

Micro-Credit Defaulter Model

Submitted by:

RAVI PRAKASH BAJPAI

**ACKNOWLEDGMENT**

In this project data source is provided by FLIPROBO Technologies. For other references I have concern to many other websites,some of them are mentioned below.

<https://towardsdatascience.com>.

<https://www.geeksforgeeks.org/>

<https://www.analyticsvidhya.com/>

I have also concern to my mentors many times for help.

**INTRODUCTION**

This project deals with a Microfinance Institution (MFI) providing the services of Microfinance using the mobile financial services (MFS). These institutions generally works with families having low income living in the remote area. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

These institutions provided small amount of credit to their customers for short period of time. Loan amount is generally provided through mobile balance .suppose it provides loan credit of 5 rupiah then customer have to pay back 6 rupiah within 5 days. If customers are unable to pay back the amount within time limit that is 5 days then he will be considered as defaulter.

The sample data is provided for improving 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.

**Analytical Problem Framing**

Importing all the necessary libraries for preparing the model. Data set has been provided in csv file that is read thorugh read function.

Droping the unnamed column for the dataset.

As seen after applying shape function data set have 209395 rows and 36 columns,there is no null value present in the data set .

After applying describe method it can be seen that data is not properly distributed and it is not in normalized form. It can be seen that

* In many features standard deviation is more than mean value.
* There is large variation between the minimum value and first quantile range.
* There is large variation between the interquantile ranges.
* There is also large variation between the third quantile and maximum range.

From above obervations we can see that skewness is present in the data and data is not normally distributed.

Droping the pcircle column as it has only one value in the column and extracting the date and months from the date column.

**Data visualization and EDA**

Ploting all the features in distplot,from distplot it is observed that nearly in all the features skewness is present and data is not in normal form.

For rechecking the data skew funtion has been applied.considering as 8 as normal value we can see that in many features values in skewness is more than 8.its means that features having value more than 8 have skewness in the data. Following are the features having skewness values more than 8 .

aon 10.392949

last\_rech\_date\_ma 14.790974

last\_rech\_date\_da 14.814857

fr\_ma\_rech30 14.772833

medianmarechprebal30 14.779875

medianmarechprebal90 44.880503

cnt\_da\_rech30 17.818364

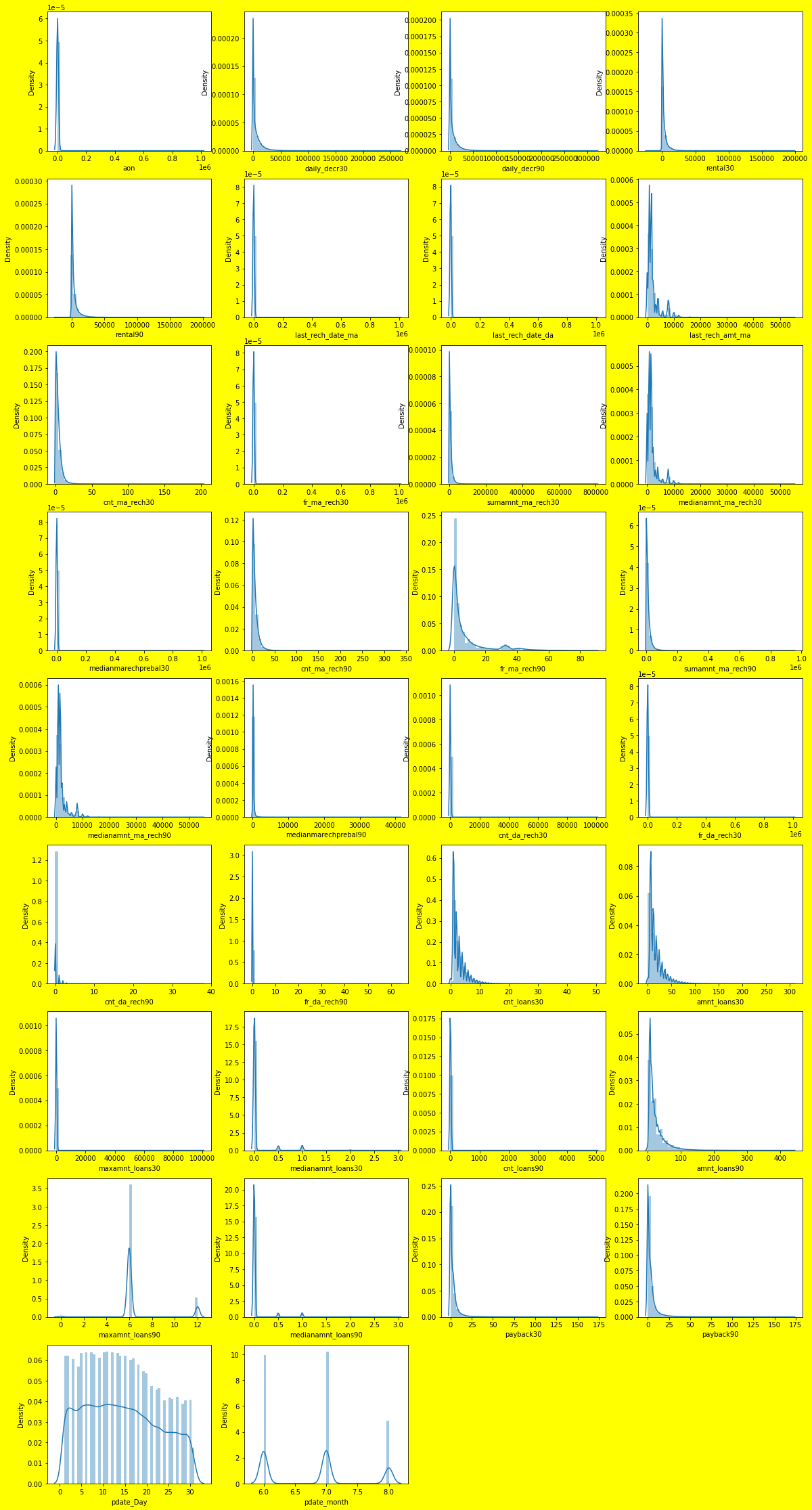
fr\_da\_rech30 14.776430

cnt\_da\_rech90 27.267278

fr\_da\_rech90 28.988083

maxamnt\_loans30 17.658052

cnt\_loans90 16.594408



**HEAT MAP**

Ploting the heat map of the given data set to find the multicolinearity in the given data set. From the given dataset following points are observed

* amnt\_loans30 and amnt\_loan90 are highly correlated with each other.
* medianamnt\_loans30 and medianamnt\_loans\_90 are corelated with each other.
* daily\_decr30 and daily\_decr90 are highly corelated with each other.
* rental30 and rental90 are highly corelated with each other.
* cnt\_ma\_rech30 and cnt\_ma\_rech90 are correlated with each other.
* cnt\_loans30 and amnt\_loans30 are correlated with each other.

So after the result of haet map some of features which are showing multicolinearity has been droped,these column are medianamnt\_loans90, daily\_decr90, rental30, cnt\_loans30 .

**Z score**

Applying z score for removal of skewness from the features of the data set. skewness has been removed as seen in results.

**Imbalance data**

From label count it can be seen clearly that label is not balanced. 1 count are 140497 and 0 count are only 22508.

**Standard scaler**

Spliting features and label and assigning them in different variables,then applying the standard scaler on the data set.

**Train test split**

Applying train test split on the scaled data, where for training the data 75 percent of whole data has been taken and for testing 25 percent data has been taken.

**Smote**

From value count function it can be clearly seen that data is imbalanced 1 count is more than 0 count, for balancing the data applying SMOTE function on data. After applying smote function we can see that 0 and 1 count are nearly equivalent.

Now data is ready for model applying different model to predict the defaulter customers.

**Model/s Development and Evaluation**

As it can be seen from label that given algorithm is classification type and it is binary classification. That is only 2 results can be seen 0 and 1. So applying different classification algorithms to predict the customers whether they are defaulter or not.

Algorithm used to predict the labels along with their classification report and results

**LOGISTIC REGRESSION**

Accuracy score- 74 %

Classification report

precision recall f1-score support

0 0.32 0.79 0.46 5641

1 0.96 0.74 0.83 35111

accuracy 0.74 40752

macro avg 0.64 0.76 0.65 40752

weighted avg 0.87 0.74 0.78 40752

**DECISION TREE**

Accuracy score- 86 %

**KNN CLASSIFIER**

Accuracy score – 77%

Classification report

precision recall f1-score support

0 0.35 0.73 0.47 5641

1 0.95 0.78 0.85 35111

accuracy 0.77 40752

macro avg 0.65 0.75 0.66 40752

weighted avg 0.86 0.77 0.80 40752

**RANDOM FOREST CLASSIFIER**

Accuracy score – 90%

Classification report

precision recall f1-score support

0 0.65 0.65 0.65 5641

1 0.94 0.94 0.94 35111

accuracy 0.90 40752

macro avg 0.80 0.80 0.80 40752

weighted avg 0.90 0.90 0.90 40752

After looking the accuracy score of all the algorithms, **RANDOM FOREST CLASSIFIER** is best algorithm for given problem. It is showing highest accuracy score that 90 % from the other algorithm.

Hyperparameter tuning

Applying hyperparameter tuning to improve the result as maximum as possible. selecting random forest classifier for hyperparameter tuning. Applying grid search cv for selecting best paramaters. Taking the best parameters and applying for Hyperparameter tuning.

From the results random forest classifier is giving accuracy of 86% which is even less than the normal parameters.(tried different probabilies to improve the results by changing different values for the different parameter but maximum is 86 % )

**CONCLUSION**

So considering the random forest classifier as algorithm for the given problem without taking the best parameter of grid search cv .