

Micro Credit Defaulter Project

Submitted by:

Mini Sharma

# 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 this 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 todays 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

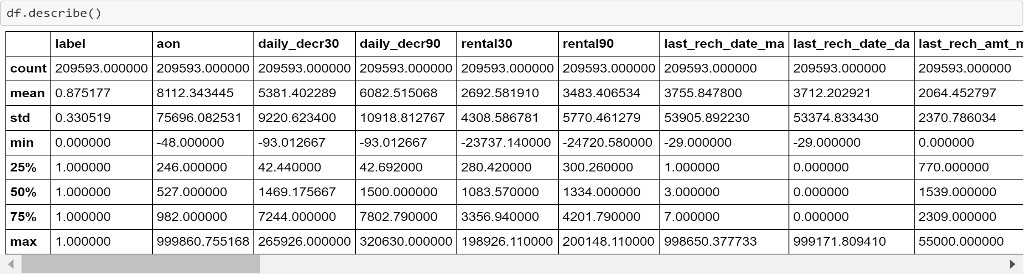
company in future weather and in which condition he should provide the customers micro credit loan.

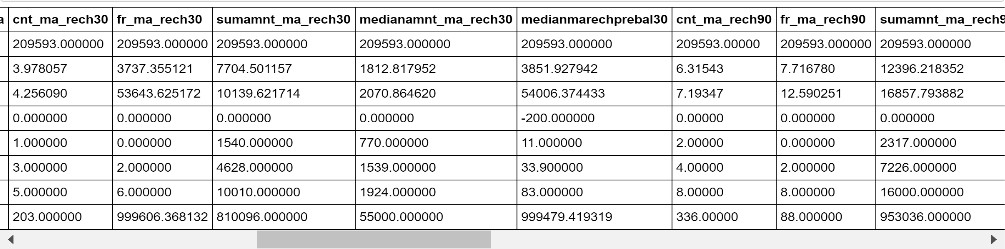
.

# Analytical Problem Framing

## Mathematical/ Analytical Modeling of the Problem

Lets view some basic statistics about the data like the percentile, mean , maximum, minimum etc.





* + There are 209593 distinct micro-credit customers.
  + The average value for Number of days till last recharge of main account is 3755.84. The standard deviation is unusually large, max value being 998650.37.
  + 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 , amx 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 is 50.

#### Snipping ToolLets See co-relation between the Columns

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

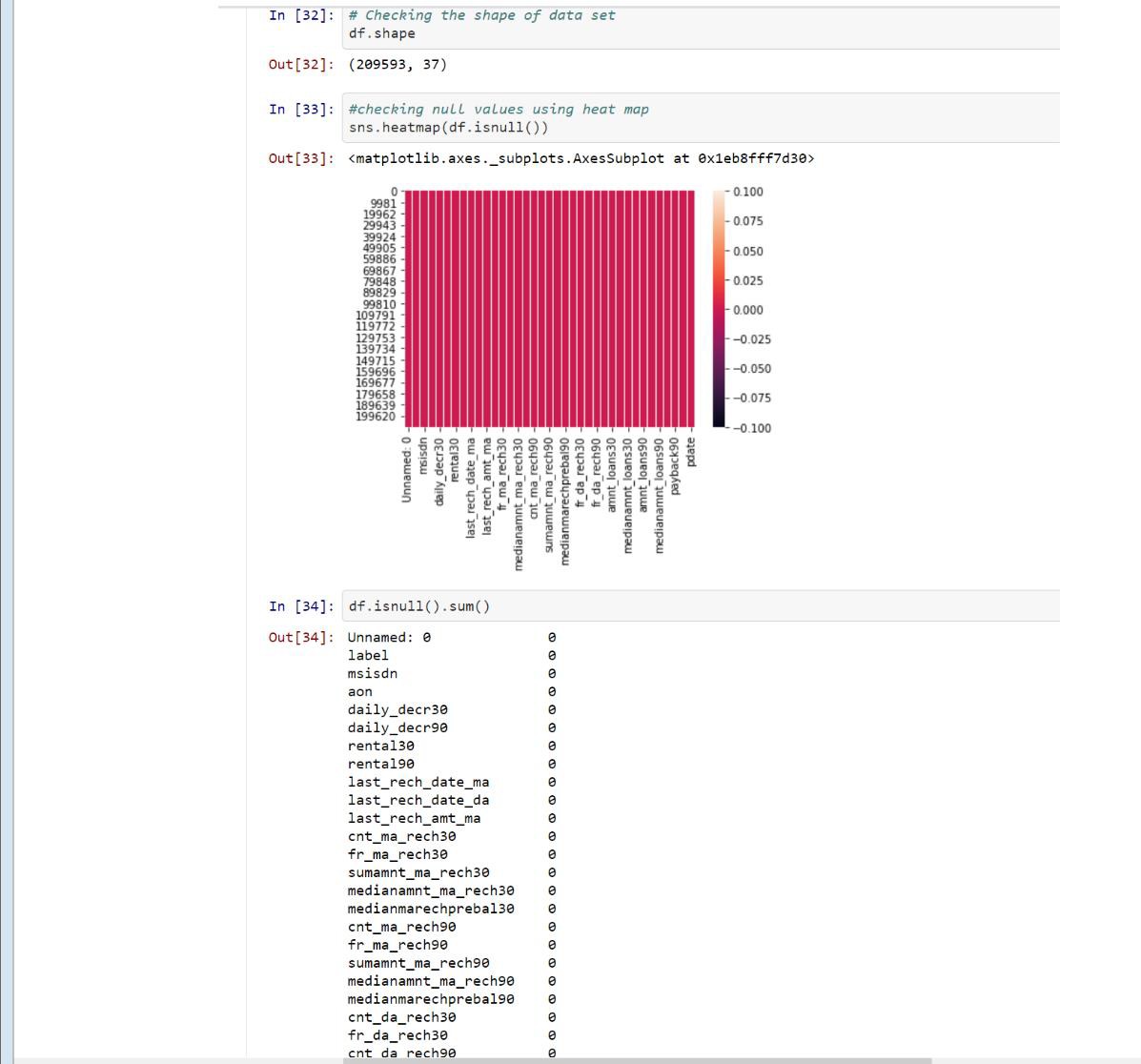
, fr\_da\_rech30 , fr\_da\_rech90 are negatively co-related and rest are positively co-related with label**.**

## Data Sources and their formats

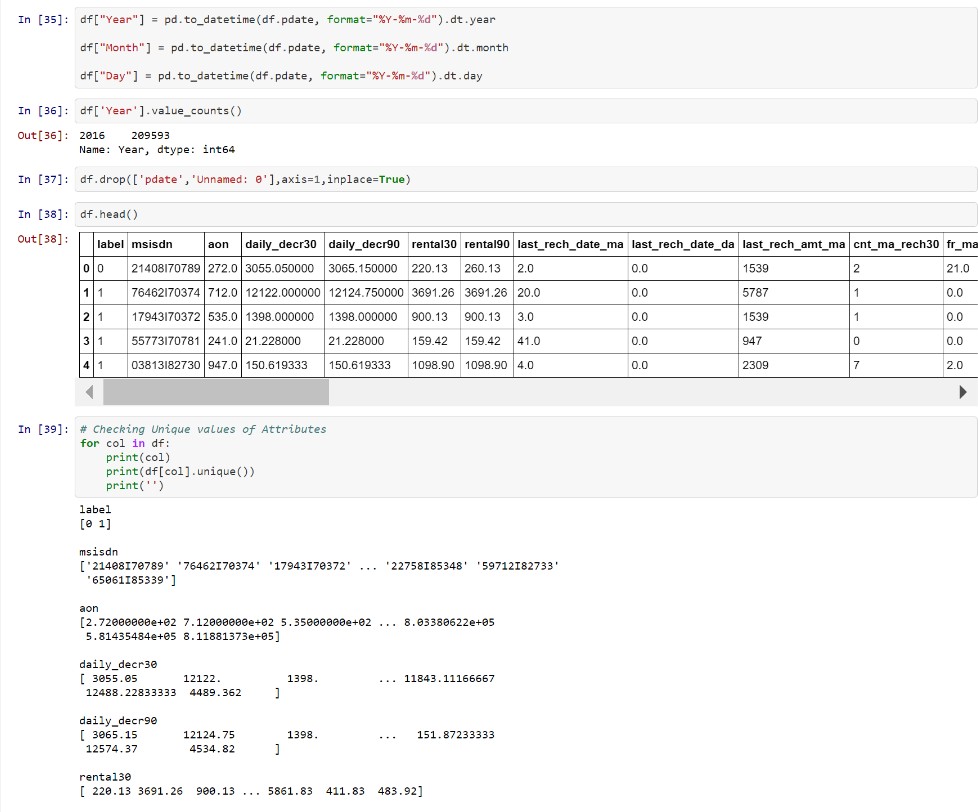
We have two excel data file one has the details of all user and their different recharges and loan taken and whether they had paid back loan or not and other file contain details of the data .

## Data Preprocessing Done

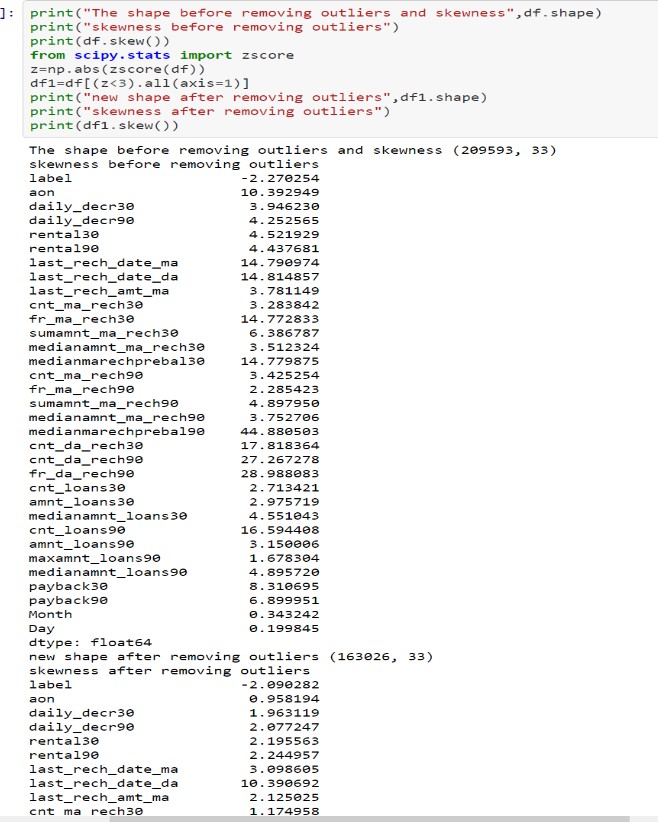
* + Lets check the shape and see count of the number of empty values in each column.



* + 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.

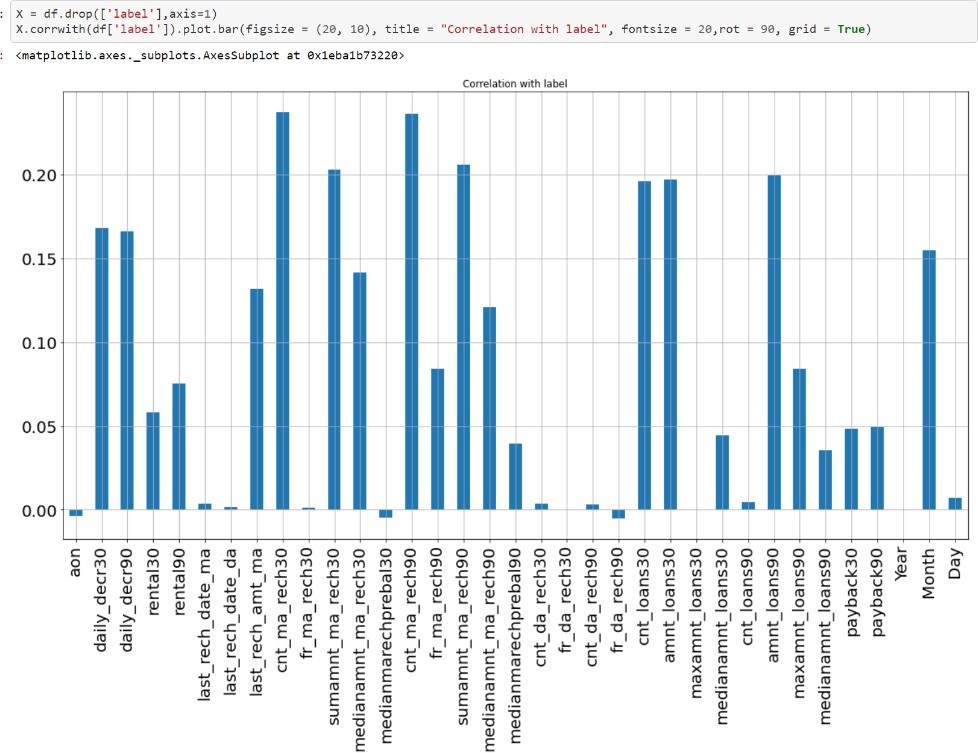


* + Data set contains pdate in format of year ,month and date. We will split the pdate column for further analysis.
  + After checking the unique values of each column we can see that year count is only one soo we will drop pdate , year and unnamed as it is of no use.



* + Data set contains many outliers so by using zscore we will remove outliers from the data set

## Data Inputs- Logic- Output Relationships



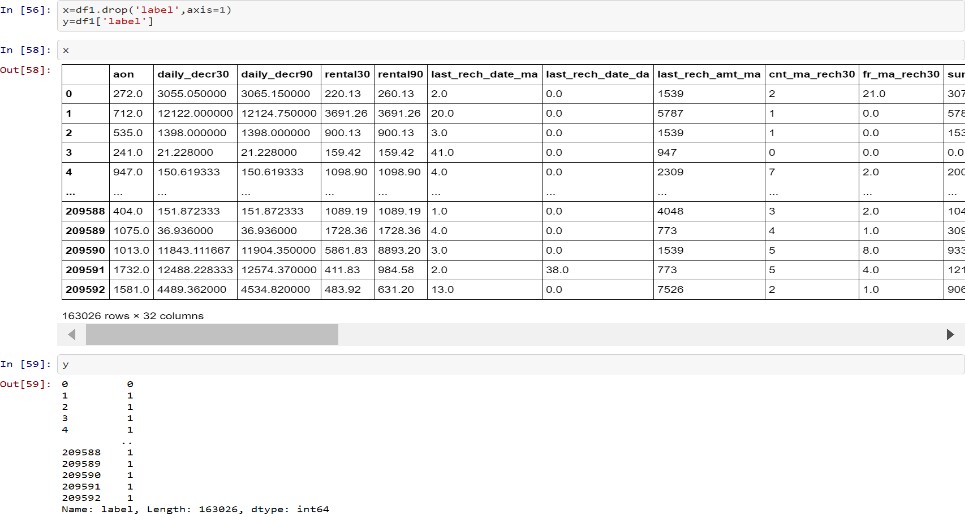
It seems from the above graph is that negatively correlated feature is age on cellular network in days, medianmarechprebal30, but we cannot blindly remove this feature because according to me it is very important feature for prediction. msisdn, year, pcircle and Frequency of data account recharged in last 30 days is unimportant and it has no role in prediction so we will remove it later.

## Hardware and Software Requirements and Tools Used

We will use here Jupyter notebook to make your Prediction Model.

# Model/s Development and Evaluation

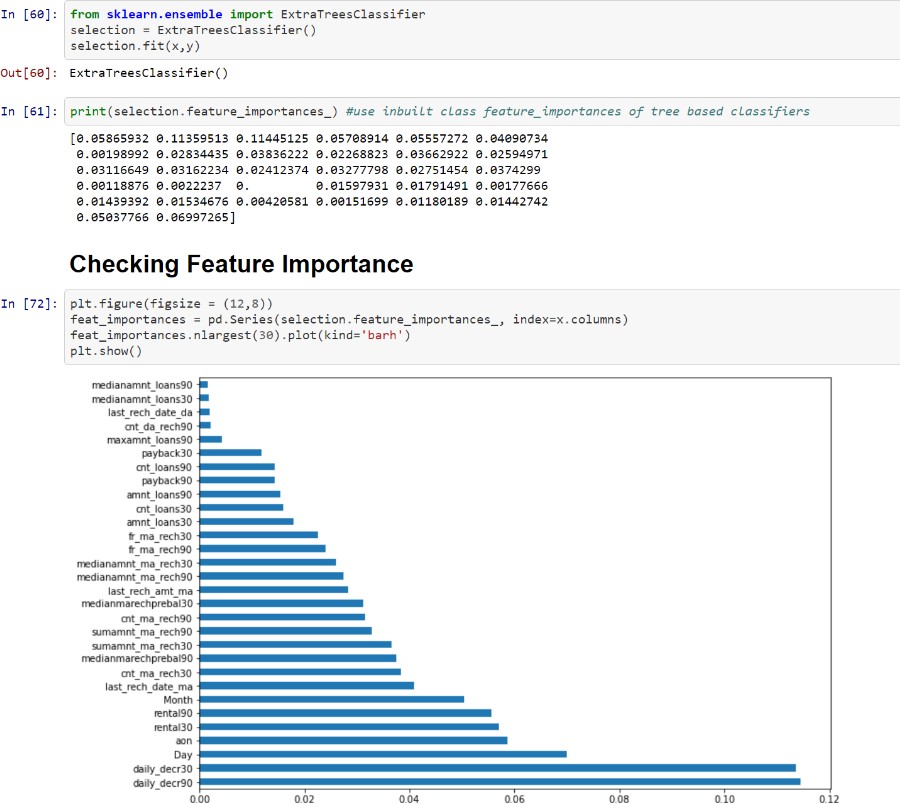
## Identification of possible problem-solving approaches (methods)



* + Now we will split the data set into input and output variable. As you can see above x is your input variable and y (label) is your target out variable.

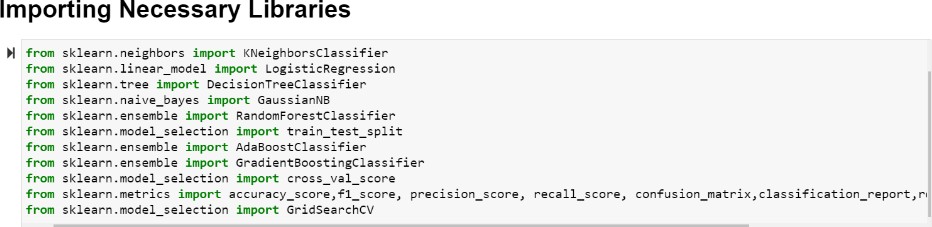
**Lets check feature importance of the Data set.**

* + 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



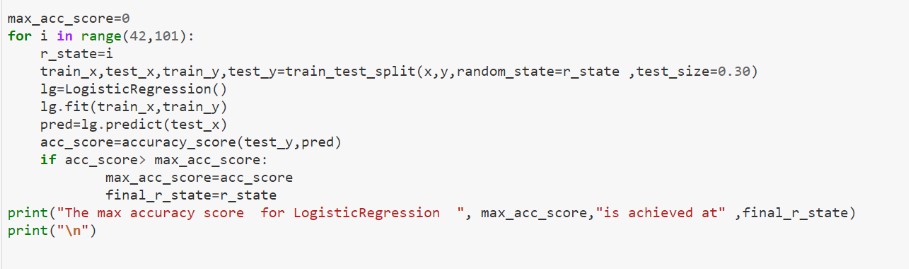
* + From the above analysis we can see that Daily\_decr90, daily\_dec30 are the month important feature for model valuation and medianamnt\_loans90, medianamnt\_loans30 are less important .

## Testing of Identified Approaches (Algorithms)

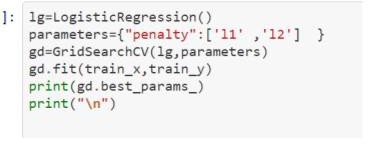


1. **Logistic Regression**

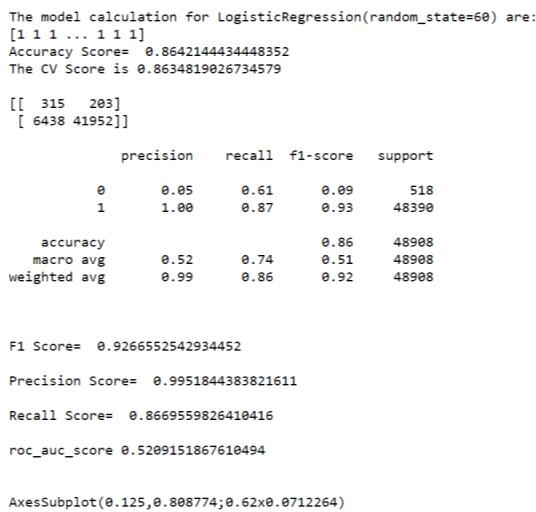
* In Logistic Regression, we wish to model a dependent variable(y) in terms of one or more independent variables(x). It is a method for classification. This algorithm is used for the dependent variable that is Categorical. Y is modeled using a function that gives output between 0 and 1 for all values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

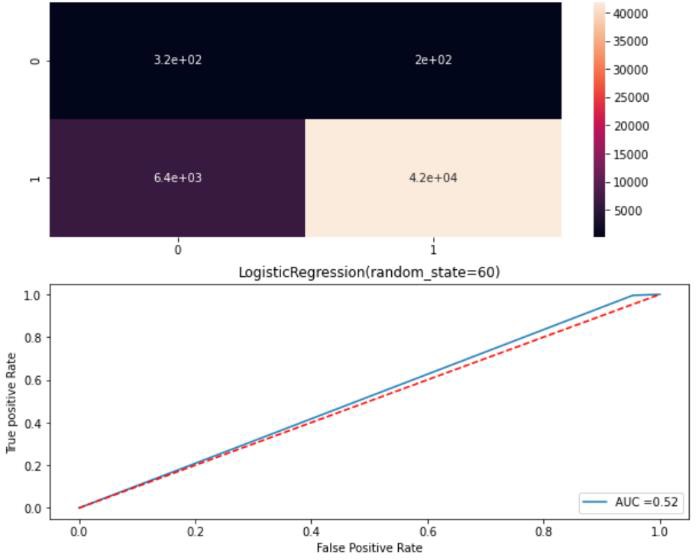


Further we will use grid search cv to find the best parameter for logistic regression.



Results:





#### 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.

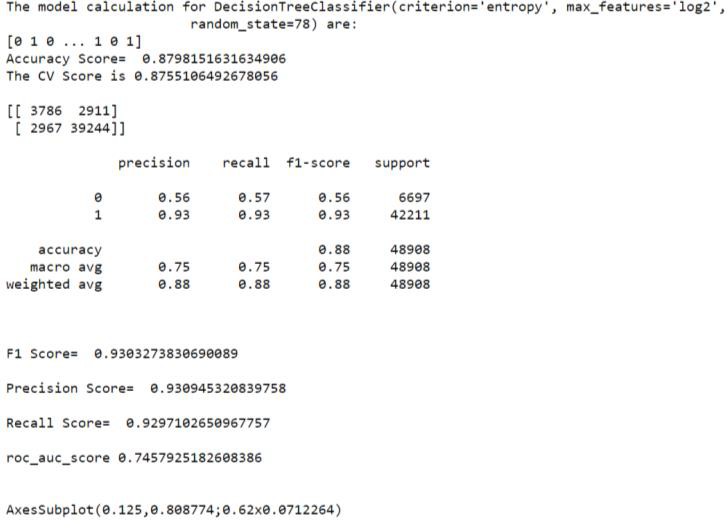
Decision trees are easy to interpret. To build a decision tree requires little data preparation from the user- there is no need to normalize data.

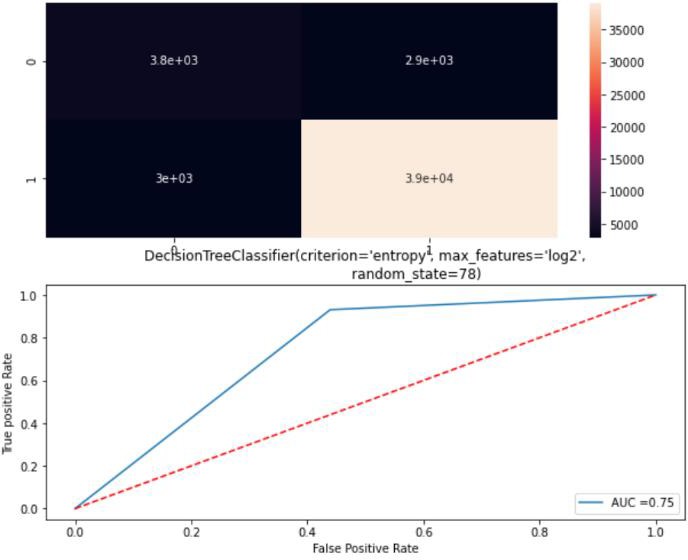


Further we will use grid search cv to find the best parameter for Decision Tree Classifier.



Results:





#### Random Forest Classification

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

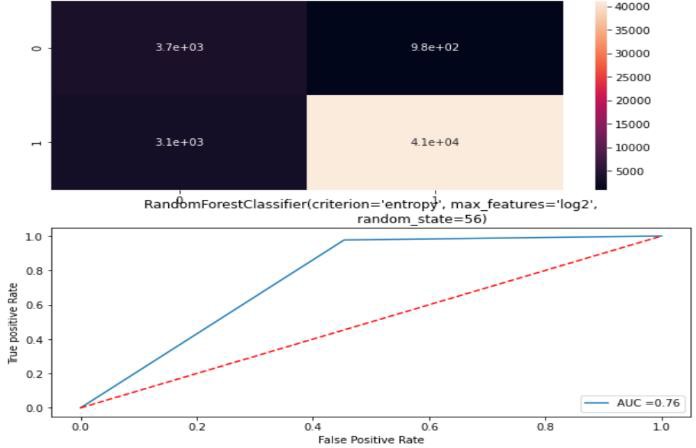
somehow random. The "forest“ it builds, is an ensemble of Decision Trees. Step-1Pick at random K data points from the training set.

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.

Further we will use grid search cv to find the best parameter for Random Forest Classifier.

Results:

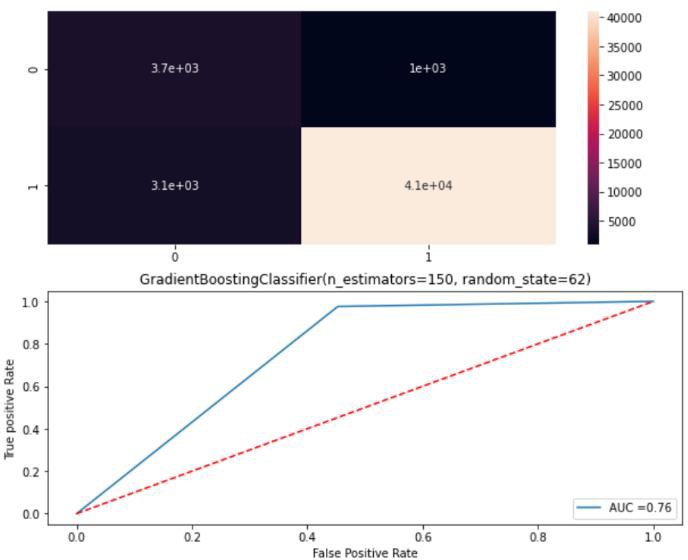
#### Gradient Boosting-

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.



Further we will use grid search cv to find the best parameter for Gradient Boosting Classifier.

Results:



#### K Neighbors Classifier-

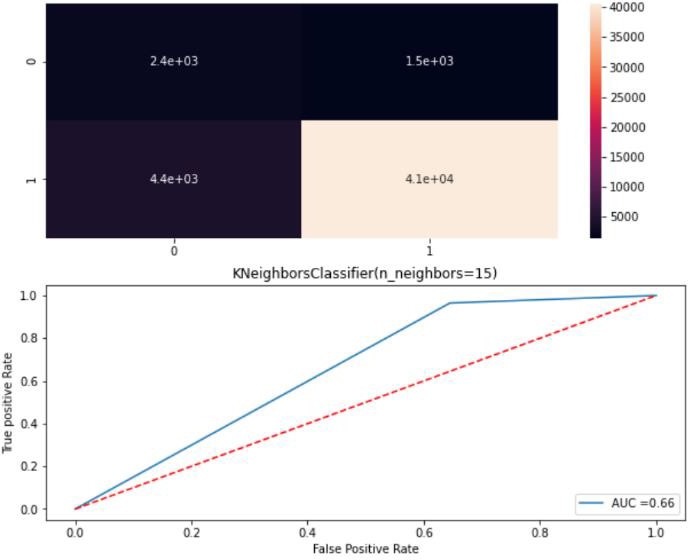
This is a supervised, non-parametric learning algorithm which classifies a given point based on its neighbours. The choice of the ‘k’ becomes very crucial since the data point is assigned to the class of the nearest ‘k’ neighbors. Once we get to know such ‘k’ nearest data points, the test data is assigned a label by taking the

majority vote from the class labels of the ‘k’ nearest data points



Further we will use grid search cv to find the best parameter for K Neighbors Classifier.

Results:



## 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

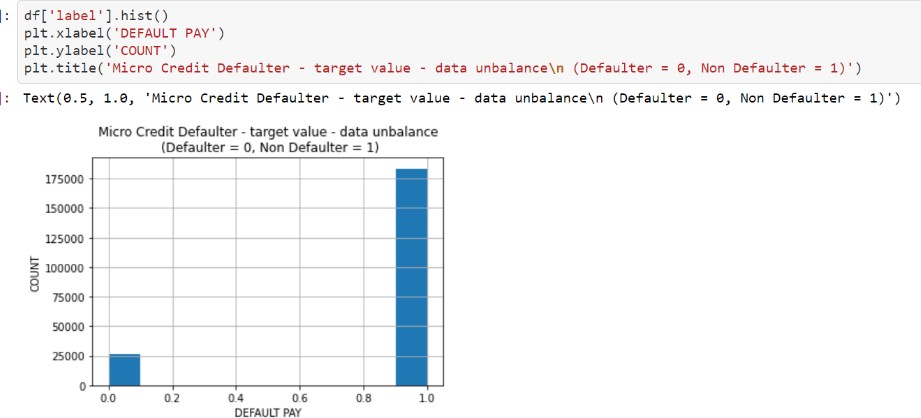
**Receiver Operating Characteristic(ROC)** summarizes the model’s performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate.

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.

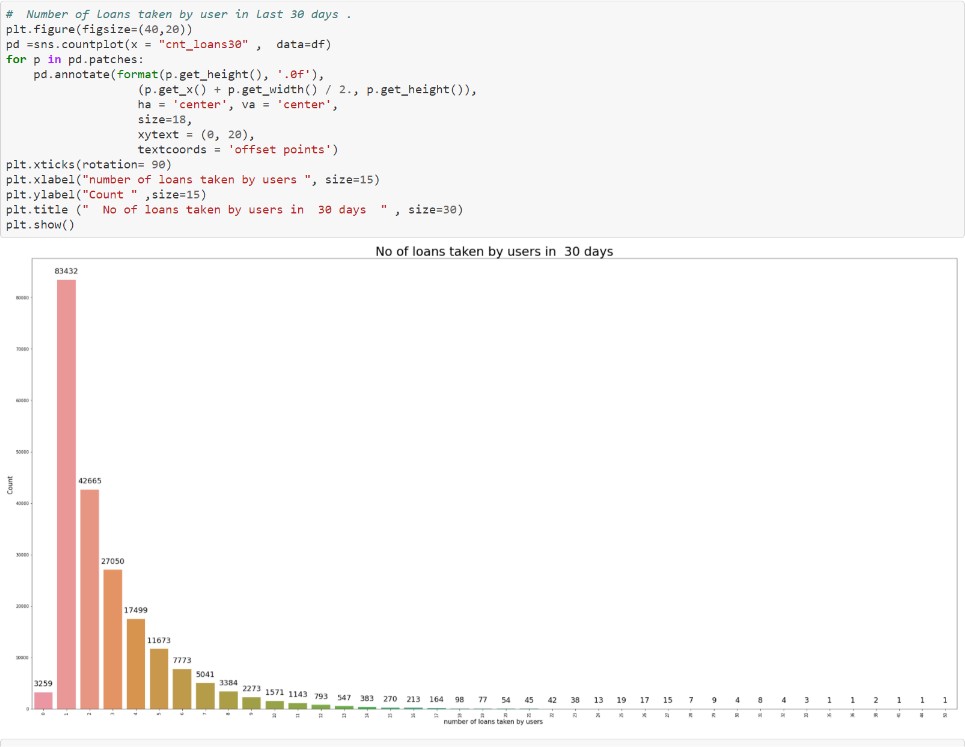
**F1-score** is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number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

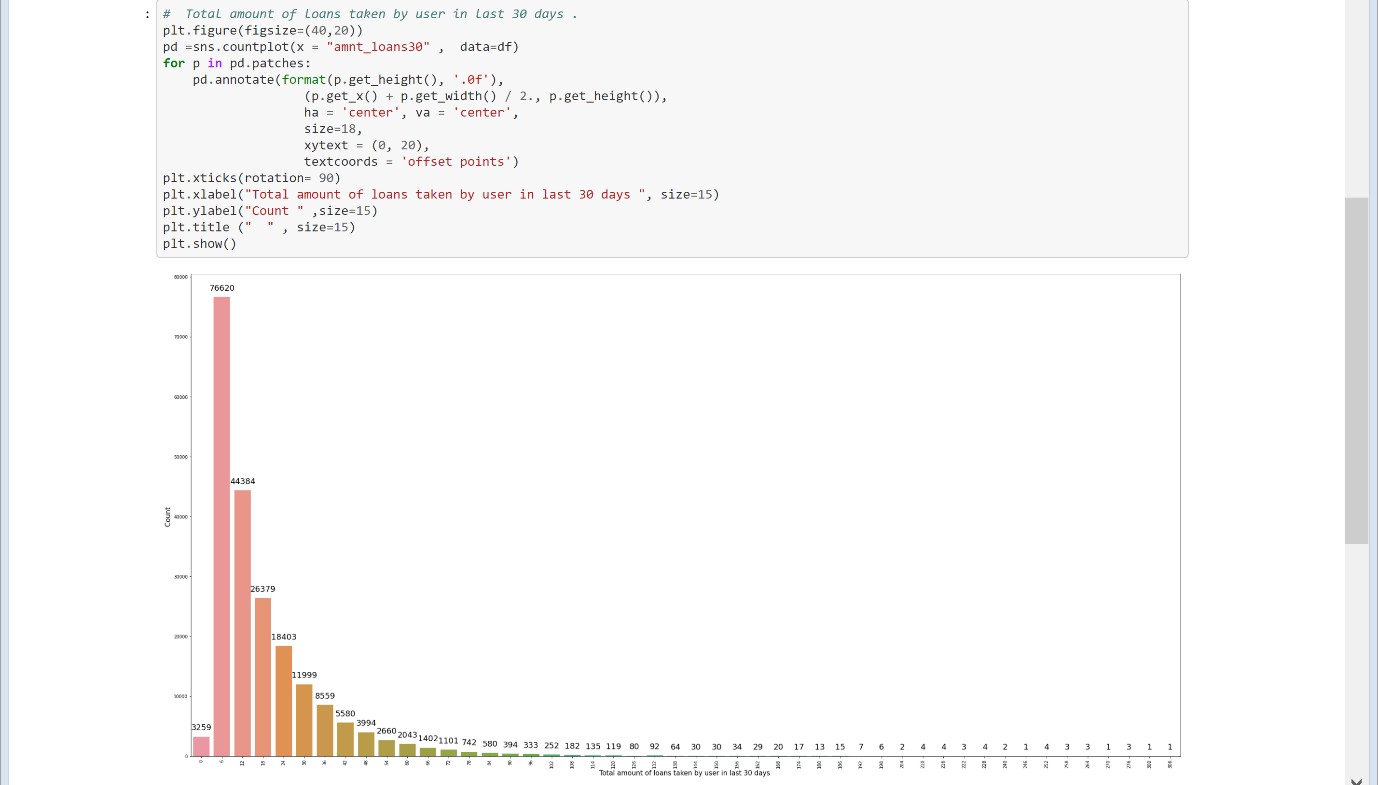
## Visualizations



As we can see in above observation the data set in unbalance. The Defaulters are very less as compared to non-Defaulters.



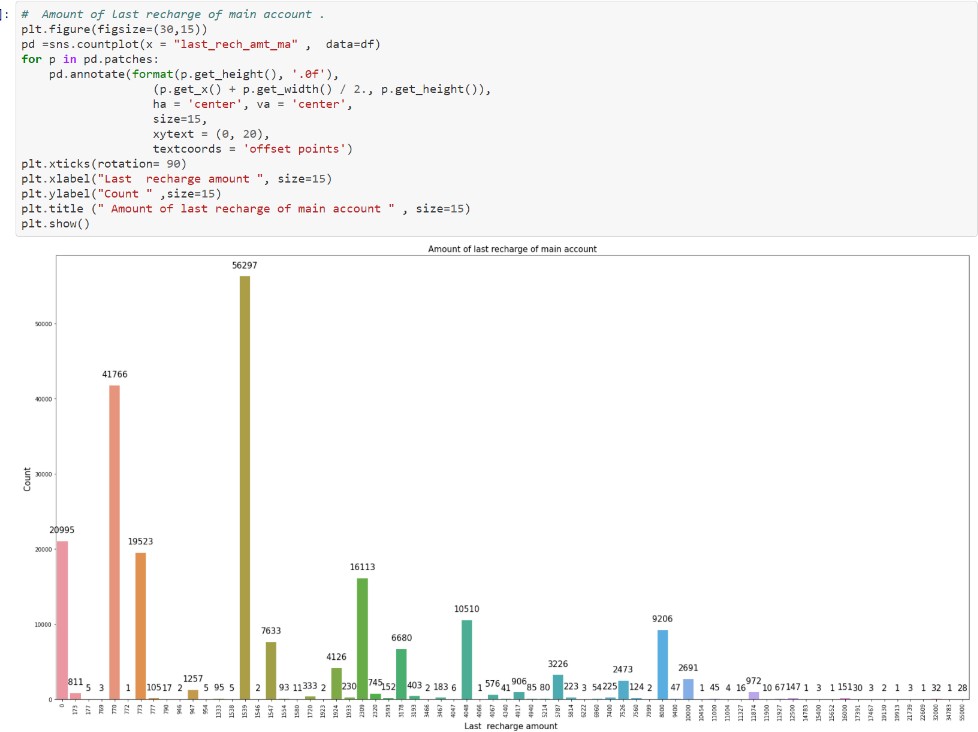
In the above observation we can see that most of the users have taken loans 1 loan in 30 days.



In the above observation we can see that most of the users have taken 6 (Indonesian rupiah) to 30 (Indonesian rupiah) loans. Only Few users have taken loan of more than 30 (Indonesian rupiah)

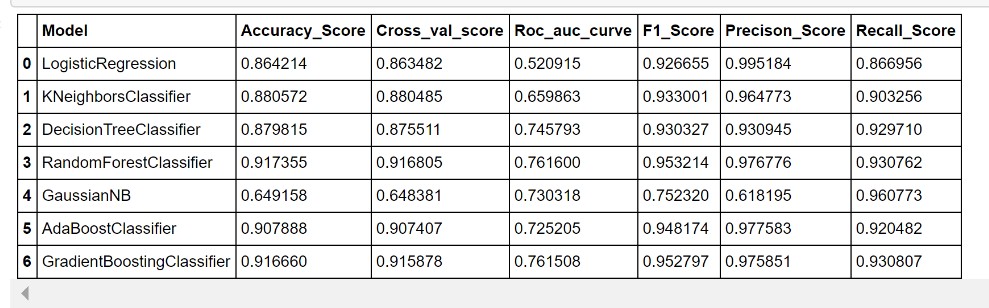


As we can see above majority of users have paid back their loan of 6 (Indonesian Rupiah) and 12(Indonesian rupiah) within 5 days of time.



As we can see above majority of people have recharge their main account with 770 and 1539 , very few users have have recharged their main account with more than 10,000 .

## Interpretation of the Results



From the above tabel we can see that Random Forest is the best performing model .So we will select Random forest classifier as our final model.

### For Model Evaluation we are referreing Confusion Matrix

After we train a Randon forest classifier model on some training data, we will evaluate the performance of the model on some test data. For this, we use the Confusion Matrix

### the accuracy of the model : - ( TP + TN ) / Total

Here, TP stands for True Positive which are the cases in which we predicted yes and the actual value was true. TN stands for True Negative which are the cases in which we predicted no and the actual value was false.FP stands for False Positive which are the cases which we predicted yes and the actual value was False.FN stands for False Negative which are the cases which we predicted No and the actual value was true

# 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 quantative model for using financial information pertaining to the loan and customer behaviour on the mobile network to predict the outcome of the loan.

Classification models are appropriate for dealing with the two distinct outcomes for customer behaviour of repayment nd deafaulter.

We have used different models for the prediction.