­

NAME OF THE PROJECT

MICRO CREDIT LOAN CASE

**Submitted by:**

Shama Tanweer

**ACKNOWLEDGMENT**

My sincere thanks to the Youtube channel CodeBasics, Blog from Medium,Stackoverflow for the completion of this projects.

**INTRODUCTION**

* **Business Problem Framing**

In this casestudy one of the MFI(micro finance institution) collaborated with one of the telecom operator to provide the small mobile recharge loans to the customer, these customer are of economical background and of remote areas .

The task is to predict on some given parameters, whether a customer will pay the micro-credit availed in a particular time.

This is a binary classification problem where 0 stands for the customer has failed to pay the credit and 1 stands for customer has successfully paid the credit back.

* **Conceptual Background of the Domain Problem**

A individual should have a domain knowledge of the financial concepts like,Credit,Debit, .Also the domain knowledge of Telecom industry like what could be the average balance in the customer’s account, Data recharge,network area circle,customer density.

A binary classification logistic regression will be the first approach.

Further we will test other algorithms to check for better accuracy and metrics.

* Review of Literature

I have examined the following article and Newspaper for the research Purpose.

**Article by**

* **Financial Inclusion-** Financial inclusion s to deliver financial services to the unbanked and under banked class of society in affordable cost.
* **Airtel payment bank**- Airtel Payments Bank, India‟s first payments bank, has enabled 100 villages across Tamil Nadu to go cashless as part of its endeavour to take its banking services deep into rural/unbanked areas and contribute to financial inclusion in the country. These villages now have access to basic banking services and the option of making digital payments, making them less reliant on cash.
* **Motivation for the Problem Undertaken**

The main motivation for the problem was to select the high valued customer which will increase the productivity and profitability of a organisation, also this model can be utilized for the fraud detection in Banking sector to,high valued bank loans.

Objective behind this project is to help Service Providers determine what amount of consumers will pay back the credit in a given amount of time.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

1.we used **describe** function,which tells us all the mathematical information about the numerical column

2. Mean,median,mode,Q1,Q2, Min, Max- methods are used to understand the skewness, , standard deviation and outliers.

3.Statistical model used- **correlation** amongst the columns and target

variable to understand the correlation between different features..

4.**Boxplot**=We have used boxplot for outlier detection virtually

5.We have removed the outliers by **Zscore which** is used to limit the data within required standard deviation

* **Data Sources and their formats**

**data sources**- Fliprobo Technology

**origin-** Microfinance bank

**formats-**Numerical, Date Format and categorical

**columns description:**

1.**label** : Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}

2.**msisdn** : mobile number of user

3.**aon** : age on cellular network in days

4.**daily\_decr30**: Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)

5.**daily\_decr90**: Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)

6.**rental30:** Average main account balance over last 30 days

7.**rental90**: Average main account balance over last 90 days

8.**last\_rech\_date\_ma**: Number of days till last recharge of main account

9.**last\_rech\_date\_da:** Number of days till last recharge of data account

10**.last\_rech\_amt\_ma**: Amount of last recharge of main account (in Indonesian Rupiah)

11.**cnt\_ma\_rech30**: Number of times main account got recharged in last 30 days

12.**fr\_ma\_rech30**: Frequency of main account recharged in last 30 days

13.**sumamnt\_ma\_rech30**: Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)

14.**medianamnt\_ma\_rech30**: Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)

15.**medianmarechprebal30**: Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)

16.**cnt\_ma\_rech90**: Number of times main account got recharged in last 90 days

17.**fr\_ma\_rech90**: Frequency of main account recharged in last 90 days

18.**sumamnt\_ma\_rech90** : Total amount of recharge in main account over last 90 days (in Indian Rupee)

19.**medianamnt\_ma\_rech90**: Median of amount of recharges done in main account over last 90 days at user level (in Indian Rupee)

20.**medianmarechprebal90:** Median of main account balance just before recharge in last 90 days at user level (in Indian Rupee)

21.**cnt\_da\_rech30**: Number of times data account got recharged in last 30 days

22.**fr\_da\_rech30**: Frequency of data account recharged in last 30 days

23.**cnt\_da\_rech90**: Number of times data account got recharged in last 90 days

24.**fr\_da\_rech90**: Frequency of data account recharged in last 90 days

25.**cnt\_loans30**: Number of loans taken by user in last 30 days

26.**amnt\_loans30**: Total amount of loans taken by user in last 30 days

27.**maxamnt\_loans30**: maximum amount of loan taken by the user in last 30 days

28.**medianamnt\_loans30**: Median of amounts of loan taken by the user in last 30 days

29.**cnt\_loans90**: Number of loans taken by user in last 90 days

30**.amnt\_loans90**: Total amount of loans taken by user in last 90 days

31.**maxamnt\_loans90**: maximum amount of loan taken by the user in last 90 days

32.**medianamnt\_loans90**: Median of amounts of loan taken by the user in last 90 days

33.**payback30**: Average payback time in days over last 30 days

34.**payback90**: Average payback time in days over last 90 days

35.**pcircle:** telecom circle

36.**pdate:** date

.

* **Data Preprocessing Done**

**Columns dropped**

Unnamed,Pcircle,msidn and PDate were the first columns to be dropped as :

1.PCircle seems to have only one unique value

2.Pdate isn’t defined well as which date it refers to.

3.Msidn .ie mobile no. of user doesn’t really affect ,if the person will be defaulter

Further checked the correlation of all columns with the label and the least negatively correlated columns were dropped.

Further negative value is present in some features,which is not even possible for most of the features.

Therefore except rental30 & rental90 ,negative value in all the features were converted to positive value by using absolute(abs) function.

Anomalies present in data was also checked & all the unrealistic data was removed

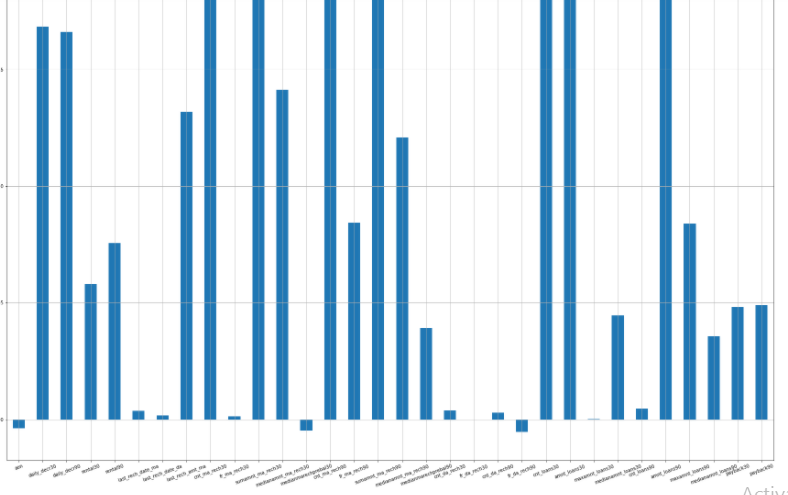
The data description revealed that there is a large difference between the 3rd quantile and the maximum, which clearly indicates the presence of outliers in the datset .

Outliers present in dataset was removed by Zscore,But when we apply zscore of 3,20% of the data was lost(due to widespread of data).We compromised by enlarging the zscore cut-off to 5 resulting in a loss of few data

**Removing of skewness**- using square root and cube root method skewness is removed.

**Scaling of the data**- standard scaler is used to scale data

**PCA technique** is used for the analysis ,to reduce curse of dimensionality & at the same time minimizing information loss

* Data Inputs- Logic- Output Relationships
* The label is the output variable here, rest all are the input variables.
* The target is imbalanced, only about 12.5% of negative values while all other are positive values
* dailydecr30,dailydecr90,rental30&90,lastreach\_dat and ma, summant\_ma\_rech30,cnt\_loans30,amnt oans40,amnt\_loans90 are higly positively correlated

Last\_rech\_date\_ma,last\_rech\_date\_da,fr\_ma\_rech30,medianmarechprebal30,fr\_da\_rech30cnt\_da\_rech90,fr\_da\_rech90 are least correlated with the target column .ie label

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

We have assumed that ,negative value were mis printed,& converted them to postive

* **Hardware and Software Requirements and Tools Used**

Framework-Annconda

IDE-Jupyter –Notebook

Coding Language-Python

Hardware used: system memory 8GB,

Processor: 5th gen core i7

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Used the following process for problem solving

Data visualization- using Matplotlib,seaborn

Data cleaning-Zscore,boxplots,percentile

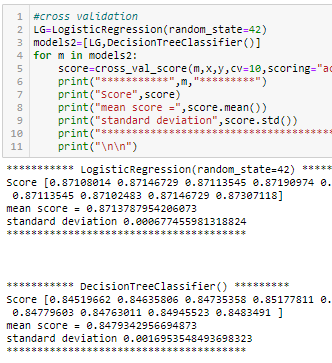
EDA-Correlation,dropping of columns,skewness treatment, PCA

Algorithm testing

Cross validation

Ensembling and boosting

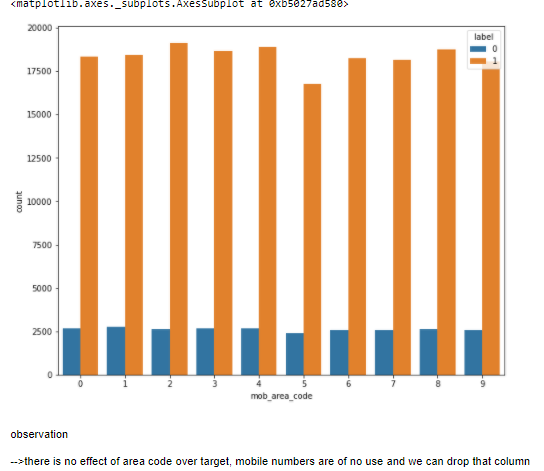
AUC ROC CURVE

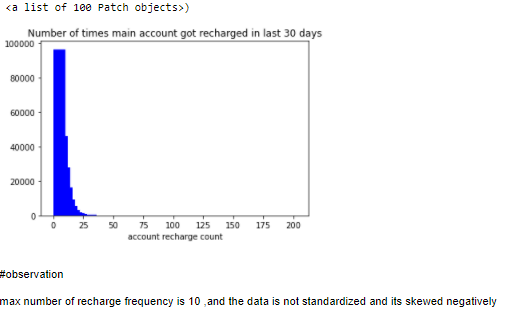
* Testing of Identified Approaches (Algorithms)
* from sklearn.linear\_model import LogisticRegression
* from sklearn.naive\_bayes import GaussianNB
* from sklearn.svm import SVC
* from sklearn.tree import DecisionTreeClassifier
* from sklearn.neighbors import KNeighborsClassifier
* from sklearn.model\_selection import cross\_val\_score
* from sklearn.ensemble import RandomForestClassifier
* from sklearn.ensemble import AdaBoostClassifier
* from sklearn.ensemble import GradientBoostingClassifier
* Evaluate selected models
* 

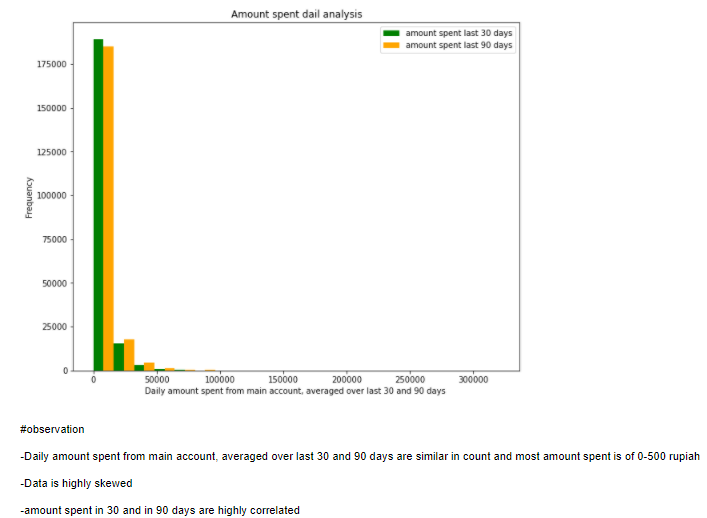
Logistic regression , Gaussian nb and knn was used to build the model, but the best result was provided by logistic regression and for further improving the result , i have used best parameters and cross validation which gave the result of 88 percent accuracy.

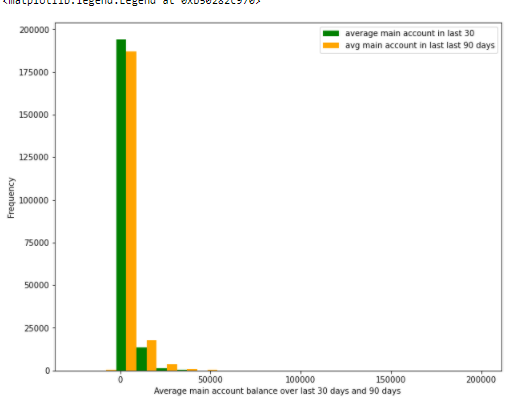
To boost the accuracy i have used boosting algorithms Randomforest classifier,GradientBoostingclassifiers, BaggingClassifier, ExtraTreeClassifier, amongst all best was randomforestclassifier which boosted accuracy to 90 percent.

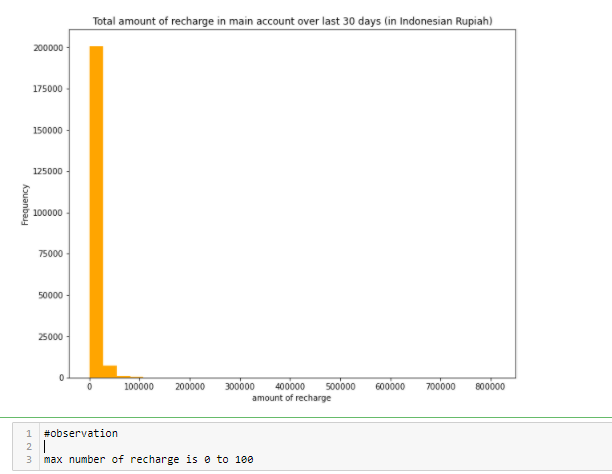
* Visualizations

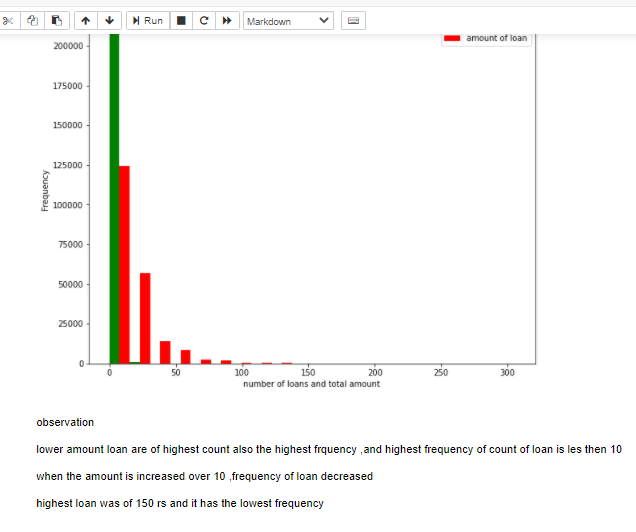


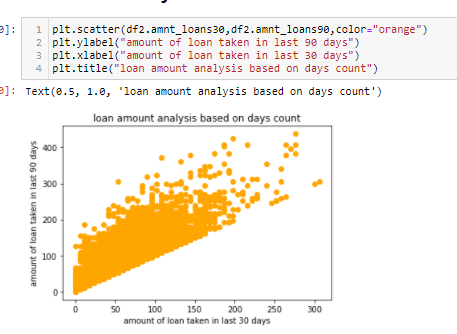


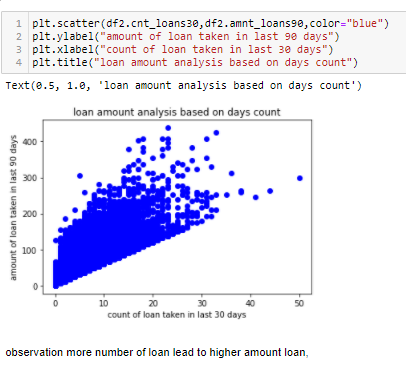


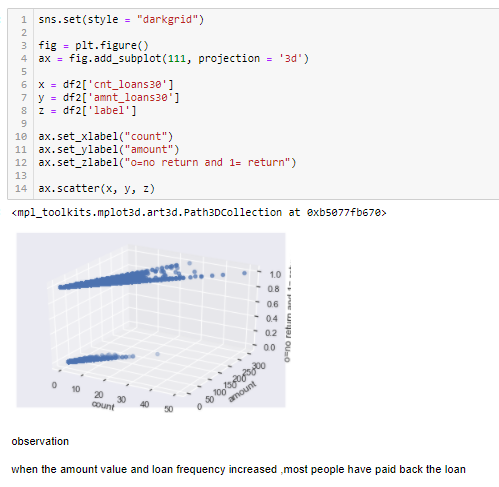


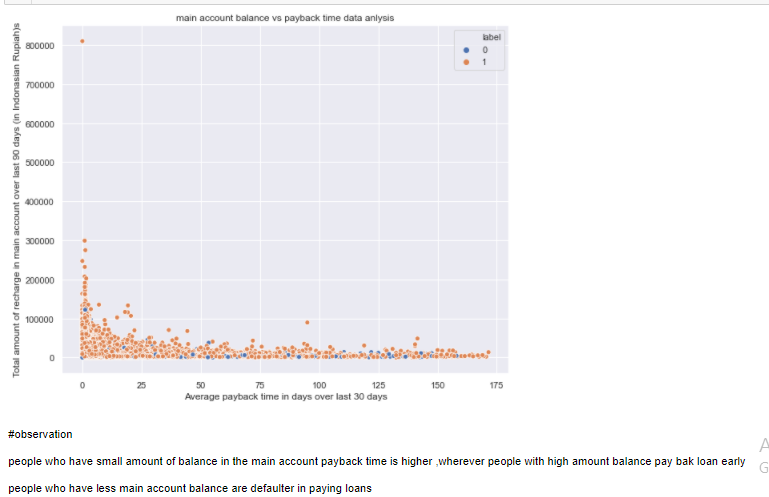


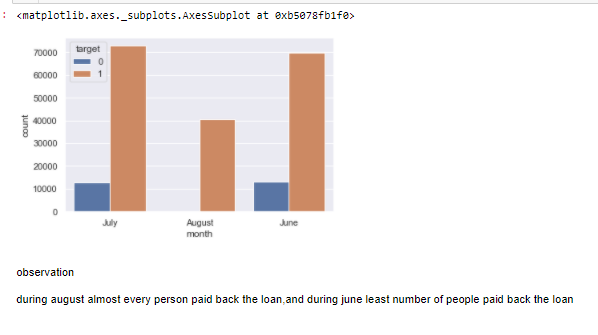
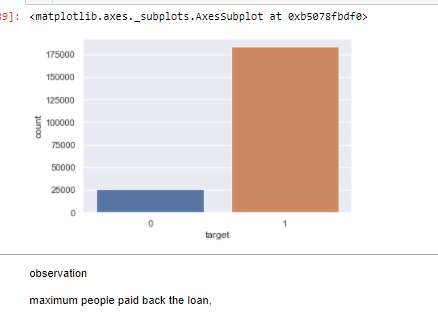










**CONCLUSION**

* Key Findings and Conclusions of the Study
* here is no effect of area code over target, mobile numbers are of no use and we can drop that column
* dailydecr30,dailydecr90,rental30&90,lastreach\_dat and ma, summant\_ma\_rech30,cnt\_loans30,amnt oans40,amnt\_loans90 are higly positively correlated
* max number of recharge frequency is 10 ,and the data is not standardized and its skewed negatively
* -Daily amount spent from main account, averaged over last 30 and 90 days are similar in count and most amount spent is of 0-500 rupiah
* -Data is highly skewed
* -amount spent in 30 and in 90 days are highly correlated
* max number of recharge is 0 to 100
* lower amount loan are of highest count also the highest frquency ,and highest frequency of count of loan is les then 10
* when the amount is increased over 10 ,frequency of loan decreased
* highest loan was of 150 rs and it has the lowest frequency
* more number of loan lead to higher amount loan,
* when the amount value and loan frequency increased ,most people have paid back the loan
* people who have small amount of balance in the main account payback time is higher ,wherever people with high amount balance pay bak loan early
* people who have less main account balance are defaulter in paying loans
* #those who take more number of loans are tend payback the loan early and less likely to be defaulter
* people taking less number of loan ,tends to take more days to return the loan also most people doesn't return loan .
* during august almost every person paid back the loan,and during june least number of people paid back the loan
* during august almost every person paid back the loan,and during june least number of people paid back the loan
* Learning Outcomes of the Study in respect of Data Science

During data cleaning i have used various data cleaning techniques like percentile and zscore and IQR and among all i choose the best for outlier removal which impact least our data.

PCA is always done after the standard scaling of data otherwise it won’t treat data accurately.

Challenges i faced was that i during outlier removal almost 50 percent data was getting removed so i used different mehod so that less data loss will happen. Also i have used clustering techniques to find best parameter so that it won’t take too much time

* Limitations of this work and Scope for Future Work

The limitation of this work that it will used for small data and won’t be used for big data.