemnly based on the Description i.e statistical summary of the data plotted above *for instance* low comes under(0-25%), average comes under(25-75%) and high comes over 75% of the data values in a given feature.**
* Using MS EXCEL I have found the maximum values a feature can have, beyond these values the values are unimaginable.
* ***\*\*(for an example beyond the value [2500], the very next value in "aon" feature comes out to be around 2379 years, which means a user is using the telephone services from 359 BCE which is clearly not possible).\*\****
* I checked the correlation of the independent and dependent features and f**rom the correlation table it is also clear that the features with time span of 30 and 90 days almost have the same correlation thus we can drop one for the same information.**
* **Data Sources and their formats**
* **label :** Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}
* **msisdn :** mobile number of user
* **aon :** age on cellular network in days
* **daily\_decr30:** Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
* **daily\_decr90:** Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)
* **rental30:** Average main account balance over last 30 days
* **rental90:** Average main account balance over last 90 days
* **last\_rech\_date\_ma:** Number of days till last recharge of main account
* **last\_rech\_date\_da:** Number of days till last recharge of data account
* **last\_rech\_amt\_ma:** Amount of last recharge of main account (in Indonesian Rupiah)
* **cnt\_ma\_rech30:** Number of times main account got recharged in last 30 days
* **fr\_ma\_rech30:** Frequency of main account recharged in last 30 days
* **sumamnt\_ma\_rech30:** Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)
* **medianamnt\_ma\_rech30:** Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)
* **medianmarechprebal30:** Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)
* **cnt\_ma\_rech90:** Number of times main account got recharged in last 90 days
* **fr\_ma\_rech90:** Frequency of main account recharged in last 90 days
* **sumamnt\_ma\_rech90 :** Total amount of recharge in main account over last 90 days (in Indian Rupee)
* **medianamnt\_ma\_rech90:** Median of amount of recharges done in main account over last 90 days at user level (in Indian Rupee)
* **medianmarechprebal90:** Median of main account balance just before recharge in last 90 days at user level (in Indian Rupee)
* **cnt\_da\_rech30:** Number of times data account got recharged in last 30 days
* **fr\_da\_rech30:** Frequency of data account recharged in last 30 days
* **cnt\_da\_rech90:** Number of times data account got recharged in last 90 days
* **fr\_da\_rech90:** Frequency of data account recharged in last 90 days
* **cnt\_loans30:** Number of loans taken by user in last 30 days
* **amnt\_loans30:** Total amount of loans taken by user in last 30 days
* **maxamnt\_loans30:** maximum amount of loan taken by the user in last 30 days
* **medianamnt\_loans30:** Median of amounts of loan taken by the user in last 30 days
* **cnt\_loans90:** Number of loans taken by user in last 90 days
* **amnt\_loans90:** Total amount of loans taken by user in last 90 days
* **maxamnt\_loans90:** maximum amount of loan taken by the user in last 90 days
* **medianamnt\_loans90:** Median of amounts of loan taken by the user in last 90 days
* **payback30:** Average payback time in days over last 30 days
* **payback90:** Average payback time in days over last 90 days
* **pcircle:** telecom circle
* **pdate:** date

****

* **Data Preprocessing Done**
* I checked the correlation of the independent and dependent features and dropped the negative and less important features with the help of correlation matrix.
* There were data for 30 and 90 days, so considering data for 90 days is adding more information rather than then data of 30 days.
* Some features can’t have any negative value, so those features were treated accordingly.
* Outliers are treated manually for the features giving some important information, and then the threshold values were set to make the data freee from outliers.
* Data lost is very less i.e less than the 7% which was stated in the documentation.
* Applied SMOTETomek, to balance the dataset as the dataset was imbalanced dataset.
* Applied StandardScaler to our dependent features.
* Applied various machine learning model and compared it.
* Applied hyper tunning several models, but couldn’t achieve much better results.
* Saving final predictions in file.csv format
* **Data Inputs- Logic- Output Relationships**

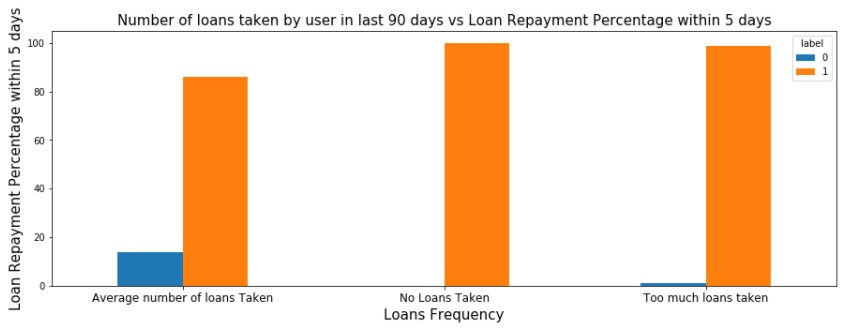
**i)**



**From the above Graph and the crosstab table it is clear that:**

1) 28% of Users having negative or zero balance are defaulters, which is very high.  
2) 10% to 12% Users are defaulters which falls in the category of Average and Low balancecategory.  
3) Users having high balance and are defaulters are very less in number

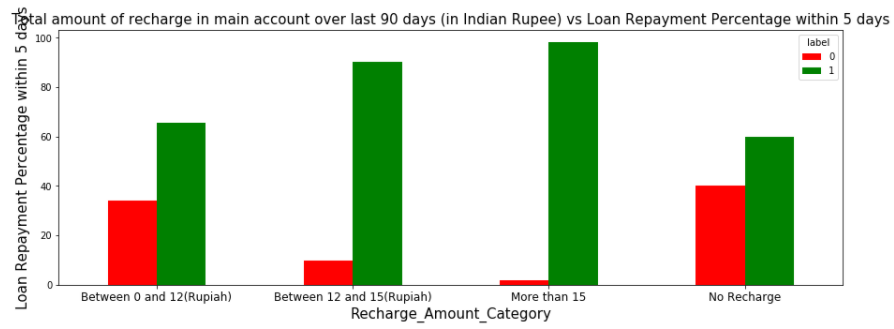
**ii)**



**From the above graph it is clear that:**

1) Users who take more number of loans are non-defaulters (i.e. 98% of the category) as they repays the loan within the given time i.e. 5 days.  
2) 14% of the Users are are among the average number of loan taken category are defaulters.

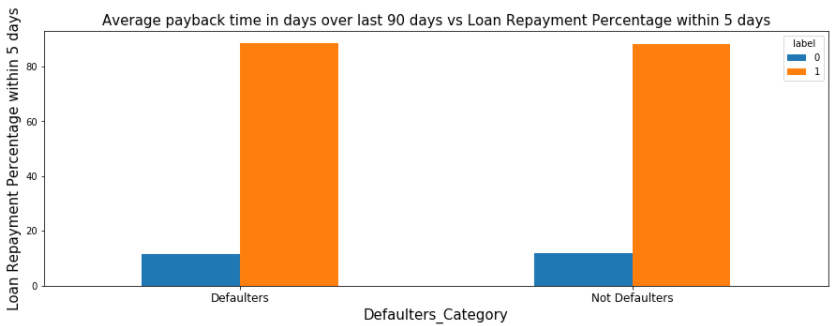
**iii)**

****

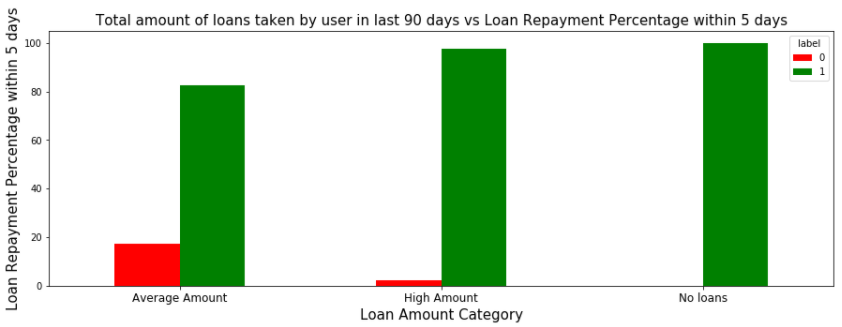
**From the above graph it is clear that:**

1) 40 % of the Users who do not even recharged in the 90 days are defaulters only.  
2) Users who do very high amount of recharge always pays their loans on time. i.e 98% of them are non-defaulters.   
3) 34% of the Users who do less amount of recharge are defaulters.

**iv)**

****

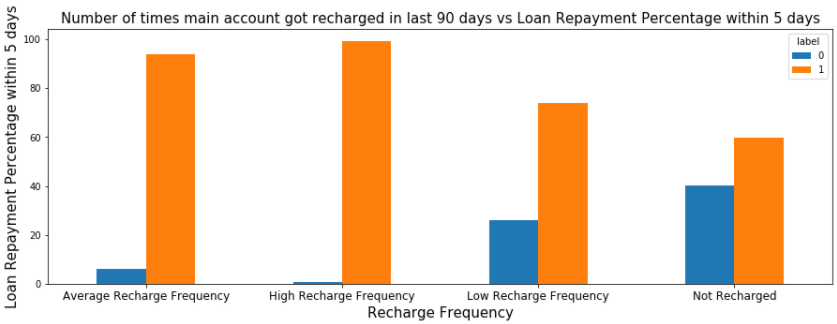
**V)**

****

**From the above graph it is clear that:**

1) Users who did not take any loans are non-defaulters.  
2) Most of the Users (i.e. 97%) who take large amount of loans comes under non defaulter category.  
3) 17% of the users who take small loans are defaulters.

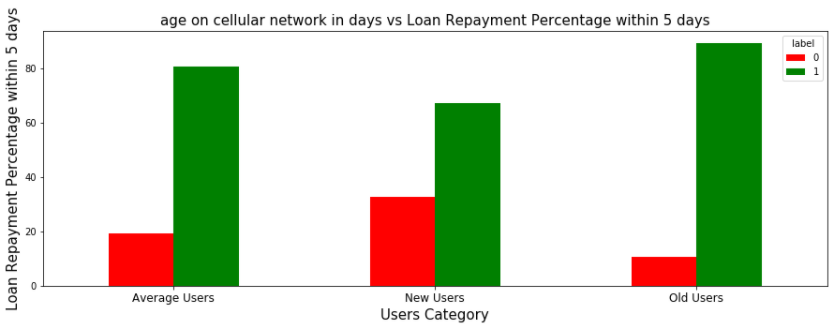
**Vi)**

****

**From the above graph it is clear that:**

1) Among the Users who have not done a single recharge in 3 months 40% are defaulters.  
2) Among the Users who are very frequent in recharging and who always pay their loans on time are more in number i.e 99% of the total category, which is a good news for the company.

**Vii)**

****

**From the above graph it is clear that:**

1) 32% of the users who are defaulters are the new users.  
2) Old Users are trusted and they are mostly non defaulters.

* **State the set of assumptions (if any) related to the problem under consideration**
* From the above statistical summary of the above part of the dataset, **the important thing is that Some features even have negative values like the age on cellular network, main account last recharge date, data account last recharge date. Negative values in these features make no sense thus these values should be removed.**
* **The Dataset we are having, consists of some features giving information about the user for the time span of 30 days and 90 days. According to me if we have data of large number of days for a particular user then we could interpret User's behavior more precisely because many users have the tendancy of repeating the same things. Thus the features having the data with a time span of 90 days gives more information about the user as compared to the features with a time span of 30 days.**
* **All the categories that is being made to make the visualizations easy are solemnly based on the Description i.e statistical summary of the data plotted above *for instance* low comes under(0-25%), average comes under(25-75%) and high comes over 75% of the data values in a given feature.**Using MS EXCEL I have found the maximum values a feature can have, beyond these values the values are unimaginable.
* ***\*\*(for an example beyond the value [2500], the very next value in "aon" feature comes out to be around 2379 years, which means a user is using the telephone services from 359 BCE which is clearly not possible).\*\****
* I checked the correlation of the independent and dependent features and f**rom the correlation table it is also clear that the features with time span of 30 and 90 days almost have the same correlation thus we can drop one for the same information.**
* State the set of assumptions (if any) related to the problem under consideration
* There were data for 30 and 90 days, so considering data for 90 days is adding more information rather than then data of 30 days.
* Some features can’t have any negative value, so those features were treated accordingly.

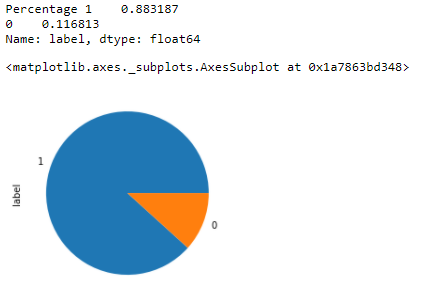
* **Hardware and Software Requirements and Tools Used**
* Hardware: 8GB RAM, 64-bit, 9th gen i7 processor.
* Software: Excel, Jupyter Notebook, python 3.6.

**Libraries used**:-



**Model/s Development and Evaluation**

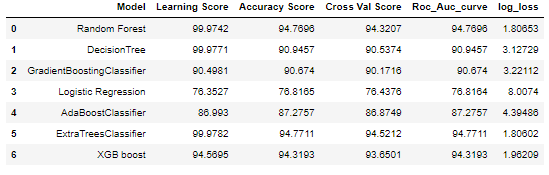
* **Identification of possible problem-solving approaches (methods).**

****

* From the above we can see that the data set is highly imbalanced dataset, so applied SMOTETomek to balance the dataset.
* **Testing of Identified Approaches (Algorithms)**
* RandomForestClassifier ()
* DecisionTreeClassifier ()
* GradientBoostingClassifier
* LogisticRegression()
* AdaBoostClassifier()
* ExtraTreesClassifier()
* XGBClassifier()
* **Run and Evaluate selected models**



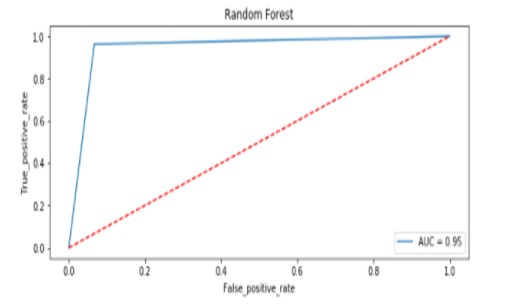
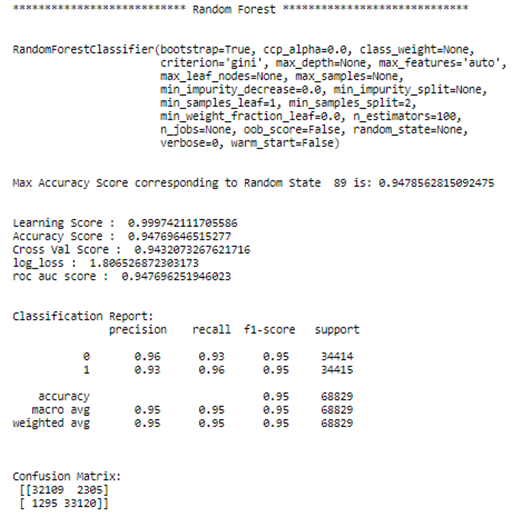
* **Key Metrics for success in solving problem under consideration**



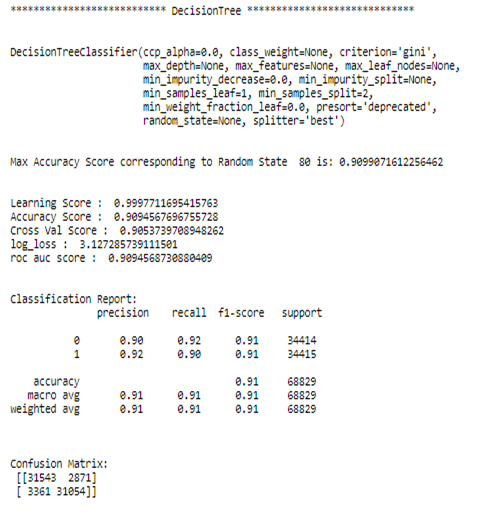
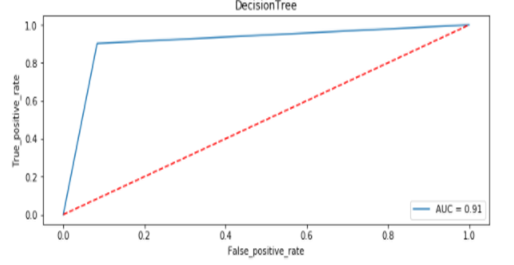
Key Metrices used were the Accuracy Score, Crossvalidation Score and AOC & ROC Curve ,log\_loss as this was binary classification problem unbalanced dataset and we focus more on AOC & ROC curve metrices to observe True Positive Rate and False Positive Rare, for users who paid the loan and falsely marked as default and will their affect the credit score and we already talked about the importance of that in financial sector, and for the users who are marked falsely marked as paid but they didn’t, can affect the company revenue.

**Visualizations**

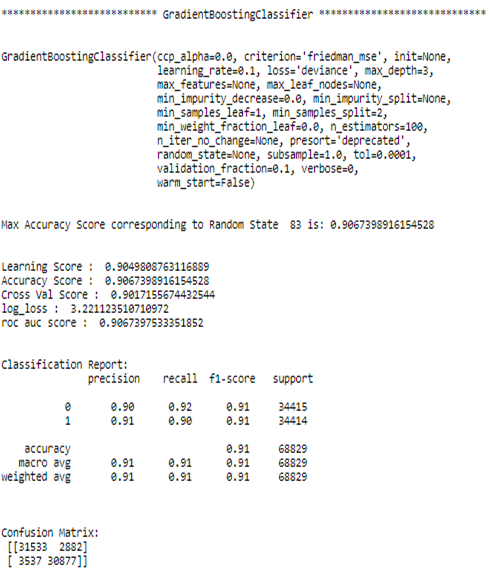
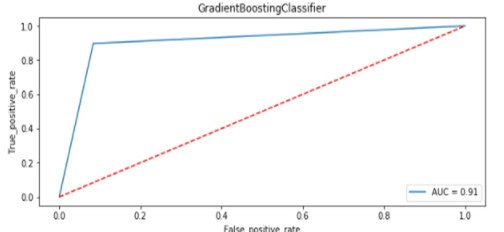
(i)Random forest



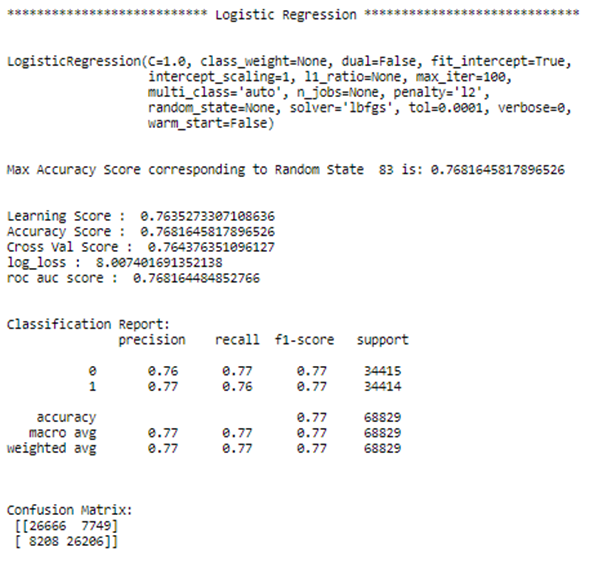
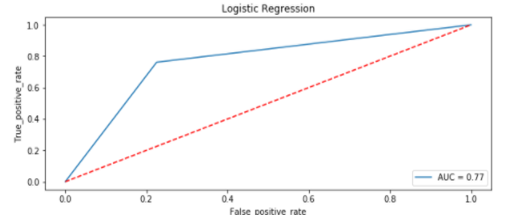
ii) Decision Tree



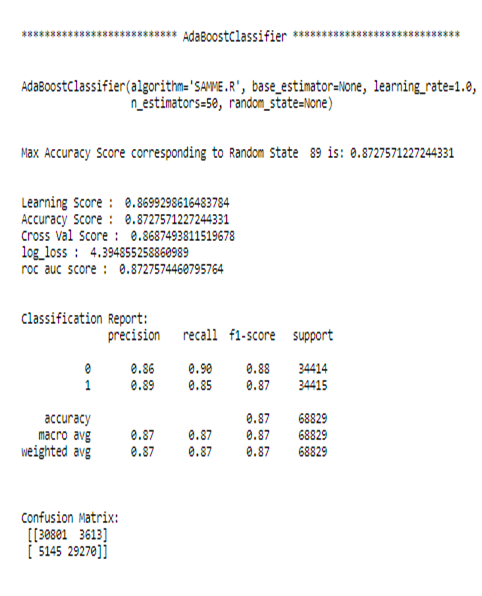
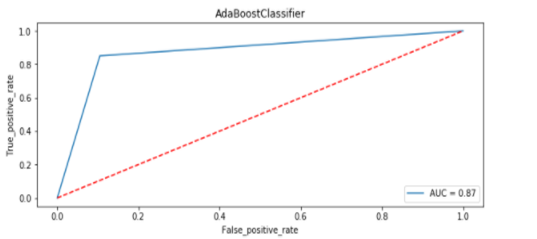
iii) Gradient Boosting Classifier



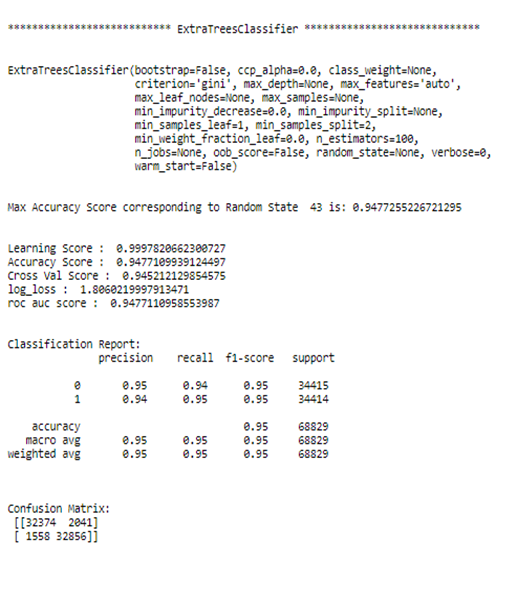
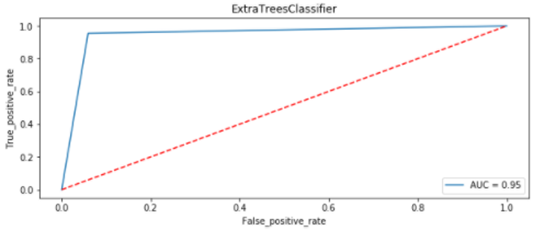
iv) Logistic Regression



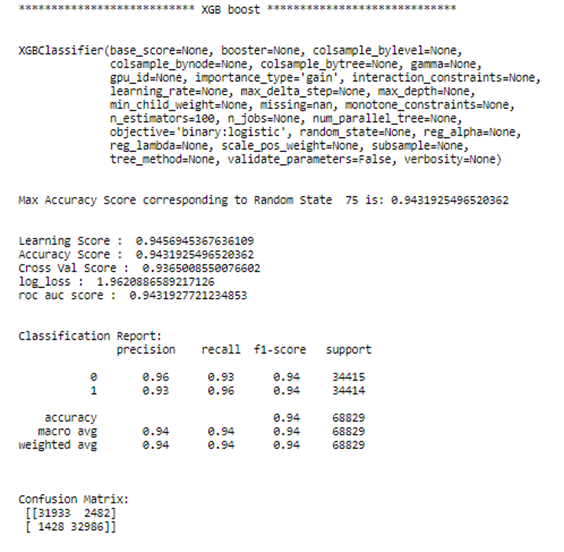
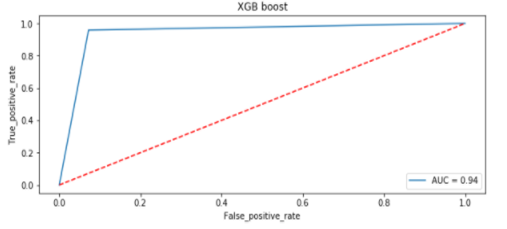
v) AdaBoostClassifier



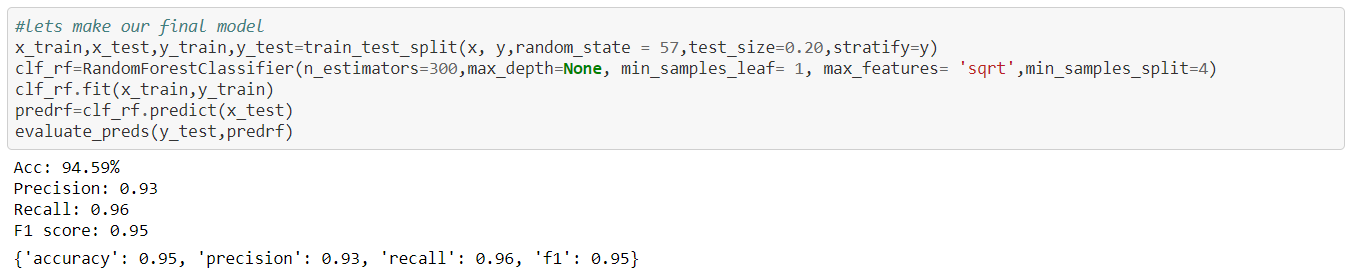
vi) ExtraTreeClassifier

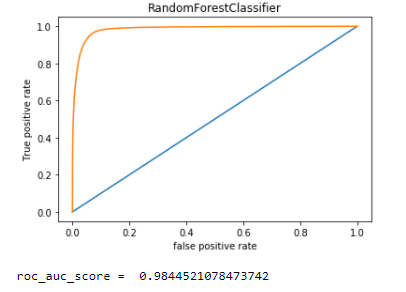


vii) XGB Boost Classifier



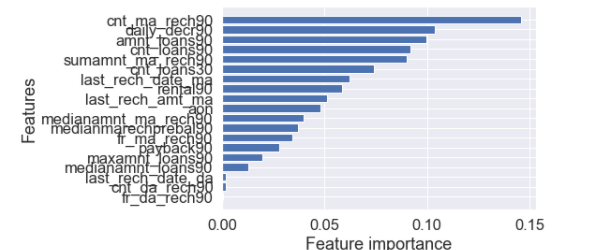
* After hypertunning random forest perfoms well as compared to other models





From the above visualization and matrices found that the RandomForest Classifier performed the best 98% AOC\_ROC\_SCORE, with precision accuracy score of 95% and recall 96%, however the max score which we were able to achieve from dataset provided.

**Feature Importance :-**  Feature importance seeks to figure out which different attributes of the data were most importance when it comes to predicting the target variable (label)

****

* Cnt\_ma\_rech90 is contribute the most as compared to other features

**CONCLUSION**

**Key Findings and Conclusions of the Study**

From the whole evaluation we found that the MFIs have provided loan to the user who have no recharge or balance in their account which needs to be stopped as 30% defaulted user are from that type, and few high frequency loan takers and among users maintaining high balances are absorbed that 8% to 10% users are defaulted and some SMS altering notification before the deadlines can play a major role, in reducing the default rate.

1) 28% of Users having negative or zero balance are defaulters, which is very high.2) 10% to 12% Users are defaulters which falls in the category of Average and Low balancecategory.3) Users having high balance and are defaulters are very less in number4) Users who take more number of loans are non-defaulters (i.e. 98% of the category) as they repays the loan within the given time i.e. 5 days.5) 14% of the Users are are among the average number of loan taken category are defaulters.6) 40 % of the Users who do not even recharged in the 90 days are defaulters only.7) Users who do very high amount of recharge always pays their loans on time. i.e 98% of them are non-defaulters. 8) 34% of the Users who do less amount of recharge are defaulters.9) Users who did not take any loans are non-defaulters.10) Most of the Users (i.e. 97%) who take large amount of loans comes under non defaulter category.11) 17% of the users who take small loans are defaulters.12) Among the Users who have not done a single recharge in 3 months 40% are defaulters.13) Among the Users who are very frequent in recharging and who always pay their loans on time are more in number i.e 99% of the total category, which is a good news for the company.14) 32% of the users who are defaulters are the new users.15) Old Users are trusted and they are mostly non defaulters.16)Random forest performs the best as compared to others models with high f1 score of 95% and roc\_auc score 98% .17)Cnt\_ma\_rech90 is contribute the most as compared to other features

**Learning Outcomes of the Study in respect of Data Science**

* Visualizations and Data Cleaning part was very crucial as without the cleaning we were not able to judge the data effectively and won’t be able to remove the outliers thus adding in to the errors.
* Visualiztions helped a lot in finding out those outliers values and helped in finding out the features having direct relation between the feature and the label.

**Limitations of this work and Scope for Future Work**

* Machine Learning Algorithms like Extratreeclassifier and XGB took enormous amount of time to build the model so that’s why I consider random forest as a final model.
* One of the limitations to my approach in comparing the models is that as it run all the models on the same time, so if we want to make changes or implement any experiments, all models runs again unnecessarily, so need to define a function for evaluation metrices and pass each algorithm, so that we can manual tune and play with it more.
* I didn’t use the pca because I just find out the importance features for the model .So If I could use PCA it might will take less for tunning the model.