

***MICRO CREDIT LOAN DEFAULTER***

**Submitted by:**

**Shubham Sahu**

# ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot. I am also grateful to Mr. Harsh Ayush for his constant guidance and support.

Some of the reference sources are as follows:

* Internet
* Coding Ninjas
* Medium.com
* Analytics Vidhya
* Secondary research papers
  + <https://www.academia.edu/44389277/Microfinance_in_Bangladesh_A_Case_Study_on_Islamic_Microfinance>
* Articles
  + <https://medium.com/digital-catapult/dealing-with-imbalanced-data-8b21e6deb6cd>
* Stackoverflow.com

*TABLE OF CONTENTS*

[**ACKNOWLEDGMENT** 2](#_Toc65131651)

[**INTRODUCTION** 1](#_Toc65131652)

[BUSINESS PROBLEM FRAMING 1](#_Toc65131653)

[CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM 1](#_Toc65131654)

[REVIEW OF LITERATURE 2](#_Toc65131655)

[MOTIVATION FOR THE PROBLEM UNDERTAKEN 3](#_Toc65131656)

[**ANALYTICAL PROBLEM FRAMING** 3](#_Toc65131657)

[MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM 3](#_Toc65131658)

[DATA SOURCES AND THEIR FORMATS 5](#_Toc65131659)

[DATA PREPROCESSING DONE 8](#_Toc65131660)

[DATA INPUTS- LOGIC- OUTPUT RELATIONSHIPS 10](#_Toc65131661)

[STATE THE SET OF ASSUMPTIONS (IF ANY) RELATED TO THE PROBLEM UNDER CONSIDERATION 14](#_Toc65131662)

[HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED 16](#_Toc65131663)

[**MODEL/S DEVELOPMENT AND EVALUATION** 19](#_Toc65131664)

[IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES (METHODS) 19](#_Toc65131665)

[TESTING OF IDENTIFIED APPROACHES (ALGORITHMS) 20](#_Toc65131666)

[RUN AND EVALUATE SELECTED MODELS 21](#_Toc65131667)

[KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION 38](#_Toc65131668)

[VISUALIZATIONS 39](#_Toc65131669)

[INTERPRETATION OF THE RESULTS 49](#_Toc65131670)

[**CONCLUSION** 50](#_Toc65131671)

[KEY FINDINGS AND CONCLUSIONS OF THE STUDY 50](#_Toc65131672)

[LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE 51](#_Toc65131673)

[LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK 52](#_Toc65131674)

# 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

Mobile financial services (MFS) are a very lucrative business as the returns are high but there is considerable risk of default involved. In our specific application, the telecom company in collaboration with a Microfinance Institute (MFI) provides loans of amount 5 and 10 (Indonesian Rupiah) for a very short period and the payback amount is 6 and 12 (Indonesian Rupiah) respectively which corresponds to a high interest rate of 20% in a very short period (usually 5 days). While the return is high, there is considerable risk of default involved, because the loan is being provided to low income populations.

Therefore it is necessary to classify all the defaulters to minimize business risk and avoid losses. The sample data is provided to us from our client database to classify defaulters which would help them in further investment and improvement in selection of customers.

We will use machine learning classification algorithms to predict the defaulters based on the sample data provided by the client.

## REVIEW OF LITERATURE

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. Microfinance services (MFS) becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The MFS provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Indonesian Telecom company is collaborating with an MFI to provide 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).

The sample data which is provided to us to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

We have used machine learning model to predict the above. Since we have categorical data so Classification model technique has been used.

We will begin our project with the sample dataset which contains loan default status along with associated features. We will look at all the features with following goals in mind:

* Relevance of the feature
* Distribution of the feature
* Cleaning the feature
* Visualization of the feature
* Visualization of the feature as per loan default status for data analysis

After having gone through all the features and cleaning the dataset, we will move on to machine learning classification modelling:

* Pre-processing the dataset for models
* Testing multiple algorithms with multiple evaluation metrics
* Select evaluation metric as per our specific business application
* Hyper-parameter tuning using GridSearchCV for the best model parameter
* And finally saving the best model

## MOTIVATION FOR THE PROBLEM UNDERTAKEN

The project was the first provided to me by Flip Robo Technologies as a part of the internship programme. The exposure to real world data and the opportunity to deploy my skillset in solving a real time problem has been the primary motivation. Further diving into the dataset, the motive is to help the poor or low-income band to have continuous access to their mobile accounts, and to make emergency calls even when they do not have account balance making use of the loan facility

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.

The objective of the project is to prepare a model based on the sample dataset that classifies all loan defaulters and help our client in further investment and improvement in selection of customers.

# ANALYTICAL PROBLEM FRAMING

## MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM

The dataset is a csv file with 37 attributes (36 features and 1 target). The target variable is either 1 or 0 which means non defaulter and defaulter respectively. The other key attributes are the account balances, days since last recharge, age on network, median recharge balance for 30 and 90 days and many more. The similar attributes for 30 and 90 days are highly correlated and conveys the same. Hence for the purpose of the project, highly correlated attributes needs to be removed

The statistical figure I get to know by the .describe() so many information the min max standard deviation the 25 percentile the 50th percentile the 75 percentile. Then by the help of correlation function I get to know the correlation of each columns with each other. From the heatmap I can visualized to see them clearly that they are positive correlated or the negative correlated the dark side is show the negative correlation among each other the lighter side represent the positive correlation among the each other

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 (Theatrically 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 from 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

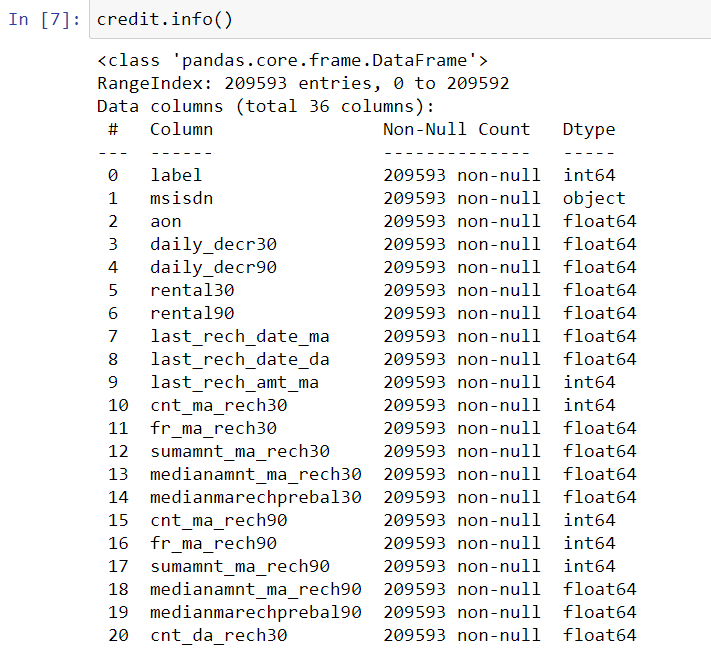
## DATA SOURCES AND THEIR FORMATS

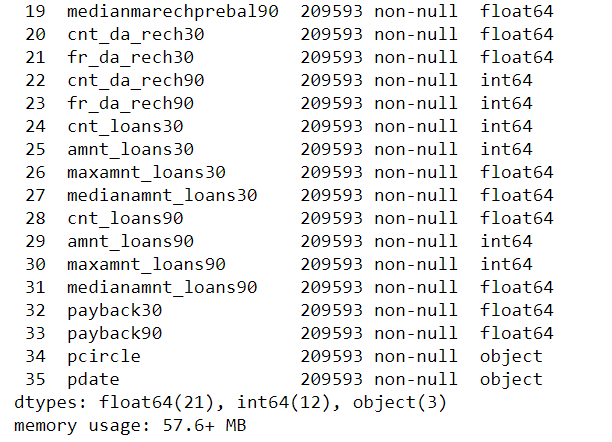
The data which I received from the Flip Robo Technologies was in CSV (Comma Separated Values) format. In the dataset there were 209593 rows and 36 columns.

The data descriptions are as follow:-

|  |  |
| --- | --- |
| ***Variable*** | ***Definition*** |
| 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 Indonasian Rupiah) |
| medianamnt\_ma\_rech90 | Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah) |
| medianmarechprebal90 | Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah) |
| 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 |

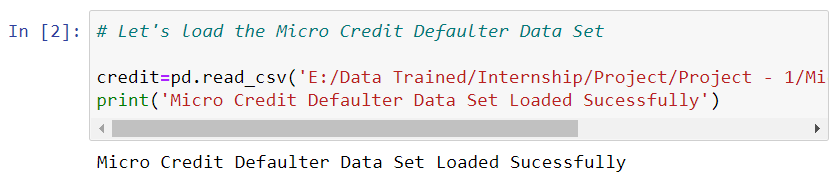
Dataset Datatypes are as follows:





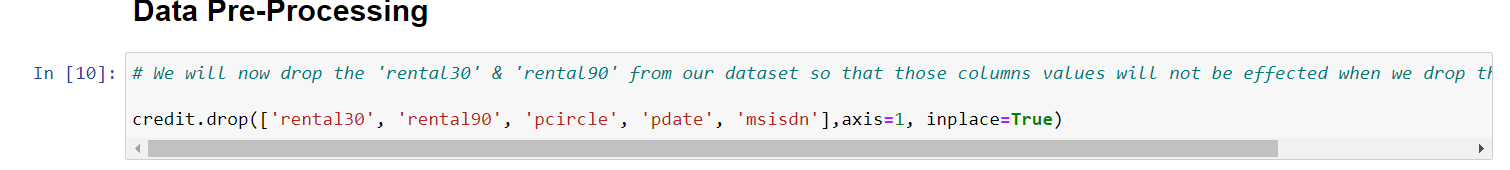
## DATA PREPROCESSING DONE

After loading all the required libraries we loaded the data into our jupyter notebook.



Feature Engineering has been used for cleaning of the data. Some unused columns have been deleted and even some columns have been bifurcated which was used in the prediction. As for example, date column has been bifurcated in days & month to complete the process. Using column aon (age of cellular network), some values have been deleted as outliers; using median to fill the minority data for improvising the outliers.

From an initial glance, it is understood that the msisdn(the phone numbers) are unique values and does not add value to our analysis. Hence the column can be dropped. Also the ‘pcircle’ which is the telecom circle, which has only one unique value across the dataset, ‘UPW’ can also be dropped.

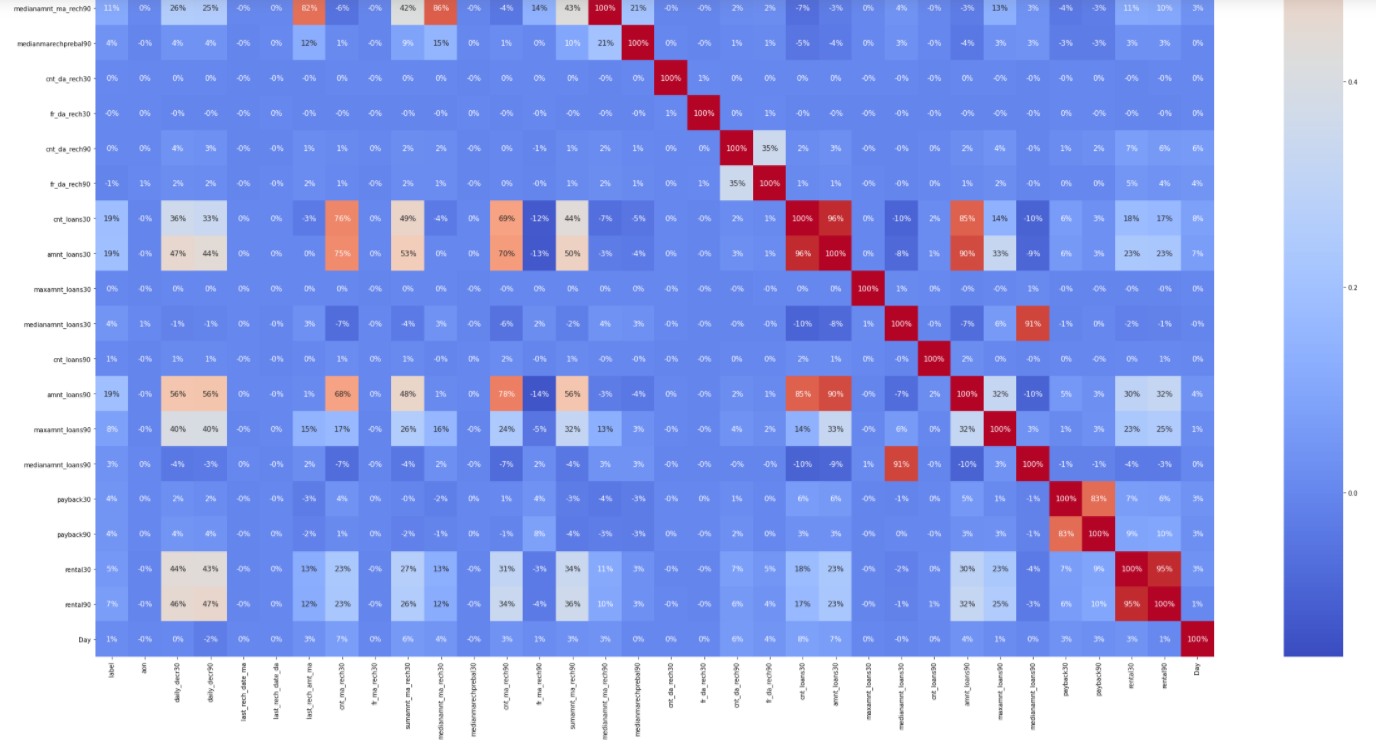


Even, using column label, we have deleted majority of rows and did the featured engineering for minority dataset.

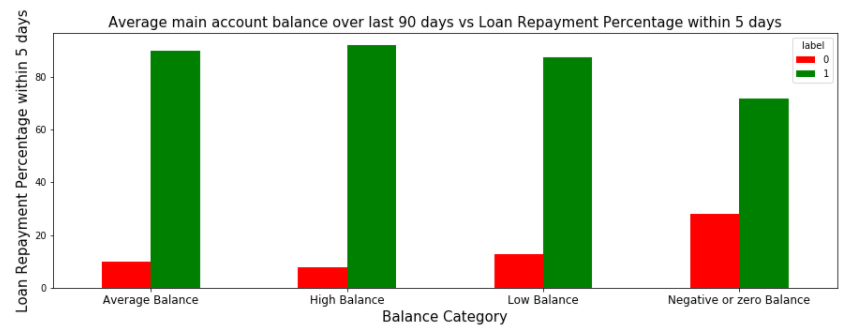
After data cleaning, visualisation is done where population of label has been checked and we found that the defaulters are less than non- defaulters. Further, we have checked the number of defaulters & non- defaulters monthly.

Correlations of columns have been checked through the heatmap plotting and found many correlated columns and deleted the columns.



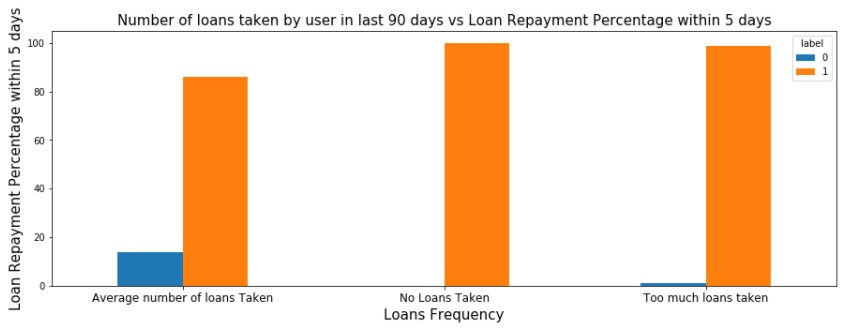


## DATA INPUTS- LOGIC- OUTPUT RELATIONSHIPS



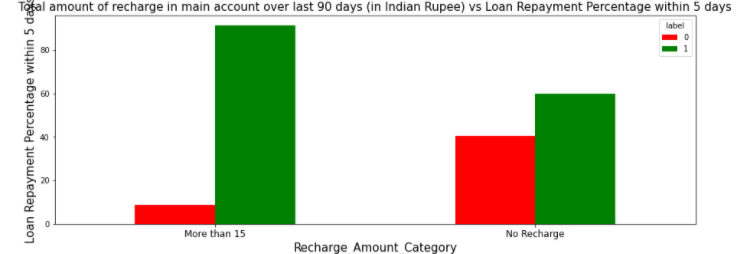
From the above Graph and the crosstab table it is clear that:

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



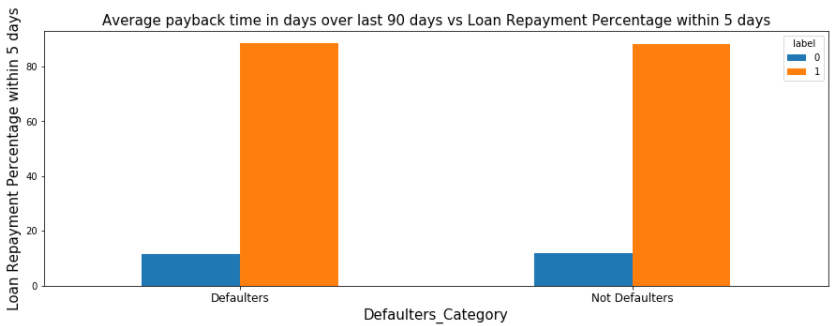
From the above graph it is clear that:

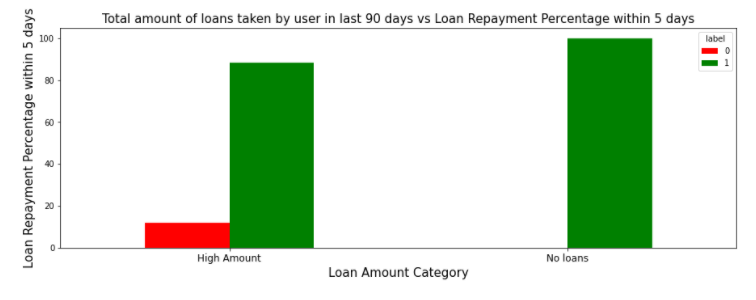
1) Users who take more number of loans are non-defaulters (i.e. 98% of the category) as they repays the loan within the given time i.e. 5 days.  
2) 14% of the Users are are among the average number of loan taken category are defaulters.



From the above graph it is clear that:

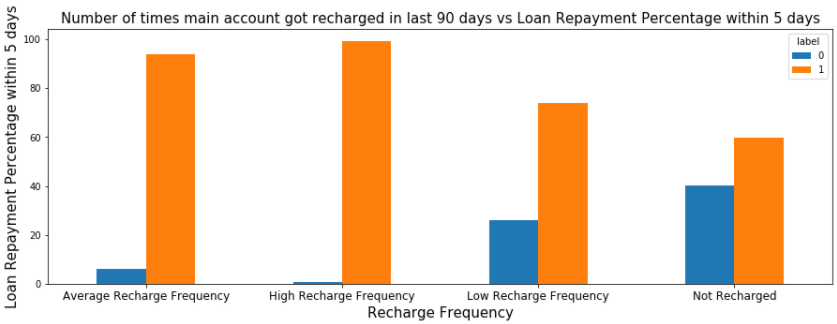
1) 40 % of the Users who do not even recharged in the 90 days are defaulters only.  
2) Users who do very high amount of recharge always pays their loans on time. i.e 98% of them are non-defaulters.





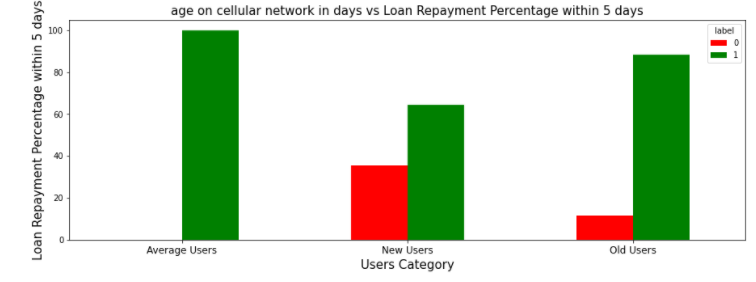
From the above graph it is clear that:

1) Users who did not take any loans are non-defaulters.  
2) Most of the Users (i.e. 98%) who take large amount of loans comes under non defaulter category.



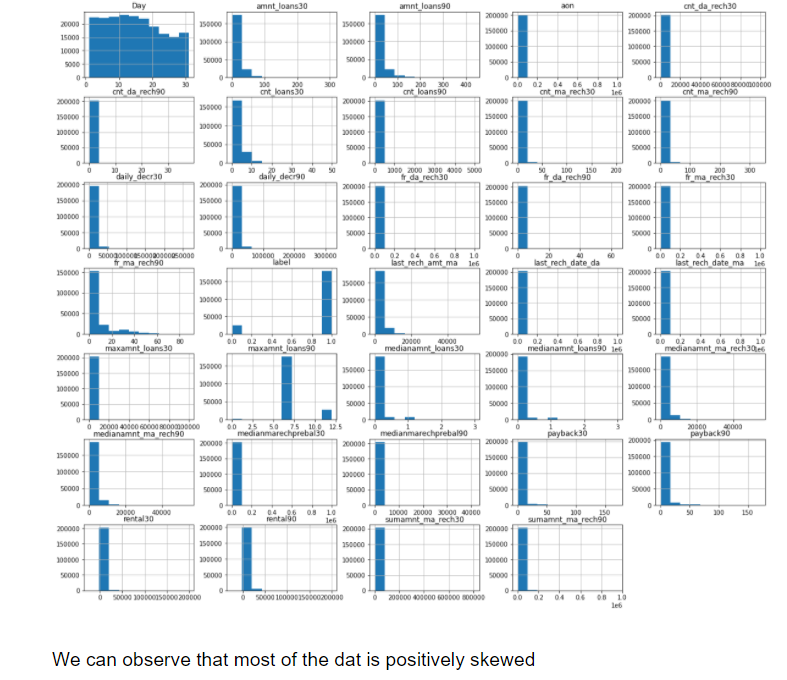
From the above graph it is clear that:

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 96% of the total category, which is a good news for the company.



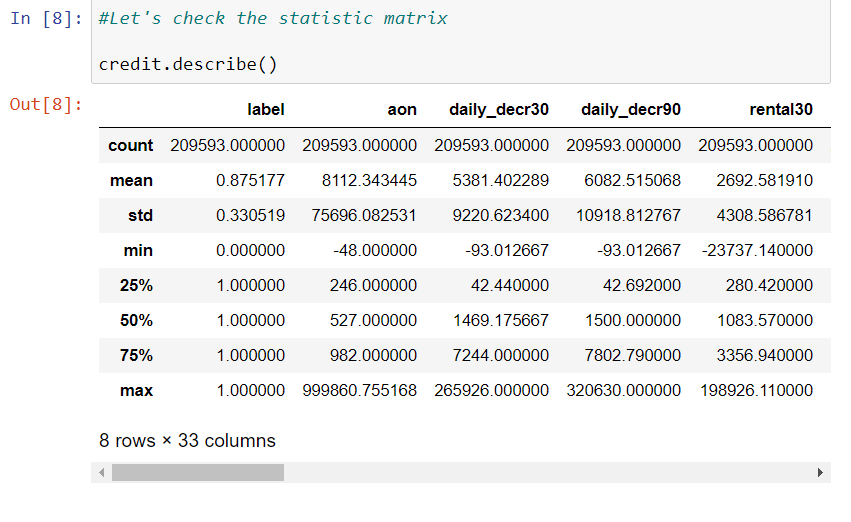
From the above graph it is clear that:

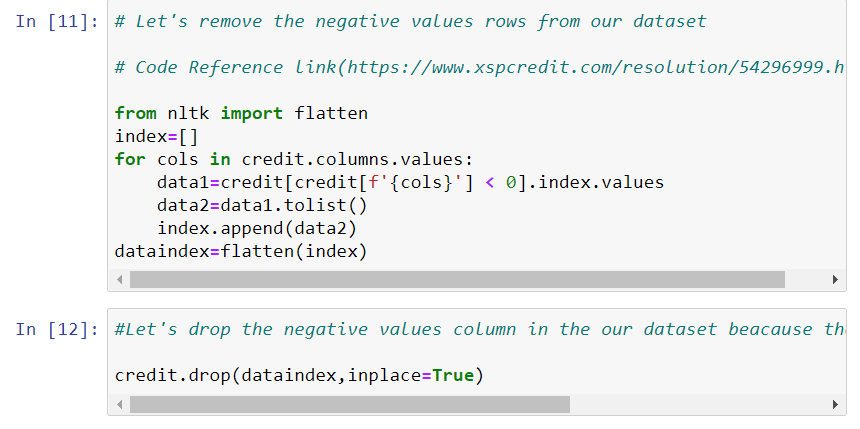
1) 32% of the users who are defaulters are the new users.  
2) Old Users are trusted and they are mostly non defaulters.



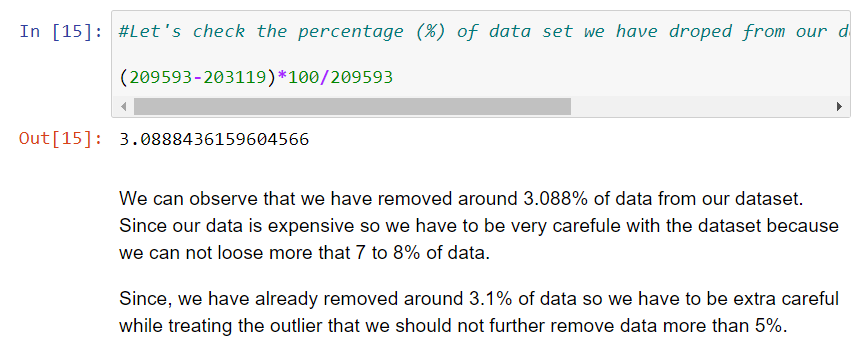
## STATE THE SET OF ASSUMPTIONS (IF ANY) RELATED TO THE PROBLEM UNDER CONSIDERATION

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.





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.

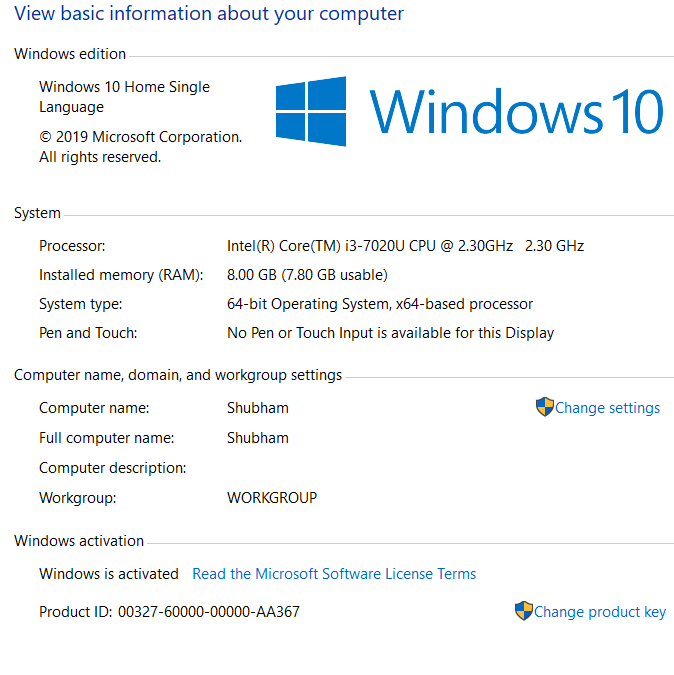


All the categories that is being made to make the visualizations easy are solemnly based on the Description i.e statistical summary of the data plotted above for instance low comes under (0-25%), average comes under (25-75%) and high comes over 75% of the data values in a given feature. Using MS EXCEL 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.

## HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED

***HARDWARE:***

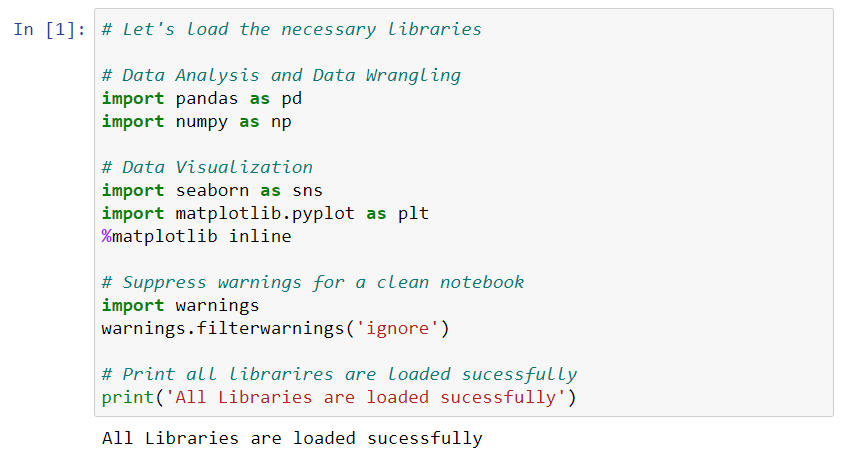


***SOFTWARE:***

Jupyter Notebook (Anaconda 3) – Python 3.7.6

Microsoft Excel 2010

***LIBRARIES:***



***From sklearn.preprocessing import StandardScaler***

As these columns are different in scale, they are standardized to have common scale while building machine learning model. This is useful when you want to compare data that correspond to different units.

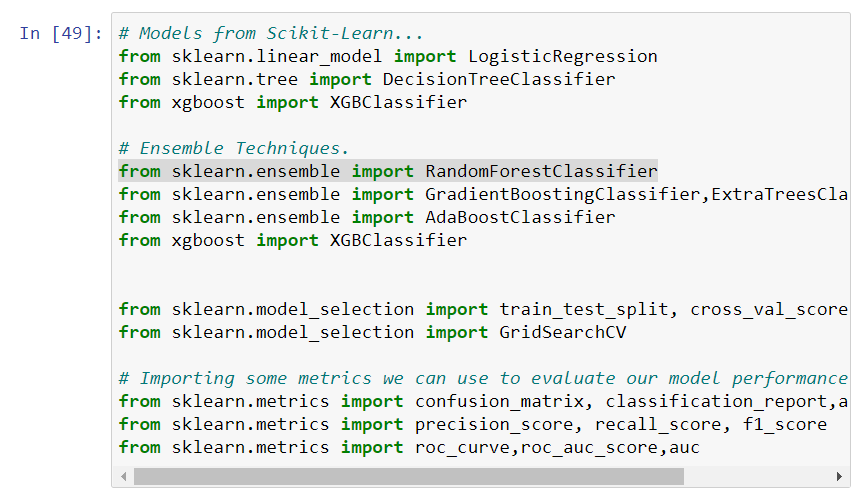
***from sklearn.preprocessing import Label Encoder***

Label Encoder  and One Hot Encoder. These two encoders are parts of the SciKit Learn library in Python, and they are used to convert categorical data, or text data, into numbers, which our predictive models can better understand.

***from sklearn.model\_selection import train\_test\_split,cross\_val\_score***

Train\_test\_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually. By default, Sklearn train\_test\_split will make random partitions for the two subsets.

The algorithm is trained and tested K times, each time a new set is used as testing set while remaining sets are used for training. Finally, the result of the K-Fold Cross-Validation is the average of the results obtained on each set.



***from sklearn.linear\_model import LogisticRegression***

The library sklearn can be used to perform logistic regression in a few lines as shown using the LogisticRegression class. It also supports multiple features. It requires the input values to be in a specific format hence they have been reshaped before training using the fit method.

***from sklearn.tree import DecisionTreeClassifier***

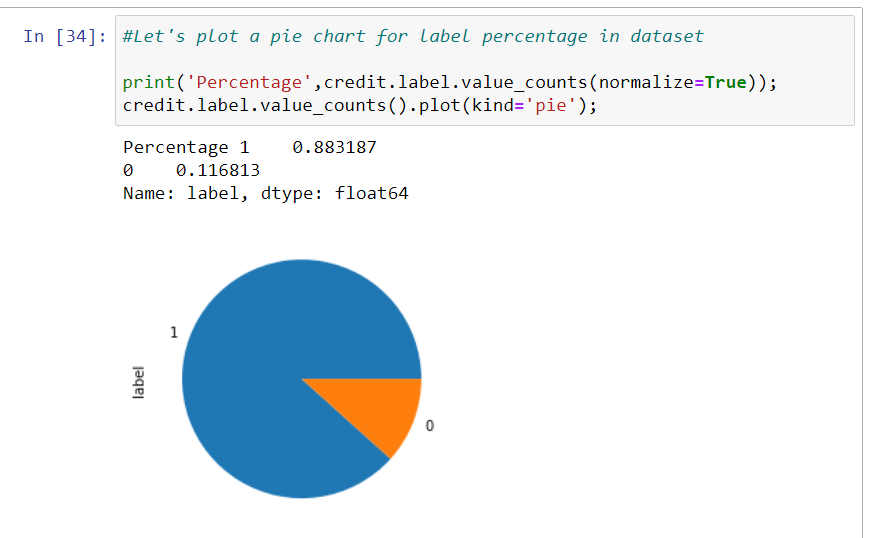
Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy

***from sklearn.ensemble import RandomForestClassifier***

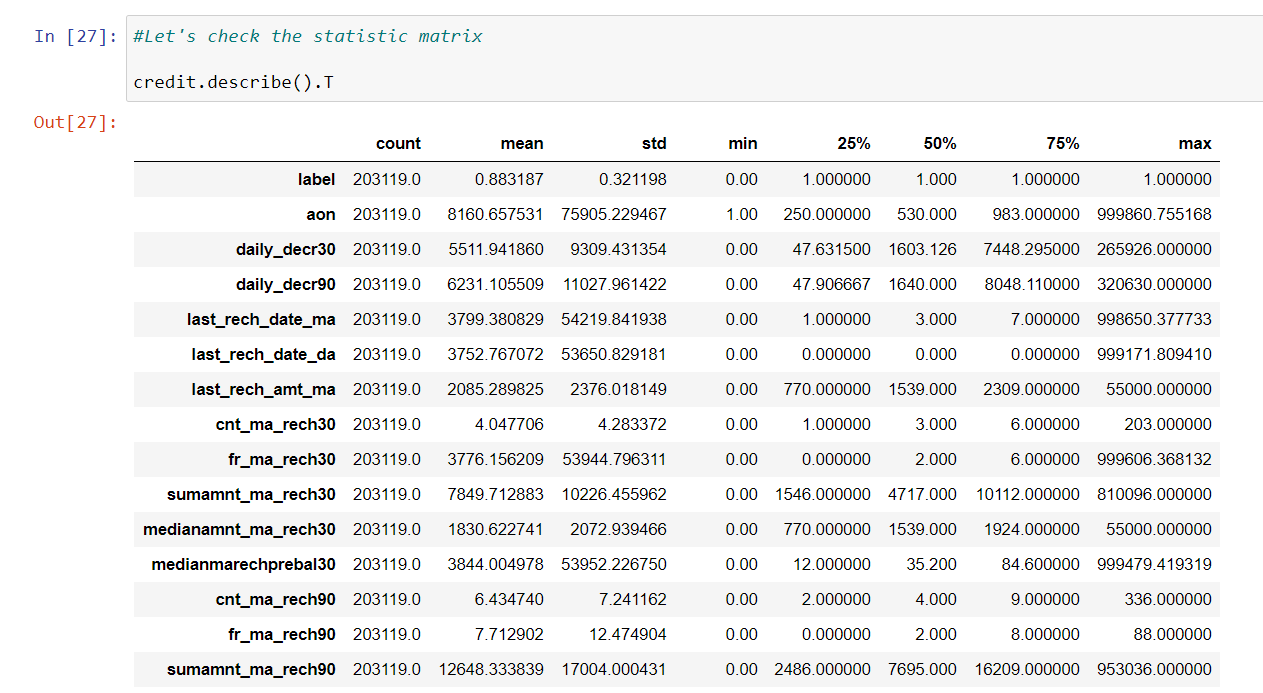
A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

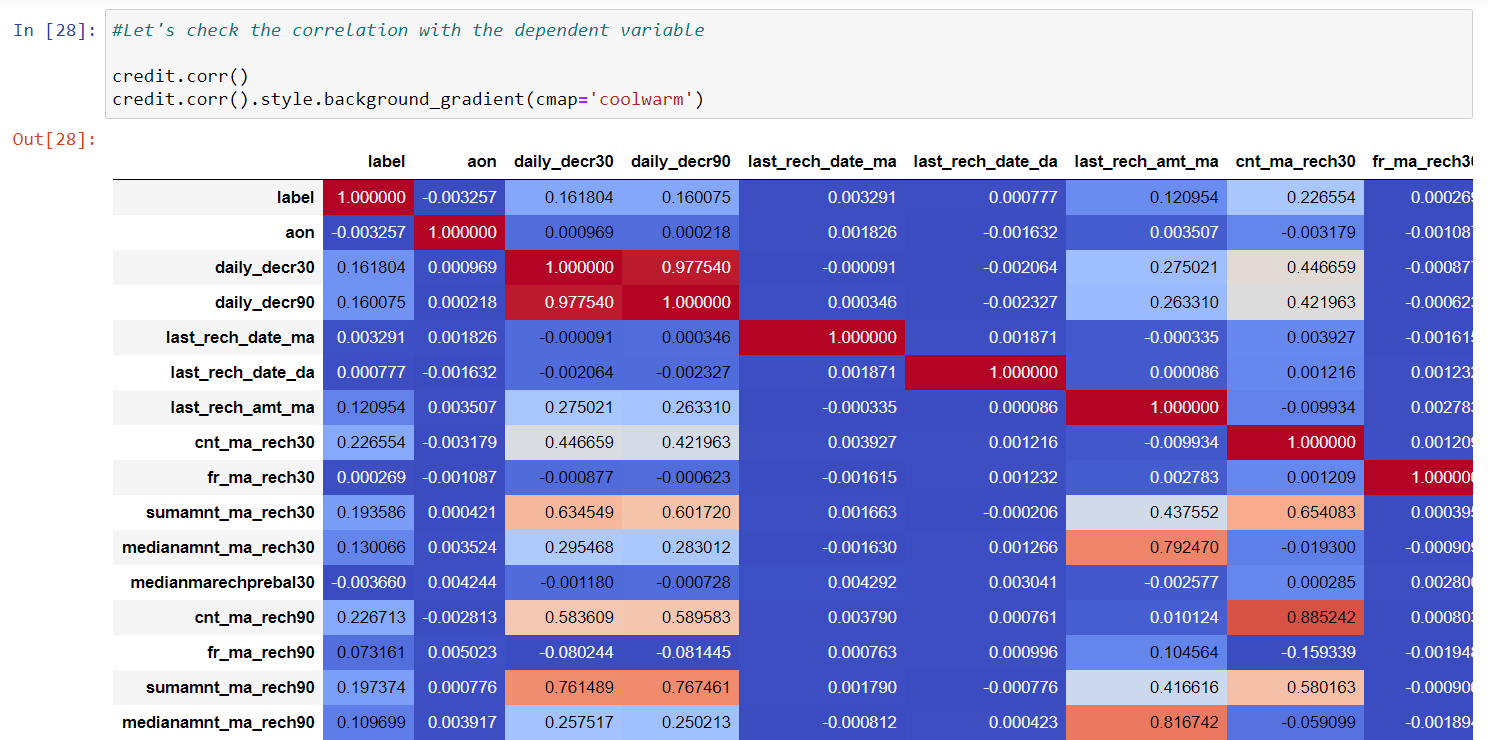
# MODEL/S DEVELOPMENT AND EVALUATION

## IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES (METHODS)



From the above we can observe that the data was highly imbalanced so we have used SMOTETomek to balance the dataset.



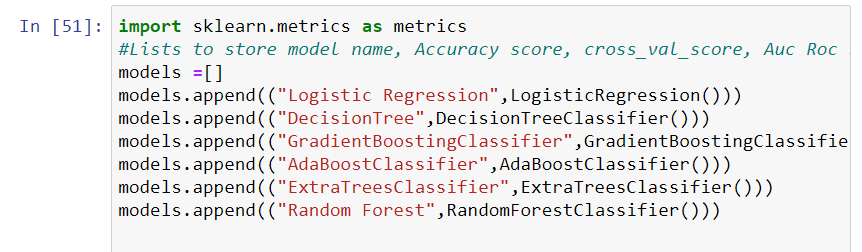


After observing the above output we came to the conclusion that we have multicollinearity which will affect our analysis we have have doped those columns which have high correlation.

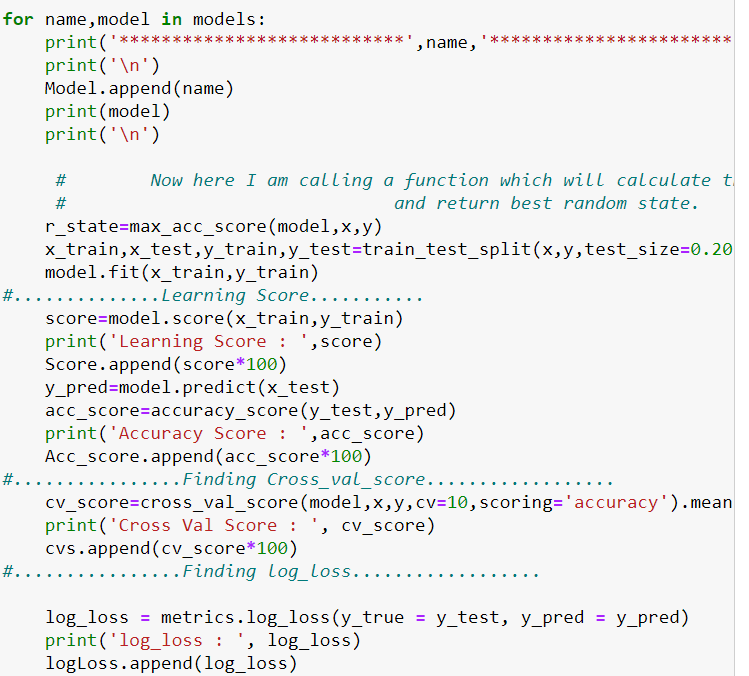
## TESTING OF IDENTIFIED APPROACHES (ALGORITHMS)

We have used the following algorithms

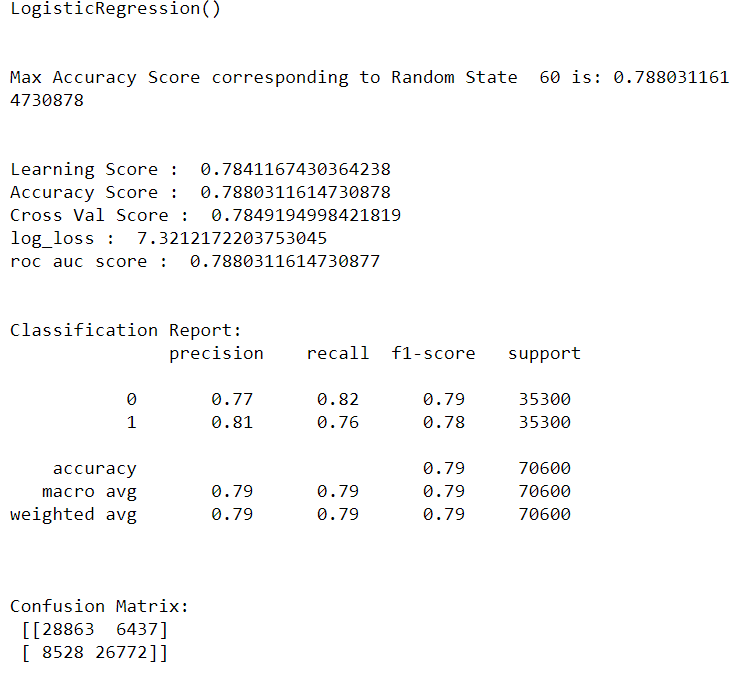
* RandomForestClassifier ()
* DecisionTreeClassifier ()
* GradientBoostingClassifier
* LogisticRegression()
* AdaBoostClassifier()
* ExtraTreesClassifier()

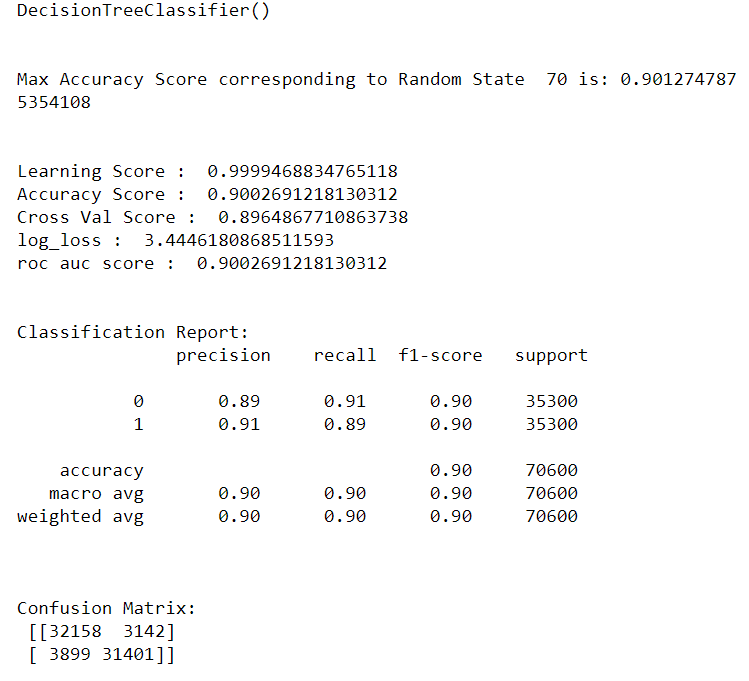


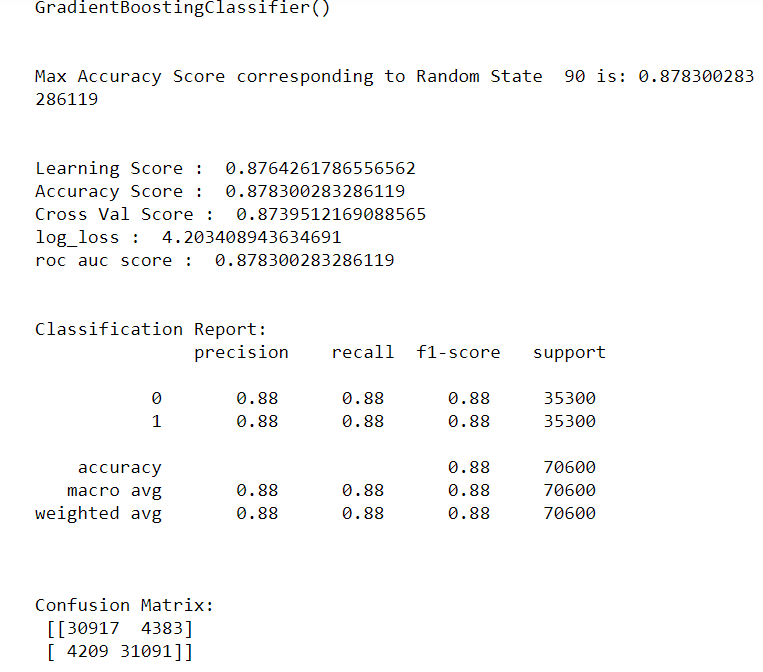
## RUN AND EVALUATE SELECTED MODELS

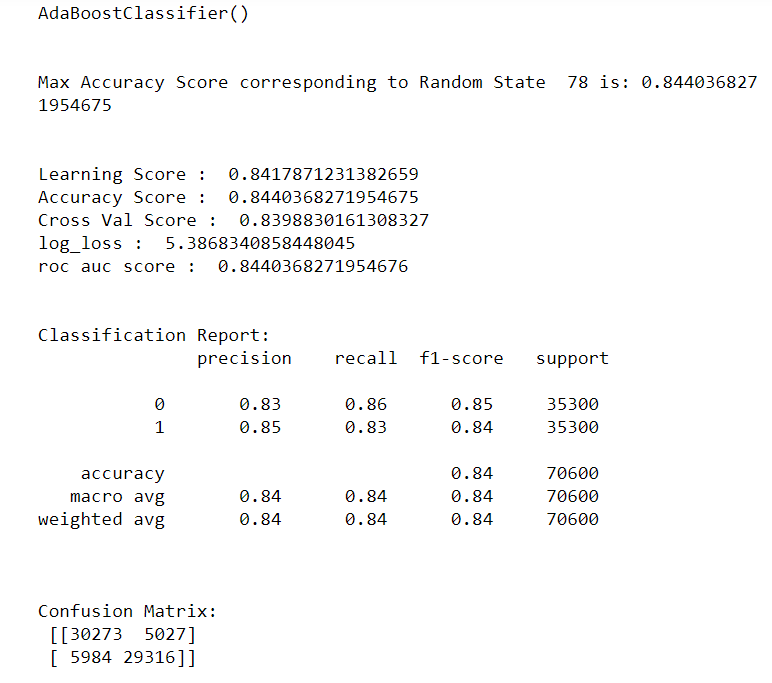


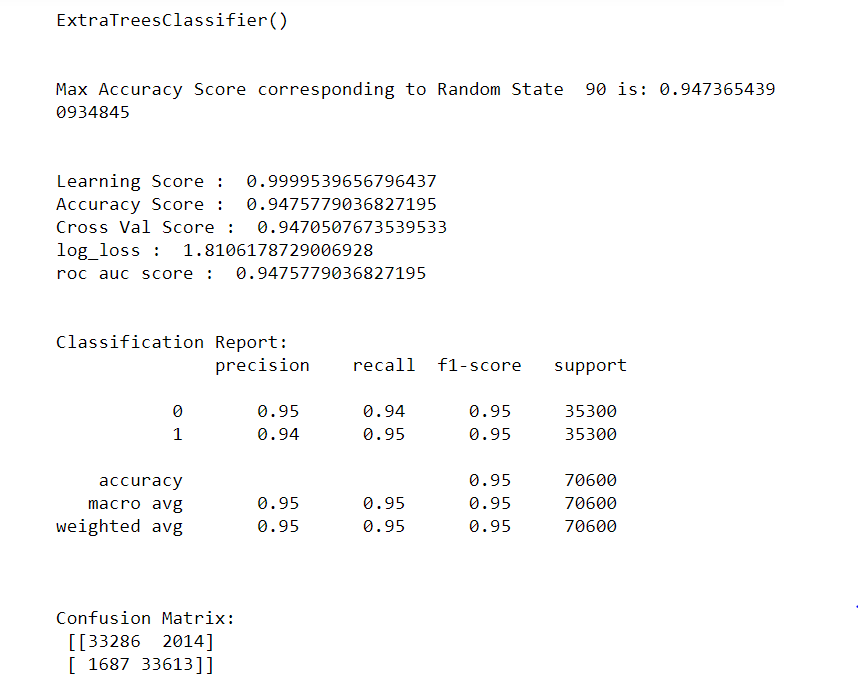


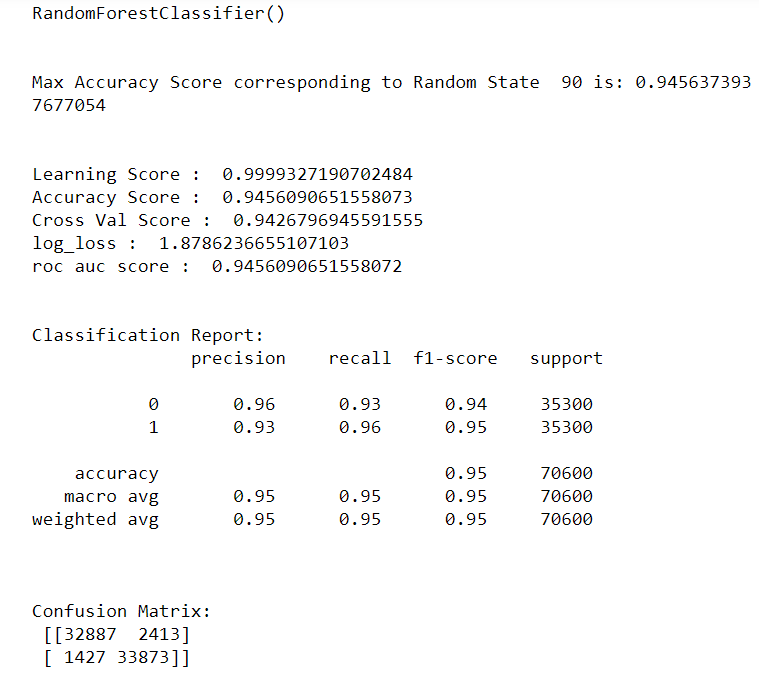


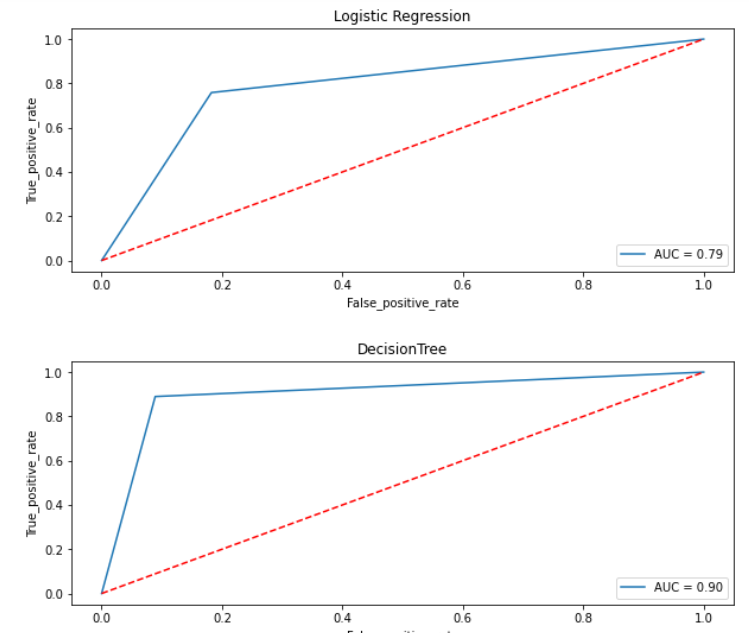


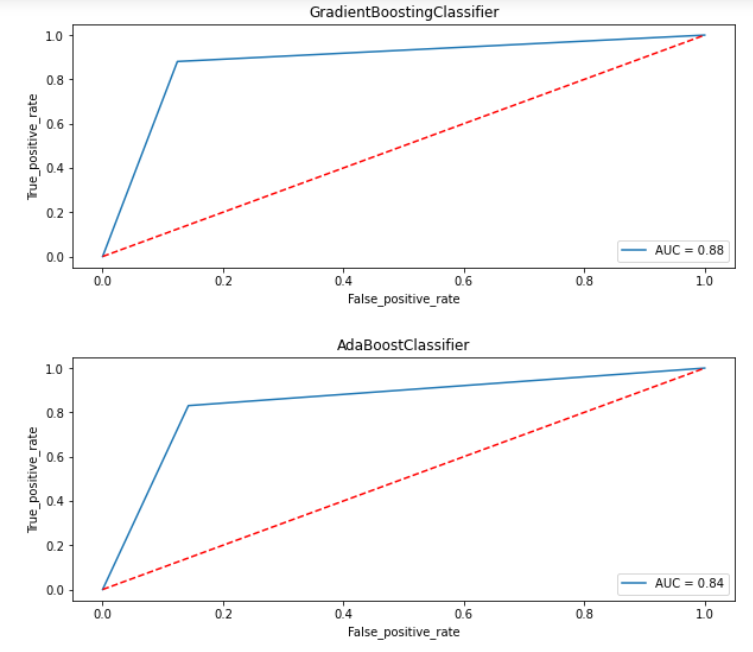


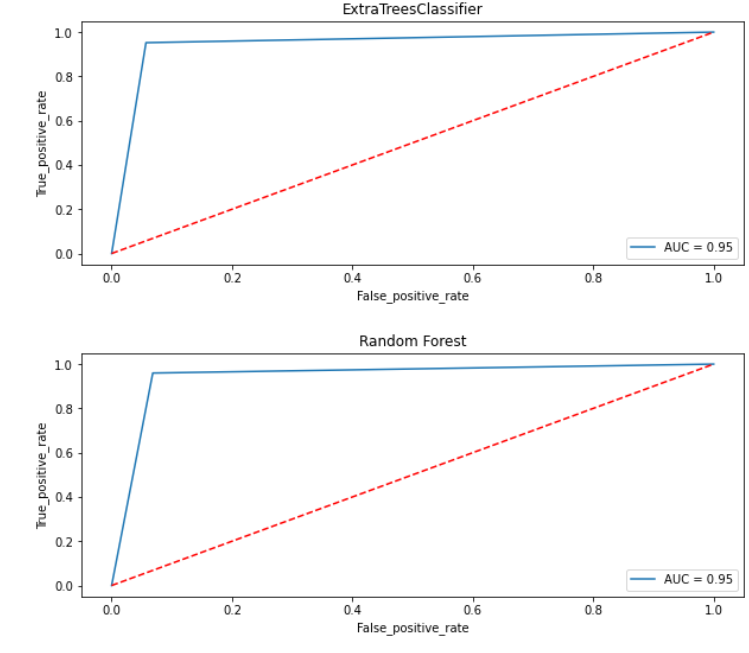


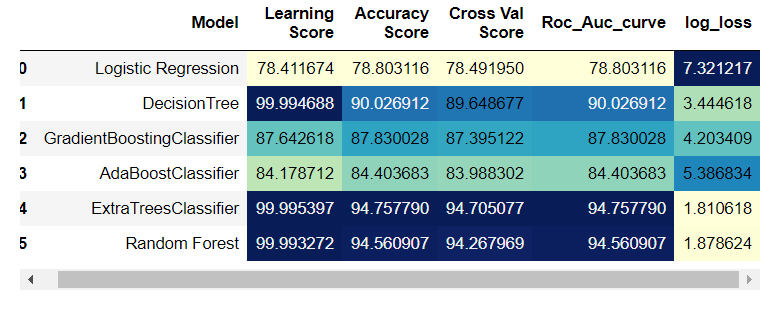


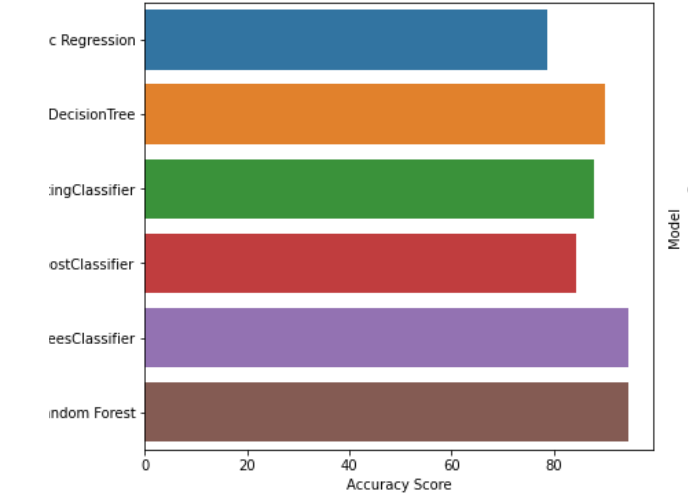


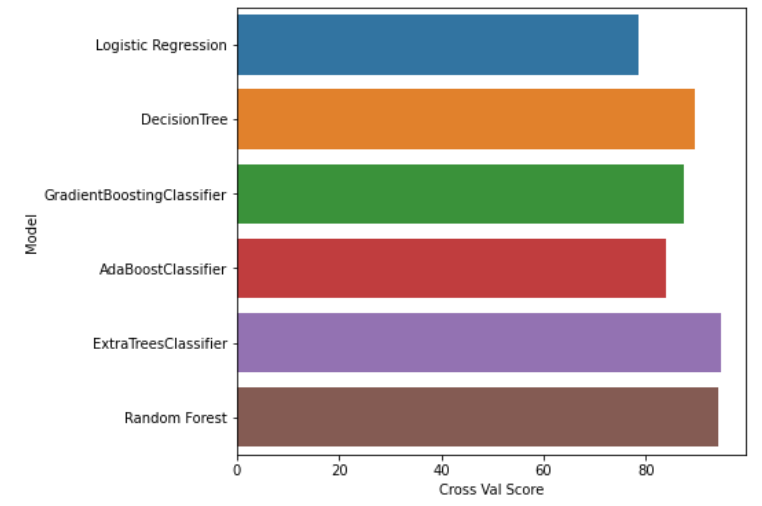


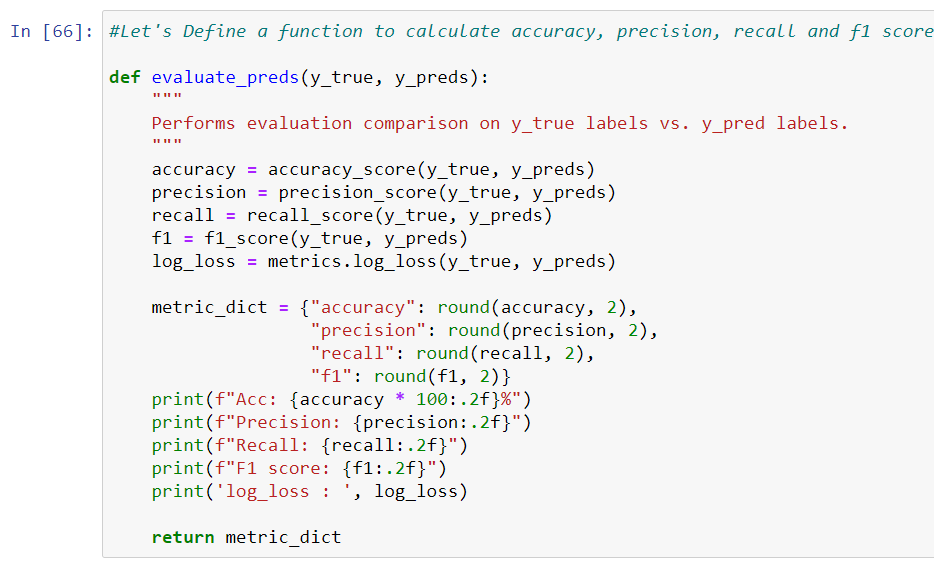


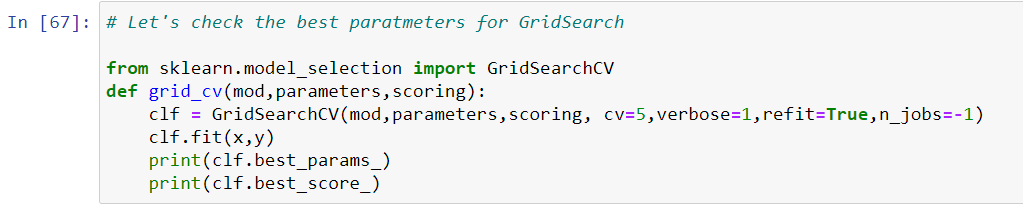


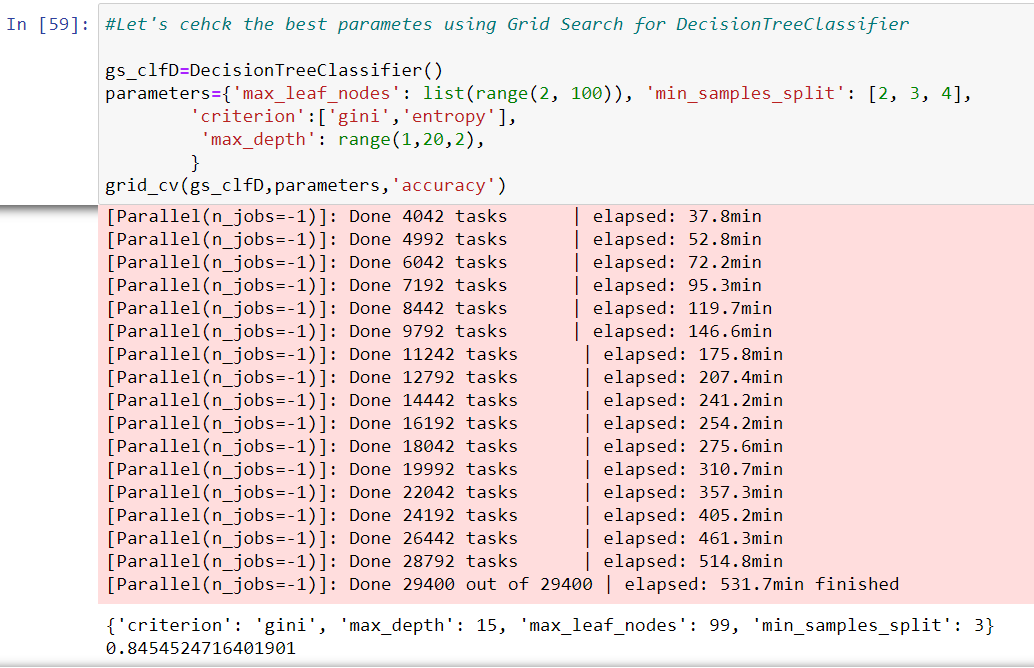


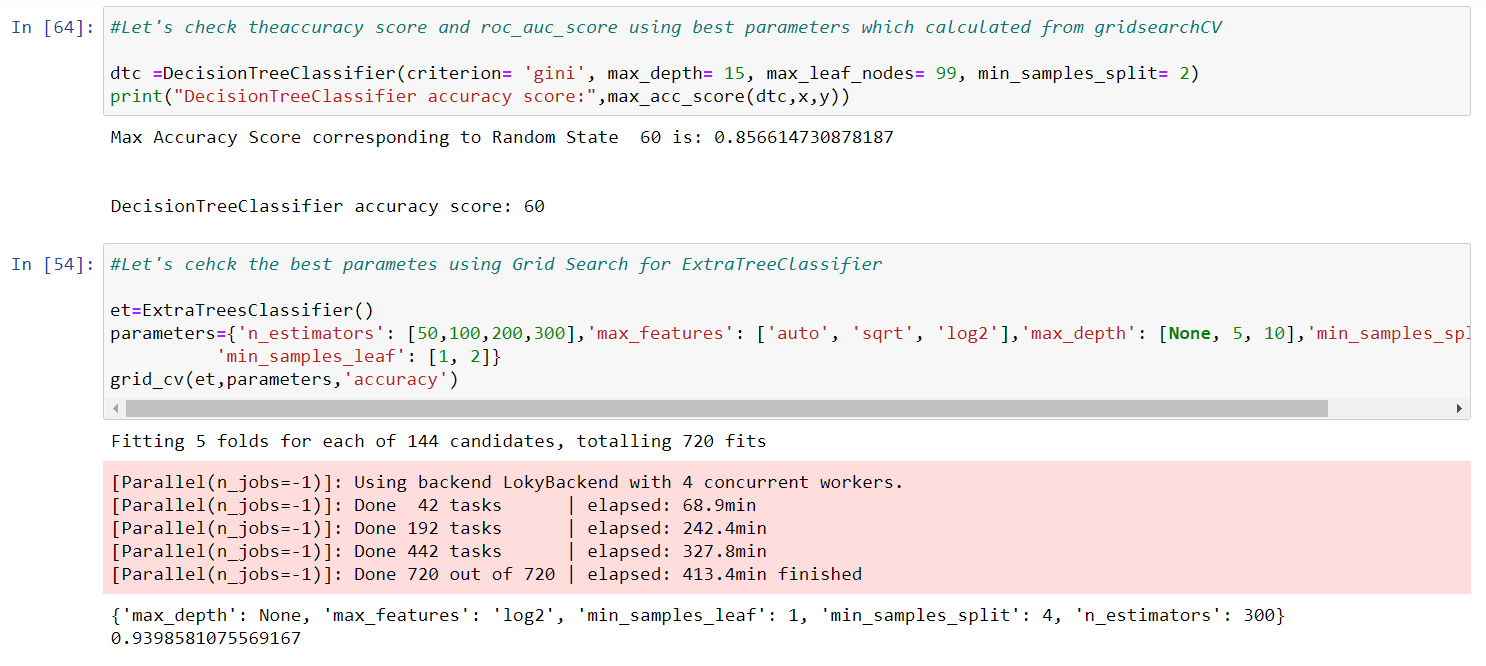




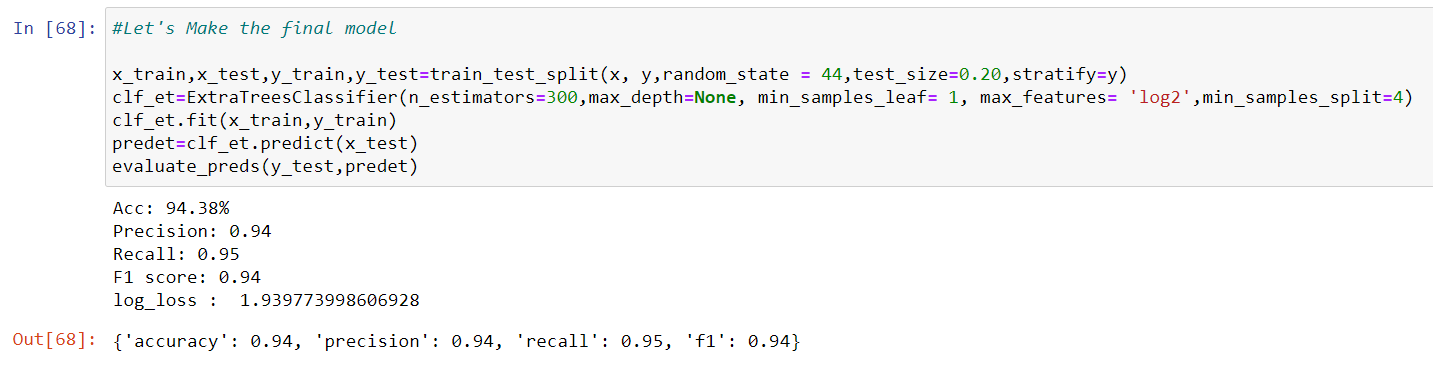


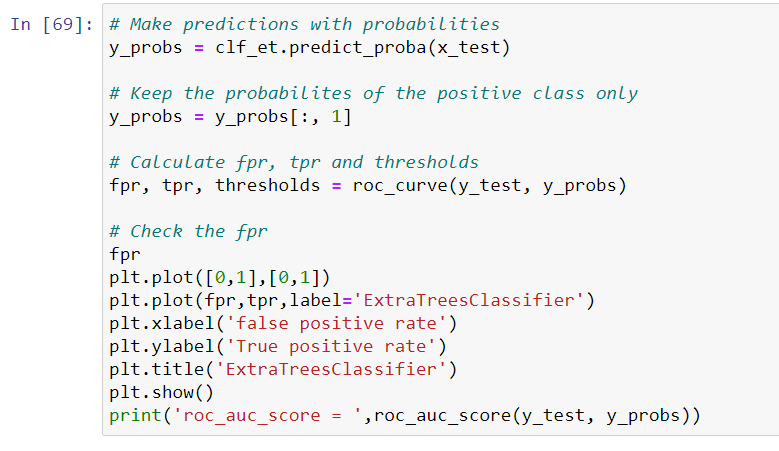




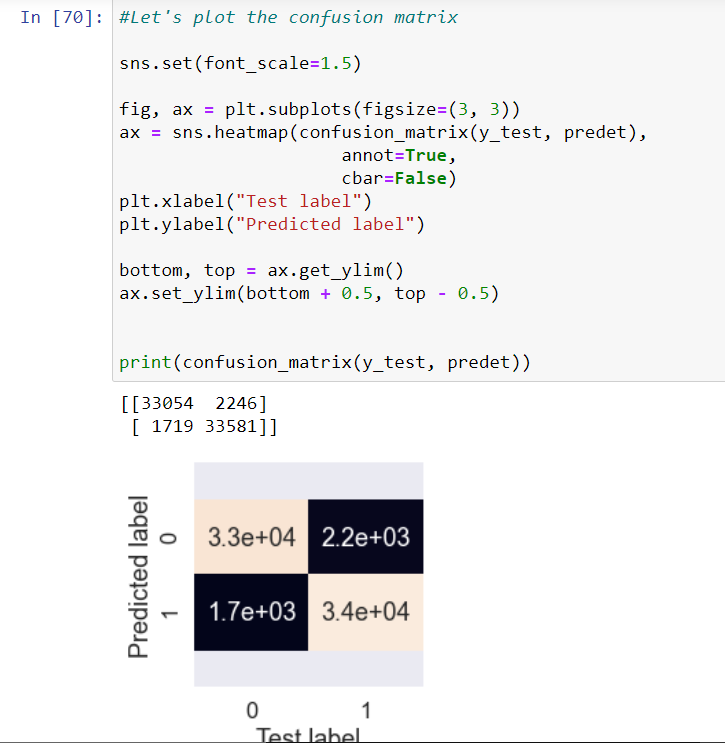


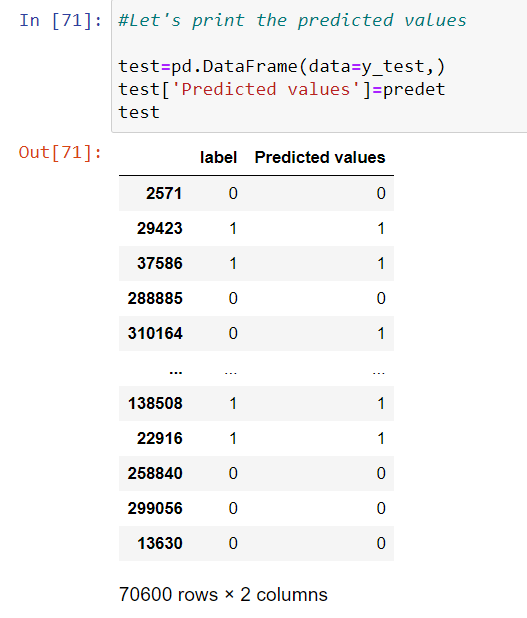


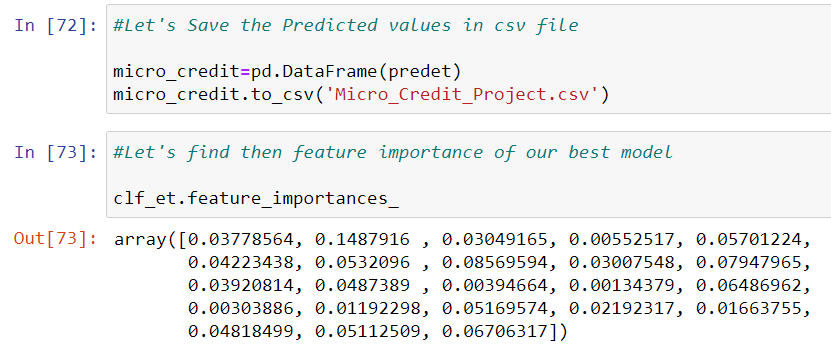


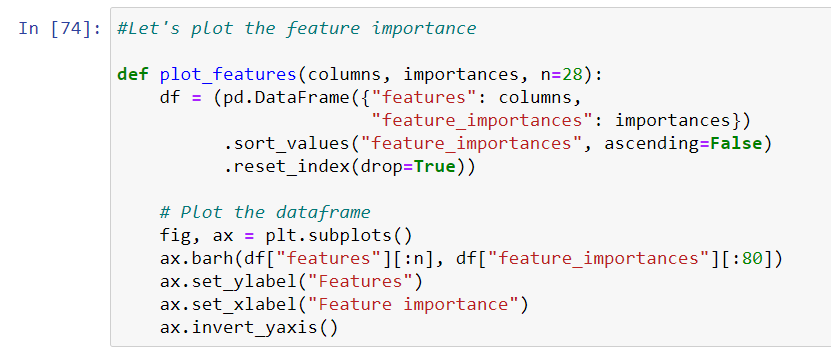


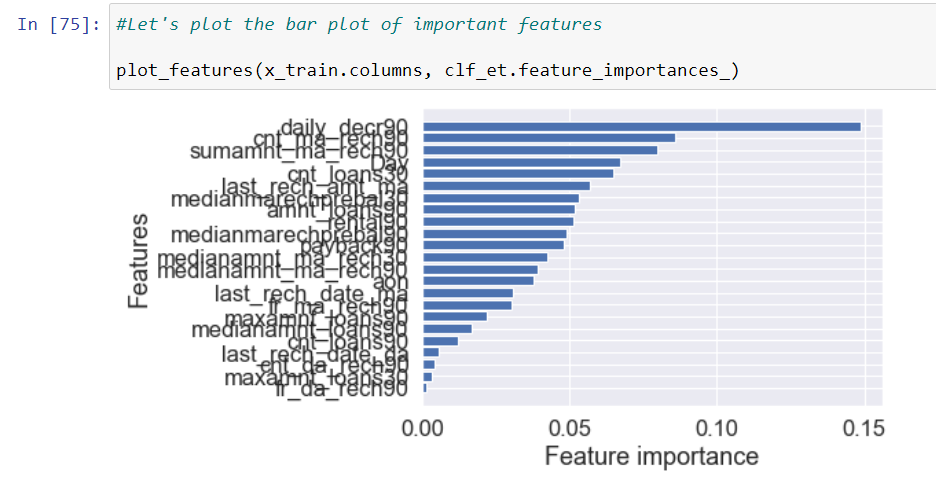


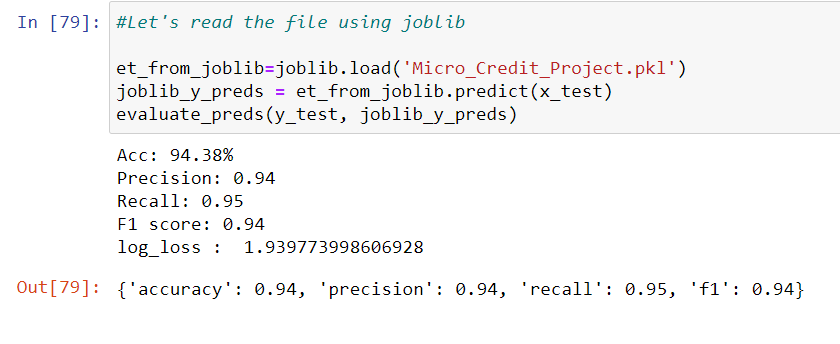
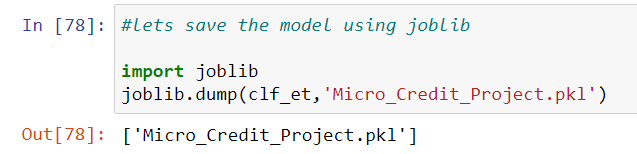












## KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

Precision: can be seen as a measure of quality, higher precision means that an algorithm returns more relevant results than irrelevant ones

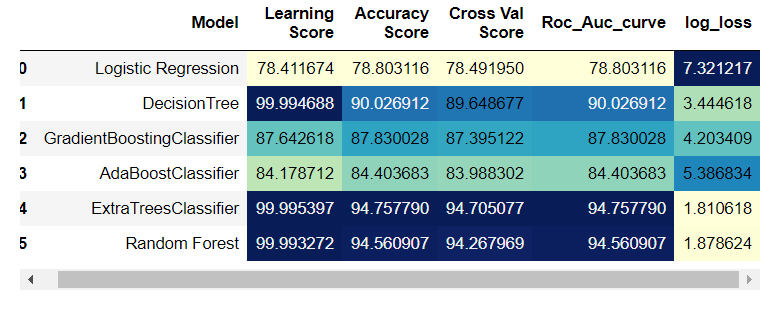
Recall is used as a measure of quantity and high recall means that an algorithm returns most of the relevant results.

Accuracy score is used when the True Positives and True negatives are more important. Accuracy can be used when the class distribution is similar

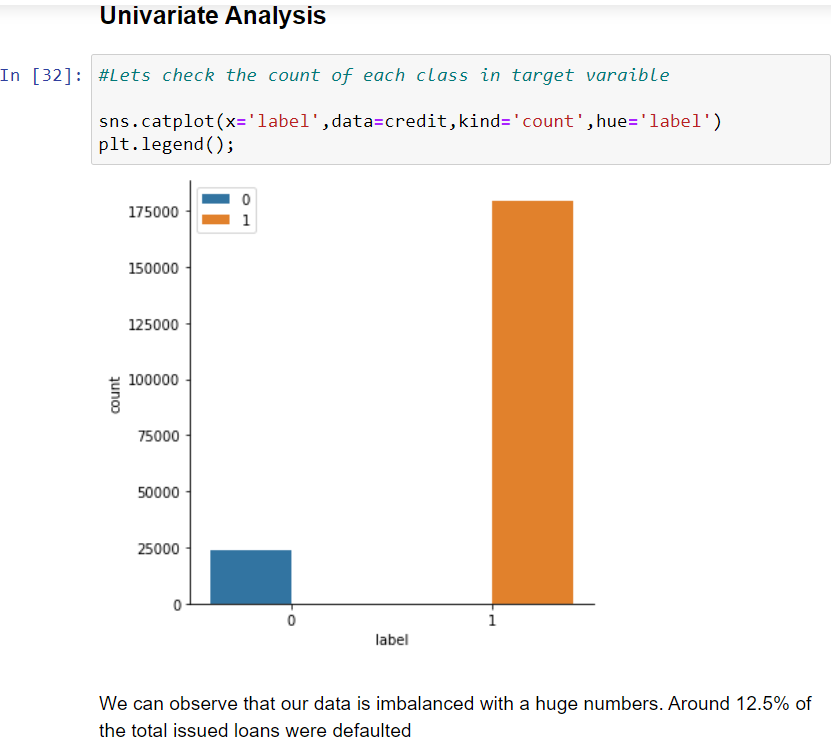
F1-score is used when the False Negatives and False Positives are crucial. While F1-score is a better metric when there are imbalanced classes.

Cross\_val\_score: To run cross-validation on multiple metrics and also to return train scores, fit times and score times. Get predictions from each split of cross-validation for diagnostic purposes. Make a scorer from a performance metric or loss function.

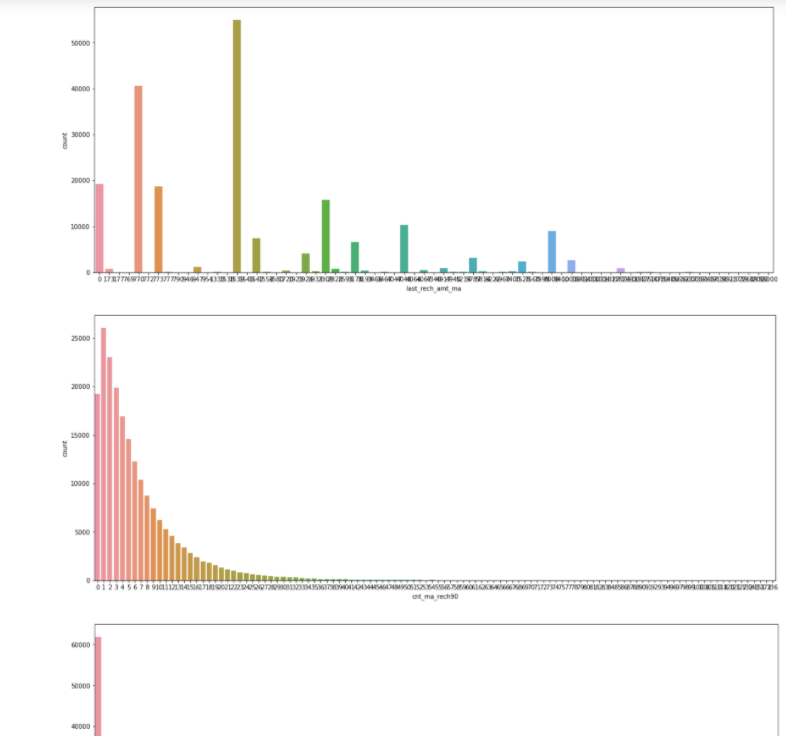
roc \_auc \_score :  ROC curve. It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0

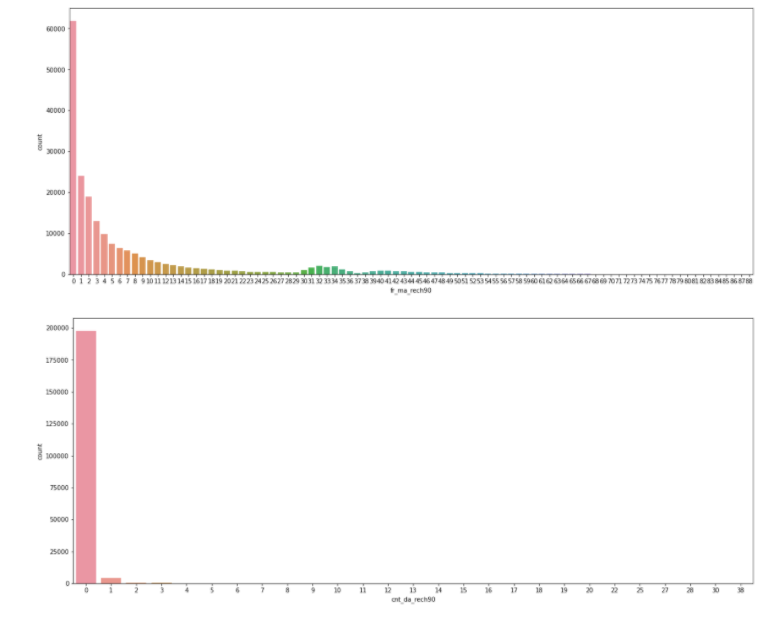


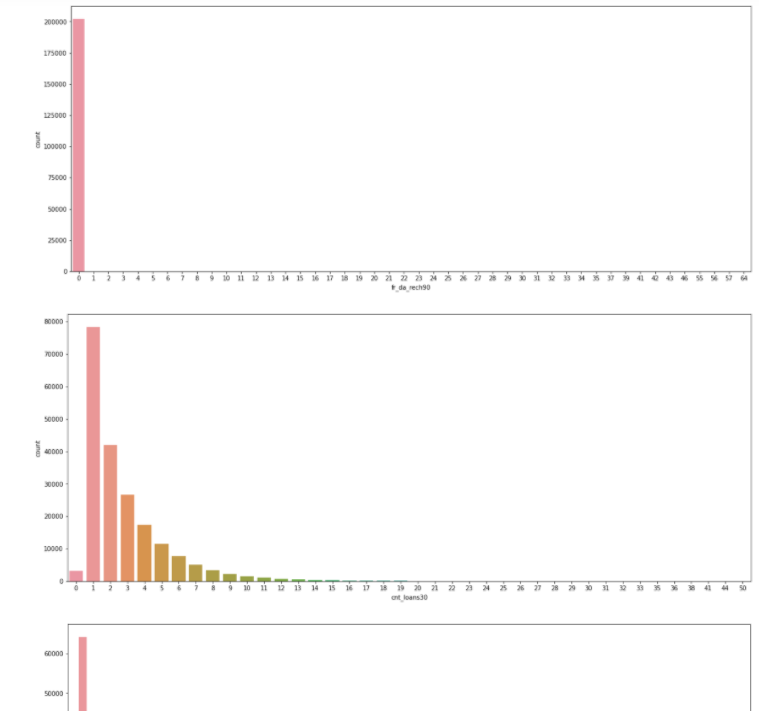
## VISUALIZATIONS

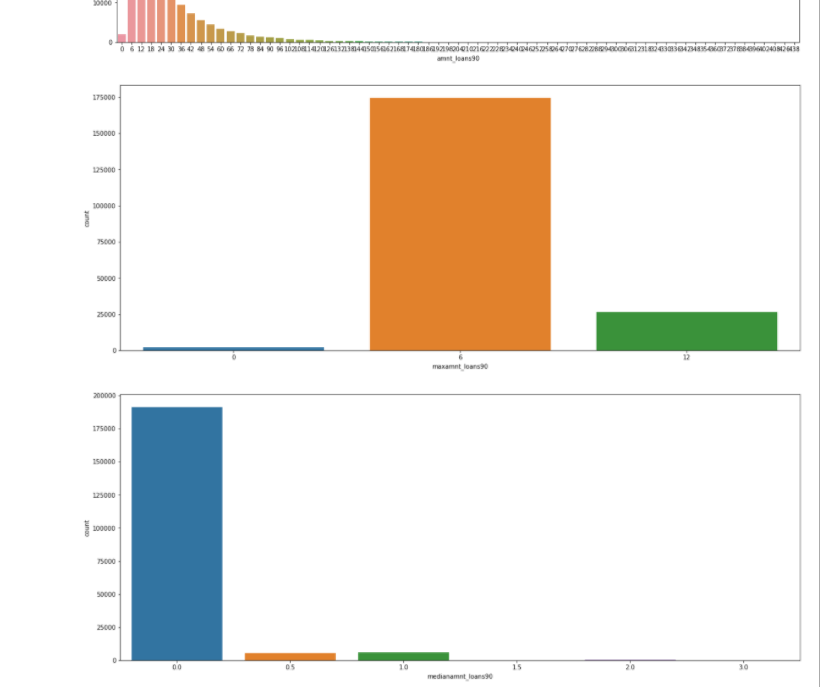


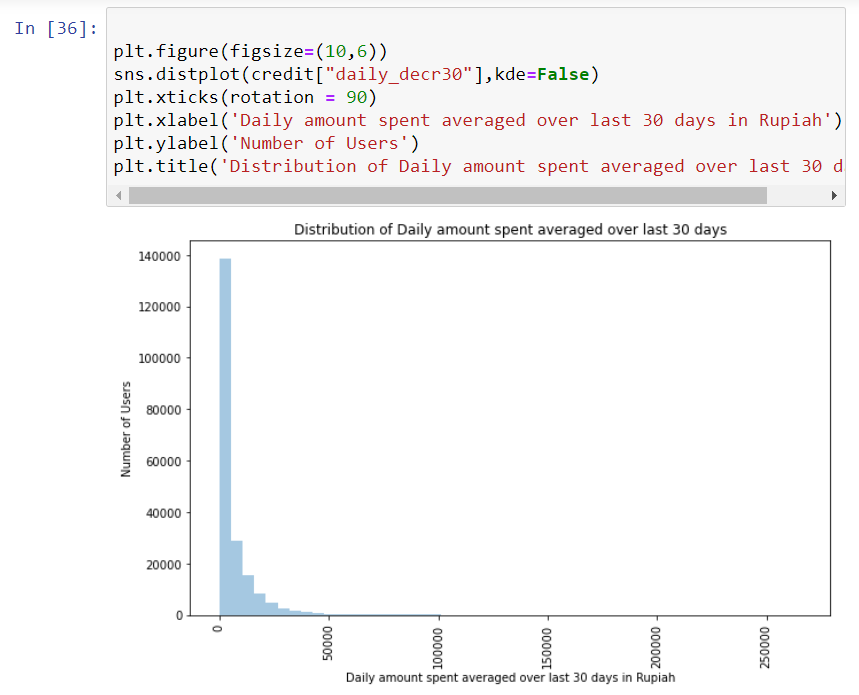


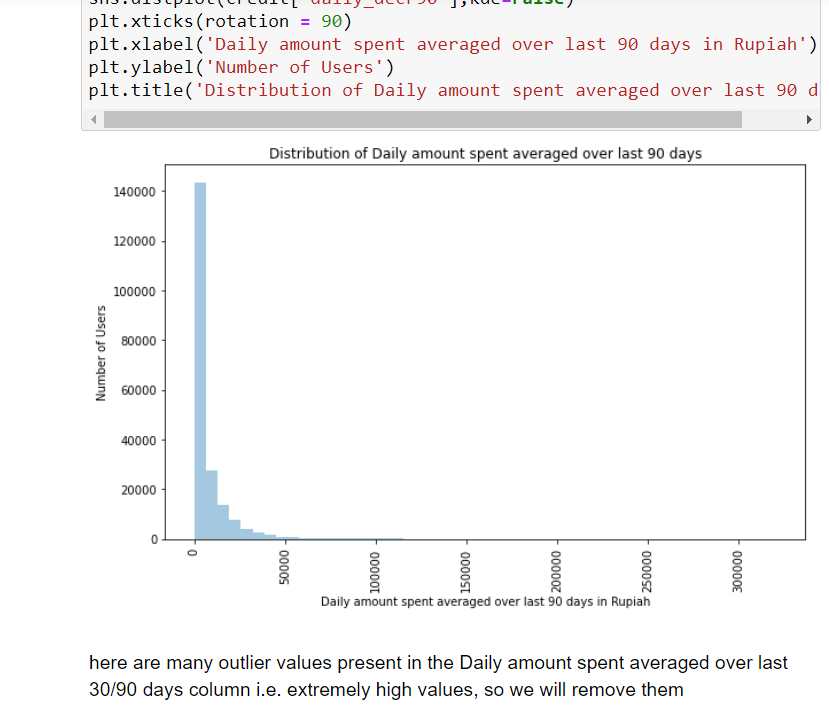


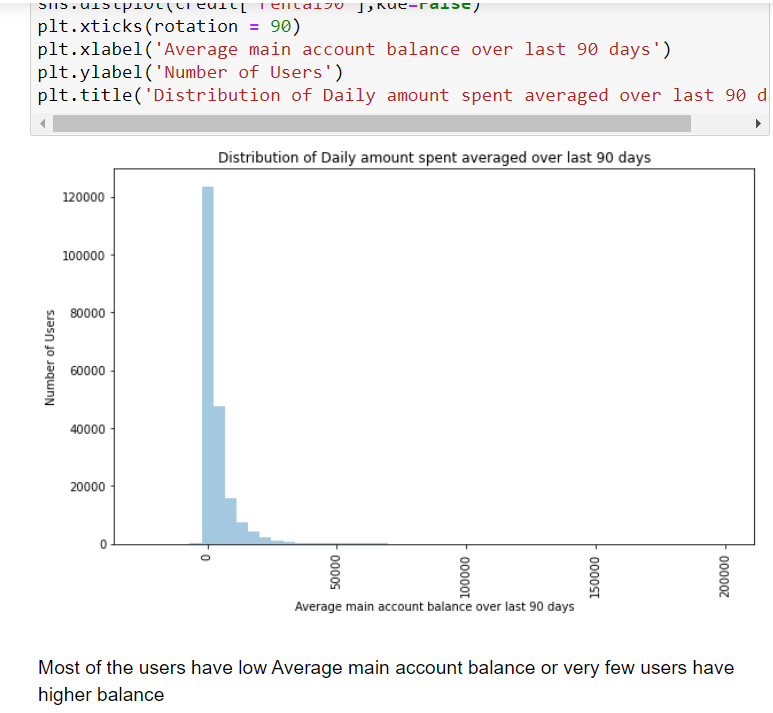


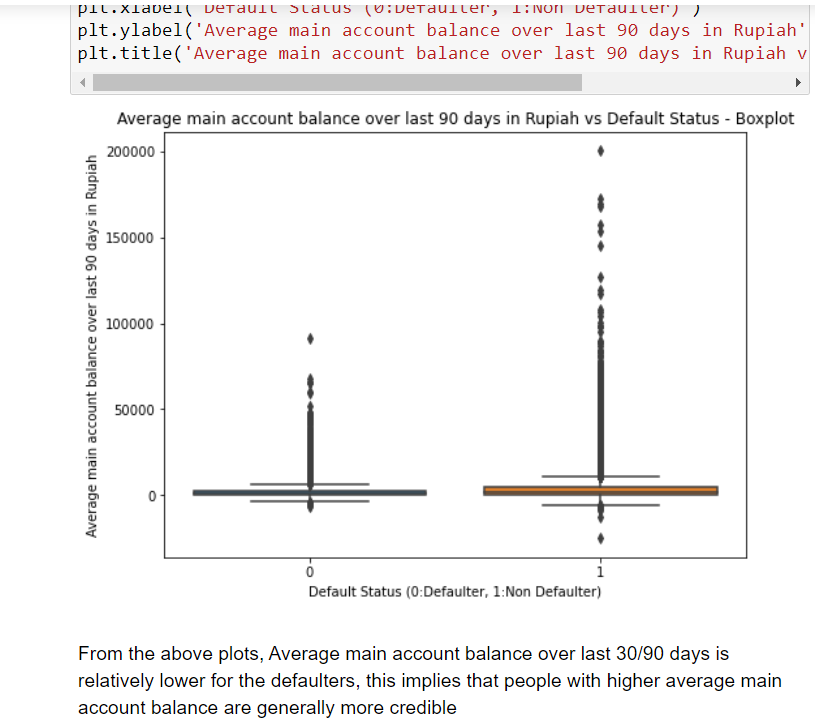


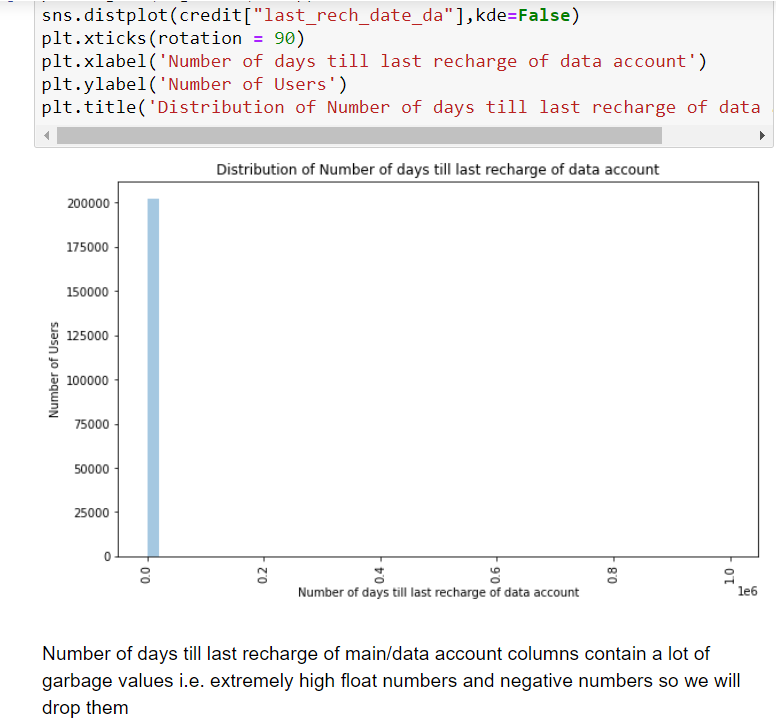


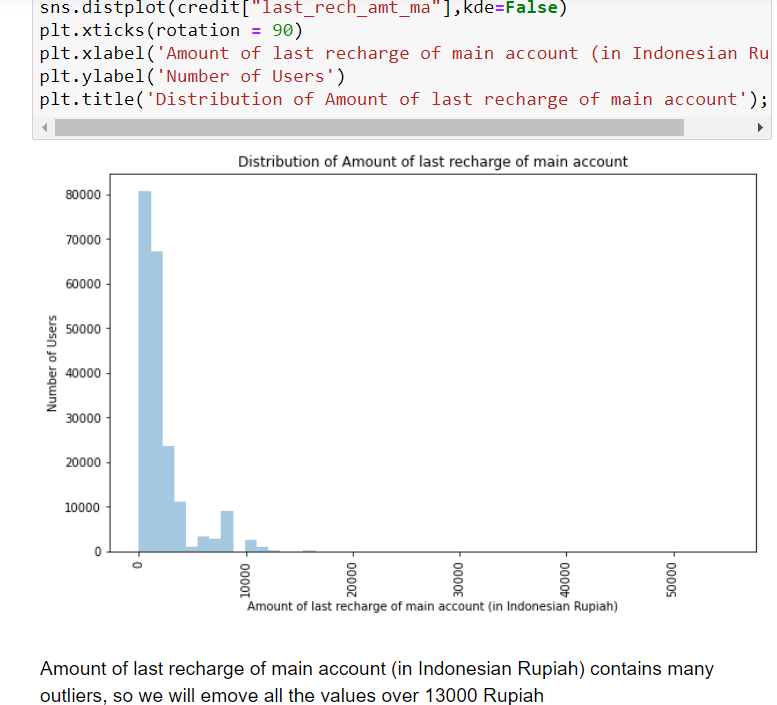




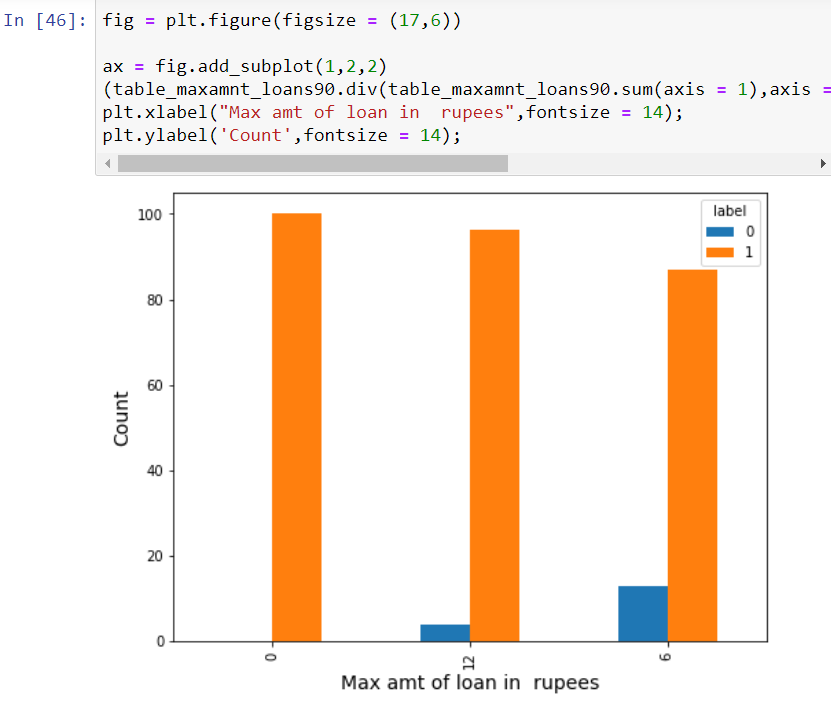


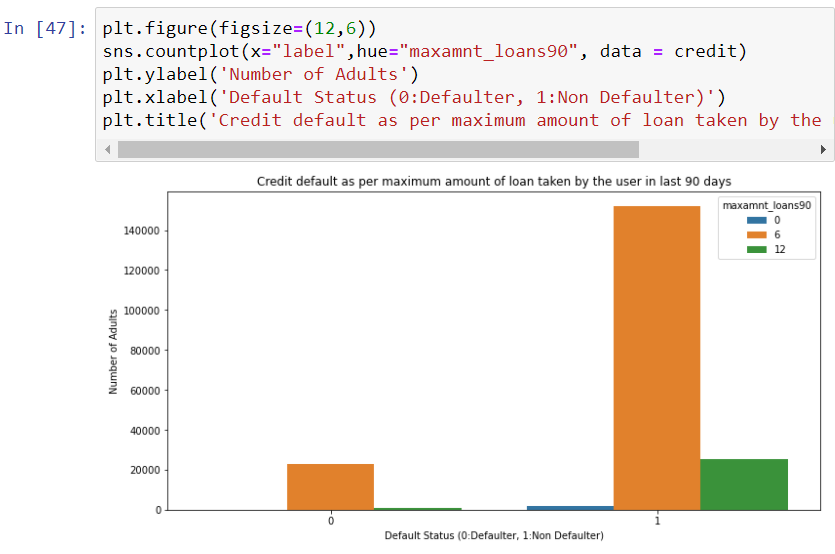


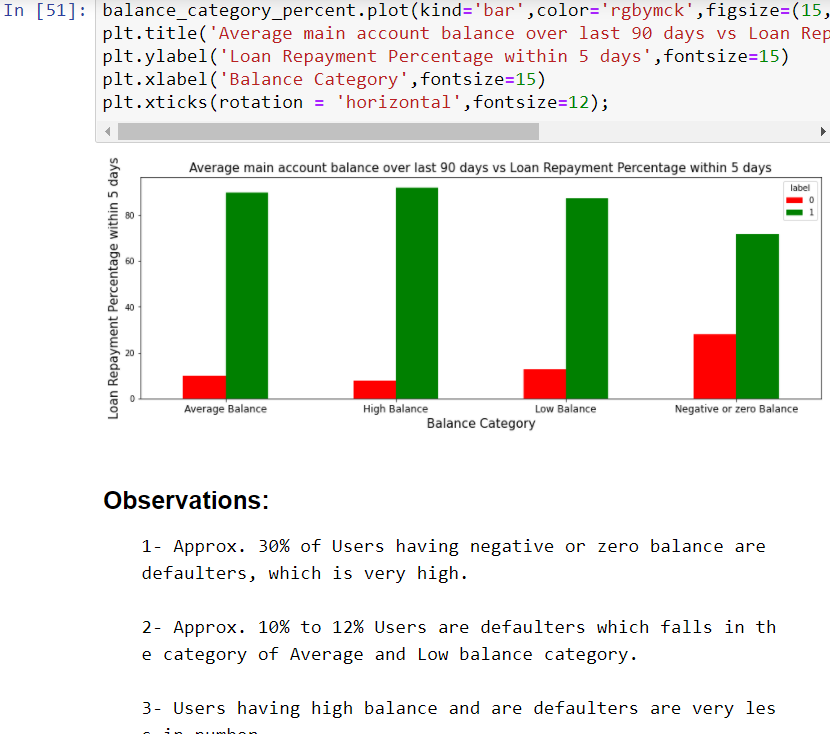


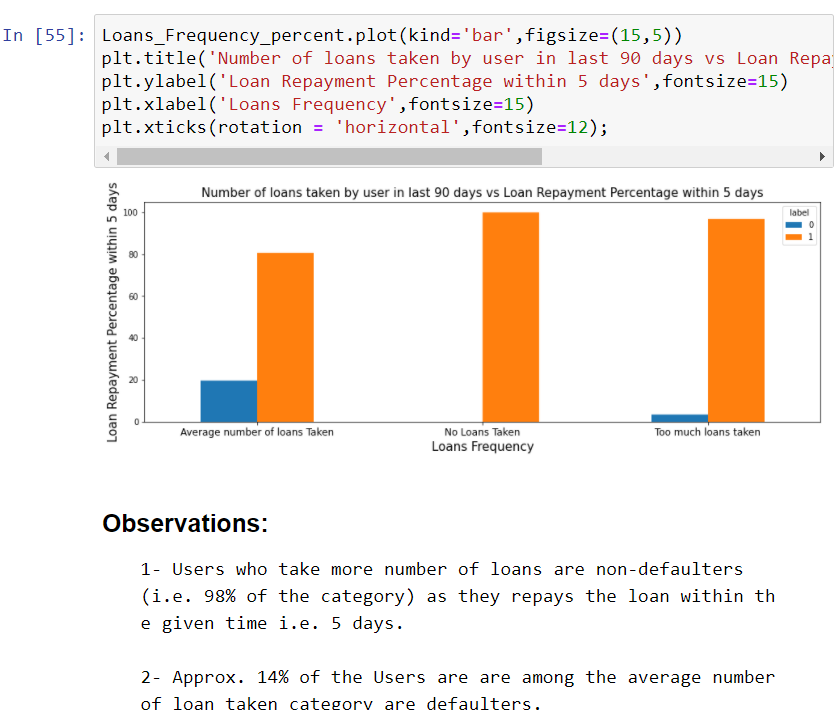


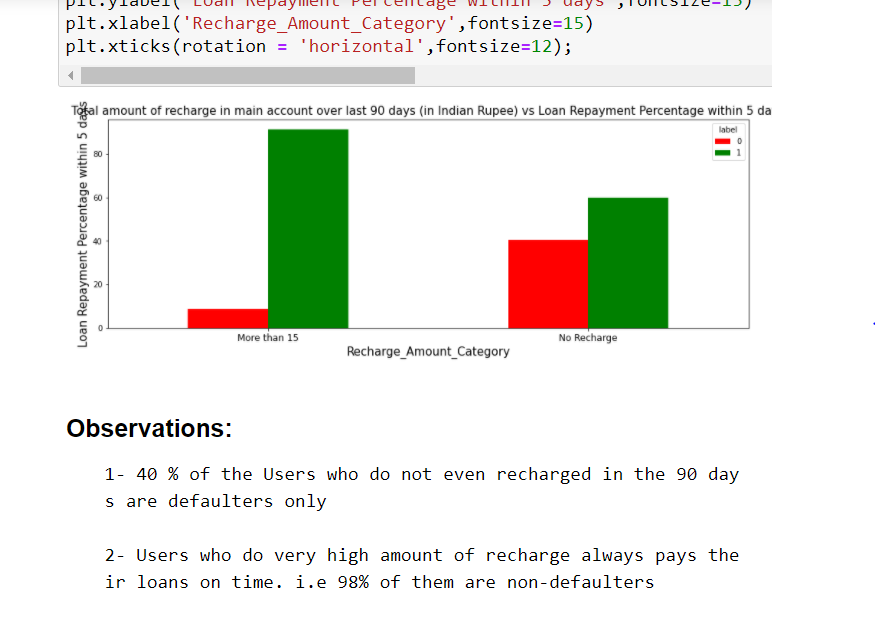


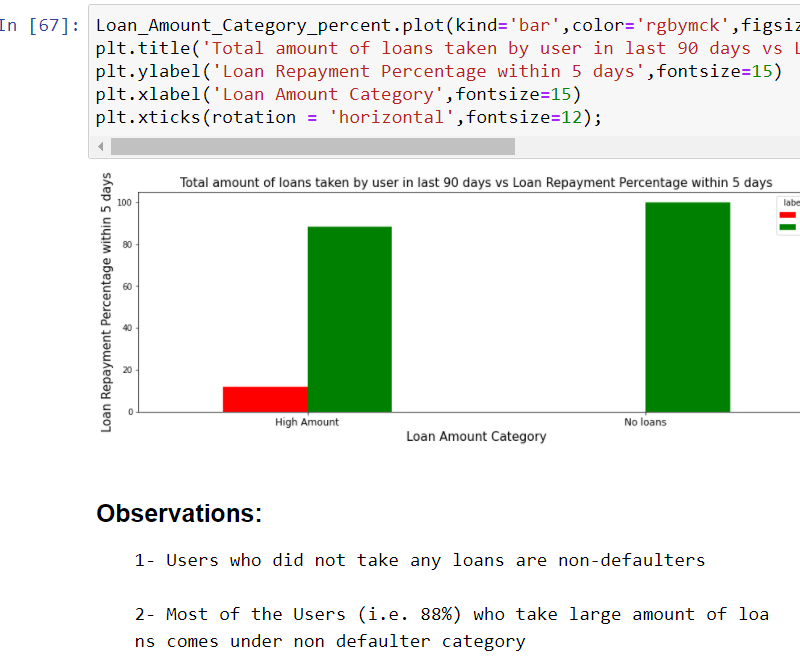


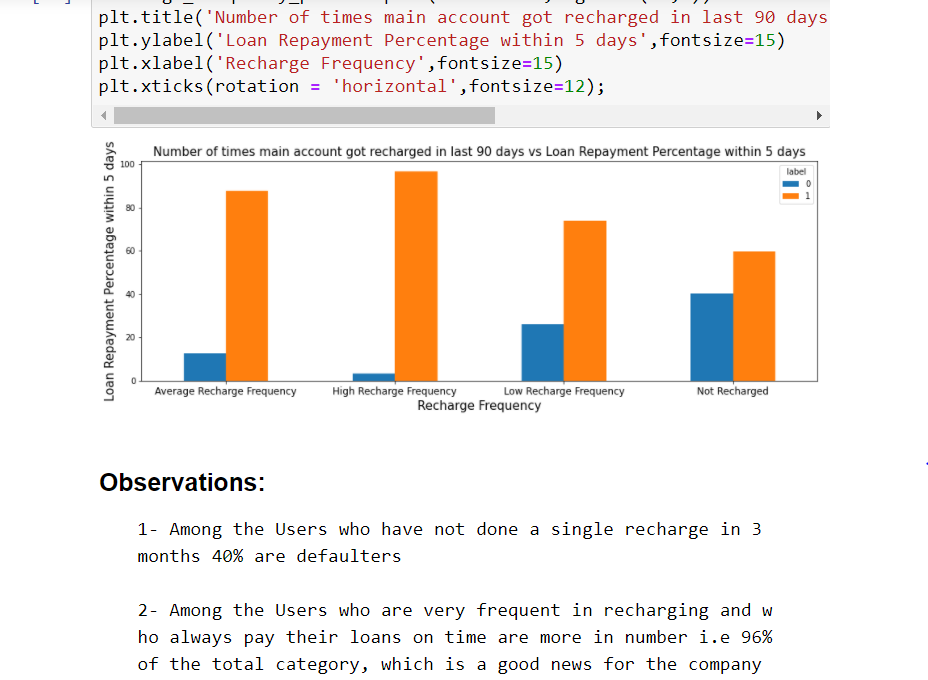


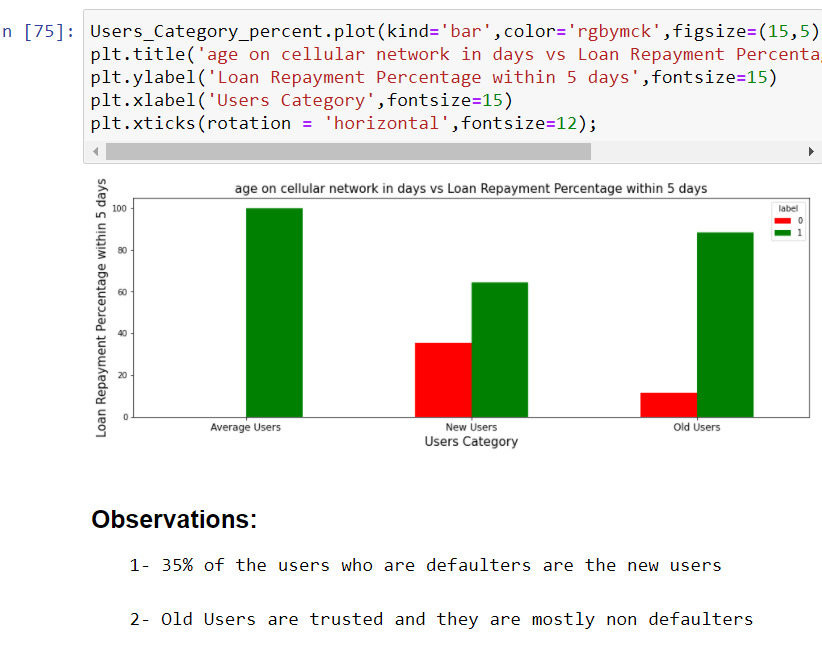




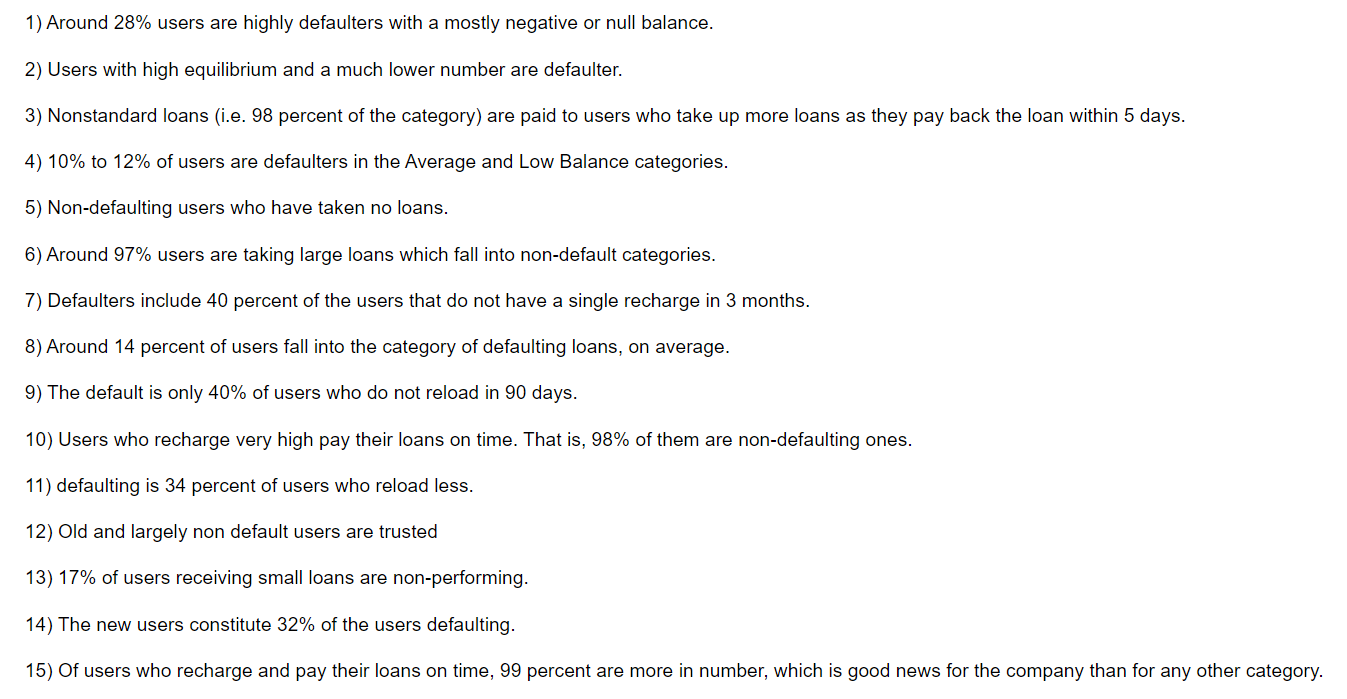








## INTERPRETATION OF THE RESULTS



# 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 28% 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.

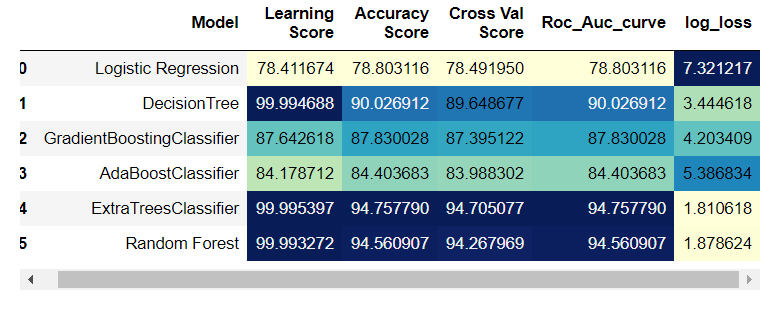
* 28% of Users having negative or zero balance are defaulters, which is very high.
* 10% to 12% Users are defaulters which fall in the category of Average and Low balance category.
* Users having high balance and are defaulters are very less in number
* Users who take more number of loans are non-defaulters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

## LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

From this problem, we have learned that through power of visualization, we can directly check the outliers through analysing the history of the dataset.

The given dataset was too large. According to data cleaning, we have learned further that the dataset was imbalanced. Even, we have found high correlated data which were deleted. Many outliers were seen in the dataset. Within the given time limit, we have tried to delete the outliers as possible as we can.

From the whole case study, we have seen the best algorithm used to train the machine according to the dataset is Random Forest Classifier as all the values along the metrics were highest.



We have found the challenge while fitting the models due to large dataset. Our laptop was not powerful enough to run the largedata set, it took me more than 60hrs just to run the whole model post correction and re checking the code. So that the major limitation I have faced.

## LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK

The only limitation was time limitation. The future scope of project is that we can train the machine identify & restrict the fraudulent through micro credit loan business.

This could further aids the telecom business income generation by restricting the fraudulent. We can use techniques like featured engineering, PCA and different boosting algorithms further to retrieve better results.

The major limitation I faced was the present of outliers as the data is expensive so we cannot drop those values and with those values our analysis is not very accurate.