## CREDIT RISK - PROJECT REPORT

## Introduction

Credit risk is a term often used in business analytics that basically involves the analysis and evaluation of the chances of a loan applicants’ failure to comply with contractual terms and requirements for example credit debts and mortgage.

Reducing the risks or chances of suffering losses as a result of applicants failing to honor the terms is a major concern for financial institutions and they invest much to evade such risks. For this reason, these institutions rely on state-of-the-art technology to forecast which clients are more likely to fail to honor their debt requirements as per the terms of contract. These companies use data mining algorithms to get optimum accuracies for their analysis and management of credit risks, an automatic means to detect possible debt-risks beforehand.

## Problem Identification

Microfinance provides individuals with initial capital to venture into business, start and or even enlarge the boundaries of their businesses. Small startup businesses or generally low scale businesses backed up by microfinances have had the chance to grow to medium scale enterprises hence filling existing business gaps and generating employment opportunities. Microfinance companies’, programs and projects have extended their way in enriching the capacity of clients in the areas of loan acquisition, customer support, price setting, promotion strategies and offering service and goods in credit terms.

Risks associated with credit terms occurs when a potential finance borrower in a contract delay or totally fails to repay the debt as a whole or a slot of it.

Like other perils, risks associated with credit terms has negative impacts in financial agreements. Exposure to such perils has continued to be the great source of difficulties in financial institutions like banks and microfinance institutions worldwide. It, further, can discourage firms from investments. Therefore, financial institutions should now be able to draw useful lessons or insights from the past experiences. This project seeks to provide a method that would help such companies evade such risks, allow them to plan strategies ahead to avoid negative outcome and support business continuity.

## Problem Formalization

This is a classification problem, the algorithms used to solve this problem extracts or mines the relationships between user attributes/features and the possibility of failing to meet the contractual loan requirements.

This project therefore seeks to classify loan applicants based on some attributes. The algorithm, given the user data, tries to find the chances of a bank borrower turning down the terms and the obligations of the contract, the expected output of the model will be categorical either if an applicant is a loan defaulter or not.

## Proposed Approach for Solving the Problem

Data mining is always are procedural task with most emphasis put in data cleaning and preparation for better results. Credit risk analysis will involve the following steps.

1. Data acquisition – the dataset is acquired from online source
2. Data analysis and cleaning – The dataset is analyzed for important insights that aid in choosing the best features for modeling, the data is observed for missing values and appropriate imputation methodologies are applied depending on the type of features such as categorical and numerical. Outliers are also treated appropriately.
3. Data transformation – The dataset is transformed in a form that is best suitable for modeling, to aid faster training and better performance.
4. Data preparation – The dataset is then split into training and test sets, that will then be used in training and evaluating the model performance on unseen data.
5. Modeling – Several machine learning algorithms are trained on the data with default parameters, the results are noted, we then use cross validation to find the best parameters for the same algorithms.
6. Evaluation – The trained models are then evaluated on test data.

## Data Sets

For this project we use credit scoring dataset that has a number of client attributes such as applicants’ income, last amount borrowed, credit limit and much more. The target variable is Boolean, `target\_default`, either True, of False. This dataset is acquired from a digital Bank, Nubank, a Brazilian digital Bank and one of the biggest FinTech’s in Latin America. The company is known to be data driven hence it relies much on data mining to make educated decisions. The company takes advantage of data mining to improve on services and maximize profits.

## Data Mining Algorithms

This project experiments three learning algorithms to determine the ones that generate best results, XGBoost, LightGBM algorithms and Random Forest.

Lightgbm and Xgboost are boosting algorithms, they are generic in their unlike ensemble models. They work by improving the predictive power of base model by learning and optimizing parameters of weak models in a sequential manner, example of weak model is regression, each model in the sequence compensating the weakness of the previous ones or its predecessors.

Random Forest is an ensemble learning algorithm, ensemble models are ‘meta’ models meaning the model is a combination of a number of different independent individual models. If a predefined condition or criteria is met, these models can learn from the knowledge of group models and minimize error rate by greater magnitudes on new and unseen data in data mining.

## Results and Discussion

Training the classifiers with default parameters as in their documentations gives the results as in the table below.

|  |  |
| --- | --- |
| ALGORITHM | RECALL |
| XBGClassifier | 0.631103 |
| LGBMClassifier | 0.648720 |
| RandomForestClassifier | 0.655921 |

The following results are obtained during cross validation and hyperparameter tuning:

XGBoost Algorithm:

|  |  |
| --- | --- |
| PARAMETERS | RECALL |
| NEstimators= 50 | 63.25 % |
| NEstimators= 50, MaxDepth= 1, MinChildWeight= 6 | 65.95% |
| NEstimators= 50, MaxDepth= 1, MinChildWeight= 6, Gamma= 0 | 65.61% |
| NEstimators= 50, MaxDepth= 1, MinChildWeight= 6, Gamma= 1, LearningRate= 0.0001 | 82.37% |

LightGBM Algorithm

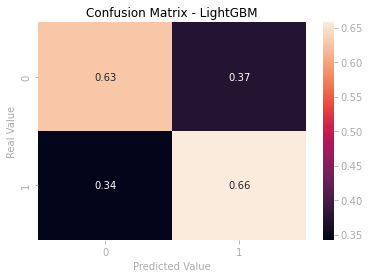
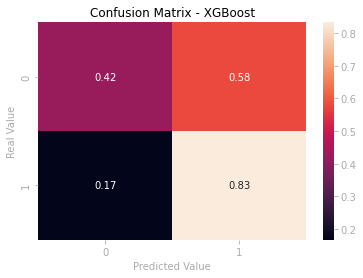
|  |  |
| --- | --- |
| PARAMETES | RECALL |
| LearningRate= 0.001, MaxDepth= 15, NumLeaves= 20 | 71.59% |
| LearningRate= 0.01, MaxDepth= 5, NumLeaves= 50, MinDataInLeat= 400 | 69.64% |

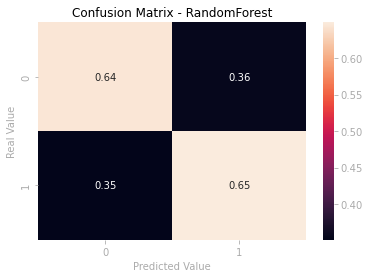
Random Forest Algorithm

|  |  |
| --- | --- |
| PARAMETES | RECALL |
| MaxDepth= 6, MaxFeatures= 120, Estimators= 150 | 68.62% |

After tuning hyperparameter tuning all three models recorded better results. XGBoost presented a great increase in recall, while LighGBM and Random Forest displayed a relative increase in performance.

The following are confusion matrices after evaluating the models on test data





## Conclusion

The primary objective of this project was to come up with a data mining model that would be able to identify potential loan defaulters and therefore minimize the tendency of a company suffering losses.

The best prediction hypothesis or algorithm would be the one that could minimize the rate of false negatives, by identifying all defaulters among the client base, while minimizing the rates of false positives, by preventing data points or clients to be misclassified as defaulters.

There always exist a tradeoff between precision and recall, increasing either of them often result in decreasing the other and vice versa, this therefore makes the task somehow challenging and tricker to attain a state-of-the-art performance, in such cases company analysts have to decide on which to sacrifice to optimize the opposite. The main target of credit risk analysis and management is to maximize and optimize a banks risk-modified ROI maintaining credit risk exposure within acceptable range of parameters this analysis would minimize company loss and maximize on profits, this project gives more emphasis on minimizing False Positives, searching through a set of parameters to identify the best hyperparameters that could possibly optimize the recall rate.

Among the three machine learning algorithms, XGBoost recorded the best performance as far as recall rate is concerned. XGBoost attained a recall rate of 82%, although recorded a poor False Positive Rate of 56%. Furthermore, LightGBM boosting algorithm and Random Forest Classifier ensemble algorithm attained a relatively better false positive rate count of 37% and 36% respectively. It should be noted that False Negatives for these two algorithms are relatively higher as compared to that of XGBoost.