

Micro-Credit Defaulter Model

Submitted by:

Amit Gawande

**ACKNOWLEDGMENT**

I have referred below resources that helped and guided me in completion of this project as below:-

[www.w3resource.com](http://www.w3resource.com)

[www.towardsdatascience.com](http://www.towardsdatascience.com)

[www.stackoverflow.com](http://www.stackoverflow.com)

**INTRODUCTION**

* **Business Problem Framing**

The telecommunication company is providing micro loan credit to their customers with primary focus on low income families. So, in this use case also there are multiple customers which taken loan and paid the loan amount and some defaulted on loan payment. So we have to predict by taking important features that whether in future customers having this type of inputs and history will able to pay back the loan or not. Hence, this a real-world use case which will help telecom industry to provide the loan to which type of customers.

* **Conceptual Background of the Domain Problem**

If a customer who has paid the loan amount is falsely marked as defaulter .This would not be much problem to the company but if a defaulter is not detected and appropriate measures were not taken by the company then it will result in loss of revenue

* **Review of Literature**

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.

* **Motivation for the Problem Undertaken**

In order to understand to whom loan to be given from lower income earning people and data from telecom industry clearly stats parameters to be taken into consideration to declare borrower as defaulter or not & amount limit also can be decide based on this.

**Analytical Problem Framing**

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

Dataset contain around “209592” rows and 36 columns which make this dataset huge. There are lot of outliers in the dataset which need to taken care

Dataset is also imbalance and mostly contain the label as 1 which defines that most of customers pay back the loan, due to imbalance dataset the model is bias towards the label as 1

* **Data Sources and their formats**

The data set is continuous and most of the features and numeric in nature and the dataset does not have any null values

* **Data Inputs- Logic- Output Relationships**

The seaborn heatmap show the correlation of each feature to each other.

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

Max value of ‘aon’ feature is very large (2739years) and this is not possible in years. I have assumed it exists and carried on.

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

**Hardware**-64bit, 4GB RAM.

**Software-**Excel, Anaconda,jupyter notebook

**Model/s Development and Evaluation**

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

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

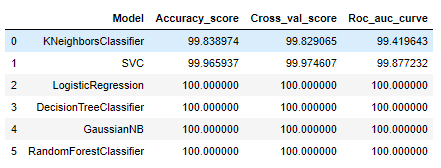
* **Testing of Identified Approaches (Algorithms)**

Below are classification algorithms used for the training and testing this dataset.

* Logistic Regression
* Random Forest Classifier
* K-Neighbors Classifier
* Decision Tree Classifier
* Gaussian NB

* **Run and Evaluate selected models**

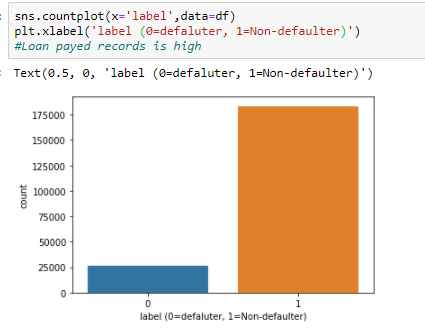
Please find below table for model statistics.



* **Key Metrics for success in solving problem under consideration**

In classification problem there are various metrics that are accuracy score, confusion matrix, classification repot, Roc Auc curve and which metrics is useful is also depend and vary on domain and data set.

* **Visualizations**



* **Interpretation of the Results**

Visualisation shows outliers which need to be removed / corrected.

Data Pre-processing done by performing EDA (Exploratory Data Analysis), checking for best accuracy score.

Modelling done based on type of data as this is categorical data, we have to go with multiple classification models & finalise the best score giving model.

**CONCLUSION**

* **Key Findings and Conclusions of the Study**

Removing unwanted column like msisdn number which is unique and columns telecom circle and date column.

As per dataset statistical analysis there is difference between the 75th percentile and max so this is an indication of the outliers in the dataset.

In this data set there are negative values which are from the total amount, age which can’t be negative

* **Learning Outcomes of the Study in respect of Data Science**

Visualize the data using univariant / multi-variant analysis.

Check the prediction score using accuracy score & get ROC-AUC score.

Train data using classification models to get the best score & finalise best score giver model for this dataset.

Save file using pickle/joblib library.

* **Limitations of this work and Scope for Future Work**

Do outlier removal, there are very few columns, and one by one analysis can also be done.

Try different Skewness removal technique and check different combinations for hyper tuning.

Check for the error occurred during PCA.