

**Micro-Credit Defaulter Model**

Submitted by:

**Dinesh Patel**

**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

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.

Now are days every company providing micro finance service , for example India Telecom provider give this facility to prepaid customer to get their mobile recharge instantly of they don’t have recharge facility. Similarly CAB providers also giving pay later account to their customer.

Using Micro credit facility consumer can keep using their services even if they have short of money. Consumer can use this amount to make payment on other services as well.

* Conceptual Background of the Domain Problem

Micro finance service give as lot flexibility to user, they can avail this service in whenever they are in need,

This facility also helping business to gain more revenue because of its convenience to use.

To understand the implications of this issue, the study details that on average 5200 unique resellers of one operator are affected every day. Due to lack of credit, these resellers were then unable to carry out airtime topup transactions, resulting into a total loss of about USD 1.98 million per annum for the operator. Such losses can be avoided if service providers can ensure service continuity despite the low stock at reseller’s end. Micro Credit is just the solution that can help them regain this lost value.

* Review of Literature

Microfinance is not a new concept. Small community-oriented microcredit operations have existed since the Franciscan monks in the 15th century. In the mid-1800s, Friedrich Wilhelm Raiffeisen founded the first cooperative lending bank to help the rural farmers of Germany. Bhaskar 4 However, the inception of the modern microfinance industry began in the village of Jobra in Bangladesh in 1976, with Muhammad Yunus establishing Grameen Bank in 1983. It is important to delve into the roots of modern microfinance to analyze and compare with any other program, as today more than 250 institutions on nearly 100 countries operate micro-credit programs based on the Grameen methodology. Borrowers of the Grameen Bank own 93% of the total equity of the bank, with 7% being owned by the Bangladeshi government. Grameen stopped soliciting and taking donor money in 1998, financing its credit program purely through existing deposits and loans from then onwards. Grameen Bank has 1,181 branches, works in 42,127 villages, and has a staff of 11,777. The total number of borrowers is 2.6 million, 95 percent of whom are women. The total amount of loans disbursed by Grameen Bank since inception is $3.9 billion. Out of this, $3.6 billion has been repaid, with the recovery rate standing at 98 percent. Grameen Bank provides three types of loans: income generating loans (with an interest rate of 20 percent), housing loans (with an interest rate of 8 percent), and higher education loans for the children of Grameen families (with an interest rate of five percent). Grameen believes education is one of the major primary components for poverty alleviation, for future generations. Grameen provided loans for higher because education culminates in more competitive and skilled workers and sustainable improvement education covering tuition, living costs, and other school expenses to 466 students studying in medical and engineering schools and scholarships for an average of 3,700 high-performing schoolchildren.

The basic mechanism that Grameen has worked into its product design to reduce tension and maintain dignity among poor borrowers is the distinction between its basic and flexi-loan products. If borrowers cannot maintain a constant stream of payments for their loans, their loans are converted from basic to flexi-loans, thus exiting the highway (metaphorically represented in the diagram above) to a higher loan ceiling typically for between six months to two years while providing respite to borrowers before subsequent re-entry. This prevents negative group peer pressure on a member struggling to pay back a loan while facilitating collection through renegotiation rather than hounding someone who is unable to pay. However, Grameen surely must manage the relative percentage of these loans on its portfolio carefully, as they carry a substantial amount of default risk and incur costs by complicating the jobs of loan officers.

* Motivation for the Problem Undertaken

Micro credit service is designed for those people who don’t have bank account or poor. Companies always worry about payback amount, they want to make sure that maximum part of loan amount is recovered, hence the try to predict their customer behaviour on the basis of different factor which are easily available at company’s data, these factors are very useful and one can predict that potential customer will be able to pay back the borrowed amount.

**Analytical Problem Framing**

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

Our method consumes raw metadata from mobile phone usage, which are already being collected at close to zero cost. These records can yield rich information about individuals, including mobility, consumption, and social networks

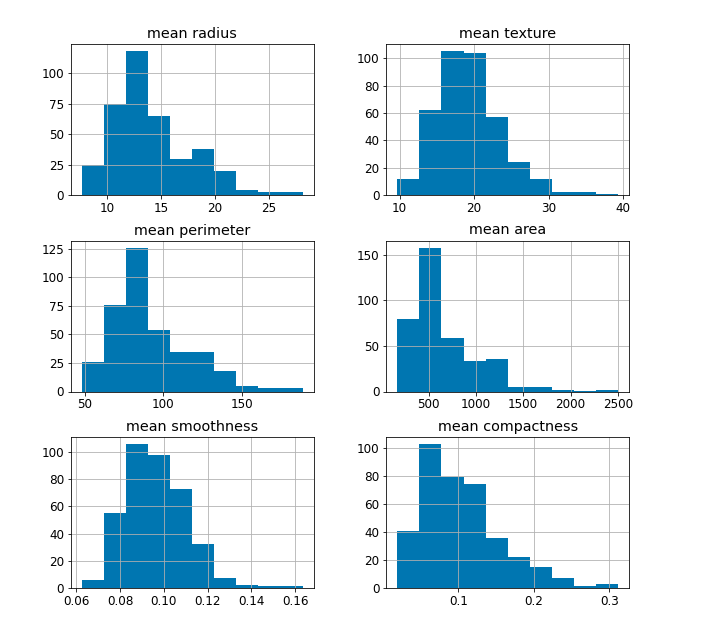
There are many straightforward indicators of behaviour that are plausibly related to repayment of credit. For example, a responsible borrower may carefully manage their balance over time so usage is smoother. An individual whose usage repeats on a monthly cycle may be more likely to have a salaried income. Or, an individual whose calls to others are returned may have stronger social connections that allow them to better follow through on entrepreneurial opportunities.

Raw data received for this analysis required cleaning and proper handling for urther process, we found that data was not balanced, having skewness and outliers.

To remove Skewness we used Power transformation and data normalization.

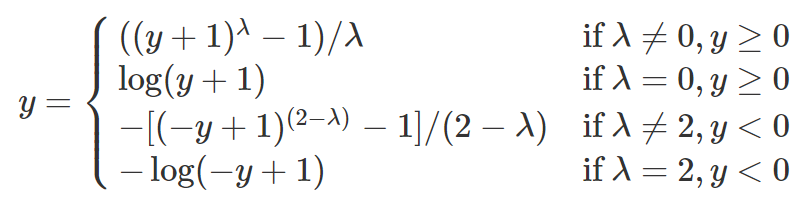
**Power Transformation:-**

Power transform is a family of functions that transform data using power laws. The idea is to apply a transformation to each feature of our dataset. What’s the purpose of a power transform? The idea is to increase the symmetry of the distribution of the features. If a feature is asymmetric, applying a power transformation will make it more symmetric. Let’s see an example using the breast cancer dataset in scikit-learn. If we draw the histogram of the first 6 features, we see that they are very asymmetric.



Some models may not work properly if the features are not symmetric. For example, models based on distances like KNN or K-means may fail if the distributions are skewed. In order to make these models work, power transformations will symmetrize the features without affecting their predictive power too much.

Yeo-Johnson transformation has this formula:

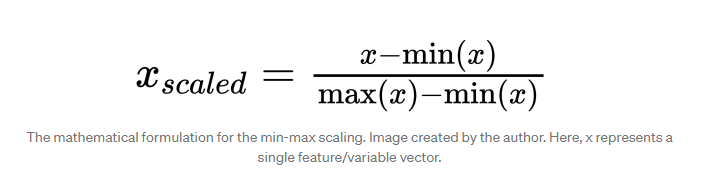


We still have a λ parameter to be estimated, but now this transformation can be applied even to negative features.

2. Outlier was also removed using IQR, but we have to roll it back because data loss was more then 20%.

3. MIN-MAX scaler:- The main idea behind normalization/standardization is always the same. **Variables** that are **measured** at **different** **scales** **do not contribute equally to the model fitting** & model learned function and might end up creating a **bias**. Thus, to deal with this potential problem feature-wise normalization such as **MinMax** Scaling is usually used prior to model fitting.

*This can be very useful for some ML models like the Multi-layer Perceptrons (MLP), where the back-propagation can be more stable and even faster when input features are min-max scaled (or in general scaled) compared to using the original unscaled data.*



logit(p)

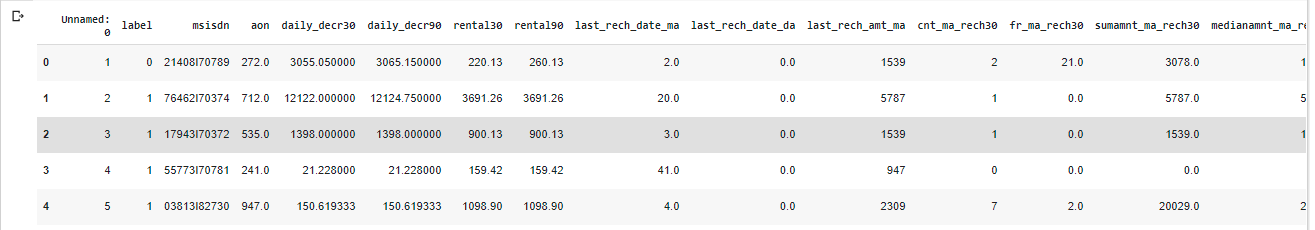
for i = 1…n .

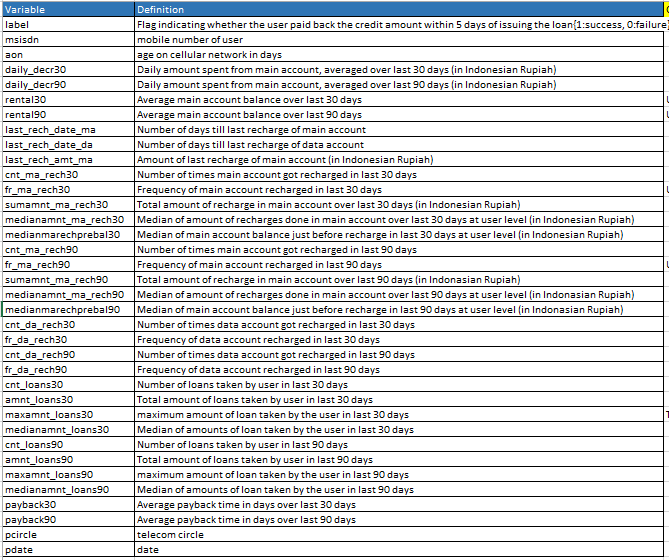
3**. Binary Logistic Regression:-**

1. The dependent variable should be dichotomous in nature (e.g., presence vs. absent).
2. There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores, and removing values below -3.29 or greater than 3.29.
3. There should be no high correlations (multicollinearity) among the predictors.  This can be assessed by a correlation matrix among the predictors. Tabachnick and Fidell (2013) suggest that as long correlation coefficients among independent variables are less than 0.90 the assumption is met.
4. At the center of the logistic regression analysis is the task estimating the log odds of an event.  Mathematically, logistic regression estimates a multiple linear regression function defined as:
5. We also used multiple model like SVM GNB, Decision Tree, Random Forest.
6. Hyper parameter tuning done on using RandomizeSearchCV.
7. Model accuracy measurement done using F1 score and Cross Validation

* **Data Sources and their formats**

Data set contains 37 feature, where column “label “ is the target variable, it has binary value , 1 stands for non-defaulter customer and 0 stands for defaulter customers .





There are no null values in the dataset.

The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while, label ‘0’ has approximately 12.5% records

* Data Pre-processing Done:

In Above data set there are 37 variable, where label was target variable and other variable was independent variable,

Below steps followed to clean data before processing.

* 1. Data contain 209593 rows and 37 columns.

(209593, 37)

* 1. Data contain extra index column which will be removed, MSISDN,pdate and pcircle column will also be removed as they are not going to help
  2. Pcircle variable has variance 0 which indicate that this column contain only 1 value and can be removed.
  3. MSISDN variable has all object type and contain mostly uniq value, we cannot encode this value and It can be removed from dataset.
  4. Pdate is also have object data type , this variable contain date of data collection, our prediction is not date depended hence we can remove them as well.
  5. Delete columns which are very less correlation with target variable, below independent variables having near to zero correlation , hence it is safe to delete these variable.

*#cnt\_loans90              0.004733*

*#cnt\_da\_rech30            0.003827*

*#last\_rech\_date\_ma        0.003728*

*#cnt\_da\_rech90            0.002999*

*#last\_rech\_date\_da        0.001711*

*#fr\_ma\_rech30             0.001330*

*#maxamnt\_loans30          0.000248*

* 1. Further we check that data don’t contain and any null or missing value (NaN Values), hence not clean-up required at this point.
  2. Outlier couldn’t be removed because it has more than 20% of data loss.
  3. Skewness removed using Power transformation and normalization done by Min Max Scaler.
* Hardware and Software Requirements and Tools Used

**Tool**: - Anaconda Jupyter Nodebook,

**Software**: - Python 3.7.0 with all required package, Jupyter Notebook already equipped with required Machine learning Packages like sklearn matplotlib, but is any package is missing then we can install it using

**!pip install <package-name>**

**Hardware: -** minimum i5 processor with 8 GB ram is required to run different machine learning classification model with this much data.

**Model/s Development and Evaluation**

* After data cleaning we need to handle imbalanced data set problem which is handled using oversampling method, below no. of sample taken in after oversampling.

**Before OverSampling, counts of label '1': 91463**

**Before OverSampling, counts of label '0': 13333**

After OverSampling, the shape of train\_X: (182926, 24)

After OverSampling, the shape of train\_y: (182926,)

**After OverSampling, counts of label '1': 91463**

**After OverSampling, counts of label '0': 91463**

Initially we used Logistic regression as target variable as binary output, we calculated best random state using logistic regression model,

Best random state applied on different classification models.listed below.

1. Logistic Regression
2. Decision tree classification.
3. Random Forest classification
4. SVM (support Vector Machine).
5. K-Nearest Neigbouer Classification.

* Testing of Identified Approaches (Algorithms)
  1. Logistic Regression
  2. Decision tree classification.
  3. Random Forest classification
  4. SVM (support Vector Machine).
  5. K-Nearest Neigbouer Classification.

Also we applied hyper perameter tuning on each model.

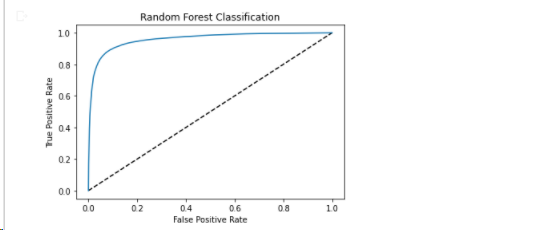
* Run and Evaluate selected models

Tested all models on the basis of of F1 score, accuracy, recall and precision, F1 is the combination of both precision.

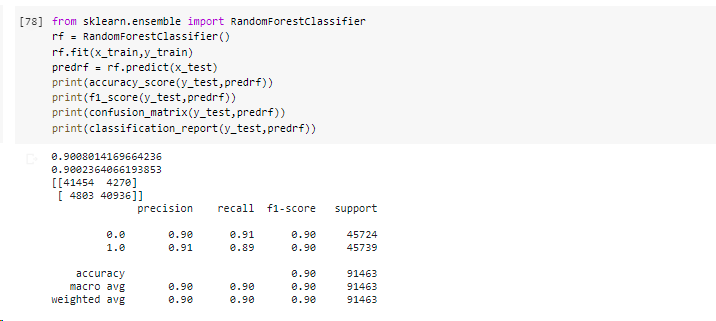
AUC-ROC curve was use to check model performance.

* Key Metrics for success in solving problem under consideration

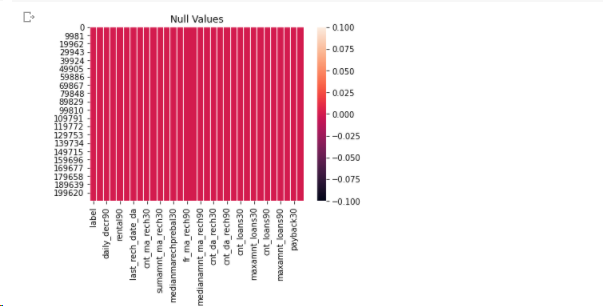
AUC ROC Curve:-



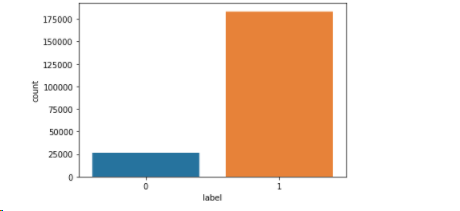
* 1. F1 score.



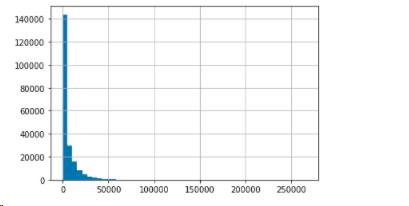
* 1. Visualizations
  2. Check for Null value:-No Null value present.

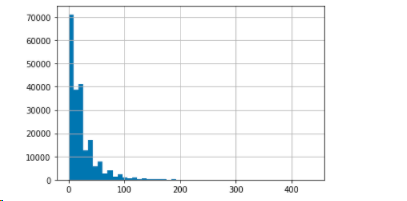


* 1. Data imbalance:- data imbalance exist in dataset.

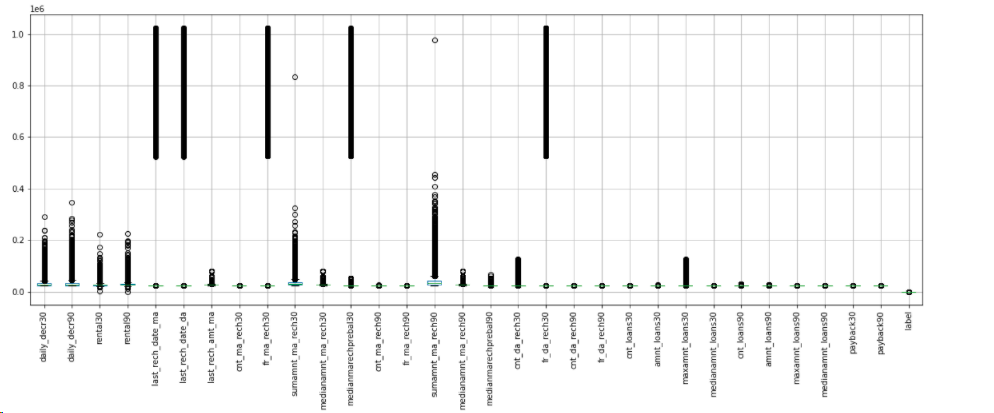


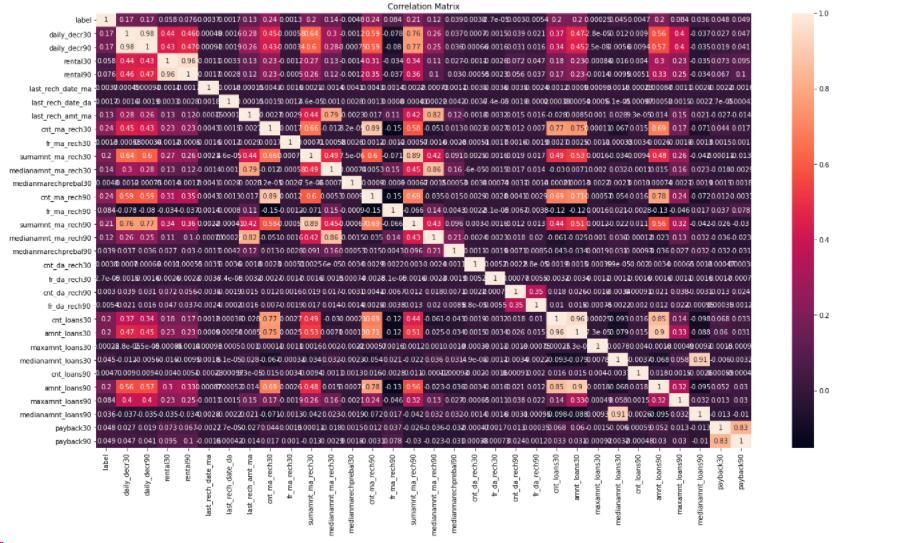
* 1. Check for skewness on different columns:- Skewness exist on multiple column.





* 1. Check for outlier:- found that multiple columns outlier presents.





* Interpretation of the Results:- We found that Random Forest is the best model which has highest F1 score(90%).

**CONCLUSION**

* Key Findings and Conclusions of the Study

Micro Credit finance on the basis of ML prediction will be very much accurate and we can reduce high no of defaulter if using Machine learning analysis to understand customer behaviour.

It will help the needy person to get facility and reducing the loss as well.

It will help to keep customer to use services even if is has lack of money for some time which will increase revenue of service providers.

* Learning Outcomes of the Study in respect of Data Science

Learn multiple thing here, how to find correct target value in data set, correlation of the different variable , outlier techniques, different transformation method and Hyper parameter tuning of the Model we have finalized.

* Limitations of this work and Scope for Future Work

Dataset was very huge , it took lots of time to get results from different model, it cause delay in our conclusion, we cannot execute other steps until one running task is completed, in this case we had to wait for long time to apply other ideas.

Hyper parameter tuning for all model was not possible with the current hardware we are using because it takes more than 3 hour to complete for one mode, sometime browser do not return result after running a test for 3 hour which as waste of time.