Micro Credit Project

Submitted by:

Saurav Kumar

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 Shwetank Mishra for his constant guidance and support .

INTRODUCTION

Business Problem Framing : - 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, 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).

Conceptual Background of the Domain Problem :- Generally, Credit Scores plays a vital role for loan approvals, and is very important in today's financial analysis for an individual, Most of the loan lending vendors rely heavily on it, so in our case users has 5 days' time to pay back the loan or else they are listed as defaulters which will impact the loan the credit score heavily, so there are few thing to lookout in this dataset as users who are taking extensive loans, user who have most frequent recharges in their main account have a good chance of 100% payback rate, and user who never recharged their main account for them loan should have never been approved as there is high chance for single user or default user taking multiple connections in name or documents of the family members.

Review of Literature :- The project objective is to find out the defaulters (i.e. the users who don't repay the loan within 5 days). Now, Using Different Mathematical and statistical tools

Many assumptions regarding the data is made and data Cleaning is done.

After the Data Cleaning part Model Training takes place in which different models like: Logistic Regression, Random Forest Classifier, Gradient Boost Classifier , Xtreme Gradient Boost Classifier . models are used for the Training of the data.

After Training of the data Hyper-parameter tuning is done and then the best model is designed.

Motivation for the Problem Undertaken :- 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.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem :- This problem is a classification problem, the target variable is itself a statistical parameter. We have to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label '1' indicates that the loan has been payed i.e. Non- defaulter, while, Label '0' indicates that the loan has not been payed .for a loan amount of 5 payback amount should be 6,and for loan amount of 10 payback amount is 12.

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	209593.0	104797.000000	60504.431823	1.000000	52399.000	104797.000000	157195.00	209593.000000
label	209593.0	0.875177	0.330519	0.000000	1.000	1.000000	1.00	1.000000
aon	209593.0	8112.343445	75696.082531	-48.000000	246.000	527.000000	982.00	999860.755168
daily_decr30	209593.0	5381.402289	9220.623400	-93.012667	42.440	1469.175667	7244.00	265926.000000
daily_decr90	209593.0	6082.515068	10918.812767	-93.012667	42.692	1500.000000	7802.79	320630.000000
rental30	209593.0	2692.581910	4308.586781	-23737.140000	280.420	1083.570000	3356.94	198926.110000
rental90	209593.0	3483.406534	5770.461279	-24720.580000	300.260	1334.000000	4201.79	200148.110000
last_rech_date_ma	209593.0	3755.847800	53905.892230	-29.000000	1.000	3.000000	7.00	998650.377733
last_rech_date_da	209593.0	3712.202921	53374.833430	-29.000000	0.000	0.000000	0.00	999171.809410
last_rech_amt_ma	209593.0	2064.452797	2370.786034	0.000000	770.000	1539.000000	2309.00	55000.000000
cnt_ma_rech30	209593.0	3.978057	4.256090	0.000000	1.000	3.000000	5.00	203.000000
fr_ma_rech30	209593.0	3737.355121	53643.625172	0.000000	0.000	2.000000	6.00	999606.368132
sumamnt_ma_rech30	209593.0	7704.501157	10139.621714	0.000000	1540.000	4628.000000	10010.00	810096.000000
nedianamnt_ma_rech30	209593.0	1812.817952	2070.864620	0.000000	770.000	1539.000000	1924.00	55000.000000
medianmarechprebal30	209593.0	3851.927942	54006.374433	-200.000000	11.000	33.900000	83.00	999479.419319
cnt_ma_rech90	209593.0	6.315430	7.193470	0.000000	2.000	4.000000	8.00	336.000000
fr_ma_rech90	209593.0	7.716780	12.590251	0.000000	0.000	2.000000	8.00	88.000000

sumamnt_ma_rech90	209593.0	12396.218352	16857.793882	0.000000	2317.000	7226.000000	16000.00	953036.000000
medianamnt_ma_rech90	209593.0	1864.595821	2081.680664	0.000000	773.000	1539.000000	1924.00	55000.000000
medianmarechprebal90	209593.0	92.025541	369.215658	-200.000000	14.600	36.000000	79.31	41456.500000
cnt_da_rech30	209593.0	262.578110	4183.897978	0.000000	0.000	0.000000	0.00	99914.441420
fr_da_rech30	209593.0	3749.494447	53885.414979	0.000000	0.000	0.000000	0.00	999809.240107
cnt_da_rech90	209593.0	0.041495	0.397556	0.000000	0.000	0.000000	0.00	38.000000
fr_da_rech90	209593.0	0.045712	0.951386	0.000000	0.000	0.000000	0.00	64.000000
cnt_loans30	209593.0	2.758981	2.554502	0.000000	1.000	2.000000	4.00	50.000000
amnt_loans30	209593.0	17.952021	17.379741	0.000000	6.000	12.000000	24.00	306.000000
maxamnt_loans30	209593.0	274.658747	4245.264648	0.000000	6.000	6.000000	6.00	99864.560864
medianamnt_loans30	209593.0	0.054029	0.218039	0.000000	0.000	0.000000	0.00	3.000000
cnt_loans90	209593.0	18.520919	224.797423	0.000000	1.000	2.000000	5.00	4997.517944
amnt_loans90	209593.0	23.645398	26.469861	0.000000	6.000	12.000000	30.00	438.000000
maxamnt_loans90	209593.0	6.703134	2.103864	0.000000	6.000	6.000000	6.00	12.000000
medianamnt_loans90	209593.0	0.046077	0.200692	0.000000	0.000	0.000000	0.00	3.000000
payback30	209593.0	3.398826	8.813729	0.000000	0.000	0.000000	3.75	171.500000
payback90	209593.0	4.321485	10.308108	0.000000	0.000	1.666667	4.50	171.500000

- From the above statistical summay we can see that we some unnecessary column like Unnamed , msisdn . So , Here we drop the columns because its unwanted column and its not contribution is Zero.

- 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 behaviour more precisely because many users have the tendency 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.

- 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**

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

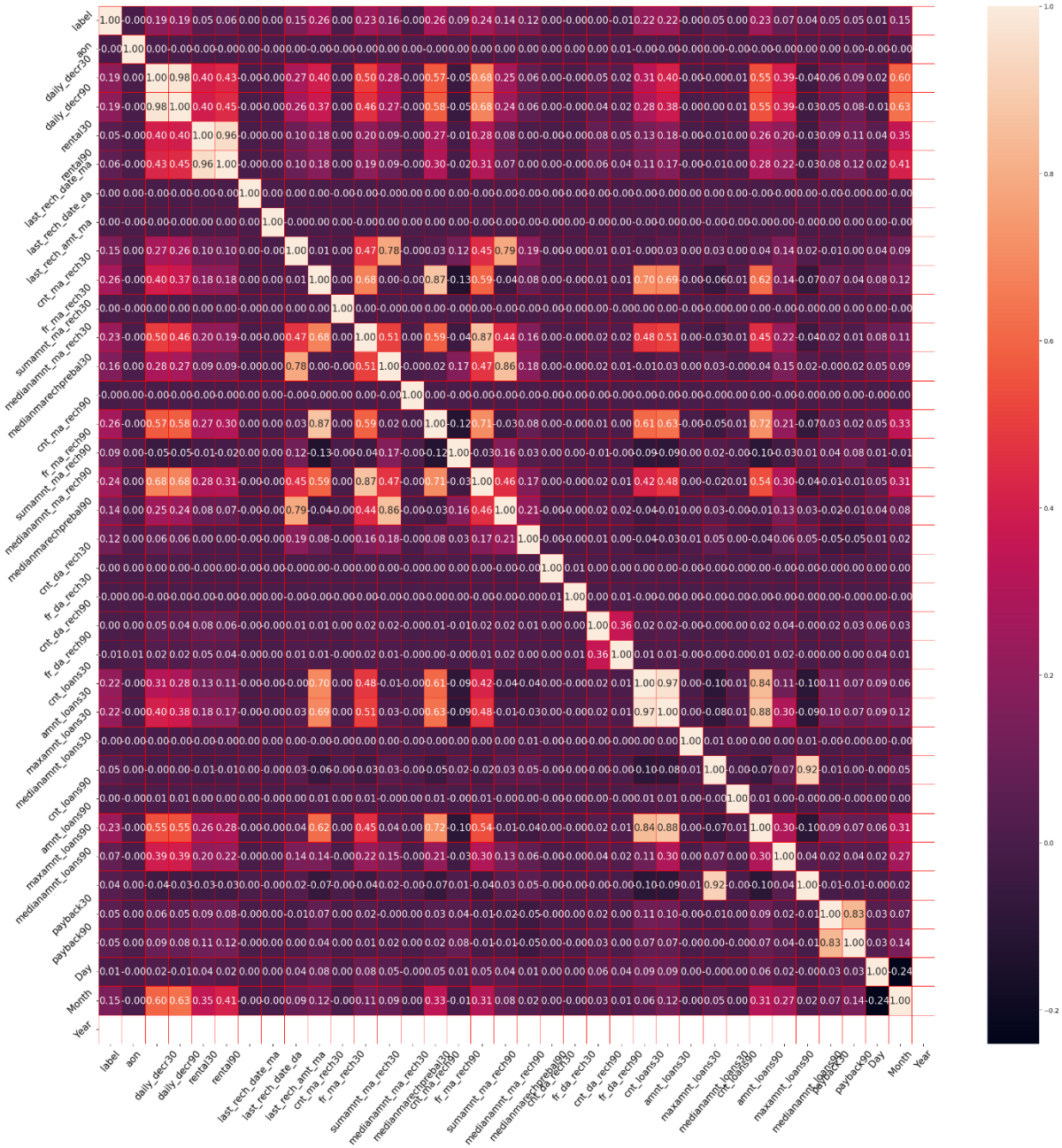
Data Type of Each Columns

```
label                   int64
aon                   float64
daily_decr30          float64
daily_decr90          float64
rental30              float64
rental90              float64
last_rech_date_ma     float64
last_rech_date_da     float64
last_rech_amt_ma        int64
cnt_ma_rech30           int64
fr_ma_rech30          float64
sumamnt_ma_rech30     float64
medianamnt_ma_rech30  float64
medianmarechprebal30  float64
cnt_ma_rech90           int64
fr_ma_rech90            int64
sumamnt_ma_rech90       int64
medianamnt_ma_rech90  float64
medianmarechprebal90  float64
cnt_da_rech30         float64
fr_da_rech30          float64
cnt_da_rech90           int64
fr_da_rech90            int64
cnt_loans30             int64
amnt_loans30            int64
maxamnt_loans30       float64
medianamnt_loans30    float64
cnt_loans90           float64
amnt_loans90            int64
maxamnt_loans90         int64
medianamnt_loans90    float64
payback30             float64
payback90             float64
pcircle                object
pdate                  object
dtype: object
```

Data Pre-processing Done :-

1 -> First we removed outliers from 'daily_decr30', 'daily_decr90','medianamnt_ma_rech90', 'medianmarechprebal90','amnt_loans30','amnt_loans90' From only these columns . Because I think these column is are contain outliers.
After Deleting Outliers we can see lost 6.5 % of data. And we use zscore method to remove outliers.

2 -> Checking Correlation With HeatMap



Observation from heatmap:

- Year is Unnecessory Columns

- daily_decr90 and daily_decr30 is highly 98% correlated with each other.

- amnt_loans30 and amnt_loans90 is also highly correlated with 88% correlation.

- medianamnt_loans90 and medianamnt-loans30 is also highly correlated with 92% correlation.

- cnt_loans30 is also 97% correlate with amnt-loans30

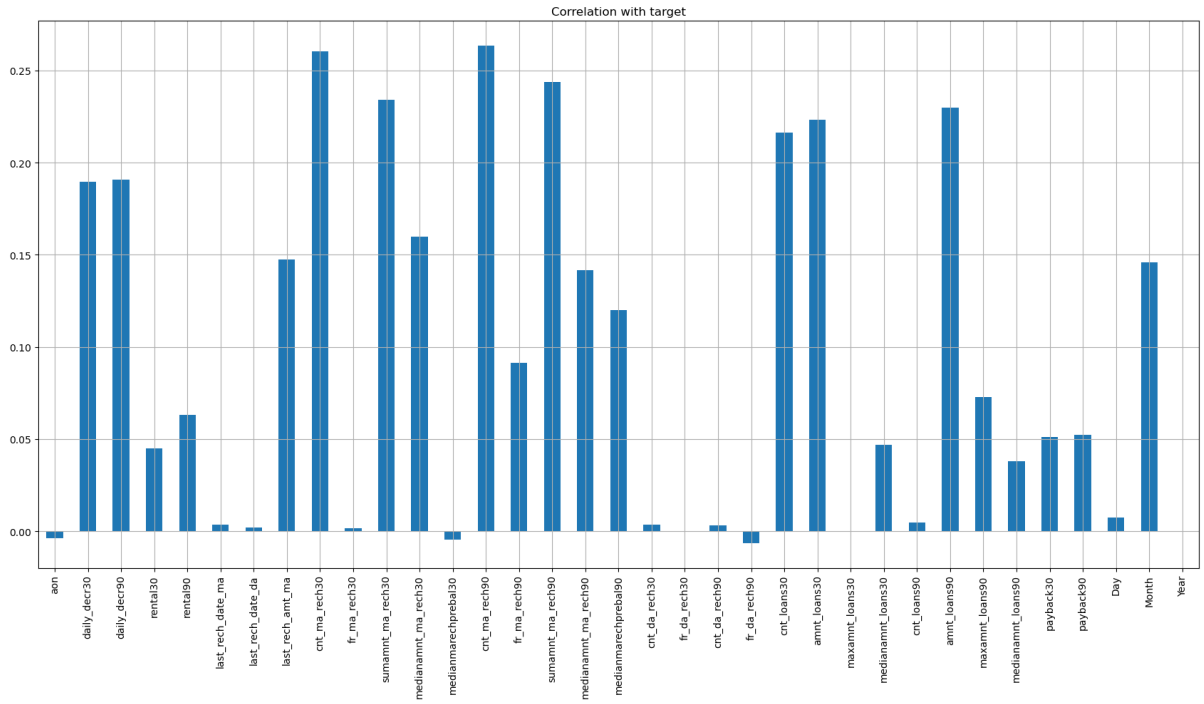
- rental90 is 96% correlate with rental30.

- cnt_ma_rech90 is 87% correlated with cnt_ma-rech30.

- medianamnt_ma_rech90 is 86% correlated with medianamnt-ma-rech30.

May be its creating the multicollinearity Problem. Lets Check The there is multicollinearity problem or not.

3 -> Relationship with Target



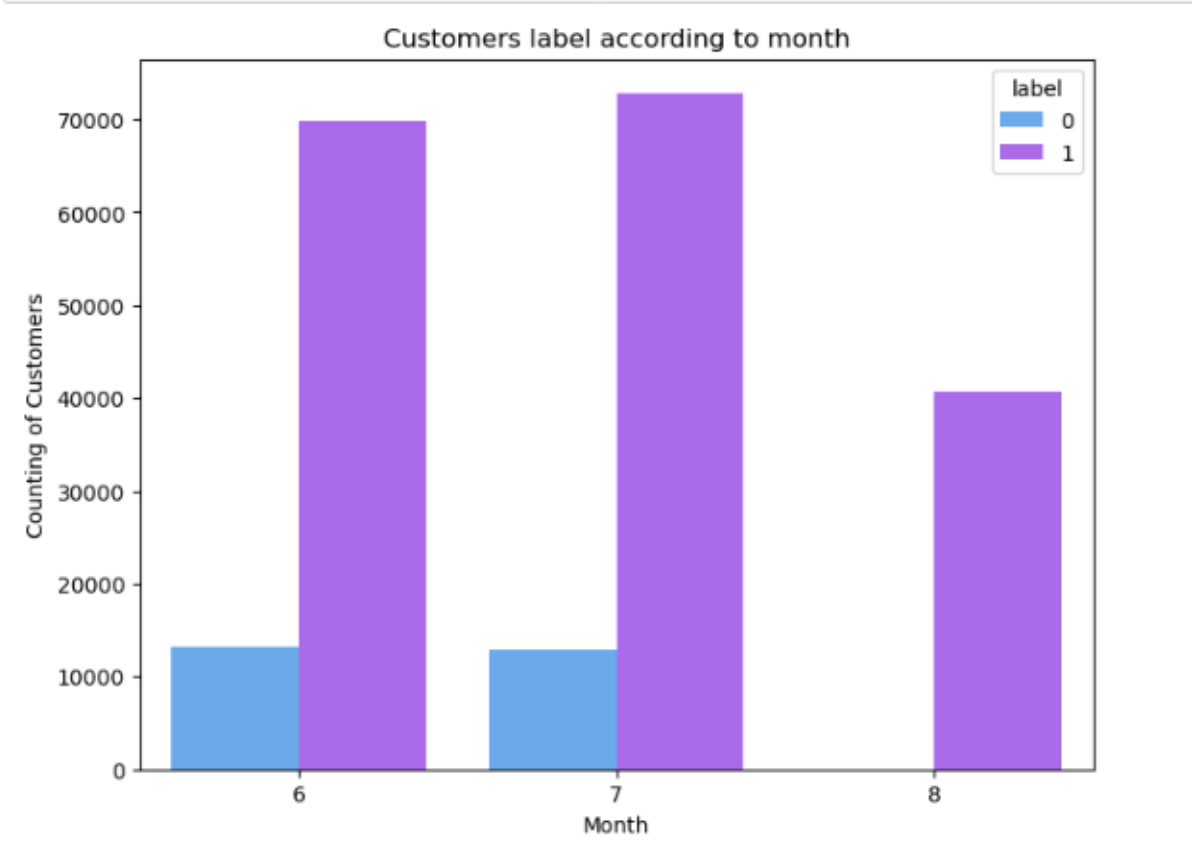
- Year have Does not having any relation . Lets drop right now.
- cnt_ma_rech90 have highly correlated
- cnt_ma_rech30 have highly correlated
- sumamnt_ma_rech90 have highly correlated
- and we have more columns which have good relation and we also have column which dont have any relation

4-> Checking Skewness

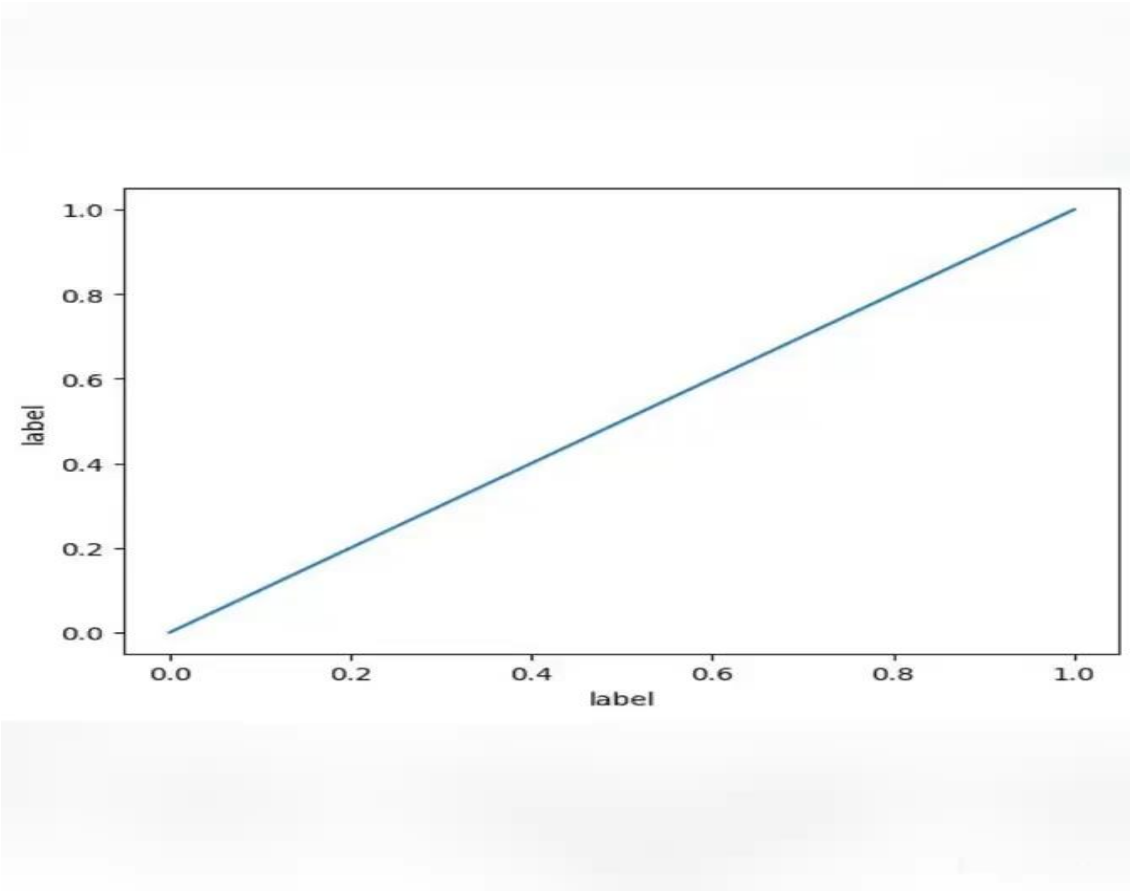
```
label                  -2.186101
aon                    10.400216
daily_decr30            1.890379
daily_decr90            1.990263
rental30                4.253425
rental90                4.380219
last_rech_date_ma      14.839638
last_rech_date_da      14.897876
last_rech_amt_ma        2.671435
cnt_ma_rech30           1.950400
fr_ma_rech30           14.764050
sumamnt_ma_rech30       2.875017
medianamnt_ma_rech30    2.384533
medianmarechprebal30   14.726477
cnt_ma_rech90           2.081323
fr_ma_rech90            2.235131
sumamnt_ma_rech90       2.397616
medianamnt_ma_rech90    2.276276
medianmarechprebal90    3.623945
cnt_da_rech30          17.840103
fr_da_rech30           14.735789
cnt_da_rech90          27.928307
fr_da_rech90           30.136185
cnt_loans30             1.588193
amnt_loans30            1.517935
maxamnt_loans30        17.670252
medianamnt_loans30      4.531760
cnt_loans90            16.609998
amnt_loans90            1.712200
maxamnt_loans90         1.844849
medianamnt_loans90      4.865652
payback30               8.115361
payback90               6.744517
Day                     0.191959
Month                   0.395402
dtype: float64
```

We are using 2 method to remove skewness cbrt method and Power Transformer

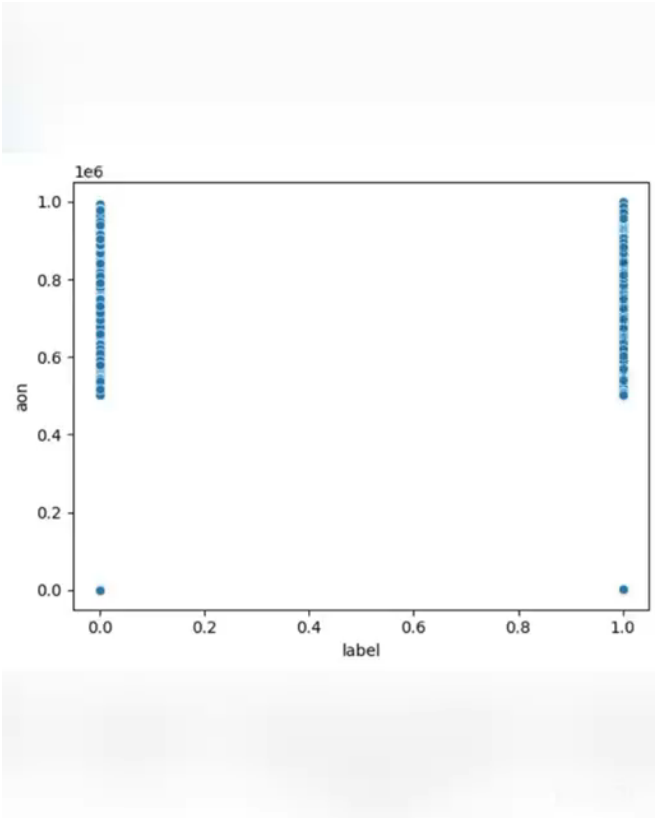
Data Inputs- Logic- Output Relationships : -



- Using Line Plot for Checking Relationship



- Using Scatter Plot to Checking Relationship



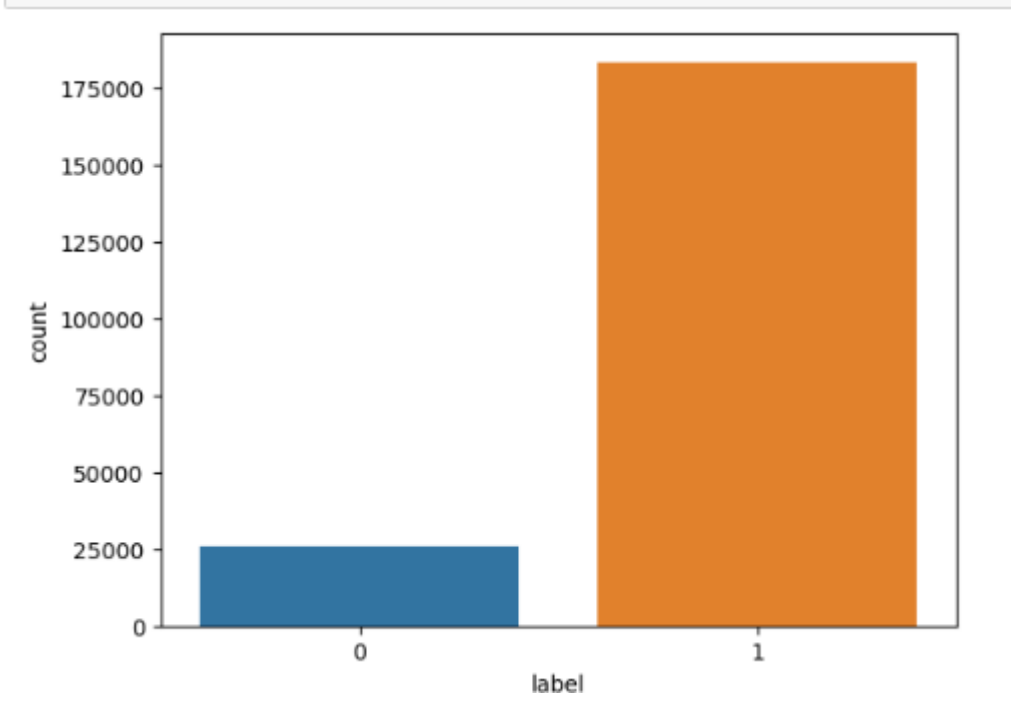
Hardware and Software Requirements and Tools Used

Hardware: 8GB RAM, 64-bit, Rygen 3 processor.

Software: MS-Excel, Jupyter Notebook, python 3.6.

Model/s Development and Evaluation :

- Identification of possible problem-solving approaches (methods)



Here , We can see that our Target Variable are imbalanced . So , Here we are using SMOTE to balace our dataset.

Testing of Identified Approaches (Algorithms)

▶ **LogisticRegression**

```
The cv score is:  0.7508120713745383
The accuracy score is:  0.755296734168325
```

▶ **Random Forest Classifier**

```
The cv score is:  0.949595998953553
The accuracy score is:  0.9492420695306099
```

▶ **Gradient Boosting Classifier**

```
The cv score is:  0.8915305113604959
The accuracy score is:  0.897825705255765
```

▶ **XGBOOST**

```
The cv score is:  0.9414549422033693
The accuracy score is:  0.9500907175465293
```

Run and Evaluate selected models

▶ **Here I am Selecting Random Forest Classifier Because its CV Score and Accuracy Score is Same. Means this model is good.**

```
The cv score is:  0.949595998953553
The accuracy score is:  0.9492420695306099
```

Confusion Matrix

```
Confusion Matrix:
[[32681  1541]
[ 1928 32194]]
```

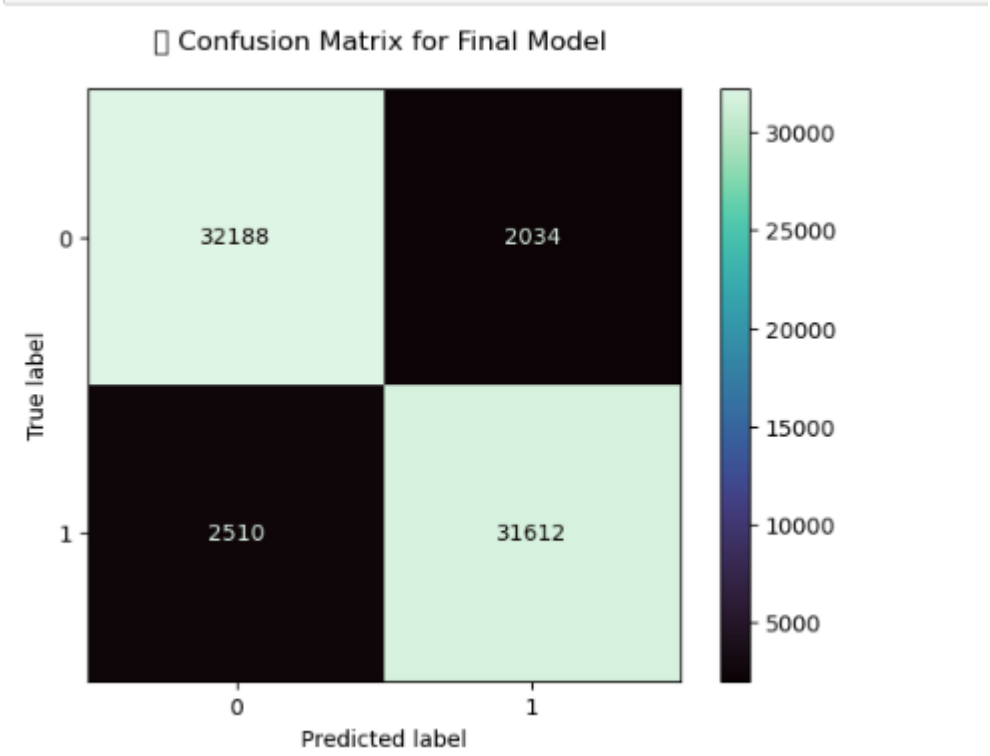
Then we are doing Hyperparameter Tuning with GridSearchCV.

In [91]:
```
rfc=RandomForestClassifier(max_depth=111 , min_samples_split=10, min_samples_leaf=7, criterion='entropy', n_estimators=300)
rfc.fit(x_train,y_train)

metric_score(rfc,x_train,x_test,y_train, y_test, train=True)

metric_score(rfc,x_train,x_test,y_train, y_test, train=False)
```

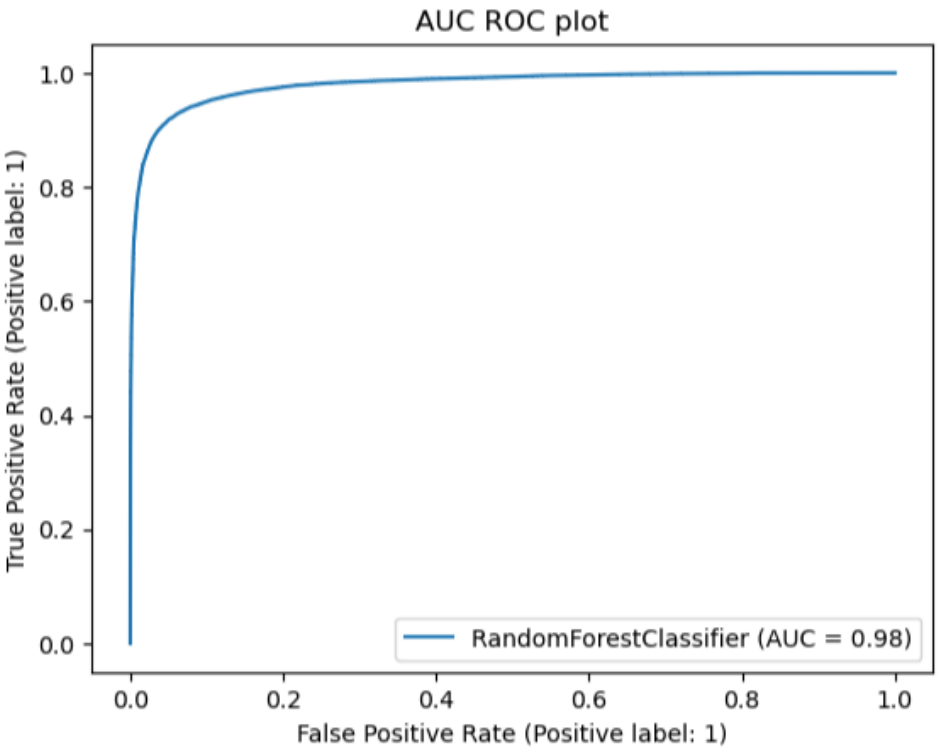
```
========================= Train Result==============================
Accuracy Score: 95.40%

=========================Test Result======================
Accuracy Score: 93.35%


 Test Classification Report
              precision    recall  f1-score   support

           0       0.93      0.94      0.93     34222
           1       0.94      0.93      0.93     34122

    accuracy                           0.93     68344
   macro avg       0.93      0.93      0.93     68344
weighted avg       0.93      0.93      0.93     68344


 Confusion Matrix:
 [[32188  2034]
 [ 2510 31612]]
```

Visualizations:

Confusion Matrix



AUC ROC Curve ->



CONCLUSION

Key Findings and Conclusions of the Study

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 balance category.
3) Users having high balance and are defaulters are very less in number.
4) 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.
5) 14% of the Users are among the average number of loan taken category are defaulters.
6) 40 % of the Users who do not even recharged in the 90 days are defaulters only.
7) Users who do very high amount of recharge always pays their loans on time. i.e 98% of them are non-defaulters.
8) 34% of the Users who do less amount of recharge are defaulters.
9) Users who did not take any loans are non-defaulters.
10) Most of the Users (i.e. 97%) who take large amount of loans comes under non defaulter category.
11) 17% of the users who take small loans are defaulters.

12) Among the Users who have not done a single recharge in 3 months 40% are defaulters.
13) 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.q
14) 32% of the users who are defaulters are the new users.
15) Old Users are trusted and they are mostly non defaulters.
16) Random Forest Classifier performed the best AUC_ROC_SCORE **i.e. 94.7%.**

Learning Outcomes of the Study in respect of Data Science

- Visualizations and Data Cleaning part was very crucial as without the cleaning we were not able to judge the data effectively and won't be able to remove the outliers thus adding in to the errors.
- Visualizations helped a lot in finding out those outliers values and helped in finding out the features having direct relation between the feature and the label.