

Micro-Credit Defaulter Model

Submitted by:

ARUMOY SAHA

**ACKNOWLEDGMENT**

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

* <https://think-asia.org/bitstream/handle/11540/2460/micro-ino.pdf?sequence=1>
* <http://www.gbgindonesia.com/en/finance/article/2016/indonesia_s_microfinance_sector_overview_key_component_for_sustainable_growth_11549.php>
* <https://core.ac.uk/download/pdf/35430822.pdf>
* <https://apfcanada-msme.ca/sites/default/files/2019-03/Micro%20and%20Small%20Businesses%20in%20Indonesia%E2%80%99s%20Digital%20Economy.pdf>
* <file:///C:/Users/sodainmind/Downloads/JOItmC-06-00050-v2.pdf>

All the other data sources were provided by the concerned enterprise for predicting the micro credit default 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 are 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. 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 choice of customers.

* **Conceptual Background of the Domain Problem:**

Here we have built a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loan 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 i.e. defaulter.

* **Review of Literature:**
* 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. I have observed them and handled them with a suitable explanation.
* I have come across outliers in some features which need to handle as per the understanding. Keep in mind that data is expensive and we cannot lose more than 7-8% of the data.

We also find out the correlation between the all the parameters in the dataset. We have proposed a random forest classifier model to predict the probability of defaulter and could help them in further investment in this particular process and improvement in choice of consumers.

* **Motivation for the Problem Undertaken:**

This project helps me understand the SME loan for mobile recharge and credit payback period and its customer behaviour. With the right set of datasets in hand I have build a model that helps the enterprise take the right decision that is whether to approve loan to a particular user or he/she will be defaulter. This also motivate learn about micro finance industry in details and how it helps build the economic development in the particular country.

**Analytical Problem Framing**

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

In this particular we need find out whether to provide micro finance loan for mobile balance that can be repaid within particular period or the particular customer will be a defaulter. We have used various parameters like Maxamnt\_Loans30(that is maximum amount of loan taken by the user in last 30 days), maxamnt\_loans90(that is maximum amount 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), Label plays a major role for understanding the dataset in details. We have used a random forest classifier model to predict the probability of defaulter and could help the client in further investment in this particular micro-finance loan for mobile balance and improvement in choice of consumers.

* **Data Sources and their formats:**

Data sources are provided internally by the enterprise.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Comment** |
| 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 | Unsure of given definition |
| rental90 | Average main account balance over last 90 days | Unsure of given definition |
| 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 | Unsure of given definition |
| 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 | Unsure of given definition |
| 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 | There are only two options: 5 & 10 Rs., for which the user needs to pay back 6 & 12 Rs. respectively |
| 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 Pre-processing:**

In the data pre-processing stage, I have found out if there is any missing data in dataset, for a particular column if there are any outliers present and how to handle the outliers. I have also dropped a few columns that are not require for model building process. I have also found the total shape of the data set. I have also found out the dataset description using describe method. So, in this pre-processing process I have mainly cleansed the data and prepared the right set of data for further processing for predicting the model.

For data pre-processing I have also used Label Encoder method to change the categorical variables into normalize labels and make it in machine readable format.

* **Data Inputs- Logic- Output Relationship:**

To find out the relationship between all the input variable I have used correlation function and find out whether there is a positive/negative relationship between a pair of variables. From this correlation function that also known as Five-point summary analysis if there are any outliers are present for a particular column.

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

Since all the dataset provided and defined properly so in this dataset, I assume label is the target variable. 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 i.e. defaulter.

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

For this particular dataset the Hardware is used Windows as operating system, and the software used are mainly Jupyter notebook for model building and various internal packages that are defined in the anaconda/jupyter notebook.

**Model/s Development and Evaluation**

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

For this particular project we have used various Classification method and Logistic Regression method to predict the outcome of this dataset. After the model RandomForestClassfier method predict the best outcome out of all the process and hence we can use this model for further evaluation.

* **Testing of Identified Approaches (Algorithms):**

Listing down all the algorithms used for the training and testing.

We have used mainly different classification method to get the outcome of the micro credit defaulter model and 80% data used for training purpose and rest 20% are used testing to predict the result for this machine learning model building process.

* **Run and Evaluate selected models:**

To predict the result of this dataset below are machine learning models used for evaluations.

|  |  |
| --- | --- |
| **ML Algorithm Used** | **Predicted Score** |
| Random Forest Classifier | 91.28% |
| Decision Tree Classifier | 86.48% |
| Gradient Boosting Classifier | 90.76% |
| Ada Boosting Classifier | 90.16% |
| Logistic Regression | 87.51% |

**Out of all the machine learning models used I have selected Random forest Classifier model for further evaluation of this dataset.**

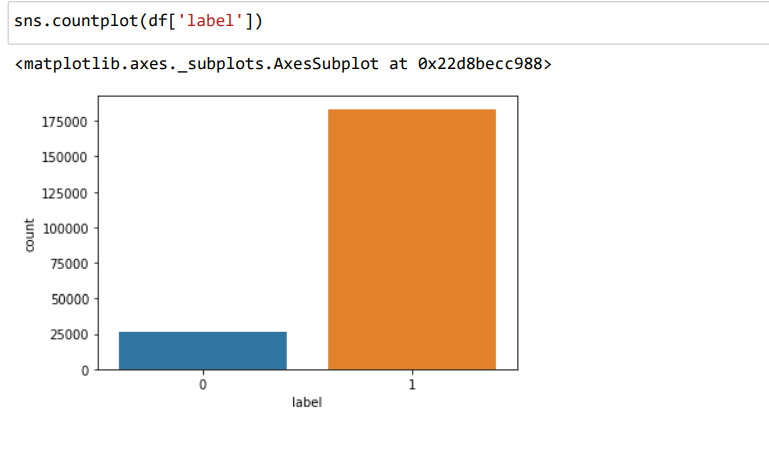
* Key Metrics for success in solving problem under consideration

The key metrics that were mainly taken into consideration were the followings:

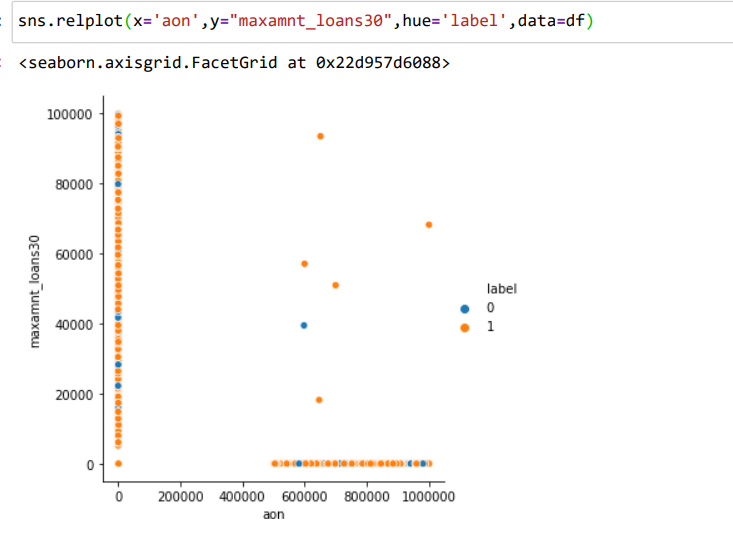
* maxamnt\_loans30
* maxamnt\_loans90
* Label
* payback30
* payback90

For this particular project label variable is target variable that is given as to whether the user paid back the credit amount within 5 days of issuing the loan {1: success, 0: failure}

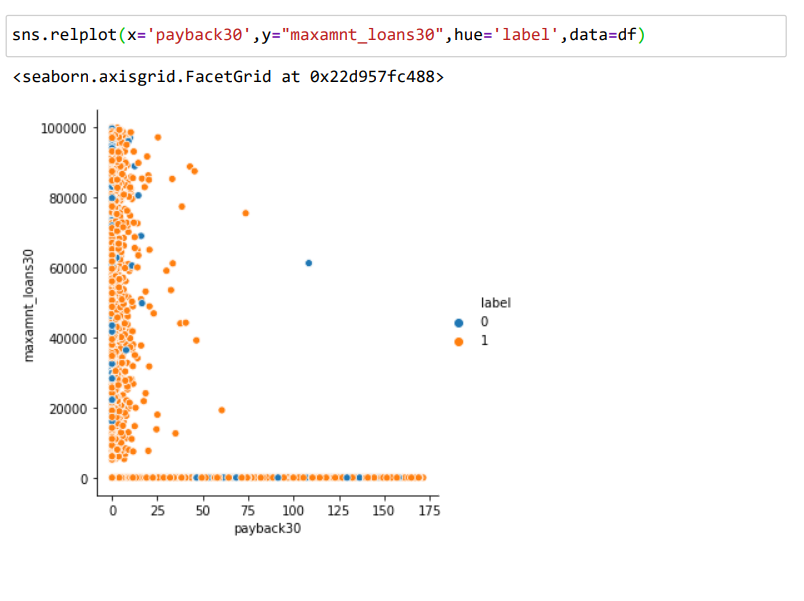
* Visualizations:



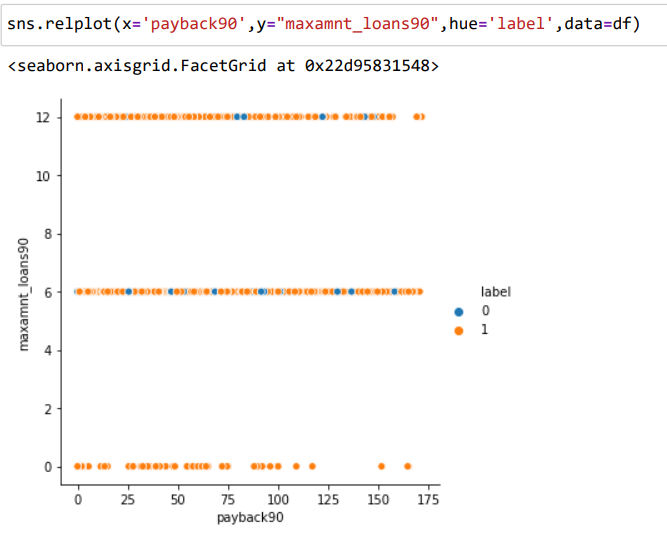
From this countplot we got know the number of defaulter and non-defaulter in mobile balance loan. In this project, Label ‘1’ defines that the loan has been payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e. defaulter.



From this relplot we got to know the relationships maximum loan taken by user in the last 30 days and whether the customer will be a defaulter/non defaulter with the age of cellular network taken into consideration.



From this relational plot we have visualize the maximum loan taken in 30 days by the customer with the average payback period over the last 30 days also we visualize whether the customer is a defaulter/non defaulter in this credit period.



From the above relational plot, we have visualized the maximum loan taken in 90 days by the user with the average payback period over the last 90 days also we understand whether the user is a defaulter/non defaulter in this particular credit period.

* Interpretation of the Results:
* More than 85% of the users were non defaulter in this micro credit process for mobile balance.
* Random Forest Classifier algorithm predict best result for this dataset.
* In case 30 days maximum loan amount data is distributed in a scatter way while taking payback period for 30 days.
* In case 90 days maximum amount loan data is distributed across multiple lines while considering the payback period for 90 days.
* I have applied Label Encoder method to change the categorical variables into normalize the input features and make it in machine standard format.

**CONCLUSION**

* Key Findings and Conclusions of the Study:
* I used various classification methods and out of all algorithm used Random Forest Classifier yields the best results.
* Since the credit period was very short so I got know the customer behaviour can be used for further research in others micro credit aspects as well.
* This micro credit process can be used rural development as well as for economic development of the country.
* Learning Outcomes of the Study in respect of Data Science:

As per as learning outcomes is concerned, we have learnt the following things:

* Algorithm need to use by understanding the dataset.
* From describe method we can get some knowledge related to outliers present in the particular columns (large difference between 75th percentile and maximum percentile)
* I also understand the visualization of related features and importance related to dataset.

Challenges:

* It was difficult to load the dataset in notebook as it took some time. (around 3 mins)
* Ruining each line code was a bit slow in notebook, possibly due to high volume of data.
* Limitations of this work and Scope for Future Work:
* Since *I* have only used a sample dataset, hence sometimes it is difficult to understand the overall impact of this micro credit process.
* *I* have not used the clustering method for predicting the outcome, otherwise *I* would group the data and create a cluster and may have analysed the outcome in better way
* *I* have not dropped too many columns so there is a possibility of outliers present in the used dataset, that sometimes may vary the results.