

MICRO CREDIT LOAN USE CASE

Submitted by:

SUDHA UDAYAKUMAR

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

The client telecom company are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The business problem is that there are customers defaulting and are not paying back the micro-credit within 5 days.

In real world as well, banks and loan institutions face a similar issue and are finding it difficult to identify the right customer to whom they can loan and collect it back with default.

The client in this case was some predictions to be made to improve the selection of the customer for credit.

Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

Domain related concepts that will be useful are as follows.

* What is the loan history of the customer?
* Who are the customers who have repaid?
* What is the average payback time in the past 30 and 90 days time frame?
* What is the mean and median number of days for them to repay?
* What is the median of loan amount for all the cases when the amount was defaulted?
* What is the median of loan amount for all the cases when the amount was paid on time?

Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

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. The client, telecom provider is 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).

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

* This project will solve the problem of identifying the right customer for microloans.
* It will ensure the right customer gets the loan and is not rejected unnecessarily.
* This is a poverty reduction tool, and the right predictive tool will help MFIs to expand their model.

**Analytical Problem Framing**

* Data Sources and their formats

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

The sample data is provided to us from our client database

**Data description**

|  |  |
| --- | --- |
| 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 Format**

Unnamed: 0                int64

label                     int64

msisdn                   object

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

**Data snapshot**



* Data Pre-processing Done

1. Checked for missing data. Although missing data is not there, was informed in the project dataset it was checked again. There was no missing data
2. Data Shape: Number of rows to column was analysed. There were enough ratio of rows to columns to continue with the data
3. Balanced or Imbalanced Dataset. The target column “label” was analysed and found it was imbalanced as mentioned. So need to think of ways to balance the data before model building

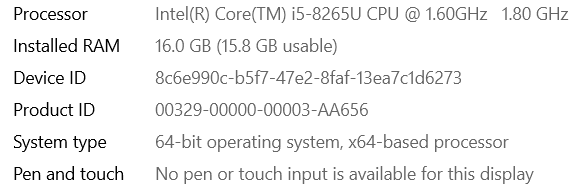
1 183431

0 26162

1. Unique count in each other column was analysed.
2. Multiple columns had huge number of unique count. So analysing these against the target variable wont be possible for bi-variate analysis.
3. Unwanted columns which don’t impact the data like the msisdn, date, p-circle etc were removed.
4. Correlation with the target variable was also done and printed in ascending order

* Hardware and Software Requirements and Tools Used

Using a good laptop



Jupyter notebook Loaded with packages and libraries from Pandas, Numby, Seaborn, Scikit Learn etc. For easy flow of the dataset used Table of Contents. Observations were highlighted in yellow and extra notes were highlighted in pink

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* IQR was done in detail and found several methods to remove outliers before model building.
* PCA was done to reduce the number of columns. We were able to retain 95% of data with 18 columns and 99% of data with 24 columns. So choose the 24 columns
* Since the data was imbalanced used the Over Sampling Method. With this method, we got the data to be balanced
* Testing of Identified Approaches (Algorithms)
* LogisticRegression
* GaussianNB
* SVC
* RandomForestClassifier
* AdaBoostClassifier
* DecisionTreeClassifier
* Run and Evaluate selected models

Since this was a imbalanced dataset , used the F1-score for model evaluation instead of accuracy. Then the precision & recall was also identified.

model=[lg,sv,gnb,rf,ad,dtc]

for m in model:

m.fit(x\_train,y\_train)

pred=m.predict(x\_test)

print("\n f1 score: of ",m)

print(f1\_score(y\_test,pred))

print("\n Classificatin report: of ",m)

print(classification\_report(y\_test,pred))

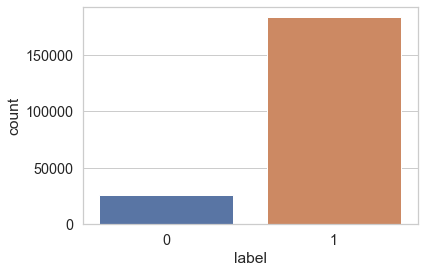
print("\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*")

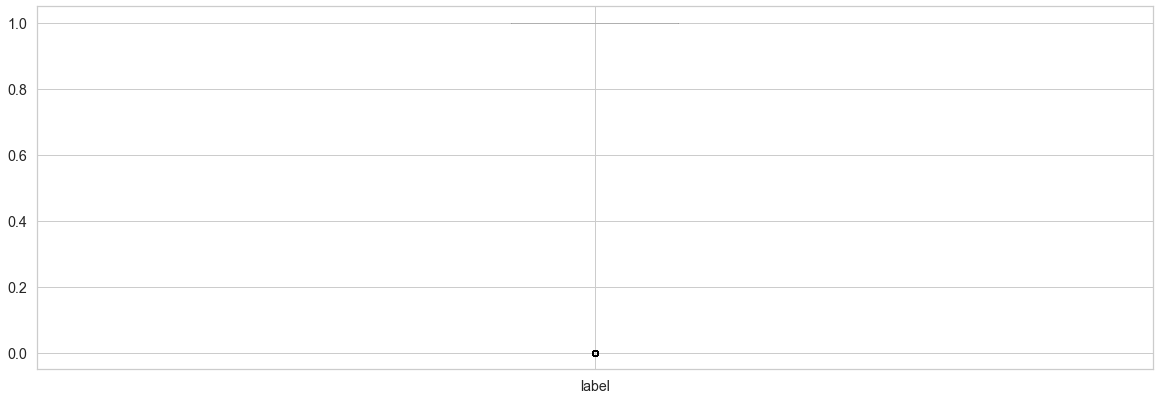
Used a for loop to run the code, then however it was taking long time. So ran each of the loop once. SVM, GNB took long time and in the practice notebook, the F1-score was less for both. So just copied the screen shots. Restarting them was an issue

Also initially had cross validation as well. Did cross validation of 10, which was high and taking time. So removed it from the for loop

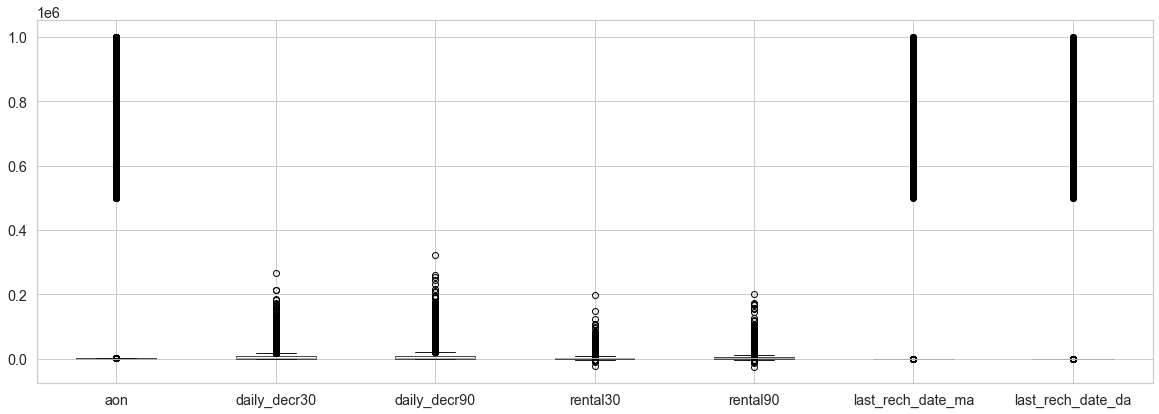
* Key Metrics for success in solving problem under consideration
* Key Metrics used : F1-Score to compare 0 & 1 value , as its an imbalanced dataset
* Also checked for precision and recall from the classification report
* Then did the cross validation with k fold of 5. The least diff between F1 score and Cross validation was taken
* Since the dataset is used, didn’t do cross validation on all models. Did only on top 2 models with high F1-score.
* Both got diff of 1 and selected RFC, as it had high F1-score
* After this Hyper Parameter Tuning was planned.
* Didn’t want to do with Grid Search, as it will take more time. Wanted to do with Random Search, but even that was not able to execute , due to issues in re-starting the notebook
* Visualizations

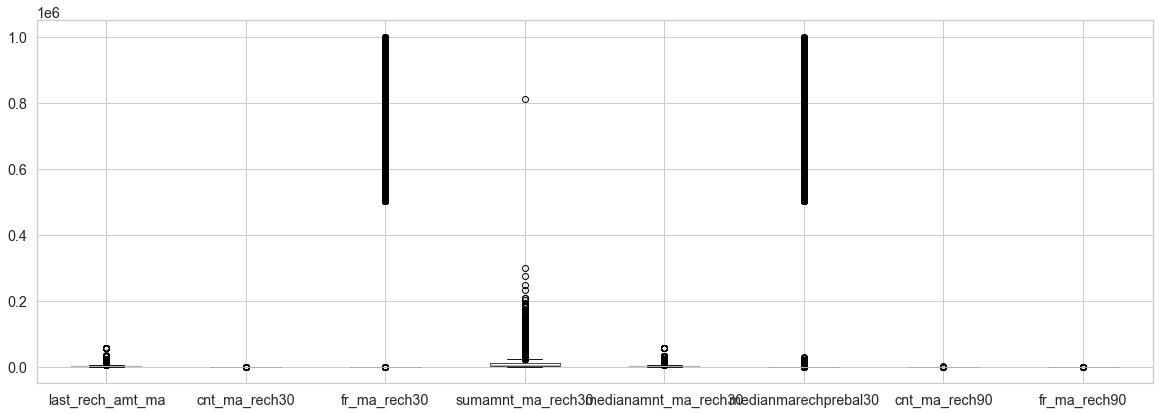
Target variable

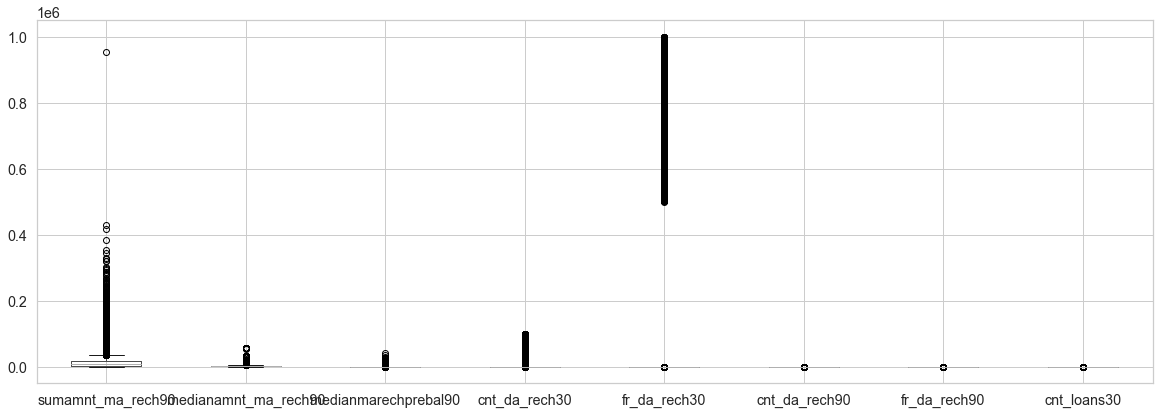


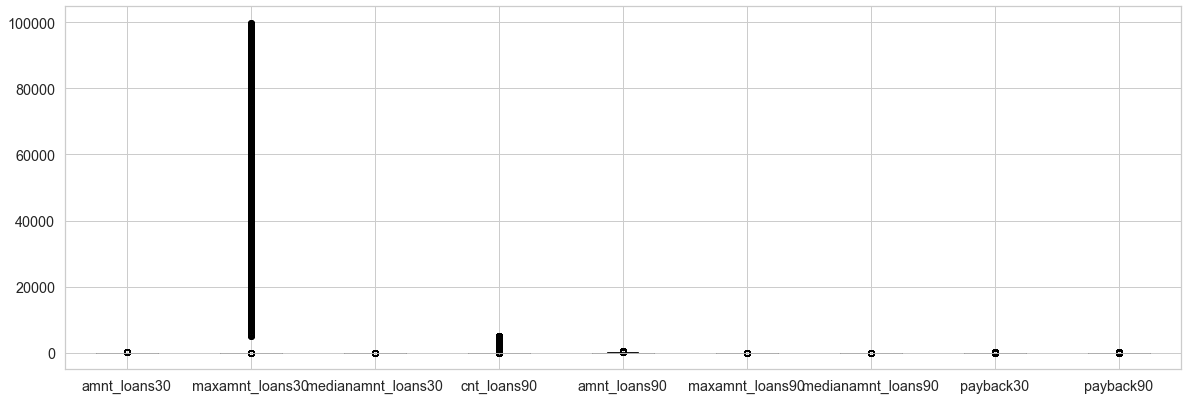


Box plot of features









We can see outliers present in the box plots

* Interpretation of the Results

Random Forest is a better Algo to use as it can handle more features, not influenced by outliers and handles missing value better.It can also identify the critical features

**CONCLUSION**

* Imbalanced Dataset, first time handling. So was very challenging to identify the workflow
* Also finding outliers was easy, but removing them was usually done with Z-score and IQR. But it didn’t work here as well
* Using model evaluation in loops was easy, but in large dataset it become a challenge by itself
* Next time for large dataset, move to Google Collab, to save time on code execution
* A feel of real life projects were given, looking forward to more challenging projects like this