

NAME OF THE PROJECT

MICRO CREDIT LOAN CASE

Submitted by:

PIYUSH THAKUR

**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 .Both of these companies want to improve the selection of customer for the credit so that maximum customer will payback the loan and minimisation of the defaulter customer.

* Conceptual Background of the Domain Problem

A individual should have a domain knowledge of the financial concepts like ROI(return on investment),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.

* Review of Literature

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

**Article by**

* **Dr Sonal Purohit & Rikke Mishra** “Payment Banks- A Revolutionary Step in INDIA for financial inclusion ”
* **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.
* Newspaper –**Economic times, Times and Hindu**
* 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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Statistical model used- **correlation** amongst the columns and target variable to understand the relation.

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

Boxplot= is used for outlier detection virtually

Zscore= is used to limit the data within required standard deviation and outlier removal

* Data Sources and their formats

**data sources**- Fliprobo Technology

**origin-** Microfinance bank

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

|  |  |
| --- | --- |
| 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 |

.

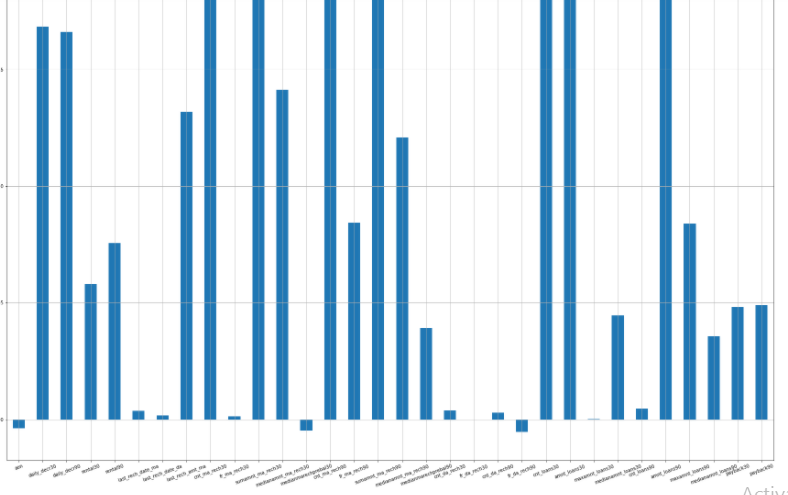
* Data Preprocessing Done

Columns dropped- rental30","rental90","fr\_ma\_rech30","fr\_ma\_rech90" these columns are dropped as these were ambiguous and doesn’t hold any logic which will help in impacting the target variable, also the **date** column which doesn’t impact target variables.

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

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

Minimising the columns number- PCA technique is used to minimise the columns.

* Data Inputs- Logic- Output Relationships
* dailydecr30,dailydecr90,rental30&90,lastreach\_dat and ma, summant\_ma\_rech30,cnt\_loans30,amnt oans40,amnt\_loans90 are higly positively correlated
* State the set of assumptions (if any) related to the problem under consideration

Here, you can describe any presumptions taken by you.

* Hardware and Software Requirements and Tools Used

Framework-Annconda

IDE-Jupyter –Notebook

Coding Language-Python

Libraries- import pandas as pd

import numpy as np

import seaborn as sns

import warnings

warnings.filterwarnings("ignore")

%matplotlib inline

import matplotlib.pyplot as plt

import scipy

from scipy.stats import zscore

from sklearn.preprocessing import StandardScaler

**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

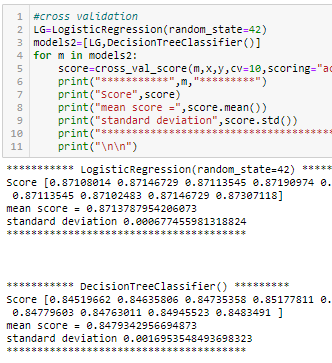
Best Parameter- finding best parameters from grid search cv

Algorithm testing

Cross validation

Ensembling and boosting

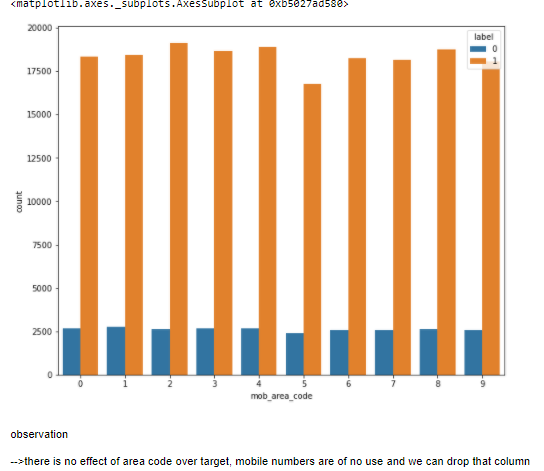
AUC ROC CURVE

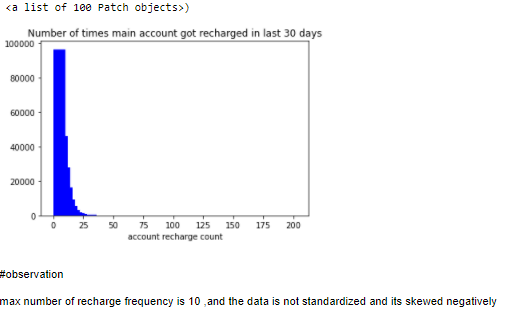
* Testing of Identified Approaches (Algorithms)
* from sklearn.linear\_model import LogisticRegression
* from sklearn.naive\_bayes import GaussianNB
* from sklearn.svm import SVC
* from sklearn.tree import DecisionTreeClassifier
* from sklearn.neighbors import KNeighborsClassifier
* from sklearn.model\_selection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

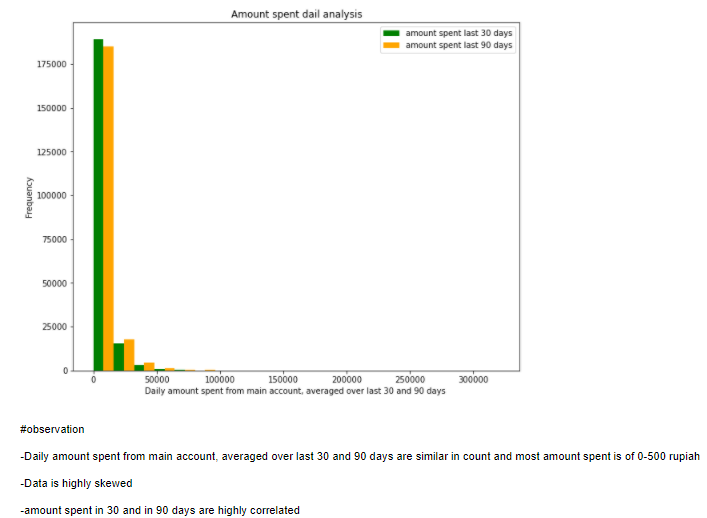
Logistic regression , Gaussian nb and knn was used to build the model, but the best result was provided by logistic regression and for further improving the result , i have used best parameters and cross validation which gave the result of 88 percent accuracy.

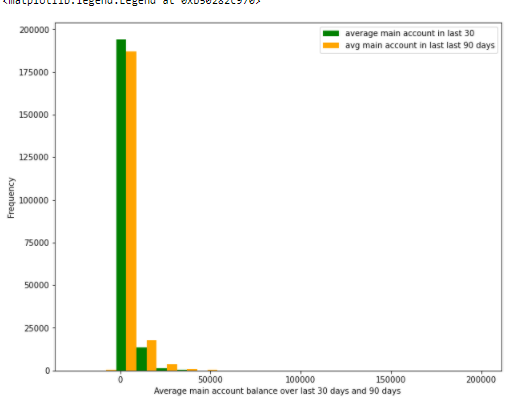
To boost the accuracy i have used boosting algorithms Randomforest classifier,GradientBoostingclassifiers, BaggingClassifier, ExtraTreeClassifier, amongst all best was randomforestclassifier which boosted accuracy to 90 percent.

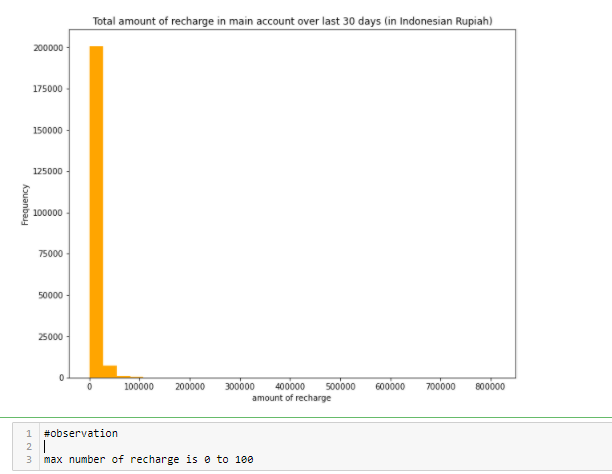
* Visualizations

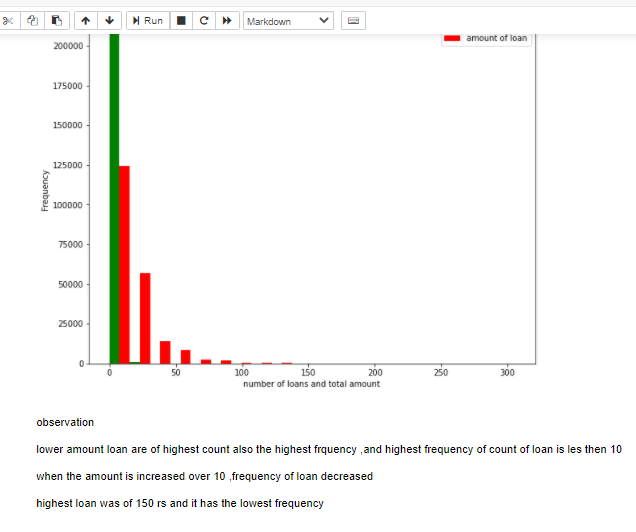


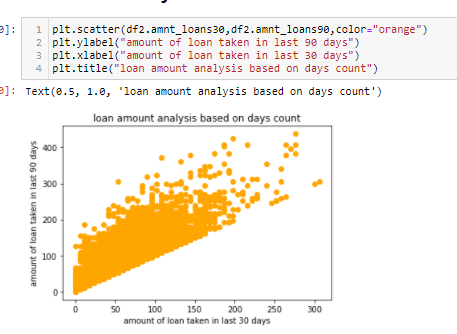


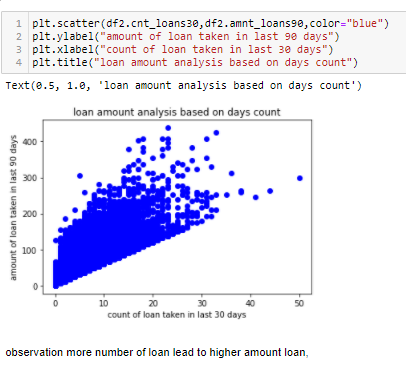


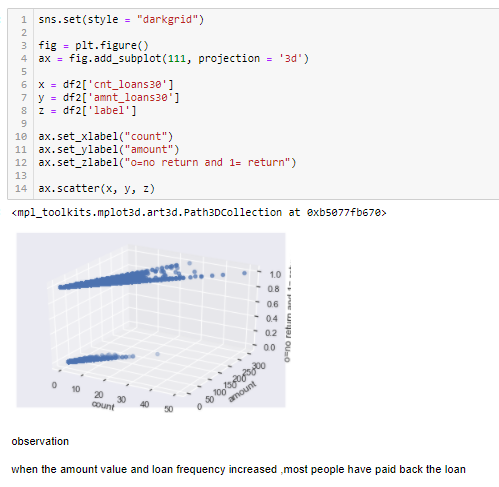


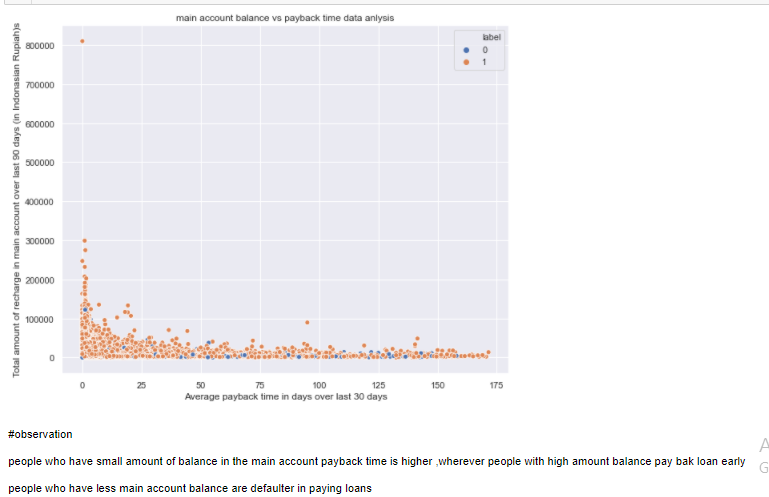


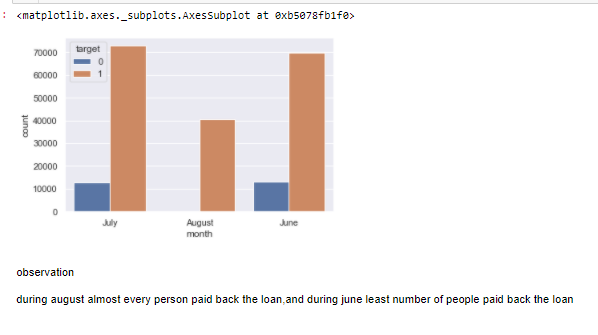
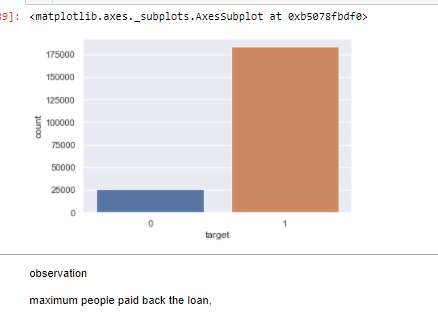










**CONCLUSION**

* 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.