

Micro Credit Defaulter Project

Submitted by:

Subhash Chandra Pal

# ACKNOWLEDGMENT

This dataset of micro credit analysis has been provided to us from a client that is in telecom industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

# INTRODUCTION

## Business Problem Framing

It is a project related to Telecom Industry. They have collaborated with Microfinance Institution (MFI) that offers financial services to low income populations. Micro Finance Service (MFS) become very useful when we are targeting unbanked poor families living in remote areas with not much sources of income.

The client is in telecom industry and they are a fixed wireless telecommunications network provider and they understand the importance of communication and how it affects a person’s life, thus focusing on providing their services and products to low income families and customers that can help them in the need of hour. They are 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. We have to build a prediction model which will tell us whether the person will become defaulter or not thus helping company in giving credit. This would help client company in further investment and improvement in selection of customers.

## Review of Literature

In this model we will study different variables and how these independent variables are related with dependent variables and how this will help us to predict whether the customer will become defaulter or not using different machine learning model and thus selecting the final model that giving us best score.

## Motivation for the Problem Undertaken

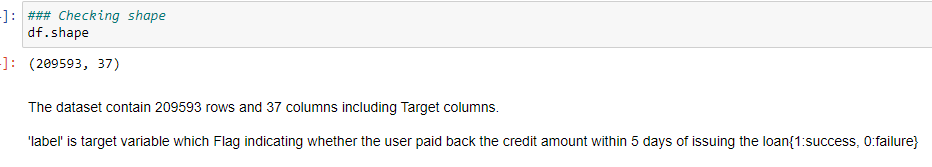
In today’s modern world scenario communication has become the backbone of every individual. The initiative of helping low-income families by proving then micro credit loans for communication has been proved very beneficial to them and building a prediction model for the company which will help them to predict whether loan provided to customer will become defaulter or not, this will help company in future weather and in which condition he should provide the customers micro credit loan.

# Analytical Problem Framing

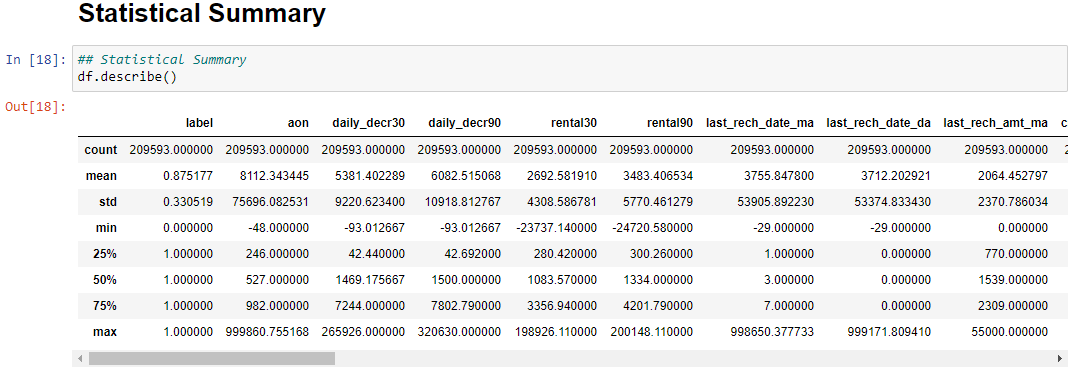
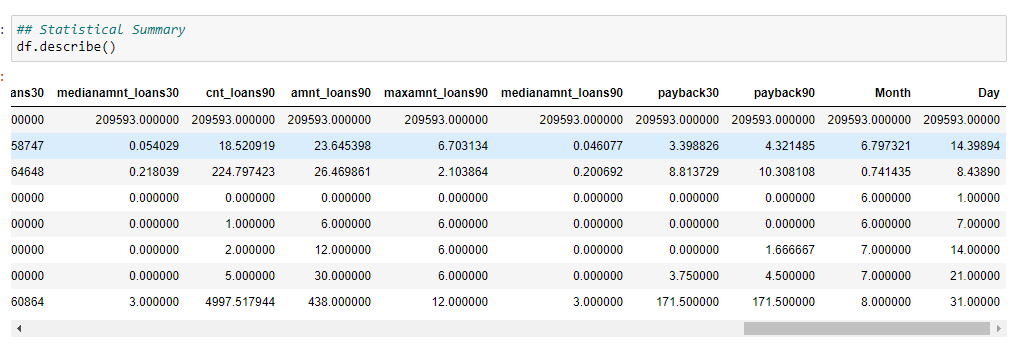
## Mathematical/ Analytical Modeling of the Problem

Let’s import our csv file into by importing some important library and loading into our jupyter Notbook



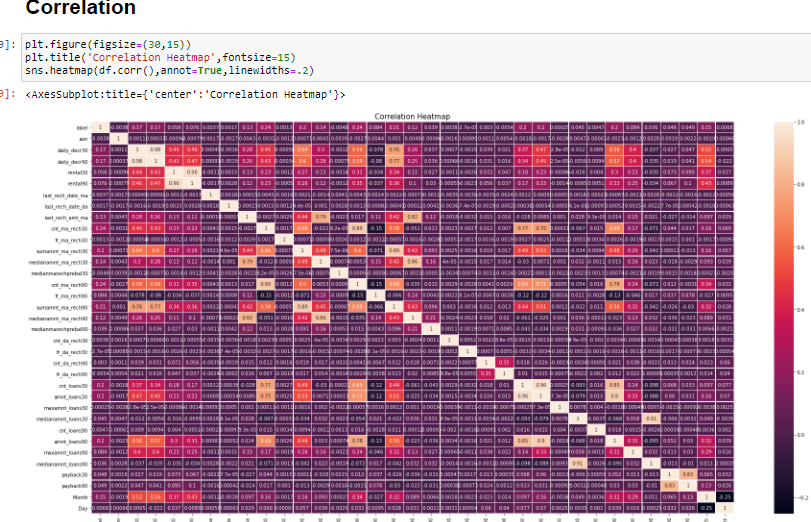


* + There are 209593 distinct micro-credit customers.
  + There are 37 attribute including target attribute.
  + ‘label’ is target which indicate whether user paid the credit loan amount within 5 days of issuing the loan(1:success,0:failure)
* Statistical Summary
* Let’s Check the Statistical Summary of our dataset.

* In most of attributes the minimum values is zero and in some attributes like **rental30 and rental90** Which seems to a erroneous data .
* Mean values are highly deviated from the median value which shows that data distributed is rightly skewed.
* The difference between 3rd quantile and maximum value are too high hence we can clearly say that our dataset have huge outliers present.
  + The average value for Number of days till last recharge of data account is 3712.20. The standard deviation is unusually large, max value being 999171.80.
  + The average value for Number of times main account got recharge in last 30 days is 3.97 and the max value of recharge is 203.
  + The average value for number of times data account got recharge in last 30 days is 262.57. The standard deviation is high , amo value being 99914.44
  + The average value for number of loans taken by user in last 30 days is 2.75 and std is 2.55, max value

#### Lets See co-relation between the Columns



* + From the above observation we can see that aon, medianmarechprebal30

, fr\_da\_rech30, fr\_da\_rech90 are negatively co-related and rest are positively co-related with label**.**

* + Feature like cnt\_ma\_rech30, cnt\_ma\_rech90, amnt\_loans90, amnt\_loan 30 have positive correlation values with target.
  + Features like last\_recharge\_date\_ma, fr\_ma\_rech30 almost have no correlation with Target variable.
  + We will not drop any feature based on correlation because data is expensive.

## Data Sources and their formats:

We have two excel data file one has the details of all user and their different recharges and loan taken and whether they had paid back loan or not and other file contain details of the data.

## 

## In this dataset there are 37 Attributes

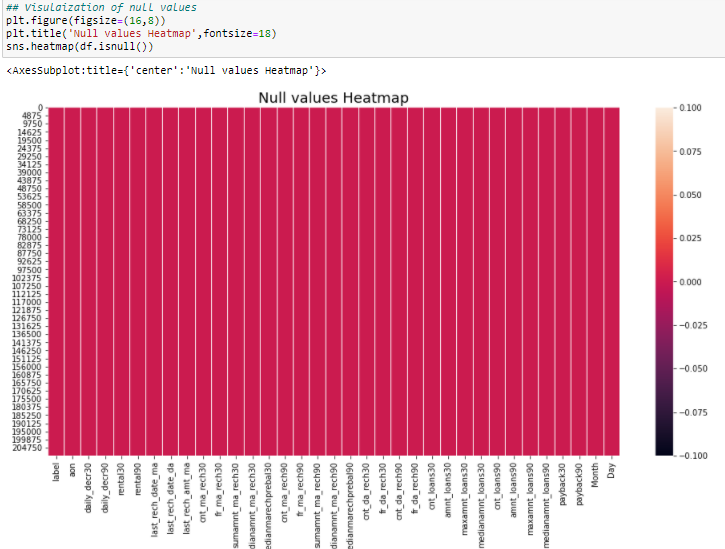
## The Whole dataset is numeric only two feature have object data type .

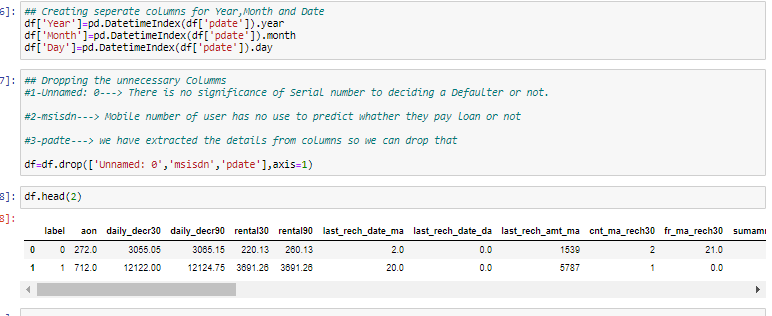
## Pdate feature has datetime datatype.

## Pcircle and msisdn have object datatype.

## Data Preprocessing Done

* + Lets check the shape and see count of the number of empty values in each column.



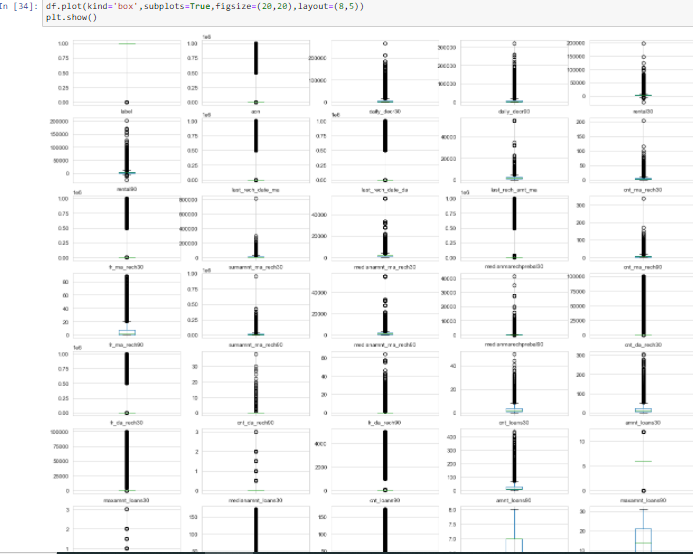


* + As we can see from above Dataset contains 209593 rows and 37 columns in which label is the dependent target column and rest are independent columns.
  + And we can see dataset contains no null values.

* + Data set contains pdate in format of year, month and date. We will split the pdate column for further analysis.
  + After checking the unique values of each column we can see that year count is only one so we will drop pdate, year,msisdn and unnamed as it is of no use.

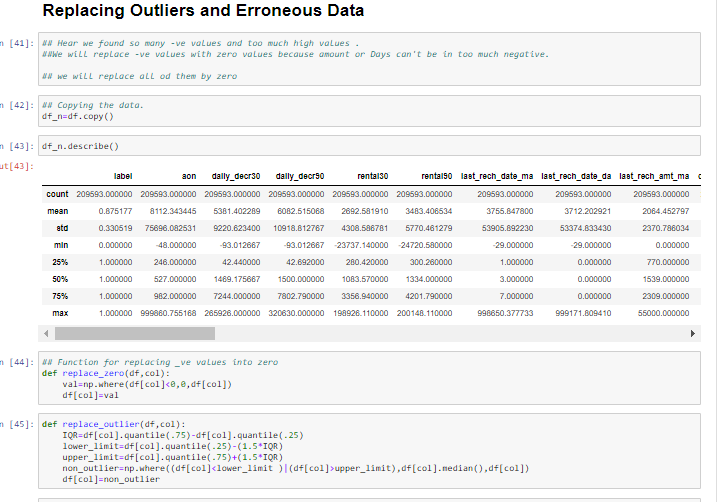
**Outliers:**

From below boxplot we can see there are many number of outliers present on our dataset.



**Outliers and Erroneous data removal:**

* There is a huge negative and imaginary data present in our dataset. But as per domain knowledge we know that that mobile recharge balance can’t be negative it could be zero only. As same we know that day can’t be negative so we replace all the negative values with zero.
* Here in our data Distribution, we can see that the outliers are present only upper to the upper whisker in box plot which shows that our data is Right skewed and we know that for skewed data we can perform IQR method to detect the outlies.
* We didn’t remove the outliers. Here instead of removing outliers cause data loss upto 10% so here we replace all the outliers with Median values.

****

****

## **Skewness Removal**:

## Our dataset had positive skewness . So after removing outliers we had to remove skewness for getting a data which is close to normal distribute bell curve. So , for getting that we had to apply some transformation methods to remove skewness. Hence, we applied here a root square method to remove skewness of a Right skewed distribution.

## 

## **Data Inputs- Logic- Output Relationships:**

## **Data Visualization:**

## **Distribution of Target variable**

## It is Clear visible that our target dataset is imbalance .

## 

## **Number of times main account get recharge in 30 days.**

## 

Observation:

1- Above graph shows that the most people recharge there phone one time in months.

2- People who recharge there phone 3-5 times in months have also very less tendency be a defaulter.

3-People who don't recharge their phone in months have very high tendency to be a defaulter.

**Number of times main account get recharge in 90 days**

## 

Observation

1- Here, we found the similar trend as of above

2- People who don't recharge their phone in 90 days have higher tendency to take micro loan and to be an defaulter by not paying within 5 days.

3-The trend (Taking loan and being defaulter) goes down as number of time account recharge increase in 90 days.

**Total amount of loans taken by user in last 30 days:**

Observation (Below Graph)

1-Mostly user recharge took loan of 6(in Indonesian Rupiah).

2-As of domain knowledge, if people recharge then 1st, he have to pay the loan then again user get chance to take another loan.

3-12,18,24(in Indonesian Rupiah) could be taken by those people who payback the multiple loan within a month and took another.

4- Gradual drop in loan rupee after 12 Indonesian Rupiah, People are also having tendency to pay back.

****

**Number of loans taken by user in last 30 days**

****

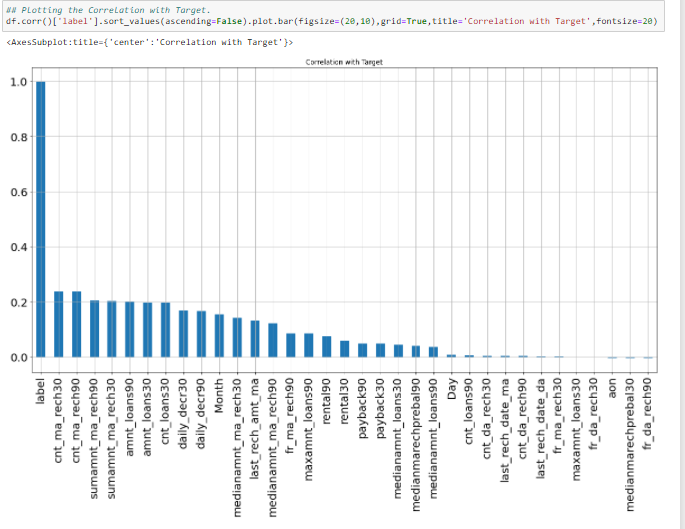
Observation:

1-The trends show, when number of loan taken by user decreases, it's tendency to be a defaulter is also goes down.

2-There is higher risk to to grant micro loan to a user who take loan once in month.

**Correlation Graph:**

This is a correlation plot of the of independent features with target features.



1-It seems from the above graph is that negatively correlated feature is age on cellular network in days, medianmarechprebal30, but we cannot blindly remove this feature because according to me it is very important feature for prediction.

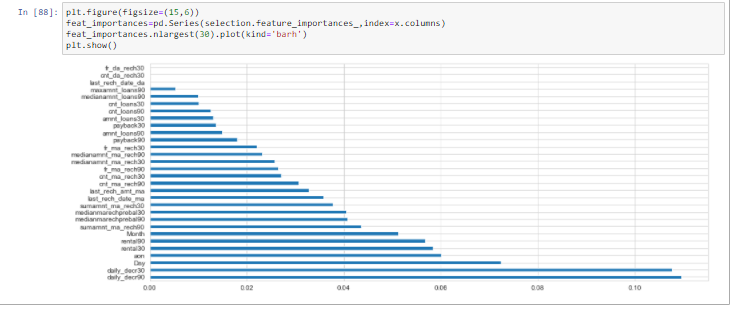
2- Features like age on age of network (aon),fr\_da\_rech30,medianmarechprebal30,fr\_da\_rech90 are negatively correlated but we won't drop these because these are important features.

3- We will perform PCA instead of dropping columns based on correlation values.

**Feature Importance**:

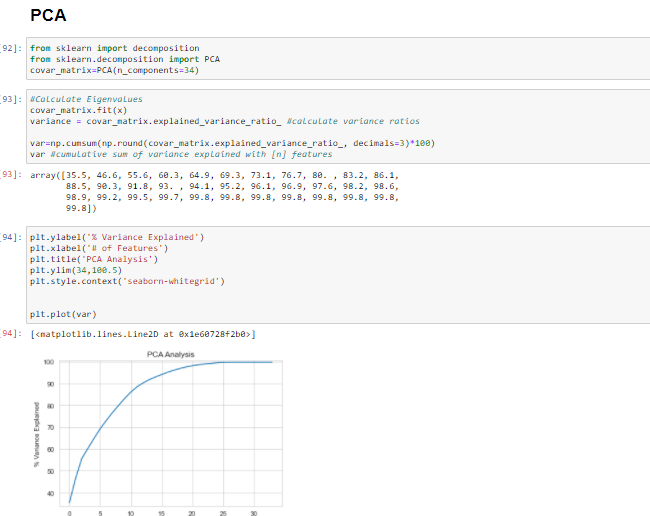
In below diagram, we can clearly see the important features for our Model. We will not remove any data columns based on this graph. We will do dimension Reduction by PCA.

Here we can see features like fr\_da\_rech30,cnt\_da\_rech30,last\_rech\_data\_da has no contribution to predict our outcome.



**PCA:**

PCA is a Dimensionality Reduction Algorithm. Here we imported PCA library from sklearn, then after successfully scaling out dataset we fit our independent data (X) and get that 99% of data can get from 20 n\_components. So , we have chosen 20 features for our model building.

****

# Model/s Development and Evaluation

## Step1: Assigning Input and Output variable

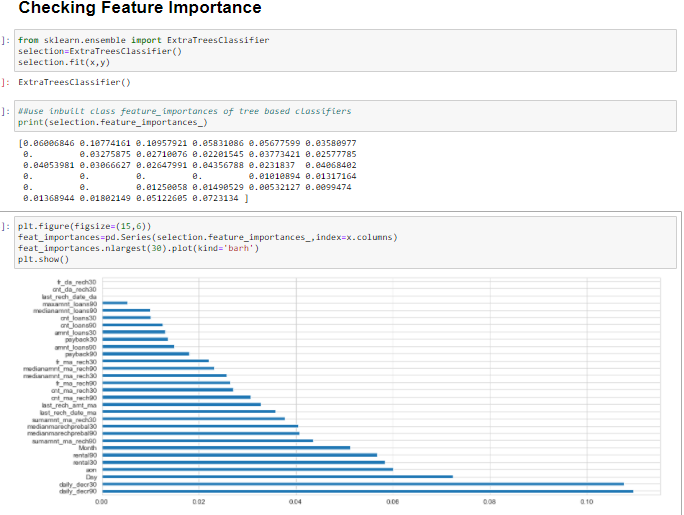
## Here we split the data frame into independent and dependent variables.

## X is the independent variable and y is dependent variable.

## 

**Lets check feature importance of the Data set.**

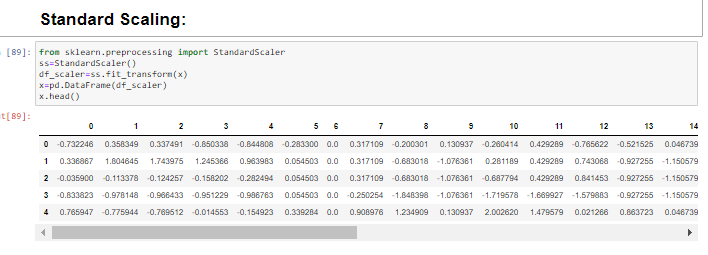
* + You can get the feature importance of each feature of your dataset by using the feature importance property of the model.
  + Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable.
  + Feature importance is an inbuilt class that comes with Tree Based Classifiers, we will be using Extra Tree Classifier for extracting the top 10 features for the dataset



* + From the above analysis we can see that Daily\_decr90, daily\_dec30 are the month important feature for model valuation and medianamnt\_loans90, medianamnt\_loans30 are less important.

**Scaling: Standard Scaling**

* + Scaling is required in distance-based algorithms like Logistic Regression, PCA, KNN and Gradient Boosting.
  + In our independent feature data have different units and variation is there.
  + So, to scale down all features we use standard scaling.

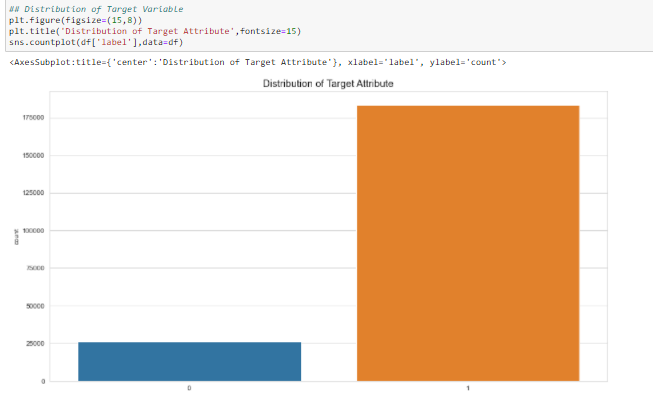
****

## Testing of Identified Approaches (Algorithms)



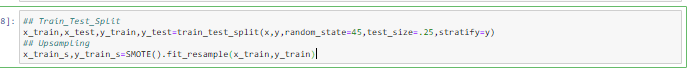
Target Distribution:

* Our target variable is imbalanced in nature. Where 1 represent that people return the loan and 0 shows they are fail to return loan.
* Due to Imbalance dataset we applied here up sampling method (SMOTE) to our training dataset.



**Up sampling:**

* We done up sampling by using SMOTE .



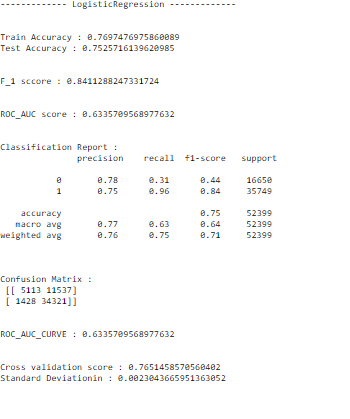
**Model Building:**

Here we made a function to perform our Training and Testing of Machine Learning Algorithms.



1. **Logistic Regression**

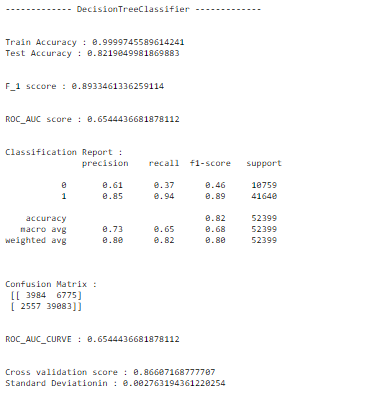
* In Logistic Regression, we wish to model a dependent variable(y) in terms of one or more independent variables(x). It is a method for classification. This algorithm is used for the dependent variable that is Categorical. Y is modeled using a function that gives output between 0 and 1 for all values of X. In Logistic Regression, the Sigmoid (aka Logistic) Function is used .



#### Decision Tree Classification

The idea of a decision tree is to divide the data set into smaller data sets based on the descriptive features until you reach a small enough set that contains data points that fall under one label.

Decision trees are easy to interpret. To build a decision tree requires little data preparation from the user- there is no need to normalize data.



#### Random Forest Classification

Random Forest is a supervised learning algorithm, it creates a forest and makes it

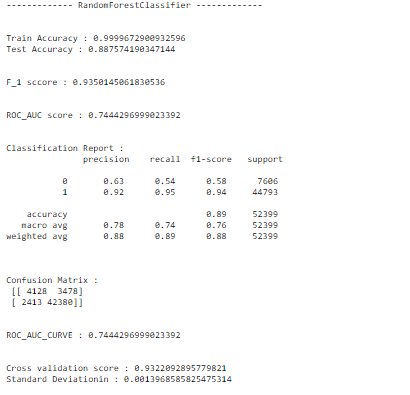
somehow random. The "forest “it builds, is an ensemble of Decision Trees. Step-1Pick at random K data points from the training set.

Step-2 Build the Decision tree associated to these K data points

Step-3Choose the Number of trees(n) you want to build and repeat Step1 and Step2

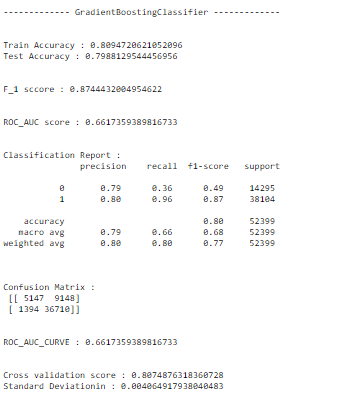
Step-4For a new data points make each one of your 'n' trees predict the category to which the data point belongs and assign the new data point to the category that wins the majority vote.

**Result** :



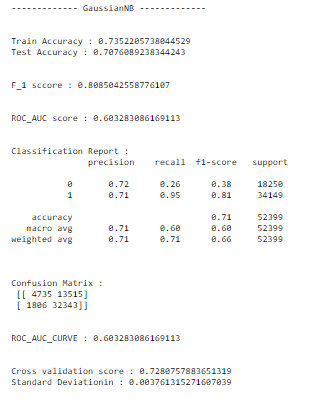
#### 4.Gradient Boosting-

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.



**5.Naive Bayes:**

In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels.



## Key Metrics for success in solving problem under consideration

**Accuracy Score** is the number of correct predictions made as a ratio of all predictions made. It is the most common evaluation metric for classification problems.

**Cross-validation** is to call the cross\_val\_score helper function on the estimator and the dataset.

To estimate the accuracy of a linear kernel support vector machine on the dataset by splitting the data, fitting a model and computing the score (n=5 or any number provided by you) consecutive times (with different splits each time):

The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes

**Receiver Operating Characteristic(ROC)** summarizes the model’s performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate.

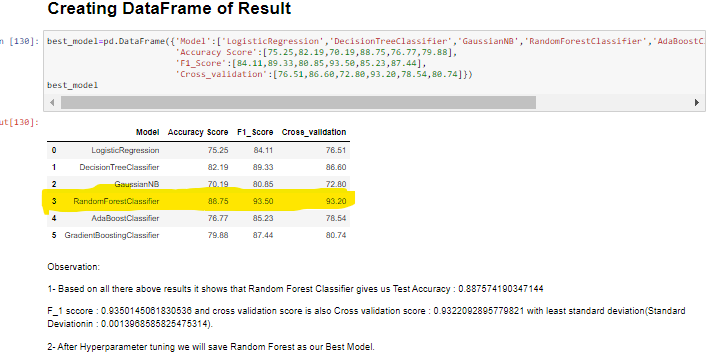
ROC summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.

**F1-score** is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive.

The F1 score is the harmonic mean of the precision and recall.

**RESULT:**

Here we have created the data Frame to compare the results of different Machine Learning Algorithms.

****

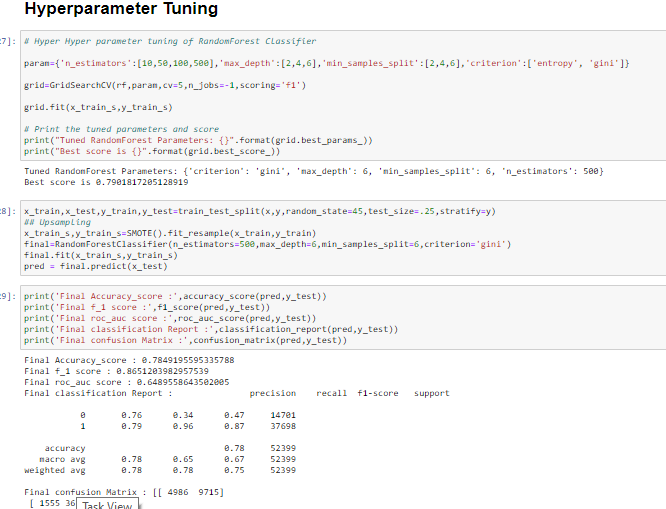
* Based of above result we can clearly see that our Random Forest Classifier has the highest accuracy and F\_1 score among all the other machine learning models.
* To check the overfitting we also find the cross validation score to compare the model result with 5 cross validation .
* We can clearly see in result that random forest classifier cross validation has the minimum standard deviation and it is also less deviated from the randomly selected random state result with cross validation score of 5 .

Best Model:

* Hence form the above analysis it is clear the our **Random forest model** is not overfit .
* So ,for more exploration we will perform hyperparameter tuning of Random Forest Model.

**Hyperparameter Tuning** :

* To get better result we will do some hyperparameter tuning of our Random Forest Model.



Observation:

* We find that with Hyperparameter tuning we gets low accuracy score and F\_1 score.
* Sometimes with our default variable we get a good score so we will go with our default parameters.

# Final Point:

# Before hypermeter tuning, our accuracy score was 88.75, f\_1 score was 93.50 and cross validation score was also 93.20 up to 5 cross validation. Some times with hyperparameter is not ideal for get improved result, As shown above we got our good accuracy and f\_1 score with default hyperparameter tuning parameters so we will use Random Forest Classifier as our best model.

# Saving And Loading the Model:

# Here we have saved our best model with having a 88.75 accuracy and 93.50 F\_1 score.

# 

# CONCLUSION

## Key Findings and Conclusions of the Study

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

The aim was to determine an appropriate quantities model for using financial information pertaining to the loan and customer behavior on the mobile network to predict the outcome of the loan.

Classification models are appropriate for dealing with the two distinct outcomes for customer behavior of repayment and defaulter.

We have used different models for the predic

**Thank You**