



# **Micro Credit Project**

Submitted by:

**ASHUTOSH MISHRA**

## **ACKNOWLEDGMENT**

I'd like to extend my gratitude to my mentor Mr. Md. Kashif for giving me this opportunity to work upon this project. Below are all the details of the project which I consumed while preparing and drafting the project –

The references, research papers, data sources, professionals, etc are majorly referred from “Flip Robo Technologies” study collaterals and data repository. Also, some of the other resources like “Research Gate” were explored to gain a deep understanding on the assigned subject.

Above stated details stands correct to the best of my knowledge and I hereby acknowledge the same.

## INTRODUCTION

Microfinance institutions (MFIs) are financial companies that provide small loans to people who do not have any access to banking facilities. The definition of **small loans** varies between countries. In India, all loans that are below Rs.1 lakh can be considered as microloans.

Microfinance institutions have been gaining popularity in the recent years and are now considered as effective tools for alleviating poverty. Most MFIs are well-run with great track records, while others are quite self-sufficient. The primary goals of microfinance institutions are the following:

- ✚ Transform into a financial institution that assists in the development of communities that are sustainable.
- ✚ Help in the provision of resources that offer support to the lower sections of the society. There is special focus on women in this regard, as they have emerged successful in setting up income generation enterprises.
- ✚ Evaluate the options available to help eradicate poverty at a faster rate.
- ✚ Mobilize self-employment opportunities for the underprivileged.
- ✚ Empowering rural people by training them in simple skills so that they are capable of setting up income generation businesses.

As per World Bank data, close to 1.7 billion people across multiple countries do not have access to basic financial services. This is where microfinance institutions play a major role.

## Business Problem Framing

See, now a day's most of the people are living below the poverty line or unemployed and out of it most of them don't have any access to any banking facilities; and in-order to fulfil the same, it creates a big opportunity amongst the Micro Finance companies.

We are working with one such client that is an **Indonesian 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.

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. 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).

Today, microfinance is widely accepted as a poverty-reduction tool, representing \$70 billion in outstanding loans and a global outreach of 200 million clients. Understanding and interpreting the same in a mathematical and graphical form can help an individual or a firm to invest in a planned manner based on mathematical statistics.

## Conceptual Background of the Domain Problem

Data science comes as a vital tool to solve business problems to help companies optimize the standards resulting in overall revenue and profits growth. Moreover, it also improves their marketing strategies and demands focus on changing trends in Microfinance Institutions (MFI).

Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques globally used for achieving the business goals for Microfinance Institutions (MFI).

Currently the assigned project utilizes **Predictive Modelling** algorithm to solve the business statement.

## Review of Literature

The research project was for a client that is in **Telecom Industry** seeking to do collaboration 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.

Being a Data Scientist, I have used the **Classification Model** using multiple algorithms to design and optimize the results.

Some of the classes which I explored from Scikit-learn Libraries were – Statistics, Analytical Modelling, Hyper Tuning Method, CV Ratings, Predictive Modelling and Regression Model, etc.

In totality, my project comprises of seven different test cases of Classification models where the objective was to train the model and test for accuracy score and different classification reports using AUC ROC Curve and Confusion Matrix to best understand the accuracy of the algorithms.

Data provided were – X\_train (variables having direct impact on label) and Y\_train (label). In the beginning, I was provided with 37, X\_train variable which I condensed to 5 columns keeping in mind the Principle Component Analysis. This resulted in condensed data being trained in a much shorter span of time with utmost accuracy.

Also, I have observed that there are no null values present in the dataset given by clients, there are also some customer who doesn't have any loan history and the target class is imbalanced in nature.

As a result of the above analytical modelling, I have managed to achieve below top three scenarios, which are given below as following—

Regression Models	Testing Accuracy (in %)
Random Forest Classifier	90
K Neighbors Classifier	89
Decision Tree Classifier	85

## Motivation for the Problem Undertaken

See, this project somehow remind me the companies that I've worked with. I'm talking out two prestigious giants in the banking industry i.e. **ICICI Bank** and **Yes Bank**. I remember the day when a customer visited the branch for some loans and other financial queries and due to no earners in their family bank refused them to proceed the loan and as a human being it hurts when you can't help that person who needs it more at that time. Also, I referred them to NFBC's but their rates were too high that's why they didn't take the loan.






Hence, this project guided me to baseline each aspect carefully and be concrete on the decision making process regardless it's an individual or an entity like a company.

In order to cater to the above project, my current knowledge and skill set has aided me a lot which I'd explore on this exponentially further.





## Analytical Problem Framing

**Mathematical Modeling-** I have used the below statistical models to find-out the corresponding mean, median, mode and relationship among the variables.

-  Descriptive Statistics- To find out the mean, median, mode, percentiles and IQR (Interquartile Range)
-  Statistical Modeling, Correlation- To find out the relationship among the variables i.e. finding out the positive, negative and zero correlated attributes.
-  Multicollinearity- To find out all those variables who are giving similar information to the target variable ('label')
-  Skewness- To check whether the attributes are the skewing left hand side or right hand side (Threshold value is=0.5).
-  Outliers- To find out variables those are having the value greater than 3.

### Analytical Modeling-

-  Label Encoder- To convert all the categorical columns into numeric categories
-  Simple Imputer-To replace all the null values present in the columns with mean or mode.

## Data Sources and their formats

The data is been given by **Indonesian Telecom Company**, they had given the data in CSV file, with data description file in excel format. Below are the attached screenshot for your reference; so that you can check out the insights that were available in the CSV file.

```
In [2]: df=pd.read_csv('Micro Credit Project.csv')
df
```

Out[2]:

	Unnamed: 0	label	msisdn	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	...	maxamnt_loans30
0	1	0	2140870789	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	...	6.0
1	2	1	7646270374	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	...	12.0
2	3	1	1794370372	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	...	6.0
3	4	1	5577370781	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	...	6.0
4	5	1	03813182730	947.0	150.619333	150.619333	1098.90	1098.90	4.0	0.0	...	6.0
...	...	...	...	...	...	...	...	...	...	...	...	...
209588	209589	1	22758185348	404.0	151.872333	151.872333	1089.19	1089.19	1.0	0.0	...	6.0
209589	209590	1	95583184455	1075.0	36.936000	36.936000	1728.36	1728.36	4.0	0.0	...	6.0
209590	209591	1	28556185350	1013.0	11843.111667	11904.350000	5861.83	8893.20	3.0	0.0	...	12.0
209591	209592	1	59712182733	1732.0	12488.228333	12574.370000	411.83	984.58	2.0	38.0	...	12.0
209592	209593	1	65061185339	1581.0	4489.362000	4534.820000	483.92	631.20	13.0	0.0	...	12.0

209593 rows × 37 columns

```
In [10]: #numeric dataframe
df_numeric= df.select_dtypes(include='number')
df_numeric
```

Out[10]:

	Unnamed: 0	label	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	last_rech_amt_ma	cnt_ma_rech30
0	1	0	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	1539	2
1	2	1	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	5787	1
2	3	1	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	1539	1
3	4	1	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	947	0
4	5	1	947.0	150.619333	150.619333	1098.90	1098.90	4.0	0.0	2309	7
...	...	...	...	...	...	...	...	...	...	...	...
209588	209589	1	404.0	151.872333	151.872333	1089.19	1089.19	1.0	0.0	4048	3
209589	209590	1	1075.0	36.936000	36.936000	1728.36	1728.36	4.0	0.0	773	4
209590	209591	1	1013.0	11843.111667	11904.350000	5861.83	8893.20	3.0	0.0	1539	5
209591	209592	1	1732.0	12488.228333	12574.370000	411.83	984.58	2.0	38.0	773	5
209592	209593	1	1581.0	4489.362000	4534.820000	483.92	631.20	13.0	0.0	7526	2

209593 rows × 34 columns

```
In [11]: #categorical dataframe  
df_categorical= df.select_dtypes(include='object')  
df_categorical
```

Out[11]:

	msisdn	pcircle	pdate
0	21408170789	UPW	2016-07-20
1	76462170374	UPW	2016-08-10
2	17943170372	UPW	2016-08-19
3	55773170781	UPW	2016-06-06
4	03813182730	UPW	2016-06-22
...	...	...	...
209588	22758185348	UPW	2016-06-17
209589	95583184455	UPW	2016-06-12
209590	28556185350	UPW	2016-07-29
209591	59712182733	UPW	2016-07-25
209592	65061185339	UPW	2016-07-07

209593 rows × 3 columns

```

0  Unnamed: 0          209593 non-null int64
1  label               209593 non-null int64
2  msisdn              209593 non-null object
3  aon                 209593 non-null float64
4  daily_decr30        209593 non-null float64
5  daily_decr90        209593 non-null float64
6  rental30            209593 non-null float64
7  rental90            209593 non-null float64
8  last_rech_date_ma   209593 non-null float64
9  last_rech_date_da   209593 non-null float64
10 last_rech_amt_ma    209593 non-null int64
11 cnt_ma_rech30       209593 non-null int64
12 fr_ma_rech30        209593 non-null float64
13 sumamnt_ma_rech30   209593 non-null float64
14 medianamnt_ma_rech30 209593 non-null float64
15 medianmarechprebal30 209593 non-null float64
16 cnt_ma_rech90       209593 non-null int64
17 fr_ma_rech90        209593 non-null int64
18 sumamnt_ma_rech90   209593 non-null int64
19 medianamnt_ma_rech90 209593 non-null float64
20 medianmarechprebal90 209593 non-null float64
21 cnt_da_rech30       209593 non-null float64
22 fr_da_rech30        209593 non-null float64
23 cnt_da_rech90       209593 non-null int64
24 fr_da_rech90        209593 non-null int64
25 cnt_loans30         209593 non-null int64
26 amnt_loans30        209593 non-null int64
27 maxamnt_loans30     209593 non-null float64
28 medianamnt_loans30  209593 non-null float64
29 cnt_loans90         209593 non-null float64
30 amnt_loans90        209593 non-null int64
31 maxamnt_loans90     209593 non-null int64
32 medianamnt_loans90  209593 non-null float64
33 payback30          209593 non-null float64
34 payback90          209593 non-null float64
35 pcircle             209593 non-null object
36 pdate              209593 non-null object
dtypes: float64(21), int64(13), object(3)
memory usage: 59.2+ MB

```

```
In [19]: df.nunique()
```

```
Out[19]: label                2
msisdn          186243
aon              4507
daily_decr30     147025
daily_decr90     158669
rental30         132148
rental90         141033
last_rech_date_ma  1186
last_rech_date_da  1174
last_rech_amt_ma   70
cnt_ma_rech30      71
fr_ma_rech30       1083
sumamnt_ma_rech30  15141
medianamnt_ma_rech30  510
medianmarechprebal30 30428
cnt_ma_rech90      110
fr_ma_rech90       89
sumamnt_ma_rech90  31771
medianamnt_ma_rech90  608
medianmarechprebal90 29785
cnt_da_rech30      1066
fr_da_rech30       1072
cnt_da_rech90       27
fr_da_rech90        46
cnt_loans30         40
amnt_loans30         48
maxamnt_loans30     1050
medianamnt_loans30    6
cnt_loans90         1110
amnt_loans90         69
maxamnt_loans90       3
medianamnt_loans90    6
payback30          1363
payback90          2381
pcircle            1
pdate              82
dtype: int64
```

## Data Preprocessing Done

I've dropped the attribute- 'unnamed: 0' as index is already there and this attribute is not making any sense there in the data frame, so I've deleted this column. The attribute 'fr\_da\_rech90' which represents the **Frequency of data account recharged in last 90 days**; indicates that there is zero recharge happened in last 90 days for almost 99.58% of the times and three recharges happened for 0.03% of the times. Hence since the occurrence of zero for this attributes is almost 100% hence have dropped this column. Also the attribute 'cnt\_da\_rech90' which represent **Number of times data account got recharged in last 90 days**; indicates that there is zero recharge happened in last 90 days for 97% of the times and one recharges happened for 2% of the times. Hence since the occurrence of zero for this attributes is very high as compare to other classes hence have dropped this column too.

Note: I've deleted all the above attributes keeping in the mind that later in the model building process these attributes can create biasness. Also, used Label Encoder to convert all the categorical variables value into numeric form.

Remove all the zero correlated variables ('daily\_decr30', 'amnt\_loans30', 'payback90', 'rental30', 'maxamnt\_loans30', 'fr\_ma\_rech30', 'last\_rech\_date\_da', 'msisdn', 'last\_rech\_date\_ma', 'cnt\_da\_rech30', 'cnt\_loans90', 'medianmarechprebal30', 'aon', 'fr\_da\_rech30') w.r.t. target variable label.

Dropped all the feature variables those were giving the same amount of information to the target variable ['cnt\_loans30', 'cnt\_ma\_rech90'].

Since data is expensive and we can't lose more than 7-8% of the data but here after removing outliers I'm losing over more than 19% data which is against the instructions of the client hence will not consider the outliers removal for the same.

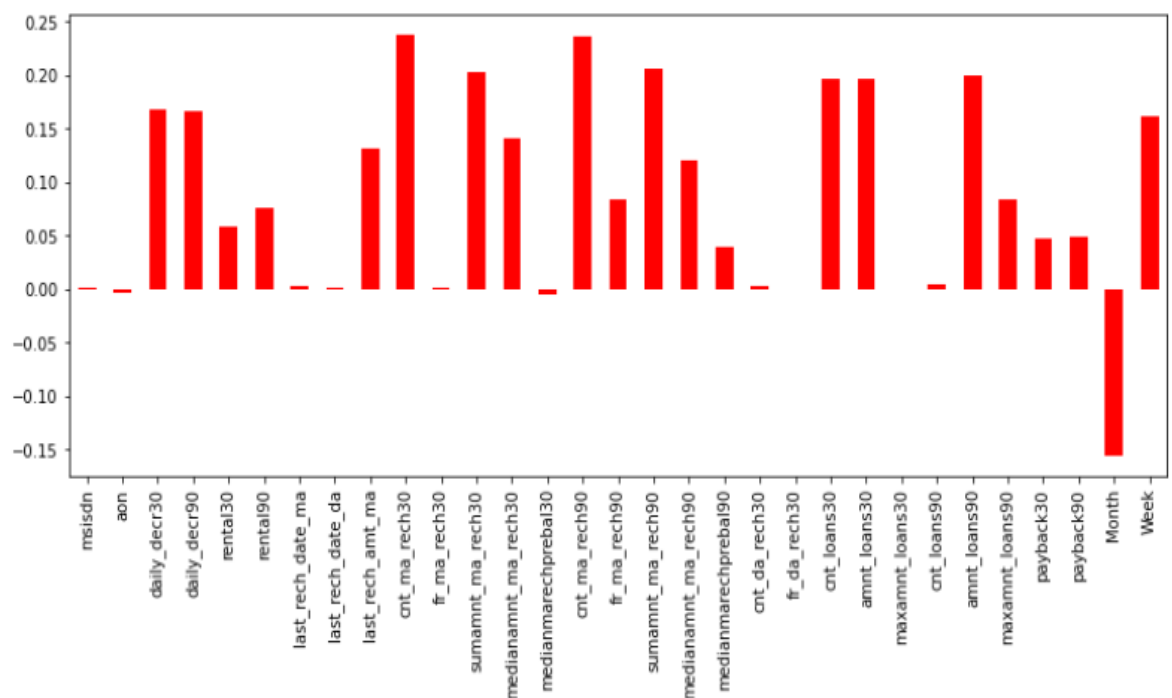
I've use Power Transformer method to remove the skewness and Standard Scaler technique to normalize the feature variable and later on I've used Principle component analysis and converted all the variables into only **Five Principle Components**.

## Data Inputs- Logic- Output Relationships

As the data given to us in CSV format and client has described it well that he need to predict the defaulters, those who are paying the amount within 5 days or not so it's very clear to us that we have to predict the binary values; which is Label '1' that indicates that the loan has been paid i.e. Non- defaulter, while, Label '0' which indicates that the loan has not been paid i.e. defaulter. So on the basis of output I have checked the correlational input data with my output, as shown in below figure.

### Multicollinearity

```
in [124]: plt.figure(figsize=(13,5))
df.corr().label.drop(['label']).plot(kind='bar',color='r')
plt.show()
```



As I checked the input and found that almost all the data is quite positively correlated with my output data, except few columns data. Last recharge date is highly negatively correlated with my output.



**State the set of assumptions (if any) related to the problem under consideration**

No.

## Hardware and Software Requirements and Tools Used

I have used **Python IDE** (Integrated Development environment) as a dedicated software throughout solving this project.

```
import numpy as np
import pandas as pd
import scipy.stats
from scipy.stats import zscore, boxcox
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

### Python Libraries that I've used throughout the process are-

- ✚ Numpy- It is use for linear algebra
- ✚ Pandas- data analysis/manipulation library o
- ✚ Scipy- Utility function for optimization o
- ✚ Matplotlib-Data visualization and plotting library o
- ✚ Seaborn- Data visualization and statistical plotting library
- ✚ Sklearn- Machine Learning Tool
- ✚ Imblearn- Deal with classification problems of Imbalanced classes o
- ✚ Statsmodels-Deal with advanced statistics

### Classes-

- ✚ Label Encoder- Encoding the categorical variables into number category
- ✚ Simple Imputer- Replacing the null values with mean, median or mode
- ✚ Variance\_Inflation\_Factor- Calculate Multicollinearity
- ✚ Power Transformer- Remove skewness
- ✚ Standard Scaler- Normalize the feature variables
- ✚ Principle Component Analysis- Reduce the dimension of the data frame
- ✚ Cross\_Val\_Score- CV score
- ✚ Grid Search CV- Find out the best parameters for the model

## Model/s Development and Evaluation

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

#### Statistical Method-

When I used the describe function then I find out that attribute **label** has more median than its mean and so it indicates that there is the possibility of left skewed data in the dataset and the interquartile range b/w the variables are varying too much hence it shows that datasets are skewing left hand side and it indicates that the variables are not normally distributed.

Also, I have used correlation method to check what are the variables that are giving strong correlation w.r.t Target variable 'label'.

#### Analytical Method-

I've uses Boxplot, Scatter Plots and Distribution Plots to check the outliers and skewness of the variables respectively through the plotting's.

## Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

1. `from sklearn.linear_model import LogisticRegression`
2. `from sklearn.naive_bayes import GaussianNB`
3. `from sklearn.tree import DecisionTreeClassifier`
4. `from sklearn.neighbors import KNeighborsClassifier`
5. `from sklearn.ensemble import RandomForestClassifier`
6. `from sklearn.ensemble import AdaBoostClassifier`
7. `from sklearn.ensemble import GradientBoostingClassifier`

## Run and Evaluate selected models

### 1st Best Model (Random Forest Classifier)

---

```
For model RandomForestClassifier(class_weight='balanced', criterion='log_loss',
                                max_features='log2')
Training_Accuracy_Score= 0.9989675933339921
Testing_Accuracy_Score= 0.9028798059231598
```

```
Classification Report-
              precision    recall  f1-score   support

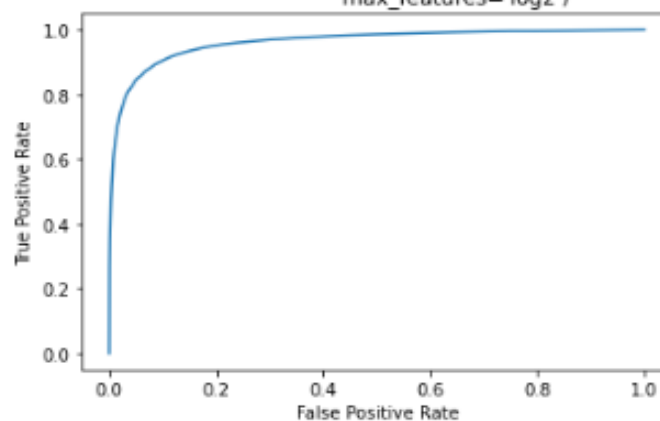
     0       0.89         0.92         0.90         36646
     1       0.92         0.88         0.90         36727

 accuracy          0.90
 macro avg         0.90         0.90         0.90         73373
weighted avg         0.90         0.90         0.90         73373
```

```
Confusion Metrix-
[[33772  2874]
 [ 4252 32475]]
```

AUC\_ROC CURVE

ROC\_Curve for the model RandomForestClassifier(class\_weight='balanced', criterion='log\_loss', max\_features='log2')



ROC AUC SCORE is- 0.9029004205655634

## 2nd Best Model (K Neighbors Classifier)

For model KNeighborsClassifier(algorithm='ball\_tree', weights='distance')  
 Training\_Accuracy\_Score= 0.9991515866012014  
 Testing\_Accuracy\_Score= 0.8899867798781568

Classification Report-

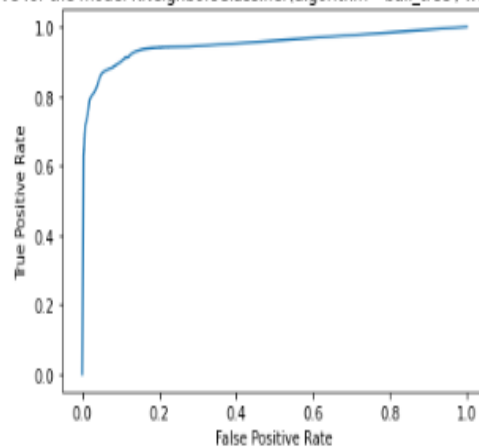
	precision	recall	f1-score	support
0	0.84	0.97	0.90	36646
1	0.96	0.81	0.88	36727
accuracy			0.89	73373
macro avg	0.90	0.89	0.89	73373
weighted avg	0.90	0.89	0.89	73373

Confusion Matrix-

```
[[35543 1103]
 [ 6969 29758]]
```

AUC\_ROC CURVE

ROC\_Curve for the model KNeighborsClassifier(algorithm='ball\_tree', weights='distance')



ROC AUC SCORE is- 0.8900749040022465

Finding out the best K-Fold Value

At the K-Fold 2 the CV score of model KNeighborsClassifier(algorithm='ball\_tree', weights='distance') is 0.8741706690799265

## 3rd Best Model (Decision Tree Classifier)

For model DecisionTreeClassifier(criterion='entropy', max\_features='log2')  
 Training\_Accuracy\_Score= 0.9989675933339921  
 Testing\_Accuracy\_Score= 0.8541425319940578

Classification Report-

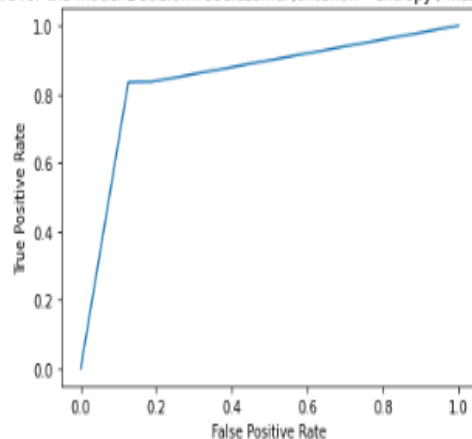
	precision	recall	f1-score	support
0	0.84	0.87	0.86	36646
1	0.87	0.83	0.85	36727
accuracy			0.85	73373
macro avg	0.85	0.85	0.85	73373
weighted avg	0.85	0.85	0.85	73373

Confusion Matrix-

```
[[32026 4620]
 [ 6082 30645]]
```

AUC\_ROC CURVE

ROC\_Curve for the model DecisionTreeClassifier(criterion='entropy', max\_features='log2')



ROC AUC SCORE is- 0.8541643510807191

Finding out the best K-Fold Value

At the K-Fold 2 the CV score of model DecisionTreeClassifier(criterion='entropy', max\_features='log2') is 0.8354340324154587

## Key Metrics for success in solving problem under consideration

Below is the best **Classification Models** where I used below metrics -

Regression Models	Testing Accuracy (in%)
Random Forest Classifier	90
K Neighbors Classifier	89
Decision Tree Classifier	85

- ✚ CV Score – Model testing Accuracy
- ✚ Hyper Tuning Method – Best parameter for the respective models
- ✚ Principle Component Analysis – Course of reduction of dimensions, etc

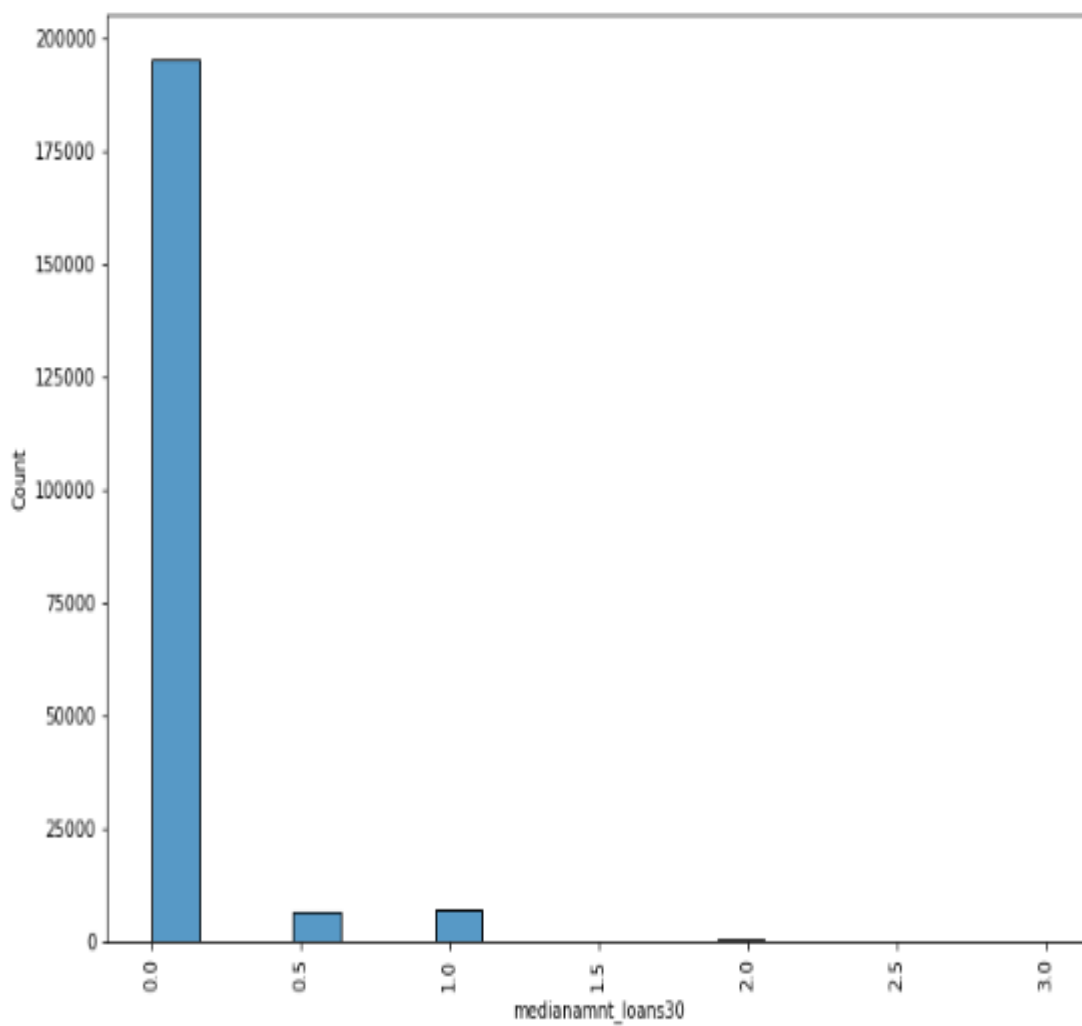


## Visualizations

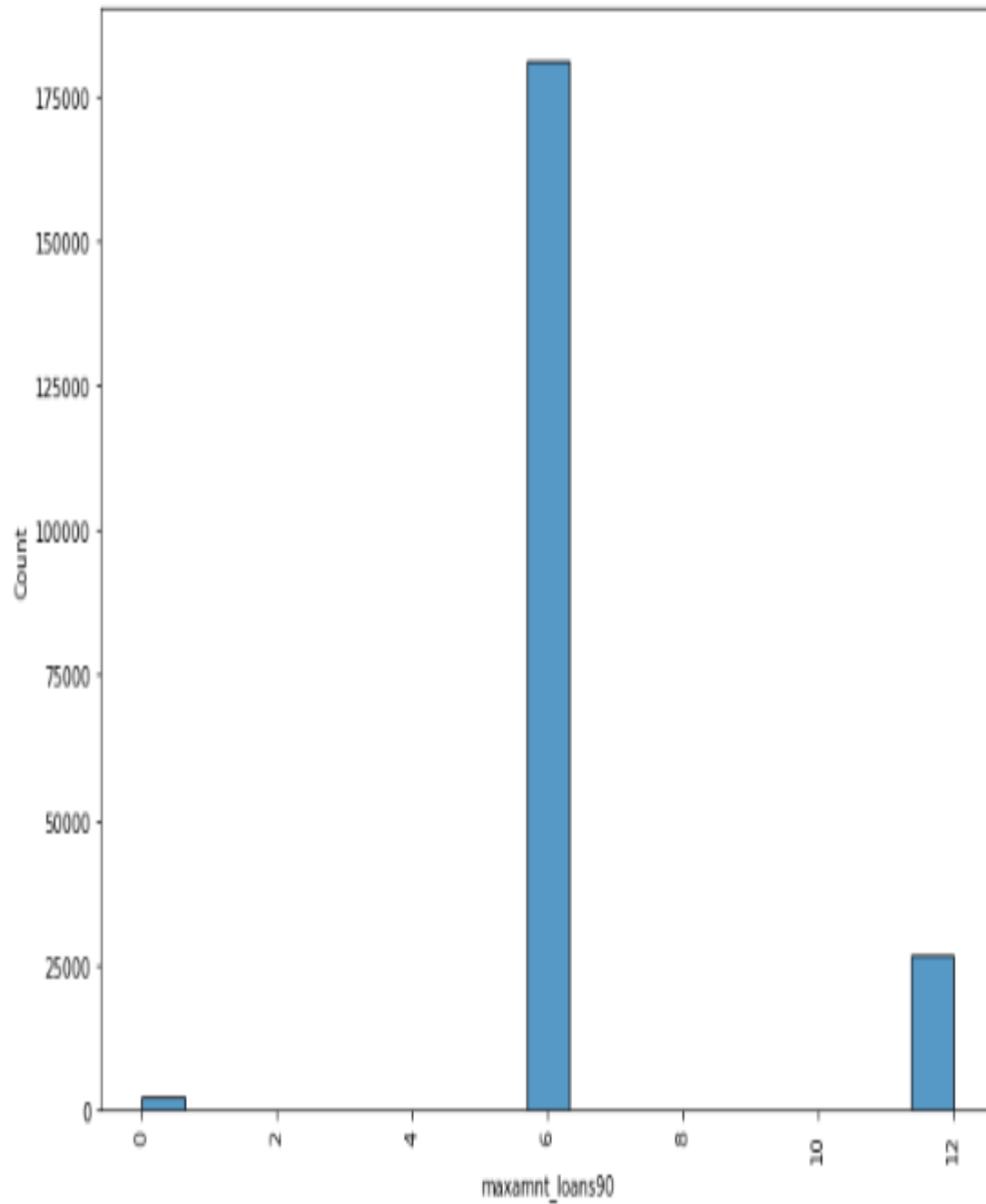
The plots are –

- Histograms
- Count Plot
- Scatter Plot
- Distribution plot, box plot, etc.

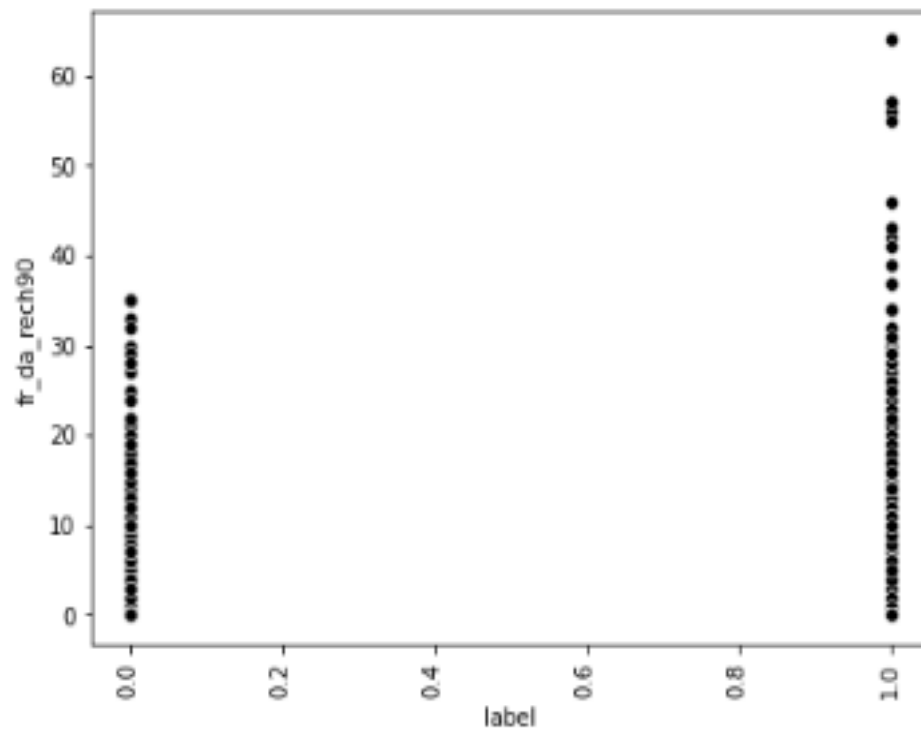
The Histplot Diagram for the attribute "medianamnt\_loans30" is  
`AxesSubplot(0.125,0.125;0.775x0.755)`



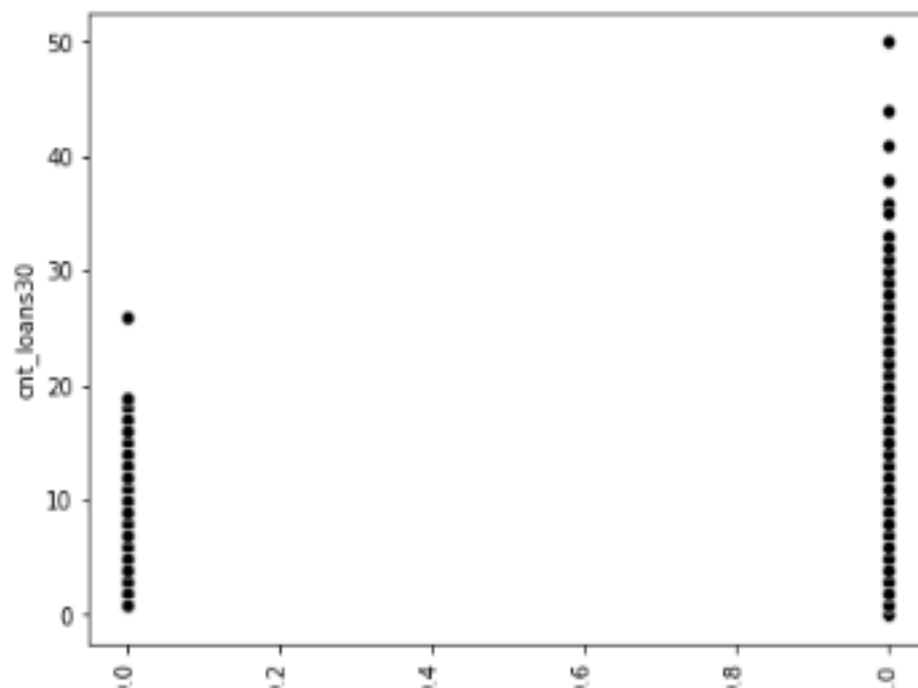
The Histplot Diagram for the attribute "maxamnt\_loans90" is  
AxesSubplot(0.125,0.125;0.775x0.755)



The Scatter Plot for the attribute "label" & "fr\_da\_rech90" is-  
AxesSubplot(0.125,0.125;0.775x0.755)



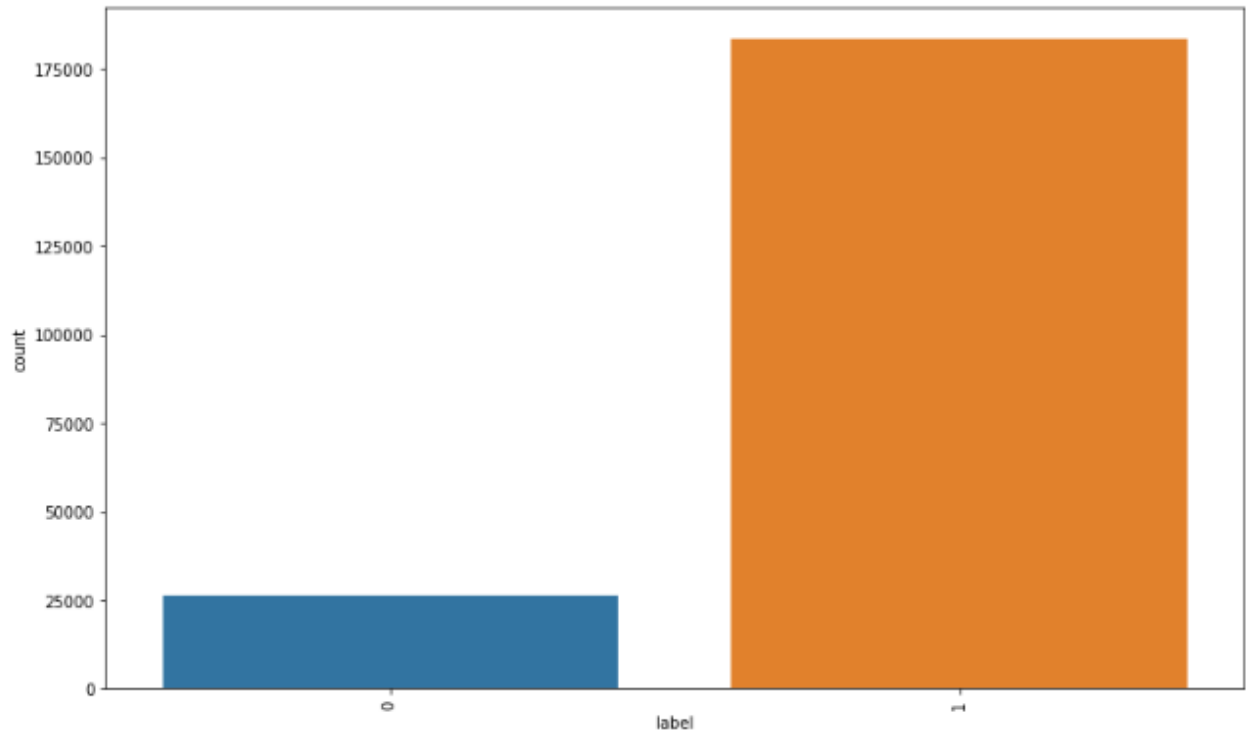
The Scatter Plot for the attribute "label" & "cnt\_loans30" is-  
AxesSubplot(0.125,0.125;0.775x0.755)



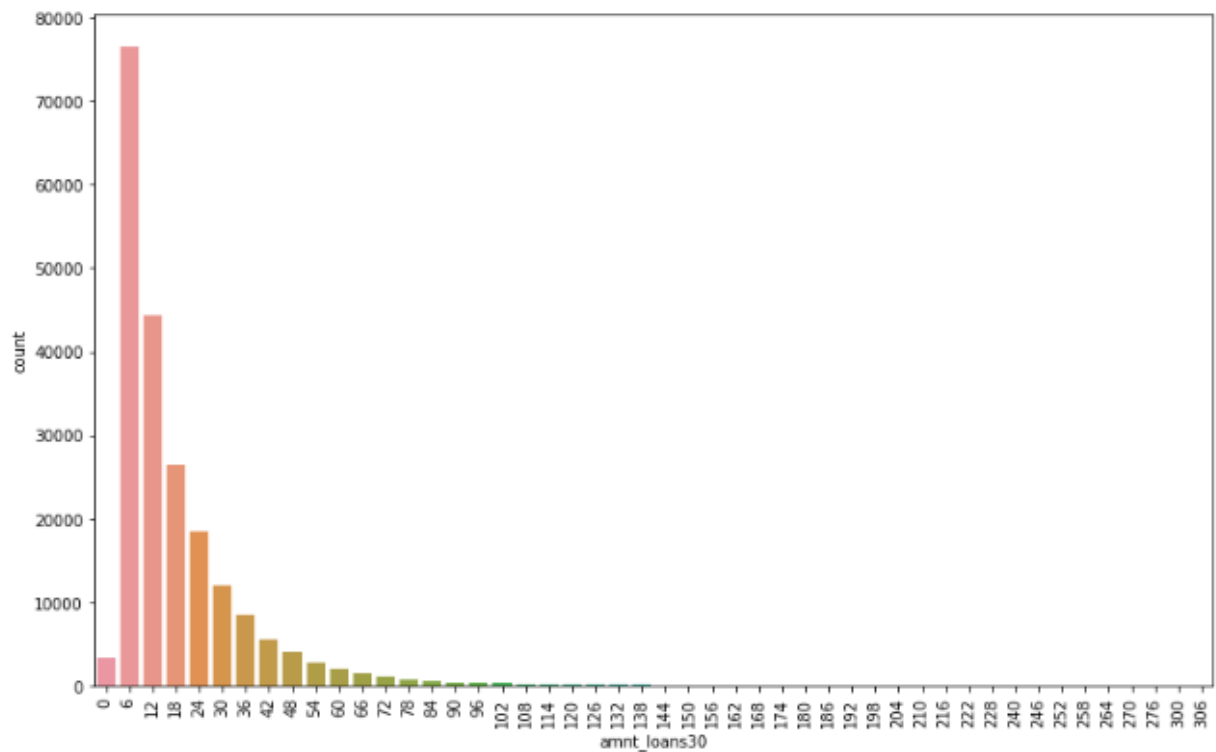
The Value Counts for the attribute "label" is

```
1    183431
0     26162
Name: label, dtype: int64
```

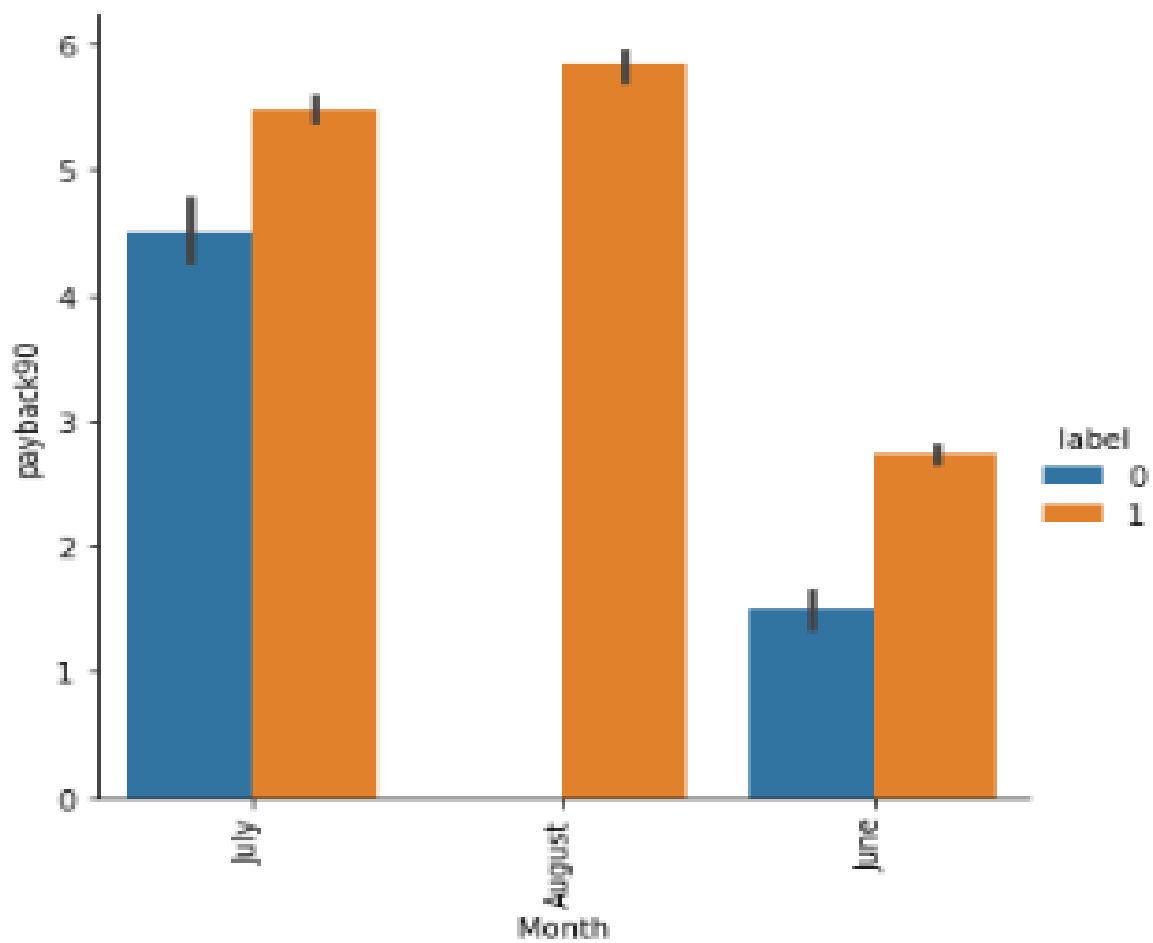
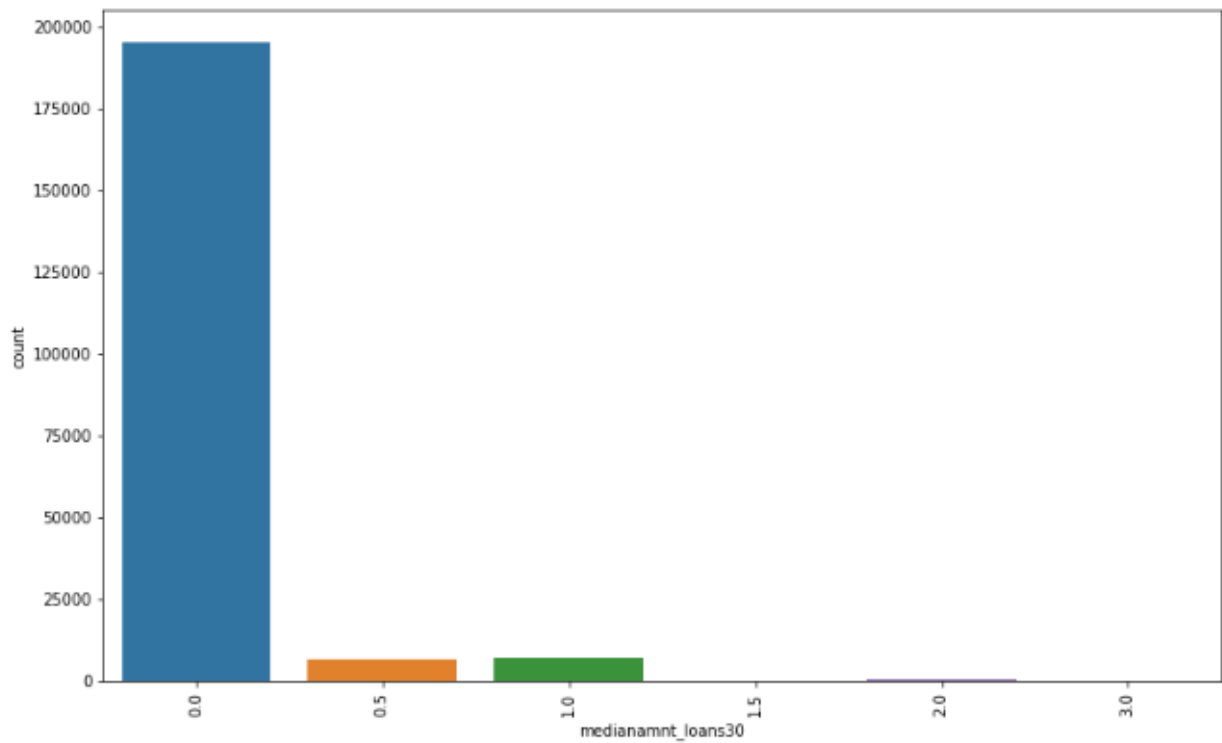
The Countplot Diagram for the attribute "label" is  
AxesSubplot(0.125,0.125;0.775x0.755)

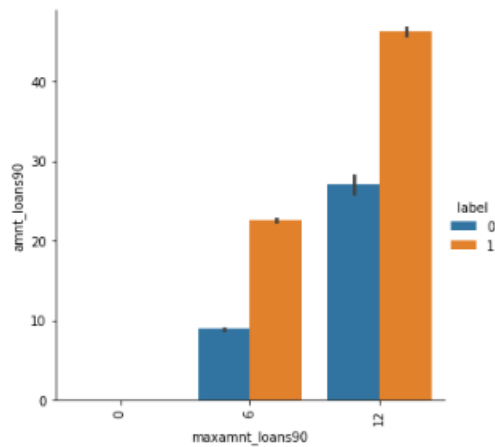


The Countplot Diagram for the attribute "amnt\_loans30" is  
AxesSubplot(0.125,0.125;0.775x0.755)



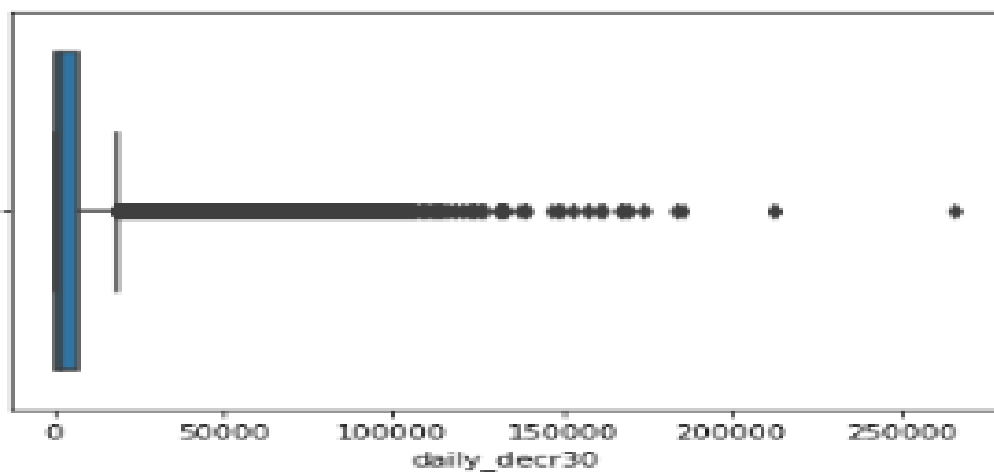
The Countplot Diagram for the attribute "medianamnt\_loans30" is  
AxesSubplot(0.125,0.125;0.775x0.755)



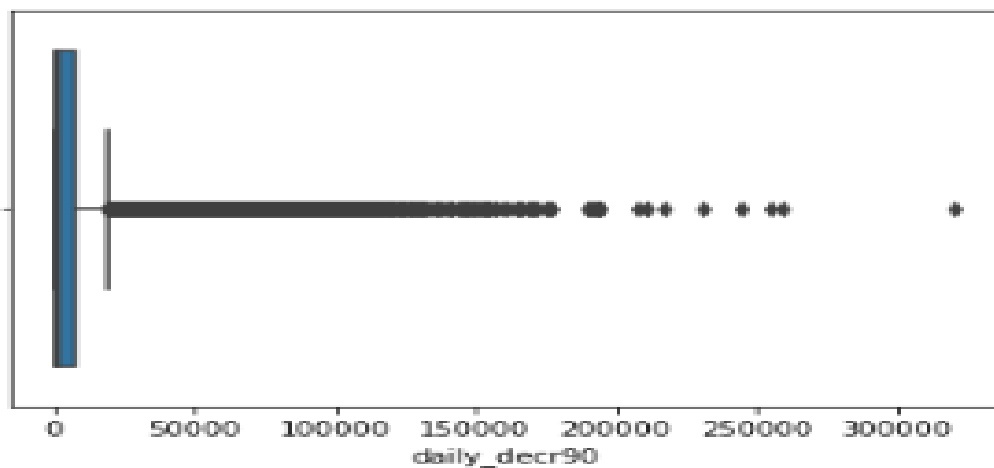


Maximum amount of loan taken by the customers in last 90 days are 0, 6 or 12 Rupiah. Total amount of loans taken by customers having maximum amount loan 12 Rupiah has higher total amount loan than that of the customers having maximum amount loan of 6 Rupiah in last 90 days.

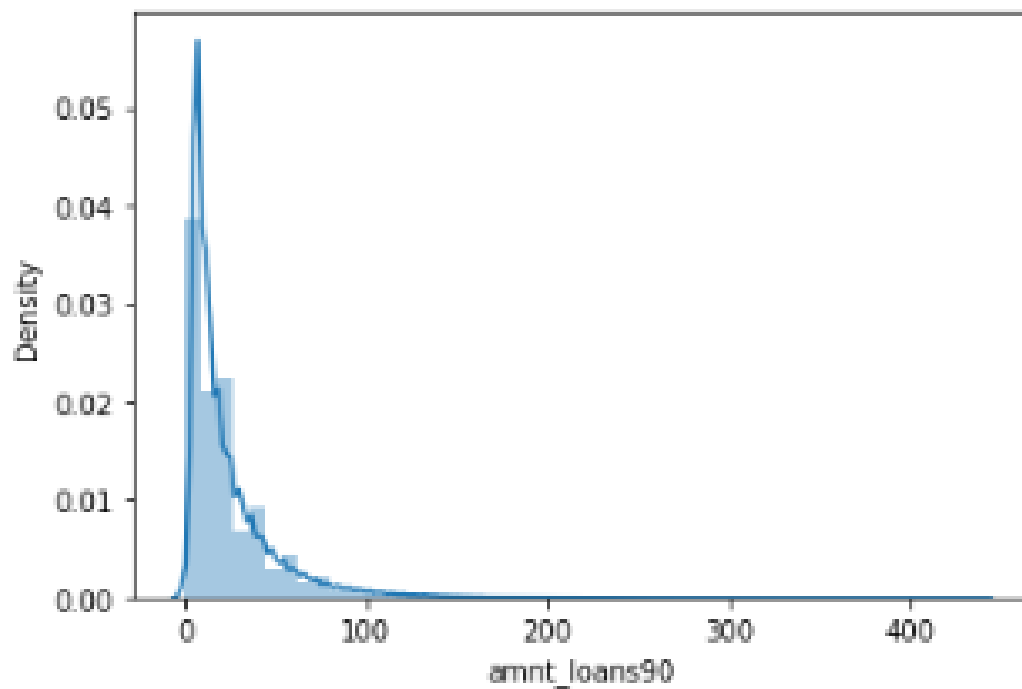
The Box-Plot for attribute "daily\_decr30" is -  
`AxesSubplot(0.125,0.125;0.775x0.755)`



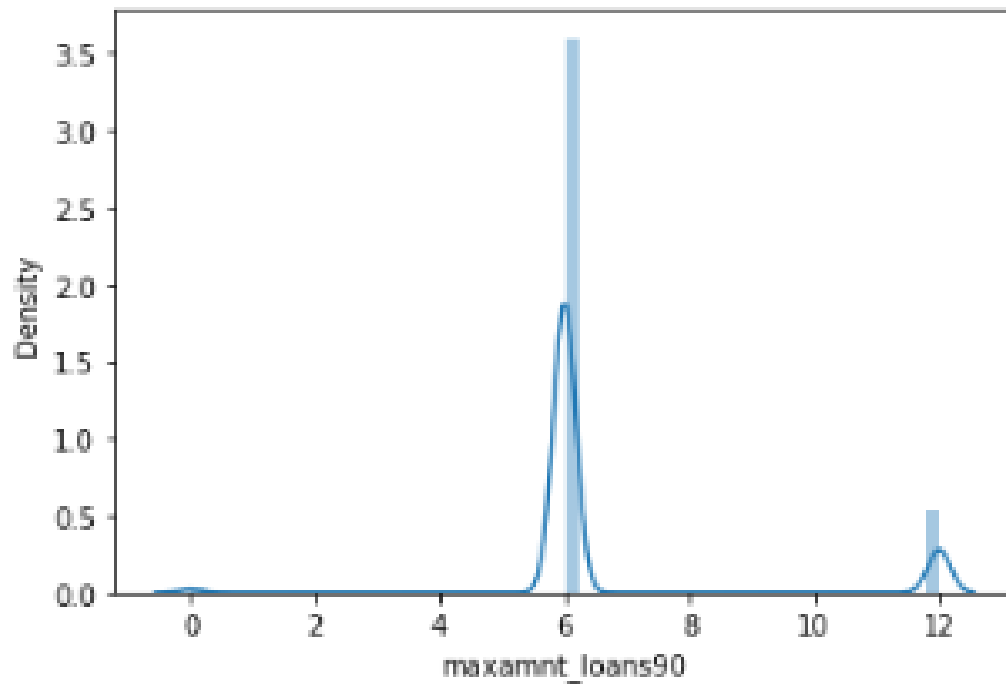
The Box-Plot for attribute "daily\_decr90" is -  
`AxesSubplot(0.125,0.125;0.775x0.755)`



The Distribution Plot for attribute "amnt\_loans90" is-  
AxesSubplot(0.125,0.125;0.775x0.755)



The Distribution Plot for attribute "maxamnt\_loans90" is-  
AxesSubplot(0.125,0.125;0.775x0.755)



## Interpretation of the Results

Below are some of the point's w.r.t. visualization, pre-processing presented above -

- ✚ As we can see that 87% of the customers have been paid their loan within 5 days of insurance of loan i.e. are Non- defaulters, while, 12% of the customers have not paid their loans back within 5 days of insurance of loan i.e. are defaulters.
- ✚ 40% of the customers have taken only one loan in last 30 days.
- ✚ 37% of the customers have taken Total amount of loans as 6 Rupiah in last 30 days.
- ✚ Median of amounts of loan taken by 93% of the customers in last 30 days is Zero and it means almost each entry of this columns is zero ergo will drop this column otherwise it will create biasness in the model.
- ✚ Maximum amount of loan taken by the 87% of the customers in last 90 days is 6 Rupiah and the same of 13% of the customers is 12 Rupiah.
- ✚ Median of amounts of loan taken by 94% of the customers in last 90 days is Zero Rupiah and it means almost each entry of this columns is zero ergo will drop this column otherwise it will create biasness too in the model.
- ✚ Maximum amount of loan taken by the customers in last 90 days are 0, 6 or 12 Rupiah. Total amount of loans taken by customers having maximum amount loan 12 Rupiah has higher total amount loan than that of the customers having maximum amount loan of 6 Rupiah in last 90 days.
- ✚ Those Customer using mobile devices since more than 15000 days have taken maximum amount of loan in last 90 days as 6 Rupiah and the Median of amounts of loan taken by the them in the last 90 days is 2 Rupiah
- ✚ Customers those have Median of amounts of loan as 1.5 or more than 1.5 Rupiah are not defaulters i.e. they use to pay back their loans on time when their Average main account balance over last 30 days is approx. 4000 Rupiah, if the Median of amounts of loan as 1.5



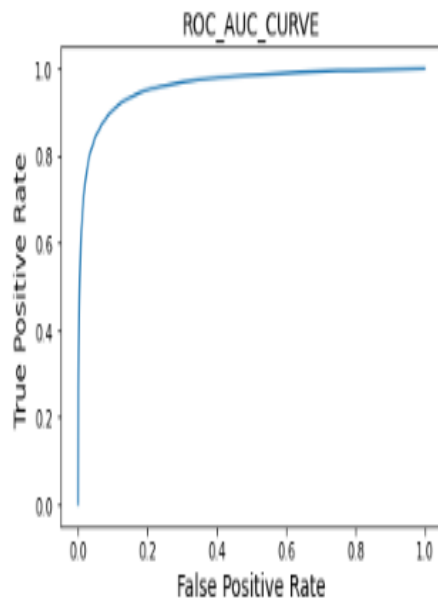
- ✚ In-case customer is taking up maximum amount of loan as 12 Rupiah the probability of returning the loan back within 5 days is 100% when Number of times main account got recharged in last 90 days is more than 10 times.
- ✚ In-case customer is taking up maximum amount of loan as 12 Rupiah then there Average payback time is 8 days over last 90 days
- ✚ As we can see that out of three months July and June are leading w.r.t. transactions
- ✚ In the month August there are no defaulters and there Average payback time in days is almost 6 over last 90 days, whereas in the month of June and July there are Defaulters present with the Average payback time for July in days is almost 5 and for June is approx. 2 days over last 90 days.

## CONCLUSION

:

	Models	Traning_Accuracy_Score_%	Testing_Accuracy_Score_%	CV_SCORE_%	ROC_AUC_SCORE
0	LogisticRegression	73	73	73	0.73
1	GaussianNB	71	71	71	0.71
2	KNeighborsClassifier	100	89	86	0.89
3	DecisionTreeClassifier	100	85	84	0.85
4	RandomForestClassifier	100	90	89	0.90
5	AdaBoostClassifier	77	77	77	0.77
6	GradientBoostingClassifier	79	79	79	0.79

As we can see in the above Data Frame; **Random Forest Classifier** model seems perfect as compare to other models as the training accuracy is 100% while testing accuracy and CV score is 90% which is excellent accuracy. Also the CV score and testing accuracy are same as 90% too and it's also indicates that our model is performing excellent by each method either random\_state or K-Fold method. The F1-score is 90% too it means that error are on lower side and ROC\_AUC\_SCORE is 0.90, which is greater than the threshold value of 0.6, which indicates that the machine probability is good while predicting 1 as 1 and 0 as 0. Also the Precision and Recalls are as- 92% and 88% respectively and this indicates that 92 times system is getting positive predictions and 88% of the times model's actual and predicted values are true positive as 1 (i.e. customer are paying the loan on time).



```
In [230]: conclusion=pd.DataFrame(data=[pred,y_test],index=['Predicted label','Original label'])
conclusion
```

Out[230]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
Predicted label	0	0	0	0	1	0	1	0	0	0	1	1	1	1	1	0	0	1	0	1	1	0	1	0	0	0	0	1	1	0	0	1	1	0	0	1	1	1
Original label	0	1	0	0	1	0	1	0	0	0	1	1	1	1	0	0	0	1	0	1	1	0	1	0	0	0	0	1	1	1	0	1	1	0	1	1	1	1

When I deployed our Random Forest Classifier Model to the `y_test` data what I found is that the testing accuracy of the model went to almost 91% which is excellent accuracy for predicting any target variable correctly. Also the ROC AUC Score is 0.90 which is greater than 0.6 of threshold value and it indicates that out of 100 times, 96 times model is predicting the right classes i.e. 1 as 1 and 0 as 0 and this is still a great accuracy. As we can see in the conclusion portion we have got almost same value in predicted Label as compare to original Label. So we can say that this model has great accuracy while predicting the status of defaulter customers.

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

I have used the Classification Model using multiple algorithms to design and optimize the results. Some of the classes which I explored from Scikit-learn libraries were – Statistics, Analytical Modelling, Hyper Tuning Method, CV Score, Predictive Modelling and Classification Model, etc.

I condensed to 5 columns keeping in mind the Principle Component Analysis. This resulted in condensed data being trained in a much shorter span of time with utmost accuracy. The visualizations helped in better and quick understanding of the outliers and skewness present in the data sets

With Mathematical Modeling helped me to find-out the corresponding mean, median, mode and relationship among the variables.

Statistical Modeling helped in Correlation for understanding the relationship among the variables

One of the key challenges which I faced was w.r.t. the Hyper Parameter Tuning step. Since there were 73000 above data went for testing phase and due to this system were taking too much time to find out the best parameters for the Random Forest Classifier, Ada Boost Classifier and Gradient Boosting Classifier and that's why I manually defined the best parameters for these classifiers otherwise it could take 2-3 days for sure.

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

As the client has instructed us not to remove the outliers which were almost 20% in the dataset could create biasness w.r.t. other company's decision making and statistical analysis, where there would not be such restrictions.

The same analysis which was done for this project can be used for another MFI related projects only if they provide the same characteristic as our Client did.