
# Micro-Credit Defaulter Model

Submitted by:

Anurag Shrivastav

# ACKNOWLEDGMENT

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

www.w3resource.com

www.towardsdatascience.com

www.stackoverflow.com

# INTRODUCTION

- ## Business Problem Framing

  This project is about providing loans (financial services) to low income populations by Micro-Financial Institution (MFI). MFI provide loan to Group Loans, Agricultural Loans, Individual Business Loans and so on. In Order to achieve this objective, MFI needs to decide criteria for customer selection.

- ## Conceptual Background of the Domain Problem

  Banking domain knowledge is required to know about generic criteria for loan giving institutions, market risk and parameters to decide defaulter, interest charges, benefits, etc.

- ## Review of Literature

  Loan giving capacity will get decided based on below parameters-Daily amount spend & average main account balance in last 30 days, Frequency of recharge for data account & main account in 30/90 days, loan taken in last 90 days & payback time for last 30 days.

- ## 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.

  In every country poor population exists to some scale and financial services to be provided to them at affordable level of loan amount to uplift their financial situation, which may reduce the vulnerability factor.

# Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- ## Data Sources and their formats

  Provided by client in excel or csv format.

- ## Data Pre-processing Done

  Dropped unwanted columns and duplicate values from rows.

  Checked correlation between columns, basis which removed Outliers.

  Post this I used label encoder 2 sets one for float data & other for string data & converted all data into integers after that verified for non-null values.

- ## Data Inputs- Logic- Output Relationships

  Input data for feature list and target is in numeric format and hence classification model (Random Forest Classifier) best suits for this dataset.

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

  I have not considered any pre-assumption, project performance from beginning to end is based on data facts only.

- ## Hardware and Software Requirements and Tools Used

**Hardware Requirement-**Laptop with below configurations-
Windows Edition-Windows 10 Pro
Processor-Intel(R)
Memory-6 GB
System Type-64 bit OS

**Software Requirement-** Anaconda 3.7 & above, Jupiter Notebook 6.

# Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
  - **Analytical Approach –**Based on type of data by performing EDA I have decided which model to be used for this data.
  - **Statistical Approach –** Data should be in scaled manner, it should not be distorted, for that I have replace all null values using mean method due to continuous data numbers.

- Testing of Identified Approaches (Algorithms)

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

  - Logistic Regression
  - Random Forest Classifier
  - Decision Tree Classifier
  - Gaussian NB
  - Bagging Classifier

- Run and Evaluate selected models

Pls find below matrix & their results also.

```
In [53]: from sklearn.metrics import roc_auc_score
         auc_score=roc_auc_score(y_test,y_pred)
         print("AUC_Score:",auc_score)

         AUC_Score: 0.7470847048940459
```

```
In [60]: from sklearn.metrics import roc_auc_score
         auc_score=roc_auc_score(y_test,y_pred_bc)
         print("AUC_Score:",auc_score)

         AUC_Score: 0.7732871413119644
```

- Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

- Interpretation of the Results88888

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

Conclusion-Loan giving capacity based on below parameters-Daily amount spend & average main account balance in last 30 days, Frequency of recharge for data account & main account in 30/90 days, loan taken in last 90 days & payback time for last 30 days.

Multi-Financial Institutions need to be taken into consideration for above parameters due to correlation & it's giving best score also.

```
In [70]: models=[]
```

```
In [*]: models.append(('LR',LogisticRegression()))
        models.append(('RFC',RandomForestClassifier()))
        models.append(('NB',GaussianNB()))
        models.append(('DTC',DecisionTreeClassifier()))
        models.append(('BGC',BaggingClassifier()))

        results=[]
        names=[]
        for name, model in models:
            kfold= StratifiedKFold(n_splits=10, random_state=1, shuffle=True)
            cv_results=cross_val_score(model, x_train,y_train, cv=kfold, scoring='accuracy')
            results.append(cv_results)
            names.append(name)
            print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))
```

```
LR: 0.864619 (0.000641)
RFC: 0.913003 (0.001837)
NB: 0.604152 (0.004011)
DTC: 0.873639 (0.002001)
BGC: 0.904983 (0.002575)
LR: 0.864619 (0.000641)
```

- ## Learning Outcomes of the Study in respect of Data Science

  This dataset is categorical in nature, we can verify data by using read method & get stats related information for each column using describe method.

  As its categorical data, classification model best suits for this.

  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.

  Get the test score for same model.

  Save file using pickle/joblib library.

  Find the prediction vs actual using distribution plot in order to get the perfect deviation if any.

- ## Limitations of this work and Scope for Future Work

  Column with no impact/no correlation have excluded as it might have reduced the performance.88888888

  It’s always good to have complete data while performing model but 7-8 % of data can be excluded based on performance impact.