FLIP ROBO

Micro Credit Defaulter Project

Submitted by:

Jessica Ghimeliya

ACKNOWLEDGMENT

This project would not have seen the light of the day without the following people and their priceless support and cooperation. Hence I extend my gratitude to all of them.

As a student of Data Trained Education, I would first of all like to express my gratitude to FlipRobo Team and seniors for granting me permission to undertake the project report in their esteemed organization. I would also like to express my sincere thanks to Miss Khushboo Mam for supporting me and being always there for me whenever indeed.

During the actual research work with FlipRobo team and other IOT that set the ball rolling for my project. They had been a source of inspiration through their constant guidance; personal interest; encouragement and help.

I convey my sincere thanks to them. In spite of their busy schedule they always found time to guide me throughout the project. I am also grateful to them for reposing confidence in my abilities and giving me the freedom to work on my project. Without their invaluable help I would not have been able to do justice to the project.

INTRODUCTION

Business Problem Framing

- A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.
- Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and is very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.
- Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.
- We are working with one such 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.
- 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 poor 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. 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).
- The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve 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.

Conceptual Background of the Domain Problem

Microfinance is a source of financial services for entrepreneursand small businesses lacking access to banking and related services.

Micro Credit is a small financial loan made to poverty stricken individuals seeking to start their own business. This type of loan typically does not exceed a couple hundred dollars, so an impoverished individual cannot solely depend on this type of loanto fund their business. It's also called micro loan.

Although often used interchangeably, microfinance and microcredit are in fact quite distinct. Microfinance is a much broader concept than microcredit and refers to loans, savings, insurance, money transfers, and other financial products targeted at poor and low-income people. Microcredit refers more specifically to making small loans available to poor people, especially those traditionally excluded from financial services, through Programmers designed specifically to meet their particularneeds and circumstances.

Review of Literature:

Muhammad Yunus, a Nobel Prize winner, introduced the concept of Microfinance in Bangladesh in the form of the Grameen Bank. The National Bank for Agriculture and Rural Development (NABARD) took this idea and started the concept of microfinance in India. Under this mechanism, there exists a link between SHGs (Self-help groups), NGOs and banks. SHGs are formed and nurtured by NGOs and only after accomplishing acertain level of maturity in terms of their internal thrift and credit operationsare they entitled to seek credit from the banks. There is an involvement from the concerned NGO before and even after the SHG-Bank linkage. The SHG-Bank link age program me, which has been in place since 1992 in India, has provided about 22.4 lakhs for SHG finance by 2006. It involves commercial banks, regional rural banks (RRBs) and cooperative banks in its operations.

"Micro Finance" is often seen as financial services for poor and low-income clients. In practice the term is often used more narrowly to refer to the loans and other services from providers that identify themselves as "microfinance institutions" (MFIs). Microfinance can also be described as a setup of a number of different operators focusing on the financially under-served people with the aim of satisfying their need for poverty alleviation, social promotion, emancipation, and inclusion. Microfinance institutions reach and serve their target market in very innovative ways. Microfinance operations differ in principle, from the standard disciplines of general and entrepreneurial

finance. This difference can be attributed to the fact that the size of the loans granted with microcredit is typically too small to finance growth-oriented business projects. Implicit guarantee of ready access to future loans if present loans are repaid fully and promptly Microfinance is seen as a catalyst for poverty alleviation, delivered in innovative and sustainable ways to assist the underserved poor, especially in developing countries.

Default in microfinance is the failure of a client to repay a loan. The default could be in terms of the amount to be paid or the timing of the payment.

Motivation for the Problem Undertaken:-

Our main objective of doing this project is to build a model to predict whether the users are paying the loan within the due date or not. We are going to predict by using Machine Learning algorithms. The sample data is provided to us from our client database. In order to improve 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. Learning about real world business problems and become Data Scientist is only motive to get hands on this project. Working on Micro Finance and Credit is a part of my Internship with FlipRobo Technologies. This project is provided by the company to get some touch of real time or real world problems, and to perform well in this project to show my capabilities is the motivation behind doing this project.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem:-

Research methodology is a way to systematically solve the problem. It is a game plan for conducting research. In this we describe various steps that are taken by the researcher.

To build a better machine learning model of predicting the defaulter case from micro credit card data set we encounter with several statistical modeling while data analysis & off course we needed mathematical modeling to build several machine learning models in this project.

“All progress is born of inquiry. Doubt is often better than overconfidence, for it leads to inquiry and inquiry leads to invention.”

Research methodology is a framework for the study and is used as a guide in collecting and analyzing the data. It is a strategy specifying which approach will be used for gathering and analyzing the data. It also includes time and cost budget since most studies are done under these two constraints. The research methodology includes overall research design, the sampling procedure, the data collection method and analysis procedure.

Some Mathematical analysis shown below screenshots:-

```
1 #checking how long our dataset is
2 df.shape
```

```
(209593, 37)
```

Our dataset has 209593 rows and 37 columns including target.

```
df.describe()
```

	Unnamed: 0	label	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	last_rech_amt_ma	cnt_ma_re
count	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.0
mean	104797.000000	0.875177	8112.343445	5381.402289	6082.515068	2692.581910	3483.406534	3755.847800	3712.202921	2064.452797	3.9
std	60504.431823	0.330519	75696.082531	9220.623400	10918.812767	4308.586781	5770.461279	53905.892230	53374.833430	2370.786034	4.2
min	1.000000	0.000000	-48.000000	-93.012667	-93.012667	-23737.140000	-24720.580000	-29.000000	-29.000000	0.000000	0.0
25%	52399.000000	1.000000	246.000000	42.440000	42.692000	280.420000	300.260000	1.000000	0.000000	770.000000	1.0
50%	104797.000000	1.000000	527.000000	1469.175667	1500.000000	1083.570000	1334.000000	3.000000	0.000000	1539.000000	3.0
75%	157195.000000	1.000000	982.000000	7244.000000	7802.790000	3356.940000	4201.790000	7.000000	0.000000	2309.000000	5.0
max	209593.000000	1.000000	999860.755168	265926.000000	320630.000000	198926.110000	200148.110000	998650.377733	999171.809410	55000.000000	203.0

 # Observations:-

- Unnamed:0 feature is a kind of serial number , so it is not related with target variable, we will delete in further steps.
- Some features like aon, daily_decr30, daily_decr90,rental30,last_rech-date_ma,last_rech_amt_ma,cnt_ma_rech30, fr_ma_rech30 and many more features have not appropriate distributed data, i.e. that the difference between min value and max value is very high. It is not in acceptable range. We will handle it later.
- Some features have high standard deviation than its mean value, which is not acceptable.

Data Sources and their formats:

After the research problem has been identified and selected the next step is to gather the requisite data. While deciding about the method of data collection to be used for the researcher should keep in mind two types of data i.e. primary and secondary.

The primary data are those, which are collected afresh and for the first time, and thus happened to be original in character. We can obtain primary data either through observation or through direct communication with respondent in one formor another

or through personal interview. Methods used in primary data collection-

- Observation method
- Interview method
- Questionnaire method

In this study data have been taken from various secondary sources like:

- Internet
- Books
- Magazines
- Newspapers
- Journals

Data Description:-

Target:

Target	Definition
label	Flag indicating whether the user paid back the credit amount within 5 days of issuin loan{1:success, 0:failure}

Features:-

Features	Definition
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 Indonesia
daily_decr90	Daily amount spent from main account, averaged over last 90 days (in Indonesia
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 Rupia
medianamnt_ma_rech30	Median of amount of recharges done in main account over last 30 days at user le Rupiah)
medianmarechprebal30	Median of main account balance just before recharge in last 30 days at user level 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 Rupia
medianamnt_ma_rech90	Median of amount of recharges done in main account over last 90 days at user le Rupiah)
medianmarechprebal90	Median of main account balance just before recharge in last 90 days at user level 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

Data Preprocessing Done:

Check Missing values

```
#checkig NaNs
df.isna().sum()
```

```
Unnamed: 0                0
label                     0
msisdn                    0
aon                       0
daily_decr30              0
daily_decr90              0
rental30                  0
rental90                  0
last_rech_date_ma         0
last_rech_date_da         0
last_rech_amt_ma          0
cnt_ma_rech30             0
fr_ma_rech30              0
sumamnt_ma_rech30         0
medianamnt_ma_rech30      0
medianmarechprebal30      0
cnt_ma_rech90             0
fr_ma_rech90              0
sumamnt_ma_rech90         0
medianamnt_ma_rech90      0
medianmarechprebal90      0
cnt_da_rech30             0
fr_da_rech30              0
cnt_da_rech90             0
fr_da_rech90              0
cnt_loans30               0
amnt_loans30              0
maxamnt_loans30           0
medianamnt_loans30        0
cnt_loans90               0
amnt_loans90              0
maxamnt_loans90           0
medianamnt_loans90        0
payback30                 0
payback90                 0
pcircle                   0
pdate                     0
dtype: int64
```

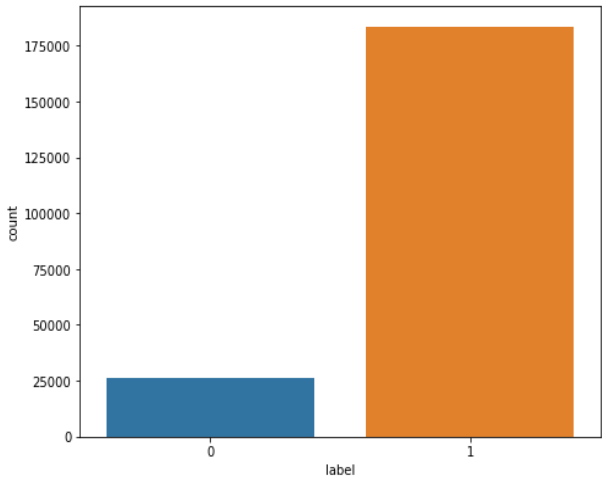
Checking Duplicates:-

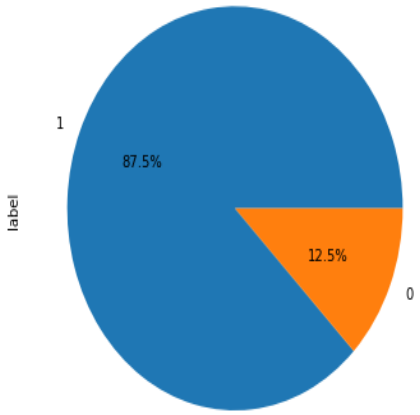
```
#checking for duplicates
df.duplicated().sum()
```

```
0
```

There is not any duplicates present.

Checking Unbalancing Dataset:-





Note:-

- We can see the imbalanced dataset here.
- As we can from above graph 87.5% customers are payed it's credit card payment but 12.5% customers are not payed.
- So that we can say that we have biased class, which is the distinctive example of Imbalanced Classification Problem. We will use Over-Sampling or Under-Sampling technique to handle this type of problem. We will take care of it before model building.

Checking Skewness:-

Checking skewness

```
1 df_new.skew()
```

```
label                -2.186101
daily_decr90          1.990263
rental90              4.380219
last_rech_amt_ma      2.671435
fr_ma_rech30         14.764050
sumamnt_ma_rech30     2.875017
medianamnt_ma_rech30  2.384533
medianmarechprebal30 14.726477
cnt_ma_rech90         2.081323
sumamnt_ma_rech90     2.397616
medianamnt_ma_rech90  2.276276
medianmarechprebal90  3.623945
cnt_da_rech90        27.928307
fr_da_rech90         30.136185
medianamnt_loans30    4.531760
amnt_loans90          1.712200
maxamnt_loans90       1.844849
medianamnt_loans90    4.865652
month                 0.395402
dtype: float64
```

Handling Skewness:-

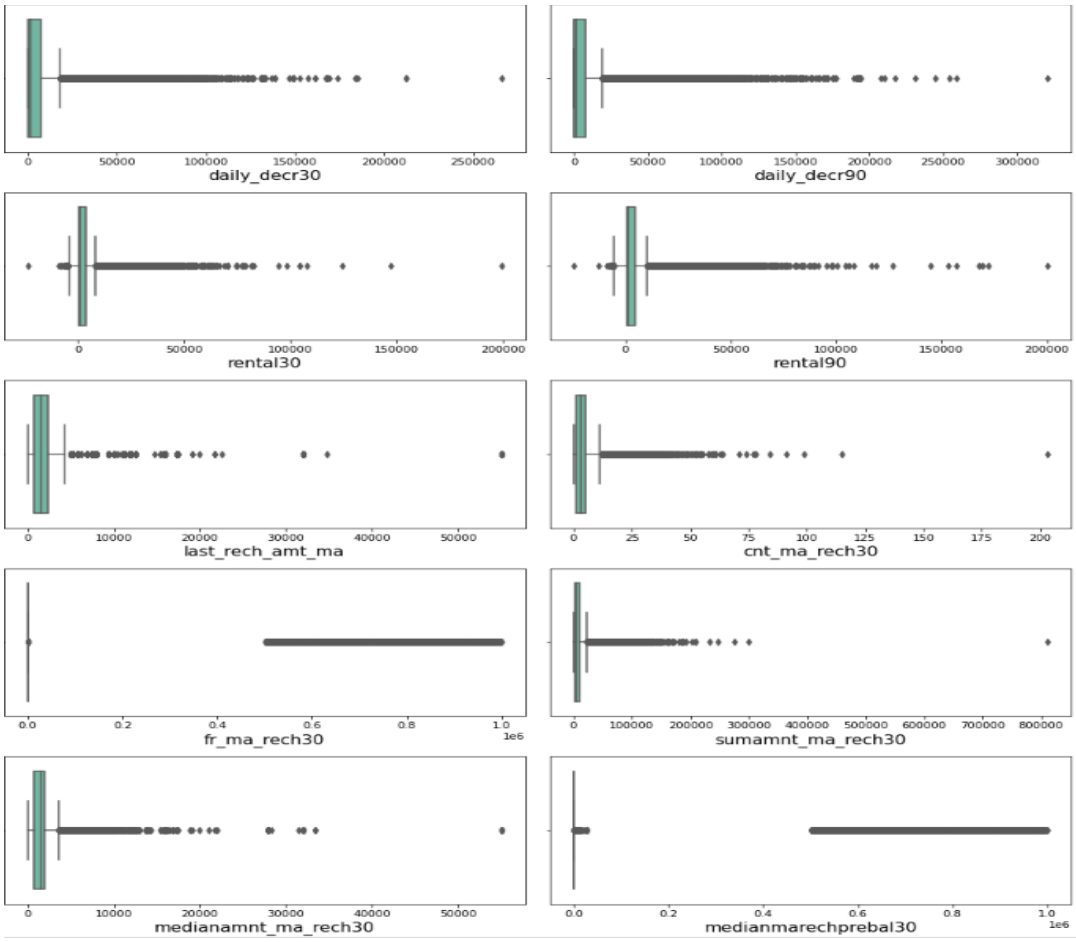
```
feature=['daily_decr90', 'rental90', 'last_rech_amt_ma', 'fr_ma_rech30',
       'sumamnt_ma_rech30', 'medianamnt_ma_rech30', 'medianmarechprebal30',
       'cnt_ma_rech90', 'sumamnt_ma_rech90', 'medianamnt_ma_rech90',
       'medianmarechprebal90', 'amnt_loans90']

for i in feature:
    df_new[i]=np.cbrt(df_new[i])
df_new.skew()
```

```
### 'fr_ma_rech30' skewness handle with power transformer
from sklearn.preprocessing import PowerTransformer
pt=PowerTransformer()
df_new['fr_ma_rech30']=pt.fit_transform(df_new['fr_ma_rech30'].values.reshape(-1,1))
df_new['medianmarechprebal30']=pt.fit_transform(df_new['medianmarechprebal30'].values.reshape(-1,1))
```

```
df_new.skew()
```

```
label                  -2.186101
daily_decr90            0.442361
rental90                0.202705
last_rech_amt_ma       -0.577155
fr_ma_rech30            0.001389
sumamnt_ma_rech30      -0.423210
medianamnt_ma_rech30   -0.715236
medianmarechprebal30    0.860463
cnt_ma_rech90          -0.557688
sumamnt_ma_rech90      -0.297338
medianamnt_ma_rech90   -0.758152
medianmarechprebal90   -0.264553
cnt_da_rech90          27.928307
fr_da_rech90           30.136185
medianamnt_loans30      4.531760
amnt_loans90            0.378695
maxamnt_loans90         1.844849
medianamnt_loans90      4.865652
month                   0.395402
dtype: float64
```

Checking Outliers:-



Handling Outliers:-

```
outlier_feature=['daily_decr30', 'daily_decr90','medianamnt_ma_rech90',
       'medianmarechprebal90','amnt_loans30','amnt_loans90']
from scipy.stats import zscore
z_score=zscore(df[outlier_feature])
abs_z_score=np.abs(z_score)
```

```
removing_outlier=(abs_z_score<3).all(axis=1)
df_new=df[removing_outlier]
print('Pervious Shape:', df.shape)
print('After removing the new shape : ',df_new.shape)
```

```
Pervious Shape: (209593, 24)
After removing the new shape :  (196637, 24)
```

Checking Data Loss

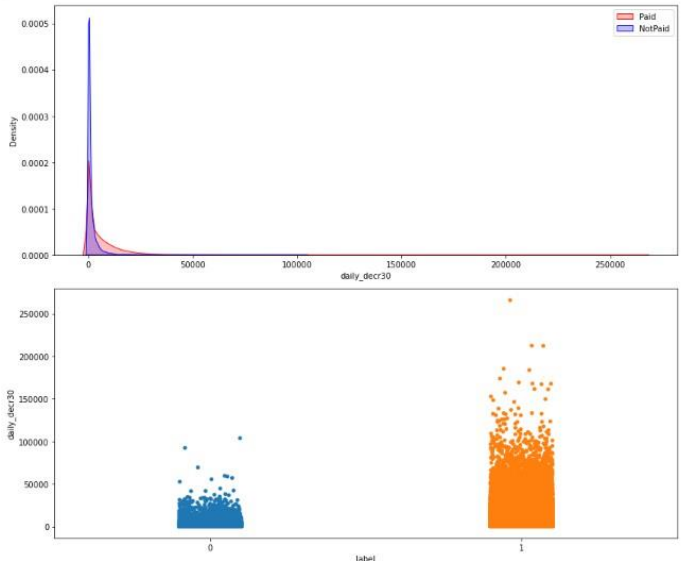
```
loss=(209593-196637)/209593*100
print('The Data Loss is : ', loss)
```

```
The Data Loss is :  6.181504153287562
```

Data Inputs- Logic- Output Relationships:-

```
daily_decr30 :- Daily amount spent from main account, (averaged over last 30 days)
```

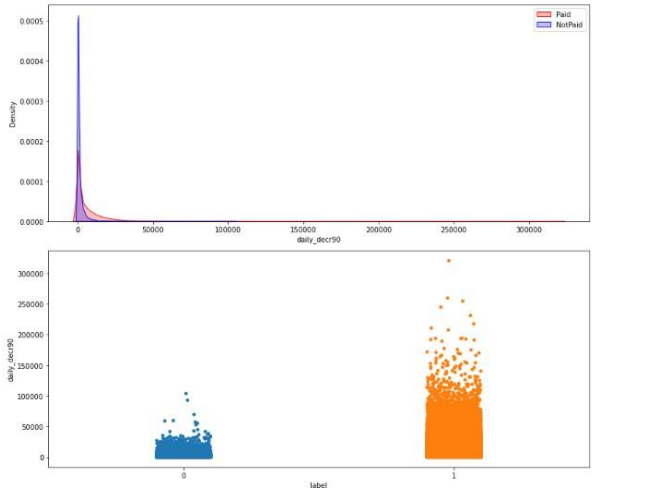
```
RelationWithTarget('daily_decr30')
```



- Daily amount spent from main account when its 0, there is high chance the user did not paid, but as daily amount spent from main account increases the user paid back the credit amount.

daily_decr90:- Daily amount spent from main account (averaged over last 90 days)

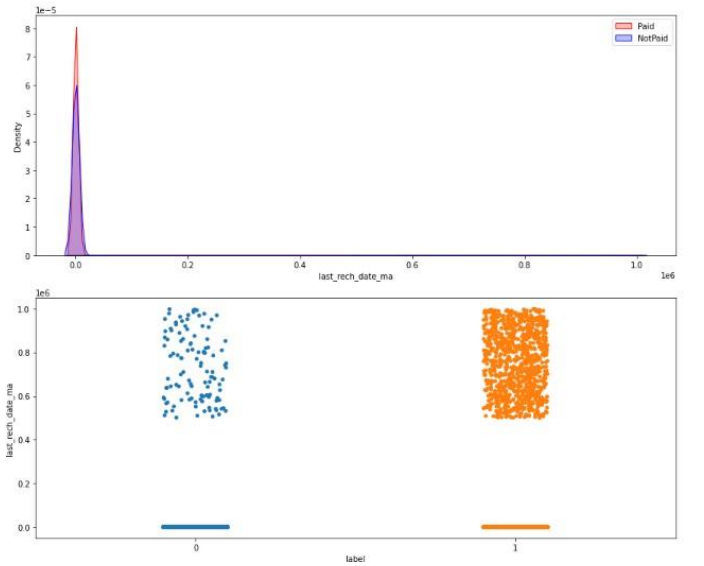
```
RelationWithTarget('daily_decr90')
```



- As we have seen previously that as Daily amount of user's mainaccount is increases than its high probability that user paid back its credit amount.

last_rech_date_ma:- Number of days till last recharge of main account

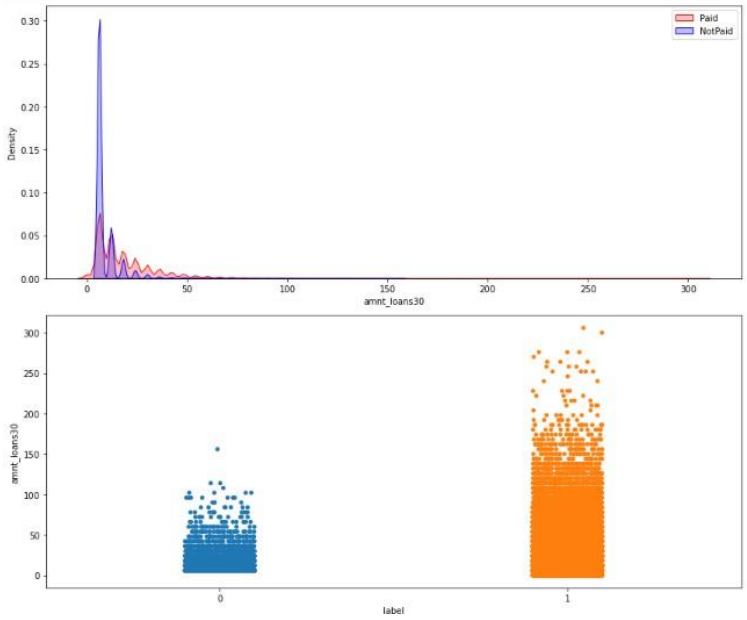
```
RelationWithTarget('last_rech_date_ma')
```



- Most of the data point fall in around 0 but.
- Here we cannot differentiate between paid back user and non-paid user. So we can say there is no relation between feature and label.

```
amnt_loans30:- Total amount of loans taken by user (last 30 days)
```

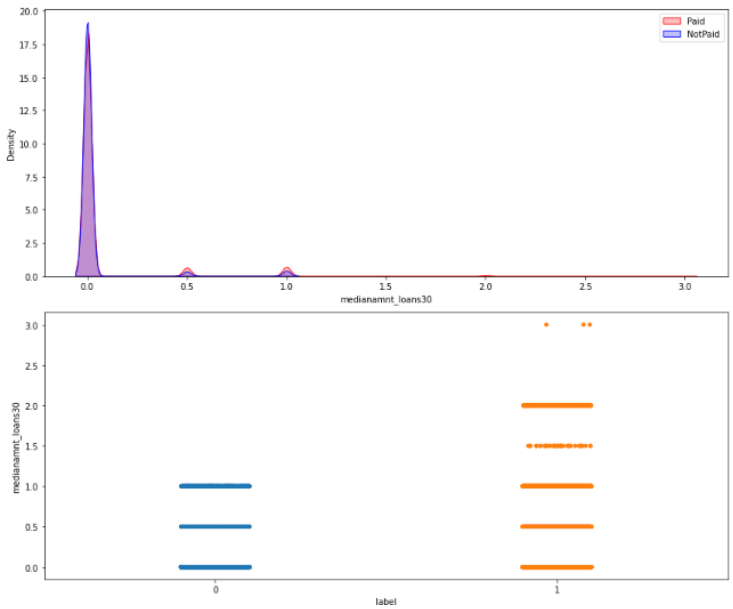
```
RelationWithTarget('amnt_loans30')
```



- As loan amount increasing the chance of cheating is decreasing.
- If Loan amount is gone up to 150 than there is no chance that user cheated with company.

```
medianamnt_loans30:- Median of amounts of loan taken by the user (last 30 days)
```

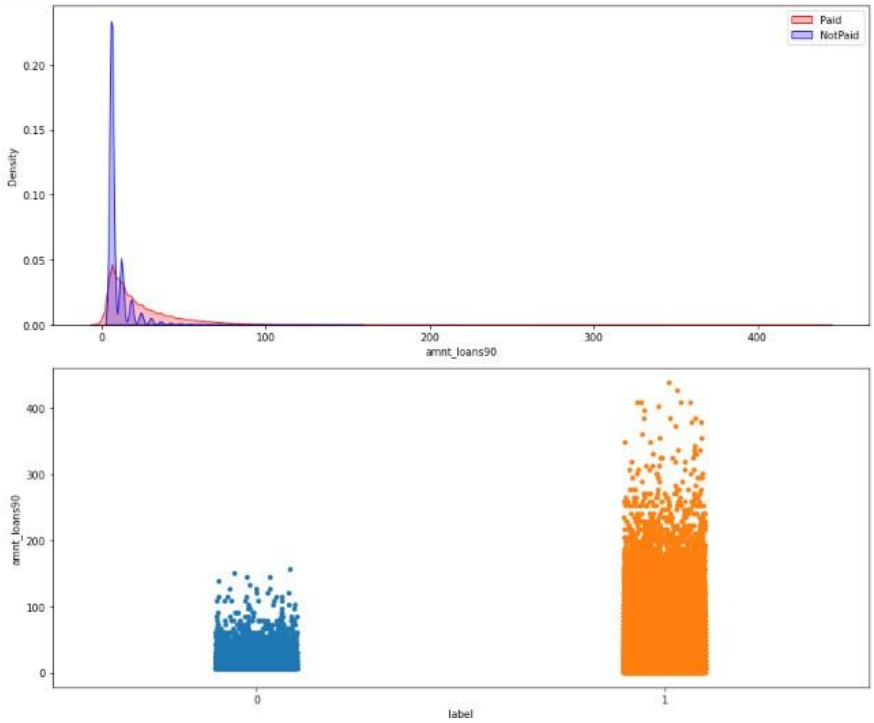
```
RelationWithTarget('medianamnt_loans30')
```



- If Median of amounts of loan taken by the user is above 1 than there is 100% chance that user will pay back to the company.

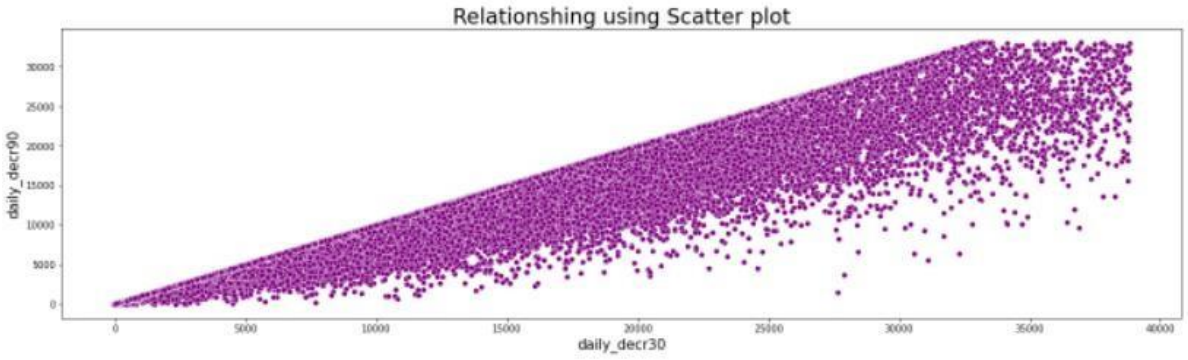
amnt_loans90:- Total amount of loans taken by user (last 90 days)

```
RelationWithTarget('amnt_loans90')
```



- We can see clear up-trend here. As loan amount increasing than chances of scam is getting down.
- And if the loan amount is above 200 then scam chances is become 0%.

```
FeatureVsFeature('daily_decr90','daily_decr30')
```

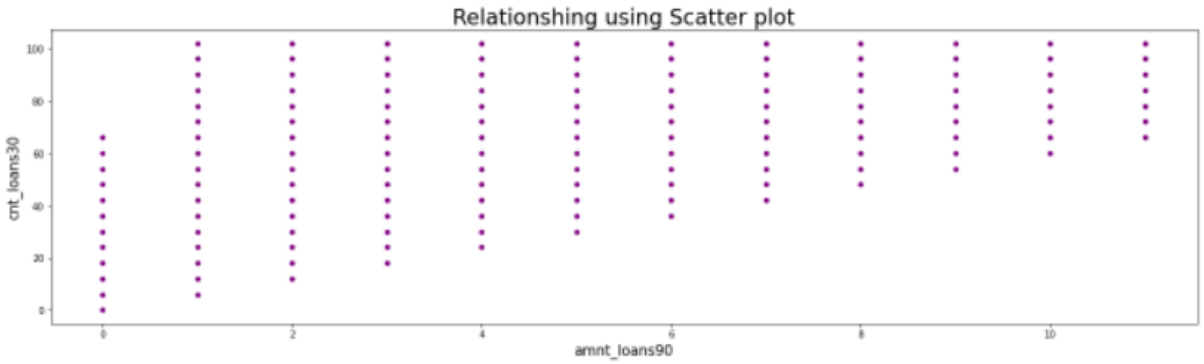


We can see the clear up trend here. Both features are tightly correlated with each other.

We can see the clear up trend here. Both features are tightly correlated with each other.

`cnt_loans30` and `amnt_loans90`

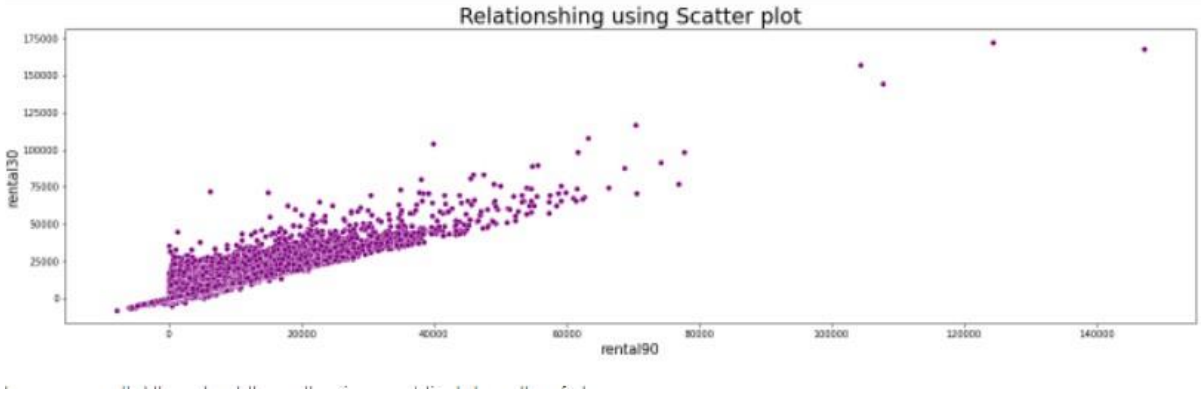
```
FeatureVsFeature('cnt_loans30','amnt_loans90')
```



Yes, here we can see the great up trend. So there is high correlation between these two features.

`Rental30 vs Rental90`

```
FeatureVsFeature('rental30','rental90')
```



Here, we can see that the up-trend. It mean there is some relation between these features.

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

We have taken many presumption in order to clean the data, pre- processing data and in machine learning. At first we found some column like unnamed:0 mssidn & pcircle and we assume based on the data inputs in it that these columns are of no use hence we dropped them.

After that I separately checked relation between each of feature with our target variable and I found some features are useless i.e. those features did not showing any relation with our target. So we are dropping allof them.

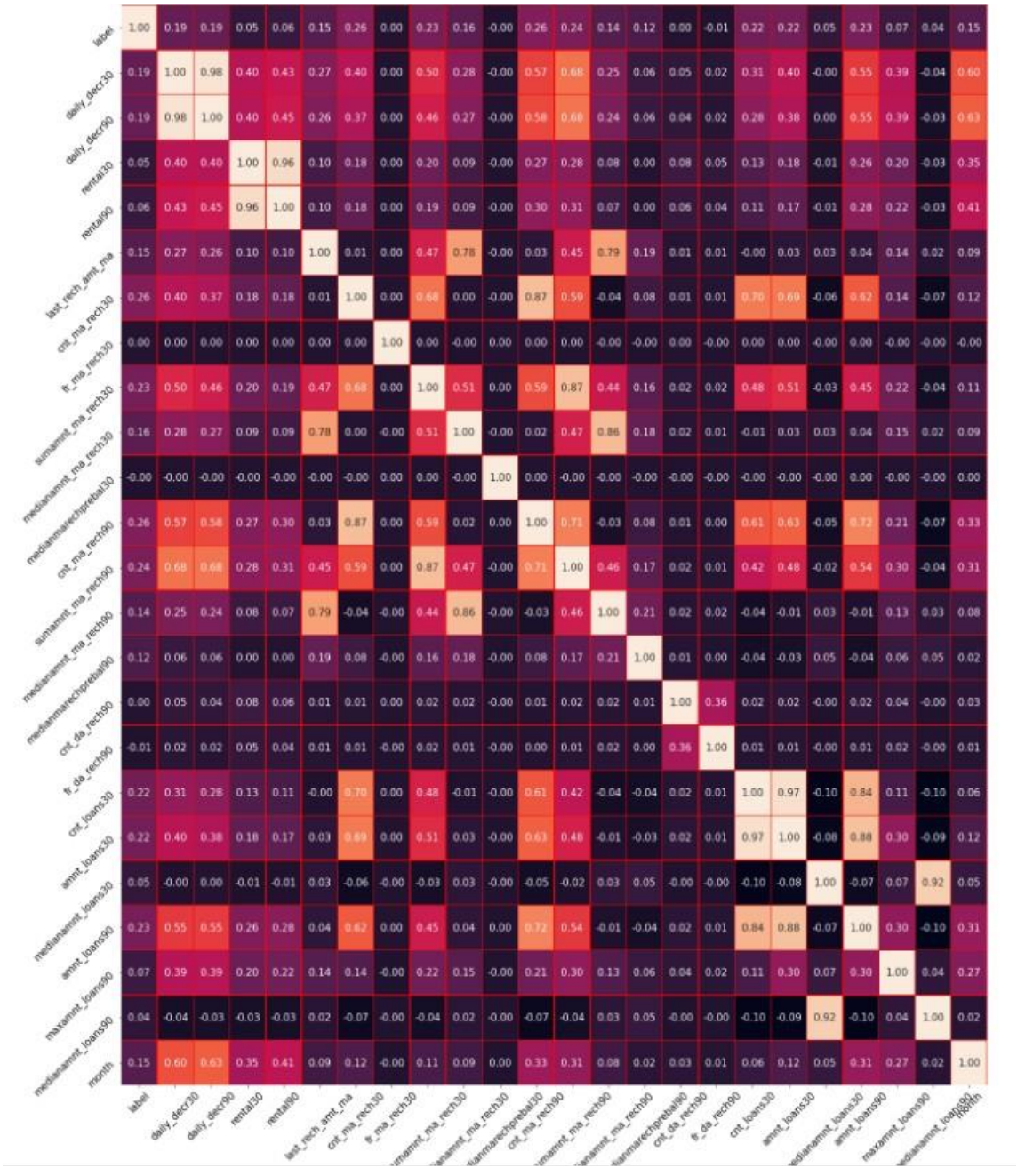
Conclusion:

based on above analysis(Relationship With Target) we find that some features did not add any value for predicting the label. So we will drop those all features. Those features are listed below:-

- aon
- last_rech_date_ma
- last_rech_date_da
- fr_ma_rech90
- cnt_da_rech30
- fr_da_rech30
- maxamnt_loans30
- cnt_loans90
- payback30
- date
- payback90

```
df=df.drop(columns=['aon','last_rech_date_ma','last_rech_date_da','fr_ma_rech90',
                    'cnt_da_rech30','fr_da_rech30','maxamnt_loans30','cnt_loans90','payback30','date','payback90'])
```

After that I plot a Heat map, to find the relation between feature andfeature.



Observation from heat map:

- we see that fr_da_rech90 is negative correlated with the target.
- daily_decr90 and daily_decr30 is highly 98% correlated with each other.
- medianamnt_loans90 and medianamnt-loans30 is also highly correlated with 92% correlation.
- amnt_loans30 and amnt_loans90 is also highly correlated with 88% correlation.
- cnt_loans30 and amnt_loans90 also 84% correlated with each other.
- 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.

Hardware and Software Requirements and Tools Used:-

The Micro Credit Defaulter Project is about built a machine learning model that could predict the default case so that the company could to whom they should provide the loan amount and whom they should not.
So for that we require lots of libraries and packages to work upon thisproject.

Hardware Required:-

- Processor:- Core i5 or above
- Ram:- 8GB or above
- SSD:- 256GB or Above

Software Required:-

- Anaconda Promt/Nevigator

Libraries Required:-

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn.preprocessing.StandardScaler
- Sklearn.preprocessing.LabelEncoder
- Sklearn.preprocessing.StandardScaler
- from sklearn.model_selection import train_test_split,cross_val_score
- sklearn.linear_model.LogisticRegression
- sklearn.neighbors.KNeighborsClassifier
- Sklearn.ensemble.RandomForestClassifier,AdaBoostClassifer,GradientBoostingClassifier

Model/s Development and Evaluation

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

During the statistical analyzation of the data distribution we found data washeavily skewed data and had lots of outliers and we solved it by applying power transformation technique of Scikit learn on the data and we were succeeded significantly in doing so. We also found some not useful column in the data set and we removed them with the help of pandas. We had sorted theDate column with the help of pandas. We had used ROC plot to evaluate the best model and its selection. To Hyper tune the model we used Scikit learns GridsearchCV method.

Testing of Identified Approaches (Algorithms):-

As per the Micro Credit default use case demanded the prediction of the default case, we analyzed the data and found that the problem is of Supervised Machine Learning Classification problem. Hence we decided to use the following algorithms to build the model for the use case:

- Logistic Regression
- K-Nearest Neighbors Classification
- Decision Tree Classifier
- Random Forest Classifier
- AdaBoost Classifier
- Gradient Boosting Classifier
- Extra Tree Classifier
- Naïve Bays

Run and Evaluate selected models:-

Logistic Regression:-

```
**********Testing Scores**********

Accuracy score for testing is :  0.7756916192026038

 F1 Score for testing is :  0.8583903735761342
Confusion Matrix :
 [[ 4712  1752]
 [ 9275 33421]]

 The Classification report for Testing
              precision    recall  f1-score   support

           0       0.34      0.73      0.46      6464
           1       0.95      0.78      0.86     42696

    accuracy                           0.78     49160
   macro avg       0.64      0.76      0.66     49160
weighted avg       0.87      0.78      0.81     49160
```

Decision Tree Classifier:-

```
**********Testing Scores**********

Accuracy score for testing is :  0.8476810414971522

 F1 Score for testing is :  0.9101058848951955
Confusion Matrix :
 [[ 3767  2683]
 [ 4805 37905]]

 The Classification report for Testing
              precision    recall  f1-score   support

           0       0.44      0.58      0.50      6450
           1       0.93      0.89      0.91     42710

    accuracy                           0.85     49160
   macro avg       0.69      0.74      0.71     49160
weighted avg       0.87      0.85      0.86     49160
```

Random Forest Classifier:-

```
**********Testing Scores**********

Accuracy score for testing is :  0.8972742066720911

 F1 Score for testing is :  0.9411394470604689
Confusion Matrix :
 [[ 3737  2777]
 [ 2273 40373]]

 The Classification report for Testing
              precision    recall  f1-score   support

           0       0.62      0.57      0.60      6514
           1       0.94      0.95      0.94     42646

    accuracy                           0.90     49160
   macro avg       0.78      0.76      0.77     49160
weighted avg       0.89      0.90      0.90     49160
```

AdaBoost Classifier:-

```
**********Testing Scores**********

Accuracy score for testing is :  0.8245321399511798

 F1 Score for testing is :  0.8920616647479854
Confusion Matrix :
 [[ 4889  1625]
 [ 7001 35645]]

 The Classification report for Testing
              precision    recall  f1-score   support

           0       0.41      0.75      0.53      6514
           1       0.96      0.84      0.89     42646

    accuracy                           0.82     49160
   macro avg       0.68      0.79      0.71     49160
weighted avg       0.88      0.82      0.84     49160
```

K-Nearest Neighbors Classifier:-

```
**********Testing Scores**********

Accuracy score for testing is :  0.8105370219690805

 F1 Score for testing is :  0.8834176137786012
Confusion Matrix :
 [[ 4557  1957]
 [ 7357 35289]]

 The Classification report for Testing
              precision    recall  f1-score   support

           0       0.38      0.70      0.49      6514
           1       0.95      0.83      0.88     42646

    accuracy                           0.81     49160
   macro avg       0.66      0.76      0.69     49160
weighted avg       0.87      0.81      0.83     49160
```

Gradient Boosting Classifier:-

```
**********Testing Scores**********

Accuracy score for testing is :  0.8620626525630594

 F1 Score for testing is :  0.9175351761543982
Confusion Matrix :
 [[ 4655  1867]
 [ 4914 37724]]

 The Classification report for Testing
              precision    recall  f1-score   support

           0       0.49      0.71      0.58      6522
           1       0.95      0.88      0.92     42638

    accuracy                           0.86     49160
   macro avg       0.72      0.80      0.75     49160
weighted avg       0.89      0.86      0.87     49160
```

Naïve Bayes Classifier:-

```
**********Testing Scores**********

Accuracy score for testing is :  0.780187144019528

 F1 Score for testing is :  0.8619359125057494
Confusion Matrix :
 [[ 4623  1899]
 [ 8907 33731]]

 The Classification report for Testing
              precision    recall  f1-score   support

           0       0.34      0.71      0.46      6522
           1       0.95      0.79      0.86     42638

    accuracy                           0.78     49160
   macro avg       0.64      0.75      0.66     49160
weighted avg       0.87      0.78      0.81     49160
```

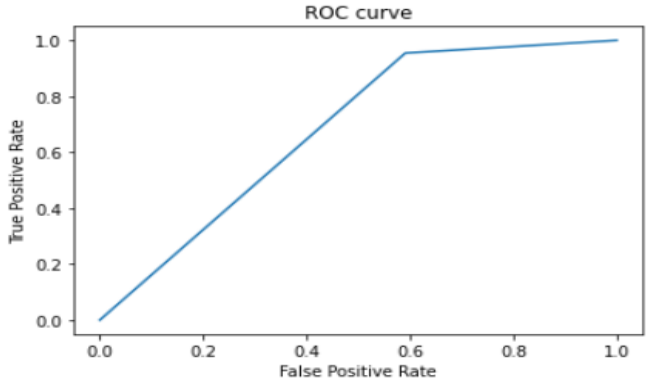
#Extra Tree Classifier:-

```
**********Testing Scores**********

Accuracy score for testing is :  0.8963791700569569

 F1 Score for testing is :  0.9408087380897049
Confusion Matrix :
 [[ 3583  2842]
 [ 2252 40483]]

 The Classification report for Testing
              precision    recall  f1-score   support

           0       0.61      0.56      0.58      6425
           1       0.93      0.95      0.94     42735

    accuracy                           0.90     49160
   macro avg       0.77      0.75      0.76     49160
weighted avg       0.89      0.90      0.89     49160
```

Key Metrics for success in solving problem under consideration:-

To evaluate the Machine Learning algorithms we mainly used almost all the classification metrics evaluation in this project. Our focus was on mainly accuracy score of the model, precision, recall, f1 score and the Roc-Auc curveof the all the models. Mainly focus was on accuracy score and we compared accuracy score with the cross validation score and prioritize the minimum difference model as a best fit model and later we evaluated the models with ROC Curve and whichever algorithm had the maximum area under the curvewe finalized that model for our project.

```
from sklearn.metrics import roc_curve,auc
fpr, tpr, threshold= roc_curve(y_pred,y_test)
plt.plot(fpr, tpr)
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC curve")
plt.show()
print("AUC value is {} ".format(auc(fpr, tpr)))
```



1. **Cross Validation:**

   Cross-validation helps to find out the over fitting and under fitting of the model. In the cross validation the model is made to run on different subsets of the dataset which will get multiple measures of the model. If we take 5 folds, the data will be divided into 5 pieces where each part being 20% of the full dataset. While running the Cross-validation the 1st part (20%) of the 5 parts will be kept out as a holdout setfor validation and everything else is used for training data. This way we will get the first estimate of the model quality of the dataset. In the similar way further iterationsare made for the second 20% of the dataset is held as a holdout set and remaining 4 parts are used for training data during the process. This way we will get the second estimate of the model quality of the dataset. These steps are repeated during thecross-validation process to get the remaining estimate of the model quality.

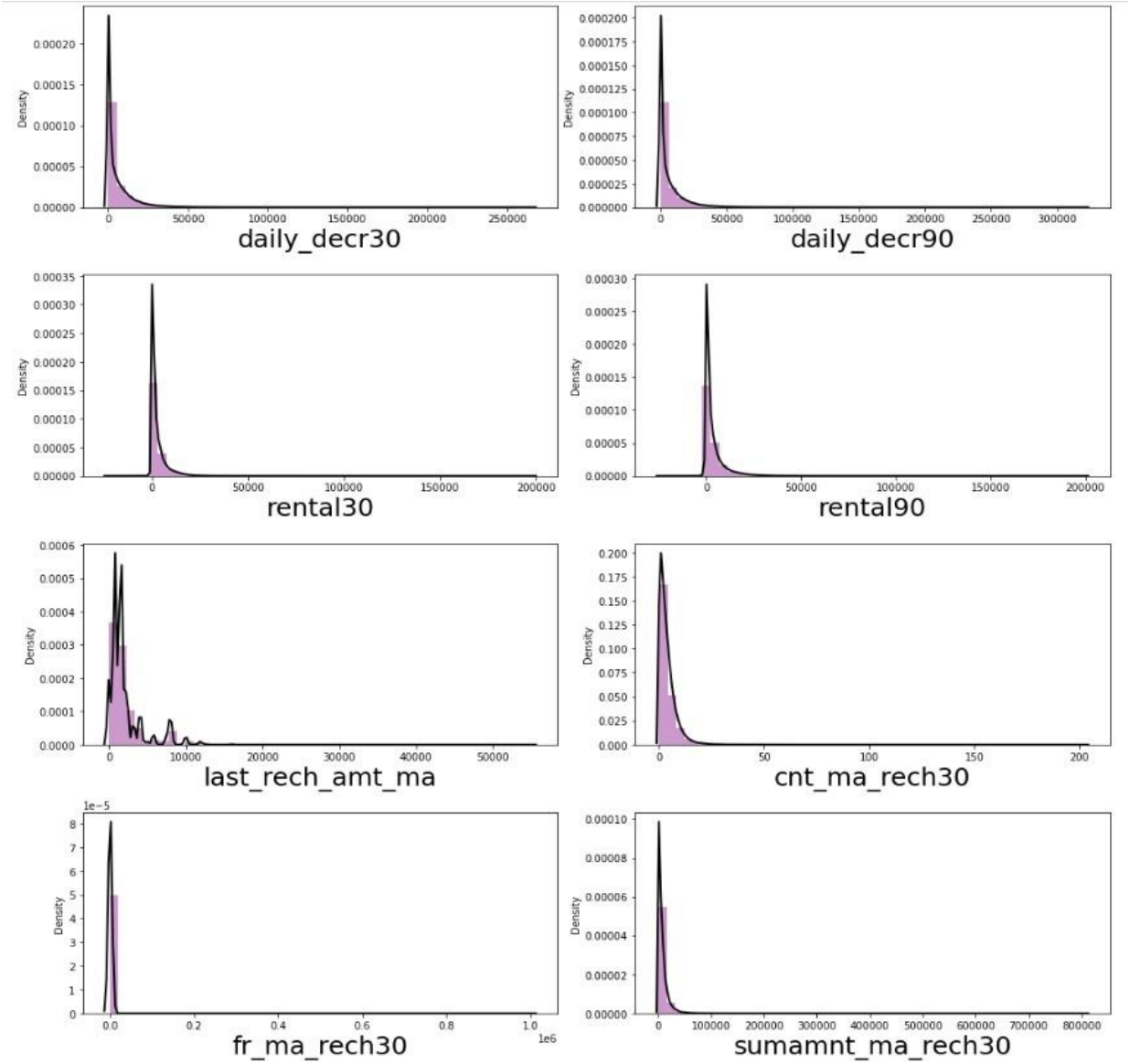
2. **Confusion Matrix:**

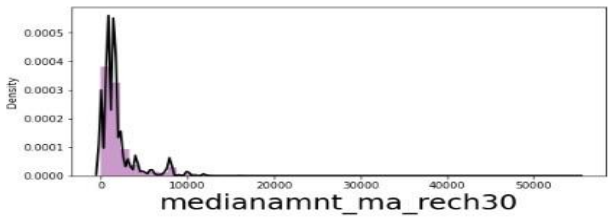
   A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one(in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class, while each column represents the instances inan actual class (or vice versa). The name stems from the fact that it makes it easy to see whether the system is confusing two classes (i.e., commonly mislabeling one as another). It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions.

```
Confusion Matrix :
 [[ 4820  1694]
 [ 6969 35677]]
True positives are :-  4820
False positives are :-  1694
False negatives are :-  6969
True negatives are :-  35677
```

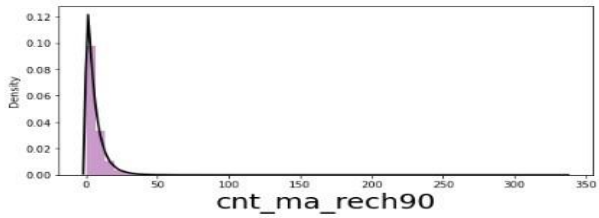
Visualizations:-

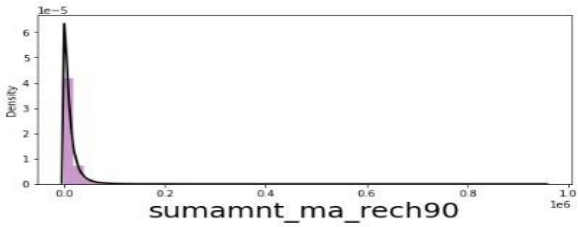
Checking Distribution Through Graph:-

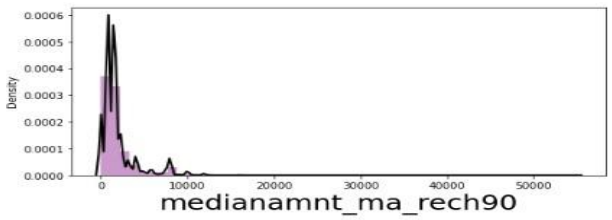



Density
medianamnt_ma_rech30

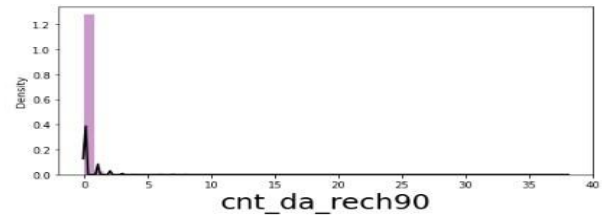

1e−5
Density
medianmarechprebal30
1e6

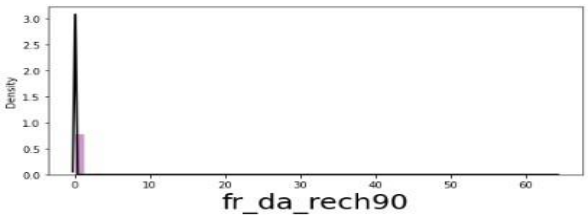

Density
cnt_ma_rech90

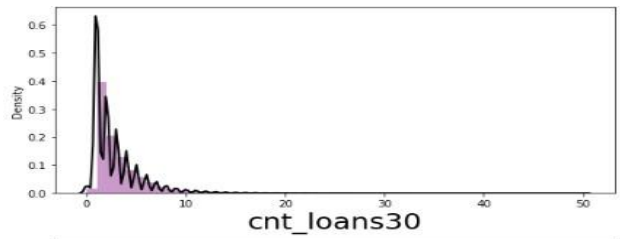

1e−5
Density
sumamnt_ma_rech90
1e6

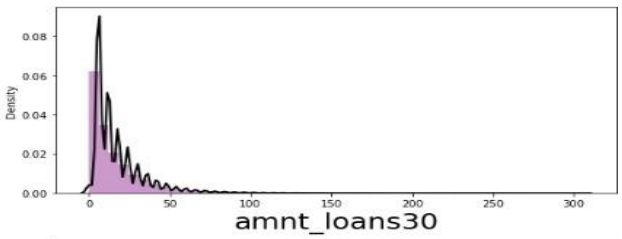

Density
medianamnt_ma_rech90

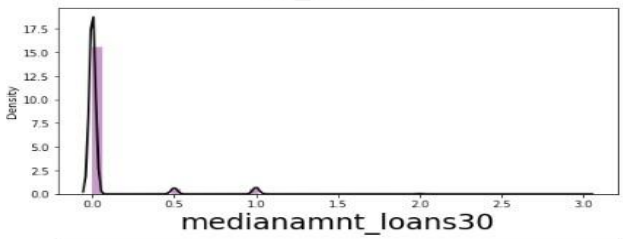

Density
medianmarechprebal90

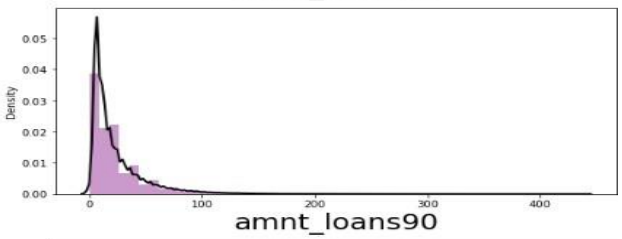

Density
cnt_da_rech90

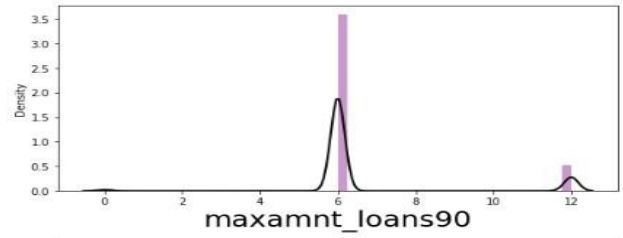

Density
fr_da_rech90

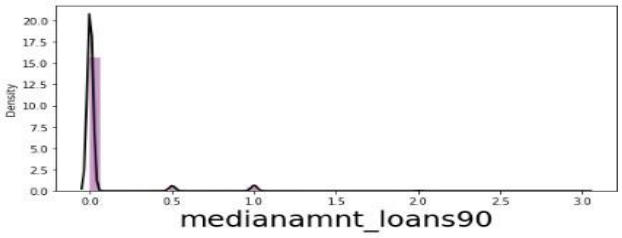

Density
cnt_loans30

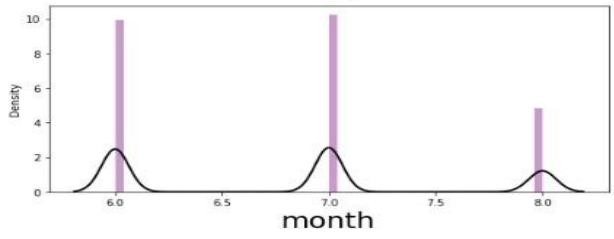

Density
amnt_loans30


Density
medianamnt_loans30


Density
amnt_loans90


Density
maxamnt_loans90


Density
medianamnt_loans90

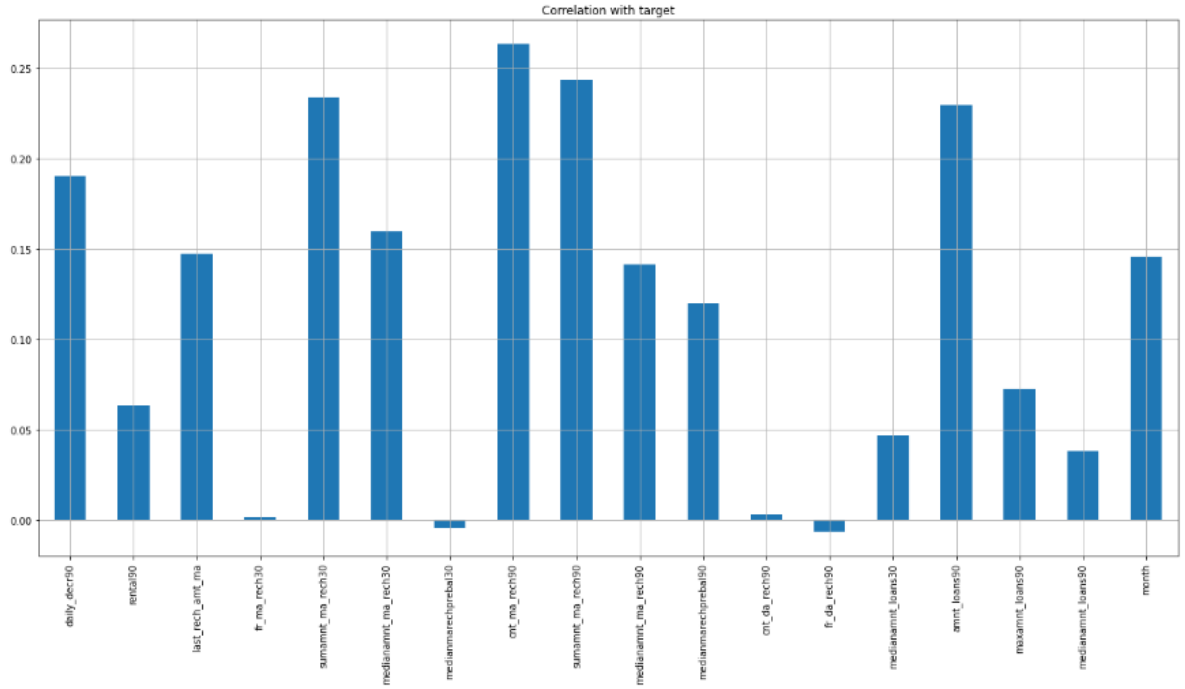

Density
month

- daily_decr30,daily_decr90,last_rech_amt_ma,cnt_ma_rech30,fr_ma_rech30,sumamnt_ma_rech30 features are highly right skewed.
- rental30,rental90 features are both right and left skewed.
- All Features are highly right skewed. There may be outliers present in my data set we will check it in further steps.
- we observe that almost all the features are highly right skewed or some are both right or left skewed. We have to handle this skewness otherwise our model become biased.
- month, medianamnt_loans90,maxamnt_loans90,maxamnt_loans30 features having nomial data points, so we cannot treat them in outliers checking.

Relation With Target:-

Relation With Target:

```
df_new.drop('label',axis=1).corrwith(df_new.label).plot(kind='bar',grid=True,figsize=(17,10),title='Correlation with target')
plt.tight_layout()
```



- cnt_ma_rech90 is high correlated with target.
- fr_ma_rech30 is least correlated with target.

Interpretation of the Results

From the Visualization of the data the results were interpreted that pcircle column has only 1 type of data so we interpreted no use of the neutral column, we interpreted that the Label column has class imbalanced data, data distribution in all the continuous column were highly skewed and have lots of outliers. In the pre-processing the skewness in the data set was detected and we interpreted that removing of outliers would lead to heavy data loss so we resolved it by the power transformation method; we used yeo-Johnson method because during the pre- processing the 0 was detected in most of the continuous column.

8 Algorithms were chosen based on the data analysis of the data to build the Machine Learning models based on the data structure and the nature of data distribution in the feature columns. After evaluating the models by metrics evaluators we interpreted all the model as very good hence we decided to choose the best model based on the minimum difference between model accuracy and the cross validation accuracy, but later we found a very much fluctuation in the precision, recall & f1 score so we decided to use Roc plot to find out the which model has maximum area under the curve and we got Random Forest Model as our best model.

CONCLUSION

Key Findings and Conclusions of the Study

From Whole Dataset and Analysis we find some conclusions:

- Dataset had 209590 records of telecom users with 36 types ofthe services related data were present in the data.
- Dataset has 3 object data type columns 12 integer data type and restof the features is float data type.
- There are no null values in the dataset.
- There may be some customers with no loan history.
- The dataset is imbalanced. Label '1' has approximately 87.5%records, while, label '0' has approximately 12.5% records.
- For some features, there may be values which might not be realistic. You may have to observe them and treat them with asuitable explanation.
- You might come across outliers in some features which you need to handle as per your understanding. Keep in mind thatdata is expensive and we cannot lose more than 7-8% of the data.

Learning Outcomes of the Study in respect of Data Science:-

In my opinion, the outcomes demonstrate that IMF is making a positive impact on the lives of the citizens in the world who are particularly from weaker sections. During the study I found out thatthe bank has been allowing and supporting Group formation, encouraging savings and monitoring the inter-loaning structure.

This model will be good way for the management to understand whether the customer will be paying back the loaned amount within 5 days of insurance of loan. The relationship between predicting defaulter and the economy in an important motivating factor forpredicting micro credit defaulter.

Limitations of this work and Scope for Future Work:-

- Due to the time limit, it is not possible to conduct a thorough study and have a deep understanding of the dataset. There are still manyfeatures in the dataset that are unused and a lot of the informationhas not been fully digested with knowledge in the banking industry.
- Only the Random Forest Classifier model is giving me the best accuracy, but there are many good ones out there, such as LogisticRegression or Extra Tree Classifier even neural networks. The models can also be improved further by finer tunings on hyper parameters or using ensemble methods such as bagging, boosting, and stacking.
- It is important to notice this fact that the default loans are only about 10% of the total loans, thus during the training process, the model will favor predicting more negatives than positive results. We have already used the F1 score and ROC AUC instead of just accuracy. However, the performance is still not as good as it could be.In order to solve this problem, other methods such as collecting or resampling more data can be used in the future.

Thank You...!!!
Jessica Ghimeliya
September-2022