MICRO CREDIT USE CASE STUDY PROJECT

SUBMITTED BY: **Sindhu Shree N**

INTERNSHIP BATCH: **19**

ACKNOWLEDGEMENT:

With great pleasure, I sincerely express my deep sense of gratitude and heartfelt thanks to several individuals from whom I received impetus motivation during the internship project work. I am very grateful to **SHUBHAM YADHAV** sir internship 19th batch guide, for the help and I do express my deep gratitude to sir for his guidance, keen interest and constant encouragement throughout the internship project work. I am thanking him for personal concern, affinity and great spirit with which he has guided the work.

PROBLEM STATEMENT

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. It is hereby given to you for this exercise. 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 selection of customers.

PROBLEM DEFINITION:

Micro credit loan default is a classic use case where micro loan models can be deployed to predict risky customers and hence minimize losses of the lenders. Financial industry is highly regulated thus any model deployed or classification of customers basis their behaviour, demographic etc. is highly regulated and must be explained to authorities to ensure unbased operations.

Credit loans are risky nut at the same time it is also a product that generates profit for the institutions through differential boring lending rates etc.

- ## Conceptual Background of the Domain Problem

  Though the term microcredit is relatively new as it was invented in 1983, the concept is to provide financial help to those of a lower socioeconomic background. It is said that lending to people of lower socioeconomic background goes as far back as the 1700s in Ireland.

- **REVIEW OF LITERATURE:**

  All possible information from all the available data tables more the information, more than for EDA and feature Engineering. its take more important to take the average during the aggregation of data from tans the table rather than taking the counts before the loan was applies no future information steps into the data to be used for modelling.

- **Motivation for the Problem Undertaken:**

  This is very useful for the organisation which are providing loans for the lower socioeconomic background people inorder to improve the selection of customers for the credit and it could help them in further investment and improvement in selection of customers.

- **Analytical problem framing:**

  This includes:

  1. Data Cleaning
  2. Selecting relevant features
  3. Exploratory data analysis
  4. Scaling techniques
  5. Training the model
  6. Hyperparameter tuning
  7. Evaluate the model
  8. Prediction of the model

- **Data Sources and their formats.**

Micro Finance Institution provided this data. The data is provided in the csv file.

Hardware and software requirements and tools used:

Hardware requirements:

PROCESSOR: Intel(R) Core(TM) i3 CPU

MONITOR : Any display unit

HARD DISK : 240GB SSD

RAM : 8.00GB

Software requirements:

OPERATING SYSTEM*:* Windows 10 Pro

FRONT END : Jupyter Notebook (Anaconda3)

BACK END : Excel 2013

Tools Used:

1) Pandas Library

2) Numpy Library

3) Seaborn Library

4) Matplotlib

5) sklearn library

- ## Data Preparation and cleaning:

  - Reading the csv file and doing initial statistical analysis (shape, values etc)

  - Data Pre-processing: Reading the unique values for each column and removing those which won't be significant in the analysis further.

  - Create a new data frame to proceed with the analysis further.

  - Partitioning and splitting that dataset account when credit loan default was applied as that dataset is skewed, stratification is

used allocate the samples evenly based on sample classes so that training set and test set have similar ration of classes.

- ## Data Inputs- Logic- Output Relationships:

- Dataset is highly imbalanced we can use our models are majority and minority logic input and outputs.
- Majority classes-the dataset is too small. Down sampling the majority class will not help, so we will up sample the minority class.

- ## Models Development and Evaluation:

The target feature is a numeric (continuous) data with only two classification we need to apply the classification models to predict the outputs. We have splitted the data into training data and testing data. Here we are using five classification algorithms and selecting the best one out of it.

Testing of Identified Approaches (Algorithms):

1) Logistic Regression
2) Decision Tree Classifier
3) Ridge Classifier
4) GaussianNB Classifier
5) XG Classifier

Run and Evaluating selected models:

1) **Logistic Classification**: It is a basic and commonly used type of predictive analysis. Logistic Classification fits a straight line or surface that minimizes the deprecancies between predicted and actual output values. Here it is giving an accuracy of 73.56%.
2) **Decision Tree Classifier:** This algorithm splits a data sample into two or more homogeneous sets based on the input variables. A part of a tree is generated with each split. As a result, a tree with decision nodes and leaf nodes is developed. A tree starts from a root node which is a best predictor. This model is giving an accuracy of 95.84%.

3) **Ridge Classifier:** It is a type of linear model used for predictive models which is giving an accuracy 73.93% for our present dataset.

4) **GaussianNB Classifier:** It is a classification machine learning algorithm which can be used as the basis for sophisticated non-parametric. For the present dataset this model is giving an accuracy of 70%.

5) **XGBoost Classifier:** Extreme Gradient Boosting classifier provides an efficient and effective implementation of gradient boosting algorithm which is used for regression predictive modeling. This model is giving an accuracy of 88.55%.

Among all the models the **Decision Tree Classifier** is giving the best accuracy score with almost similar cross validation score, so we considered this model for the tuning and further evaluation in which it has performed well.

CONCLUSION:

Most classification problems in the real world are imbalanced. Also, almost always data sets have missing values. Here covered strategies to deal with both missing values and imbalanced data sets. We also explored different ways of building ensembles in sklearn.

- Learning Outcomes of the Study in respect of Data Science

  - There is no definitive guide of which algorithms to use given any situation. What may work on some data sets may not necessarily work on others. Therefore, always evaluate methods using cross validation to get a reliable estimate.
  - Sometimes we may be willing to give up some improvement to the model if that would increase the complexity much more than the percentage change in the improvement to the evaluation metrics.
  - In some classification problems, false negative is a lot more expensive that false positives. Therefore, we can reduce cut-off points to reduce the false negative.
  - Missing values sometimes add more information to the model than we might expect. One way of capturing is to add.

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

1) Limited Access to information

2) Time Limits
3) conflicts on biased views and personal issues
4) How to structure my project research limitation correctly
5) How to set my project research limitation.
6) Formulation of my objectives and aims
7) Implementation of my data collection methods
8) Scope of discussions
9) Finding my error on the codes
10) Concluding thoughts.