FLIP ROBO

# MICRO CREDIT DEFAULTER

Submitted by:

Nirav Mehta

# INTRODUCTION

- Business Problem Framing

A microfinance institution (MFI) is a company that provides financial services to those with limited resources. Targeting impoverished, unbanked families in rural locations with few sources of income makes MFS highly effective. The MFI offers group loans, agricultural loans, individual business loans, and other microfinance services (MFS). The usage of mobile financial services (MFS), which many microfinance institutions (MFI), experts, and donors believe to be more practical, effective, and cost-effective than the conventional high-touch strategy employed for many years to deliver microfinance services, is becoming increasingly popular. Although the MFI sector focuses mostly on low-income families and is tremendously helpful in these areas, MFS implementation has been mixed, with both major setbacks and victories. With 200 million clients worldwide and $70 billion in outstanding loans, microfinance is now widely recognised as a strategy for reducing poverty.

- Review of Literature

We are now working with a client in the telecom sector. Their business is one that offers fixed wireless telecommunications networks. They have introduced a number of products and built their company and organisation around the budget operator model. They do this by using a disruptive innovation strategy to provide its subscribers better things at lower prices.

They concentrate on offering their services and goods to low-income families and underprivileged consumers in order to aid them when they are in need since they recognise the value of communication and how it affects a person's life. They are working with an MFI to offer microcredit on cell phone balances with a 5-day repayment period. If the Consumer does not repay the lent money within the allotted five days, he is considered to be in default. For a loan of five dollars (in Indonesian Rupiah), the payback amount should be six dollars, and for a loan of ten dollars (in Indonesian Rupiah), the payback amount should be twelve dollars (in Indonesian Rupiah).

We receive the sample data from our client database. You have it by way of this activity. The client needs some predictions that might aid them in future investments and better customer selection in order to increase the selection of clients for credit.

## Analytical Problem Framing

Our client database provides us with the example data. You have it for this exercise as of this moment. The client requests some projections to aid in future investments and better customer selection in order to increase the quality of consumers chosen for credit.

Let's examine the null count in the features now that we are aware of the features that are present in the dataset.

The heat map in the aforementioned graph indicates that all of the features have zero counts.

By examining the graph, we can see that this dataset does not contain any null values.

The data are now being analysed by plotting a histogram to characterise the target variable, "Label."

## Model/s Development and Evaluation

Asymmetry exists in the dataset. Label "1" has about 87.5% of the records, whereas label "0" has about 12.5%. At the model-building stage, this data imbalance can be further adjusted.

The data must then be encoded in order to analyse the model in the most effective manner possible. Here, all of the object data is encoded to numeric characteristics using an ordinal encoder.

The concept of correlation describes the connections between one or more variables. These factors could be characteristics of the raw data that were used to forecast our target variable.

A statistical approach called correlation shows how one variable changes or moves in connection to another variable. It provides us with a general understanding of how closely the two variables are related. This bi-variate analysis measure explains the relationship between many variables. In the majority of business situations, it is helpful to discuss a subject in terms of how it relates to other issues.

Heat map and other plot graphs can be used to determine the correlation in this case.

Check to see if there are any multi-collinearity issues.

Since there are numerous features, determine the correlation between each feature and other is quite challenging, so we go on to boxplot and skewness

Looking at the aforementioned graphs, it appears that there may be some outliers and skewness in the dataset.

We use the Feature Selection approach to select the optimal characteristics for model construction in order to avoid them. This approach uses f classif methods to frame a new dataset using the top 30 characteristics that are best for modelling.

Using the Zscore method, outliers and frames from another dataset with dimensions are removed (164646,31)

The percentage of data loss is determined by the difference between the data before and after using the Zscore algorithm.

An algorithm for supervised machine learning is K nearest neighbours (KNN). Learning a function such that f(X) = Y, where X is the input and Y is the output, is the aim of supervised machine learning algorithms. KNN has applications in both classification and regression. All of the discussion in this article will centre on classification. Although there is hardly any change for regression.

The characteristics of KNN include being a nonparametric approach and a slow learning algorithm.

The algorithm learns slowly since it just stores the data from the training phase, which takes essentially no time at all (no learning of a function). The evaluation of a new query point will then be conducted using the stored data.

The term "non-parametric method" describes a method that makes no distributional assumptions. As a result, KNN need not discover any distributional parameters. While using the parametric approach, the model discovers new parameters that are then used to make predictions. K, the number of points to be taken into consideration for comparison, is the only hyper parameter (given by the user to the model) KNN possesses.

The model will save the data points during the training phase. To categorise each point in the test dataset, the distance between the query point and the training phase's points is determined. There are many different distances that can be calculated, but the Euclidean distance is the most common one (for data with lesser dimensions).

The steps to determine the class of query point are as follows:

- From the query point, the distances to all 500 points are calculated.
- K nearest neighbours are chosen for comparison purposes based on the value of K.
- Assuming K=7, 3 of the 7 points are in class 1, and 4 of the 7 points are in class 0. Then the query point p is classified as class 0 based on the majority.

Decision Trees are a sort of supervised machine learning in which the training data is continually segmented based on a particular parameter, with you describing the input and the associated output. Decision nodes and leaves are the two components that can be used to explain the tree.

The choices or results are represented by the leaves. The data is divided at the decision nodes.

Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance.

Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.

In comparison to the other models, "Random forest Classifier" has an accuracy score of 91.38%, making it the best model that is saved and can be used to make predictions in the future.

# CONCLUSION

After hyper parameter adjustment, the anticipated accuracy score of 89.99% is forecasted, which is regarded as an excellent model. Now that the model has been trained, we may use it to predict both the test data and any other future data.