UNIVERSITY OF RWANDA
COLLEGE OF BUSINESS & ECONOMICS

AFRICAN CENTER OF EXCELLENCE
IN DATA SCIENCE

ACE-DS

Bank Loan Approval Prediction Using Machine Learning Techniques

By

Theoneste NDAYISENGA

Registration Number: 215042343

A Dissertation submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science in Actuarial Science at the African Center of Excellence in Data Science in University of Rwanda.

Supervisor: Dr. Charles RURANGA

Kigali, June 2021

Declaration

I declare that this dissertation entitled “**Bank Loan Approval Prediction Using Machine Learning Techniques”** is the result of my own work and has not been submitted for any other degree at the University of Rwanda-African Center of Excellence or any other institution.

Date: June 2021 . . . .

Theoneste NDAYISENGA

Date: June 2020 . . . .

Dr. Charles RURANGA

Approval sheet

This dissertation entitled "**Bank Loan Approval Prediction Using Machine Learning Techniques**: Application of data science in financial industry" written and submitted by Theoneste NDAYISENGA in partial fulfilment of the requirements for the degree of Master of Science in Data Science in Actuarial Science is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 16% which is less than 20% accepted by African Center of Excellence in Data Science (ACE-DS).

…………………..

Supervisor: Dr. Charles RURANGA

……………………..

Head of Teaching: Dr. Ignace KABANO

Dedication

Almighty God who created and kept me healthy during my research.

To my beloved Father and Mother for your motivation and encouragement.

To all my relatives and dear friends.

To my brothers and sisters (all).

To my classmates especially those who helped in my studies,

This work is dedicated to you.

Acknowledgment

I am grateful to Almighty God for his love, protection, mercy and blessings throughout my life.
I am deeply grateful to my supervisor Dr. Charles RURANGA for his great effort and kindness; without it, this work would have never been completed. I thank him a lot for having shown me this way of research.

He could not even realize how much I have learned from him. He always monitored. I am particularly grateful to my fellow students, my brothers and sisters for their immeasurable support rendered to me during the study period.

I would like to thank my mother, father for their endless moral and economic support. I would like to express my sincere appreciation to my brothers and sisters for their considerable valuable and moral support and my classmates, friends and relatives.
Please, consider my thanks to you all.

Abstract

Loan is the essential product of banks and other financial institutions. As a big number of people go to banks to borrow money for different activities, the number of customers have increased and some banks expect to earn a lot of money as a result of interest paid on loans. However, loans are associated with risk of defaulting, i.e. the possibility that some borrowers may not be able to pay back their loans. Thus, high levels of non-performing loans can be a source of instability of the banking sector and lead to bankruptcy. One of the important steps for banks to decide if a loan has to be authorized is to ensure that the candidate to borrow has the capacity of paying back the loan in the proposed terms. The advancement of technology like machine learning, computer science and other science is playing an important role by supporting banks to predict the probability of defaulting for a given customer based on his past behavior.

In this research, we contribute to work by commercial banks to predict the behaviors of borrowers by developing and testing the accuracy of different models using data from Bank of Kigali. Collected data was divided into training dataset and test dataset where the train dataset was made by 70% and 30% was for test. After training the machine by using the training dataset, then we used the test dataset to check the accuracy of different models. By running ensembles, combinations of different machine learning techniques are used to find the best to use while predicting bank loans default prediction. The results of our analysis show that the Gradient Boosting is the best model to predict bank loan default, followed by XGBoosting while others like decision trees, random forest, logistic regression performed poorly.

I would recommend financial institutions to use machine learning techniques because it saves money and time for both sides. Moreover, Findings show that the customers with Credit Score B will have low probability of defaulting. In this work we used data from one financial institution and we would recommend anyone who might want to further this study to consider using data from different financials institution across the region to capture the insight

Keywords: Bank Loan, Classification, Creditworthiness, Machine Learning, and Prediction

Table of Contents

List of figures

List of Tables

List of Abbreviations

ACE-DS: African center of Excellence in Data Science

UR: University of Rwanda

SVM: Support Vector Machine

CBE: College of Business and Economics

KNN: K nearest neighbors

XGBoosting: Extreme Gradient boosting

Chapter I: General Introduction

1.1 Background and rational of study

Banks' Loans have become an important source of external financing for firms and households' due financial constraints to develop firms and business. For commercial banks, the activity of lending to the economy is very beneficial and loans represent a big part of banks' assets. However, the increase of loan lending is associated with a number of risks, such as risk of defaulting or credit risk, which is linked to the inability of the borrower to pay back the loan at the agreed time and conditions. If the debtor pays back the loan, then the creditor would earn the profit from the borrowed money. However, if the debtor fails to pay back the borrowed money, then the creditor loses his/her interest and the invested money. Therefore, creditors are facing the problem of forecasting the risk of a debtor being unable to pay back a loan.

Credit risk is known to be among main determinants of financial instability. Because the activity of lending is not only seen by commercial banks as source of profit but also associated with high risks, commercial banks try to minimize defaulting risks by assessing the capacity of the borrower to pay back the loan and request for collateral prior to the supply of the loan (Calcagnini, G., Cole, R., Giombini, G. and Grandicelli, G., 2018). This exercise is achieved by employing highly qualified professionals in commercial banks to evaluate whether a candidate was eligible for receiving a loan verifying the worthiness of a candidate for loan approval or rejection based on different criteria and leading to a numerical score. Very recently, with the development of technology, machine learning algorithms and neural networks were developed to automatically predict the credit score of an individual based on their historical data and sift through credit defaulters from the lot before the loan is approved (Perera, H.A.P.L., at.al, 2016; Marqués, A.I., at. al 2012; Adewusi, A.O., at.al, 2016; Choudhary, G., at.al. 2019; Atiya, A.F., 2001).

To better understand the development in the use of technology to predict the credit score, a comprehensive literature review was done to study the key factors (features) to be taken into consideration in this kind of work by focusing more on the application of data science techniques for the prediction and classification of the loan defaults.

Loan default prediction is one of the most critical and important problems faced by lending institutions like banks and other financial institutions as it has a huge effect on profit and development. Even if many traditional methods exist for mining information regarding loan application, most of these methods seem to be underperforming as there have been reported increases in the number of bad loans.

Throughout the years the machine learning techniques were used in calculation and prediction of credit risk by evaluating the historical data of a single person. This study will play a great role for financial institutions to evaluate the creditworthiness of the borrowers and by taking calculated risks by taking into consideration different Variables to predict the risk associated with the borrowers. Acquiring loans for different purposes such as the home loan, education loan, car loan, business loans etc., has become part of our day-to-day life from different financial institutions like credit unions and banks. However, many people are unable to determine the total amount of credit that they can afford to pay back.

Analysis of creditworthiness is one of the most important for banks and other financial institutions to stay working in the highly competitive market and for their profitability. They must set clear and defined criteria for lending. These criteria must be sufficient and adequate to provide the required information about the structure of credit, borrowers, and mode of payment.

Everyday Financial institutions (Bank, Microfinance and so on) receive a huge number of credit applications from their different customers, however, all loan applicants will not get the approval of the banks or financial institutions. Most of the banks use their own criteria of credit scoring models and risk assessment techniques while analyzing the loan application aiming at making decisions on whether to approve the application or not. In this research project, Machine learning techniques will be used to study the historical credit data of customers to extract patterns from their repayment data, which would help to predict the likelihood of defaulting, thereby helping the financial institutions in Rwanda for making better decisions in the future regarding the loan applicants.

Credit has been playing an important role in the financial world since many years ago, and it is quite profitable and beneficial for both the creditors and the debtors. However, it carries a high risk. With this, financial specialists and Researchers around the whole world are trying to make it less risky by leveraging technology like machine learning, neural networks, Artificial intelligence and so on.

Machine learning has been used to allocate people with numerical values known as credit scores to measure the risk of defaulting and creditworthiness. Also, it has been used to predict and evaluate the credit risk associated with an individual by referring to his/her historical data. This project reviews the present literature on different techniques and work related for predicting risk assessment of loan default that use supervised machine learning algorithms and will help to predict the likelihood that companies or individuals will be unable to make the required payments on their debt commitment or obligations. In other words, it is the probability that if you lend money, there is a likelihood that they will not be able to give the money back on time with agreed interest on a given loan.

This research has two objectives. The first objective is to find a classification solution that will be as accurate as possible at predicting whether a borrower will be paid off or default.

This will help potential investors during evaluations of loan applications to decide whether borrowers are worth credits or not. The second goal of this work is to find and investigate relationships and associations among the attributes that contribute to the repayment that can be used to to discover possibly hidden knowledge that can be valuable to potential investors. For finding classification models for default prediction we compared the performance of classifiers with various tuning parameters: Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), k-Nearest-Neighbors and Artificial Neural Network (ANN). For finding hidden associations between various attributes of loan applications we used the Apriori algorithm which helped to find interesting and meaningful rules.

1.2 Problem Statements

The importance of developing machine learning techniques has been recognized by the industry and different models have been built on to support different industries.

This research will review different models and propose the one to be used to predict the risk of defaulting from borrowers in the Rwandan banking sector.

My project intends to develop supervised machine learning models using binary classification techniques to calculate the probability or likelihood and risks associated with each borrower to payback the given loan with agreed interest. Limitation in my research project is that the dataset has many variables and the relationships between variables in the dataset are not linear which will result in some errors and lower the accuracy. In this research, I will assume that the relationships between the variables are linear, which is not true a hundred per cent because my dataset is not limited to being linear.

1.3 Objectives:

This research aims to solve the problem of identifying loan defaulting based on the different factors and will help financial institutions in Rwanda to reduce the risks of defaulting from the borrowers. This idea comes from the fact that different banks in Rwanda launched projects of data scientists and data engineers. I intend to develop a prediction model using binary classification techniques for defaulting borrowers. This method will help banks to find the easy way of classifying their customers due to different variables provided by customers and their history. We will use the case of the Bank of Kigali which is the biggest commercial bank in Rwanda and the first to launch data scientist and data engineer projects. Using the dataset of the bank of Kigali, we discovered that more than 23% of customers failed to pay back the loan on time.

1.3.1 Main objectives

The long-term goals of this research are to develop loan default prediction classification models. The objective of the current study is to provide a comprehensive review of literature and industry practices in relation to machine learning and outline a conceptual framework for loan analysis. Particularly, the study has the following sub-objectives:

1. To provide a comprehensive review of sources and benefits of using machine learning in finance 2. To develop a classification model which will help the financial to evaluate and predict the defaulters 3. To review current industry practices and researches in regards to risk modelling 4. To outline a conceptual framework for Bank loan default prediction by using Machines learning; The result of this study will be valuable to the industry practitioners as well as related software providers in developing better practice and tools for loan default prediction related projects

1.3.2 Specific objectives

Specific objectives are detailed objectives that describe what will be researched during the study:

1. To find the features that contribute to lowest probability of default in Bank of Kigali clients
2. To find the loan grade that is associated with highest default rate in Bank of Kigali clients
3. To assess the creditworthiness of customers of Bank of Kigali

1.4 Research questions

This research, will respond to the following questions

1. What model can be used for loan default prediction in the Rwandan banking sector?
2. What are the types of data that are very correlated to the low probability of default?
3. What are the benefits of using machine learning in terms of time consuming, and accuracy?

1.5 Significance of the study

As for other academic studies, once successful completed, the report of this research will be available in UR's libraries, and it can be used by any students or academicians or professionals as reference document about the application of machine learning and big data in financial industries include; but not limited to Bank loan prediction using machine learning techniques and related work.

1.6 Structure of the dissertation

Bank Loan Approval Prediction Using Machine Learning dissertation is comprised of V chapters and references: Chapter I provides a general introduction of the study including the background of the study, problem statement, significance of the study, research objectives, research questions and structure of the study; chapter II provides a literature review; chapter III describes the methodology used in the study; chapter IV shows the Data analysis and discussions and the last one which is chapter V provides conclusion and recommendation.

Chapter II: Preliminary Literature Review

Classification: is the one of the machine learning models where the model tries to predict the out label as discrete. In this work I will use a binary classification model which predicts the binary outcomes like (0 or 1, Yes or Not, Default or Not) and so on. Machine learning is the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. With this in mind, I will use supervised machine learning techniques to draw the patterns and insights from the dataset, and this will help algorithms to predict the new dataset without a label.

As has mentioned (Moin & Ahmed 2012), Machine learning can learn to classify the data items into different groups in classification methods. Classification process includes two key steps: first is the learning step which is building the classification model by using a small portion of the dataset called training dataset and second is the classification step itself which is using a classifier for classification. For this step, you use the remaining portion of the dataset called test dataset in order to test the performance of built algorithms. In learning, the classification algorithms build the classifier. The classifier is built from the training dataset and their associated class labels (variables). In classification, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification models. The classification modules can be applied to the new datasets tuples if the accuracy is considered acceptable, and this will depend on the confidence level or tolerance level. The classifier-training method uses these pre-classified examples in order to determine all the variables (parameters) necessary for proper judgment or analysis. The algorithm then encodes these parameters (variables) into a model called a classifier.

From a machine learning outlook, the loan default prediction (or credit scoring) can be viewed as a binary classification problem. Previous approaches focus on prediction using ensemble methods or fuzzy systems. Neural networks, successfully applied to various fields include; financial institutions, data mining, cybersecurity, also has application in the default prediction problem. More recently, there have been many machine learning techniques-oriented approaches to solving this problem of loan default prediction. A significant proportion of these approaches are based on Random Forests, Linear Regression, classification and combination of other ensemble methods.

Data mining was proposed in Customer Relationship Management (CRM) (Raju, Bai & Chaitanya 2014). The supervised learning technique, Decision Tree, implemented using the Classification model and Regression Trees (CART) algorithm is used for customer continuation possession (retention). One of the sectors which applied the data mining is Customer Retention in banking which also depends on two methods: Classification Methods and Value Prediction Methods. In Value Prediction Methods, there is another way to classify the application of new loans, it attempts to predict conventional amounts for new loan applications by using the Neural Network.

Preliminary literature review shows the past studies focused on the loan default of online lending clubs and online lending institutions for credit risk assessment by using historical data of customers and FICO score to measure the risk associated with the borrowers so that the data collected from online was not enough and very complex to clean.

Abdou & Pointon (2011) orchestrated an in-depth analysis of credit scoring in various fields and finally concluded their thorough research by saying that there is no single overall best classification technique for credit scoring models. However, they were in accordance with Hand and Henley (1997), who said that the performance of classification methods precisely depends on the convenient size and variables on the datasets, data structures or just the objective classification.

Loan default trends have been long studied from a socio-economic viewpoint. Most economics papers believe in empirical modelling of these complex systems in order to be able to predict the loan default rate for a particular person. The use of machine learning for such tasks is a trend which we are perceiving now.

Blanco, Mejias, Lara & Rayo (2013) develop many credit scoring models that are based on the multilayer approach. The work proves it performs more than the other models that use logistic regression techniques. The outcomes show that the neural network model works better than the other three techniques.

T. Harri compares support vector machine-based algorithms for credit-scoring developed using the various default definitions. The work concluded that the wide definition models are better than the narrow definition models in their performance. Financial data analysis was done in using techniques such as Decision Tree, Bayes classification, Random Forest, Bagging Boosting, algorithm and others. Perceptron model, Support Vector Machine, Logistic Regression, Decision Tree, Neural Network, all these techniques are combined in this model. The effectiveness of applying the above methods on credit scoring is studied. The analysis results show the performance is outstanding based on desired accuracy.

Zhou and Wang (2012), proposed improvements for the Random Forest models and it shows efficacy for different customers' loan datasets include; imbalance and big datasets.

Kurapati and Bhansali (2018), their work shows that loan default prediction models can be used by many people as possible if the model has high accuracy. And showed that the Random Forest algorithm performed better than other models like Decision Tree, Gradient Boosting to identify the credit defaulters.

Addo, Guegan, and Hassani (2018) Work shows that the choice of variables to respond to business objectives and the choice of the algorithms used to make a decision are two important key aspects in the management processing when issuing the loan.

Credit scoring has become an essential tool in the highly competitive financial world, which has brought more focus on research on credit risk assessment in the recent year by Sarma (2013).

Abdou and Pointon (2011) work shows that there have been a multitude of ways of carrying out a particular task which was used to allocate credit scores and much research has been done on the topic since many years ago. Different from before, where the initial models were determined by professional opinions for assessing the loan worthiness of an individual person, recently focus has shifted towards applying advanced machine learning techniques and neural networks for credit scoring and risk assessment. These techniques can be further classified into two key categories: traditional statistical techniques and advanced statistical techniques.

Previously lenders and financial institutions employed highly professional people to evaluate a borrower's worthiness before approving or rejecting a loan application. However, recently these institutions have started employing various models for loan evaluation in order to decide whether to reject or approve a loan to a borrower based on their credit score and ability to repay a given loan. The models that they have picked out are based on machine learning models along with artificial neural networks for accurately predicting credit defaulters among the borrowers. These models typically assign a numerical score (1 or 0), which represents the creditworthiness of the borrowers based on historical data (Nalica & Švraka, 2018).

Many statistical models are used to credit scoring, such as logistic or linear regression, probability analysis, linear discriminant analysis, and naïve Bayes (Hand & Henley, 1997). However, these methods often perform unsuccessfully when dealing with nonlinear relationships. It is difficult for them to meet the given statistical assumptions in an empirical application. Therefore, artificial intelligence methods and many machine learning have been applied to credit scoring in different sectors including banks, and these algorithms have performed better than statistical analysis. These methods include the Support Vector Machine (SVM), Random Forest and Artificial Neural Networks (ANN) Harris (2015).

The financial institutions rely on three kinds of data to build credit scoring: customer's transaction history data, demographic information, and credit history data Thomas (2010).

The social media boom has generated an enormous amount of social network data. The data gathered from social media may be considered an important origin of information. applied social network data to empirical research. They observed social network data from calling conduct to predict product/service adoption. In order to reduce information asymmetry, many peers to peer lending platforms motivate borrowers to build online groups. The practical results of research reveal that the nearby the location of the borrowers in the group, the lower the default probability. In addition, if they have a definite connection in real life, joining the group will significantly reduce the borrower default probability. According to social media disclosure by the borrower, borrowers with many friends are more likely to obtain loans, and the default probability is very low (Lin, Prabhala, & Viswanathan, 2013).

In general, there has been significant interest in loan default prediction, and many algorithms have been tested. Ensemble algorithms have been recommended for performance. In order to ameliorate the performance of the classification model, related research verified "hard information" such as income, age, FICO score, and debt information, and they have also focused on the "soft information" of borrowers. In this aspect, social network information has gained much attention. Some intellectuals have investigated the role of social network information, but most only explored the correlation between online social network information and loan default. Although (De Cnudde, Moeyersoms, Stankova, Tobback, Javaly, & Martens, 2019) extracted social network data from Facebook to increase the predictive capability, this method cannot be applied in many countries. For example, Facebook is closed in some countries. In addition, many people are reluctant to give online social network information or do not have social media accounts, and some people provide poor data and their social network information cannot be obtained, which leads to the failure of the method and limits the improvement of its risk prediction capability.

Due to tremendous extension in data the banking industry deals with, analysis and transformation of the data into useful knowledge has become a job beyond human ability. Data mining techniques can be adopted in answering business problems by finding patterns, associations and correlations which are masked in the business information stored in the databases. By employing data mining techniques to investigate the patterns and trends associated with them, bank executives can predict, with high accuracy, how customers will respond to adjustments in interest rates, which customers are likely to receive new product offers, which client (customers) will be at a higher risk for defaulting on a loan, and how to make client (customer) relationships more profitable by Jesse (2018).

In the lending industry and financial institutions, the lenders normally assess the repayment ability of the loaners and the risks of lending money to borrowers (clients). Based on the repayment ability and risks associated, the lenders, especially the banks, can balance the interest rates of the loans which are issued to the client or borrowers (Gorton, Gary, and James, 2000).

Chapter III: Methodology and Modeling

3.0. Machine learning

Machine Learning is a kind of algorithm that permits software applications to become more accurate in predictability without being explicitly programmed. a subset of AI supports the thought that a system can learn from data, identify the pattern and make decisions to urge optimal solutions with minimum human intervention. There are two sorts of ML algorithms, supervised machine learning algorithms and unsupervised machine learning algorithms.

3.1 Logistic Regression

Logistic Regression model is a Machine Learning classification method (algorithm) that is used to forecast or predict the probability of a categorical dependent factor. In a logistic regression model, the dependent variable is a binary that contains data coded as 1 (yes, etc.) or 0 (no, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

Logistic Regression is one among most popular useful models for categorical data, especially for binary response data in data modelling. Unlike rectilinear regression models, logistic regression models can directly predict probabilities (values that are restricted to the (0,1) interval); furthermore, these probabilities are well-calibrated in comparison to the possibilities predicted by other classifier models, like Naive Bayes.

Logistic regression preserves the marginal probabilities of the training data. The multiplier of the model also gives some hints about the relative importance of every input variable. Let us consider again the bank of Kigali Loan dataset where the payment status falls in two categories (completely paid or other) as the figure below shows that the probability of defaulting falls between 0 and 1.

$P(X) = Pr(X = 1/X) \ldots\ldots\ldots\ldots\ldots\ldots Equation\ 4.a$

Also we can use Linear Regression as equation

$p(X) = Bo + B_1 \quad X \ldots\ldots\ldots\ldots\ldots\ldots Equation\ 4.b$

By using prediction approach (Completely paid=0, Default=1)

While working with Logistic Regression we use the logistic function in order to avoid that the probability of $P(X)$would go beyond 0 and 1.

$p(X) = \frac{e\beta0+\beta1X}{1 + e\beta0+\beta1X} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots Equation\ \ 4.c$

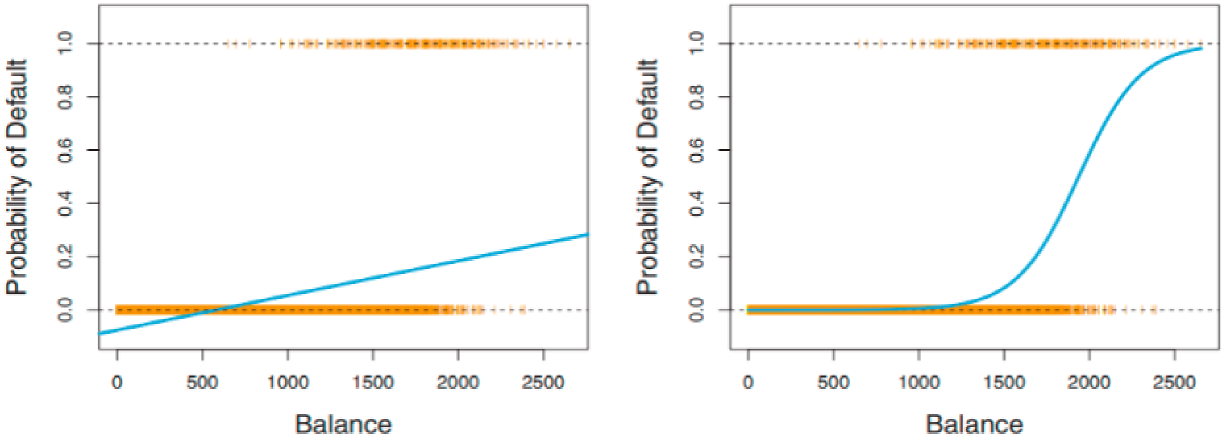


Figure 1. Probability distribution

3.2 Decision Tree classifier

Decision trees work as classification models and they have two steps; the first step is called the learning step while the second step is called prediction steps. In the learning step is where the machine learns from the given dataset and develops based on it. The prediction step is where the machine uses the patterns from the learning step to predict the response of the given dataset.

Decision trees are one the most useful types of supervised learning algorithms that are mostly used for classification problems. Furthermore, it works for both continuous and categorical dependent variables. In this algorithm, we split the dataset into two or more homogeneous sets. This is done based on most significant independent/attribute features to make as recognizable groups as possible. Decision Tree is a tree where each node represents a feature, each branch left or right represents decision (rule), and each leaf represents an outcome (categorical or continuous value). There are quite a handful of articles about decision trees. Some articles give you a detailed explanation about decision trees, including information on what's a decision tree, how to generate trees, how to do pruning, and why we should use decision trees. Decision trees model in the form of tree structure and work as classification or regression models. Decision trees break the given dataset in smaller and smaller subsets.

3.3 Support Vector Machine

It is a classification method. In this model, we plot each data item as a point in n-dimensional space where n is the number of independent factors you have with the value of each factor being the value of a particular coordinate. Support

Vector Machine (SVM) is a supervised learning method with the goal of constructing a hyper-plane in a high-dimensional space, which could be used to segregate different populations. A very good support vector machine needs to create multiple hyperplanes for multi-classification or hyperplane which maximize the distance to the nearest training data points of any margin support vectors or class as this would reduce (lower) the generalization error of the classifier.

This could be expressed with the following optimization problem: In this research I will use the supervised machines, where we have an input variable and an output variable; Algorithms are used to learn the mapping function, from input to output. Supervised machine learning techniques mainly classified in two sub-groups, classification and regression. Classification deals with discrete outputs and regression model runs with continuous outputs. Support vector machines (SVM), Random Forest (RF), Logistic Regression, Decision Tree are the most popular and widely used supervised algorithms. For example, if we only had two features like Hair length and weight of an individual, we'd first plot these two variables in 2-dimensional space where each point has 2 coordinates (these coordinates are known as Support Vectors)

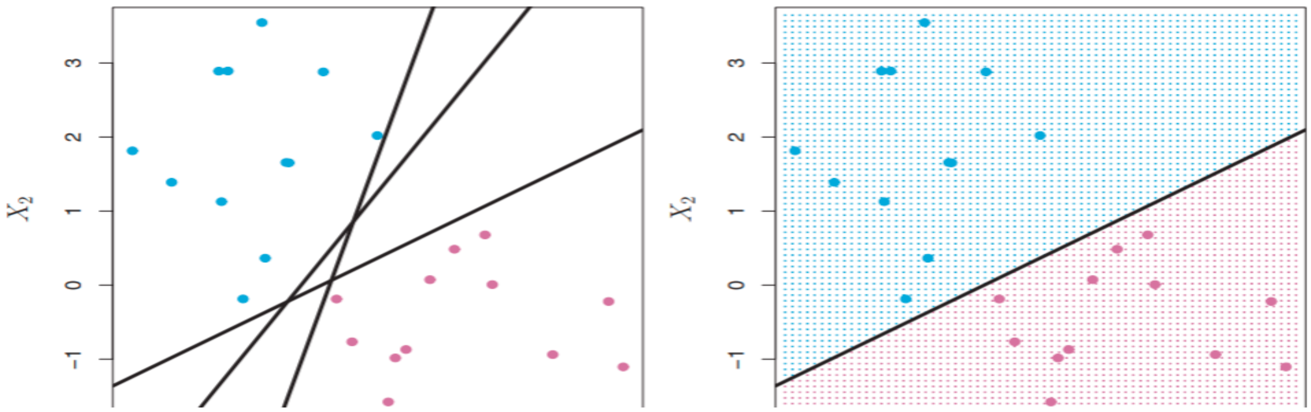


Figure 2.Variables in Support Vector Machine

3.4 Random Forest

Random Forest model is a trademark term for an ensemble of decision trees. In Random Forest, we have a collection of decision trees (so known as "Forest"). To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification having the most votes. Each tree is planted & grown as follows

If the number of cases in the training dataset is P, then a sample of P cases is taken at random but with replacement. This sample will be the training dataset for growing the tree

If there are N input features, a number n<<N is specified such that at each node, n features are selected at random out of the P and the best split on these m is used to split the node. The value of n is held constant during the forest growing

$MSE = \frac{1}{N}\sum_{i=1}^{N} \quad (fi - yi)^2$

Where N is the number of data points in from given dataset, $fi$is the value returned by the model and $yi$is the actual value for data point i

3.5 K Nearest Neighbors classifier

Is a simple machine learning algorithm that stores all available variables and classifies new variables based on a similarity measure (distance) **KNN** has been used in statistical estimation and pattern recognition already in the since1970's as a non-parametric technique. KNN classifiers use the distance to classify to class with its neighbors and this depends on the value of K. If K=1 means that the class is simply assigned to its neighbor by using distance function.

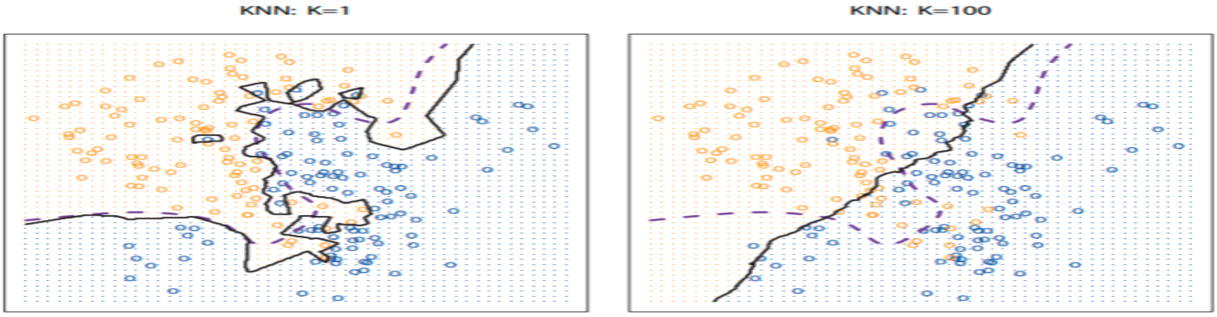


Figure 3.K Nearest Neighbors

Knowing the correct optimal value K is best by first controlling the data. In general, a high value of K is more precise because it reduces the overall noise in your data but there is no guarantee. Cross-validation is another way to consider a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. Distance function

$Euclidean = \sqrt{\sum_{i=1}^{k} \quad (x_i - y_i)^2}$ ……..….*Equation 4.5.a*

$Manhattan = \sum_{i=1}^{k} \quad |x_{i-}y_i|$ ……..….*Equation 4.5.b*

$Minkowski = \left(\sum_{i=1}^{k} \quad (|x_i - y_i|)^q\right)^{\frac{1}{q}}$ ……..….*Equation 4.5.c*

3.6 GaussianNB classifier

GaussianNB classier is the oldest machine learning algorithm which is in the form of two parts; The Naive Bayes formula (Theorem) and a distribution (in this case Gaussian one). We use Naive Bayes formulas in probability especially when we are calculating the probability of event A when we already know that event B happened.

GaussianNB helps to calculate the probability of framework for fitting the model for training dataset, referred to as maximum posterior and in development of models for classification predictive issues such Bayes Naive and Bayes Optimal classifier.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\delta_y{}^2}} exp\left(-\frac{(x_{i-}\mu_i)^2}{2\delta^2{}_y}\right) \ldots\ldots\ldots\ldots \textit{Equation 4.6.a}$$

3.7 Gradient Boosting

Gradient boosting is one machine learning technique used with decision trees of fixed size as base learners, and this method improves the quality of fit each base learner.

Gradient Boosting method is a classification method where trees are built in series and compared to each other based on mathematically scores of splits.

$$g_t(x) = E_y\left[\frac{\partial\psi(y,f(x)}{\partial f(x)}|x\right]_{f(x)=f^{t-1}(x)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots Equation\ 4.7.a$$

3.8 XGBoost

XGBoost is one of the decision-tree-based ensemble Machine Learning algorithms that uses a gradient boosting framework. It is used in classifications and predictions problems involving unstructured data. However, when it comes to small or medium tabular/structured data the decision trees are the most effective rather than Extreme Gradient Boost. It has many applications because it can be used in regression, ranking, user defined predictions problems and classification. XGBoost dominates tabular or structured datasets on regression and classification predictive modeling problems.

3.9 Data preprocessing

The Bank of Kigali dataset was used in this study and was collected internally with the Bank of Kigali, it contains all data for loan applicants from 2013 to 2018, and was used as a secondary data.
The Bank of Kigali dataset contains more than 58,000 observations and 29 variables. After cleaning data using python, I remained with more than 56, 000 observations and 24 variables. The independent variables are: 'Province', 'District', 'Customer Branch', 'Principal amt', 'Paid principal', 'Paid Interest', 'Paid Penalty' 'Total Remaining principal', 'Remaining principal', 'Amount due', 'Due Principal', 'Due interest', 'Due pen interest', 'Due Fee', 'Paid Fee', 'Due TAX', 'Paid Tax', 'Previous Days', 'Overdue Days', 'Effective Date', 'Maturity Date', 'CreditScoreGroup', 'Duration'

The dependent variable is: 'PaymentStatus'

In data analysis, different Machine learning techniques will be used to build models like Classification method, linear regression, Logistic regression, Random Forest, Machine Vector Support and Ensemble…

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from given data, identify patterns in data and make decisions with minimal human intervention.

Data preprocessing is an important task to be done prior to analysis to get the data ready for analysis. As good data can only provide better results. In data preprocessing, the proposed system performs data cleaning, data imputation, data normalization, and transformation.

Data cleaning process removes null values and redundant attributes from the dataset. Implementation of the proposed model applies the sampling technique (PCA) on the preprocessed dataset to balance it. On this sampled data, the proposed system implements the Machine Learning Algorithms to check which algorithm suits better, which algorithm is suitable for prediction. This system also compares the accuracy of algorithms before and after feature selection to select the best algorithm that predicts the defaulters effectively. The architecture of the proposed system is given below.
Python as a tool for data analysis will be used to clean data, split data into training set and test dset, training set is 70% while the test set is 30 %. I will use different models for the purpose of getting the best model to use, which will give high accuracy and less error.

Machine learning processes

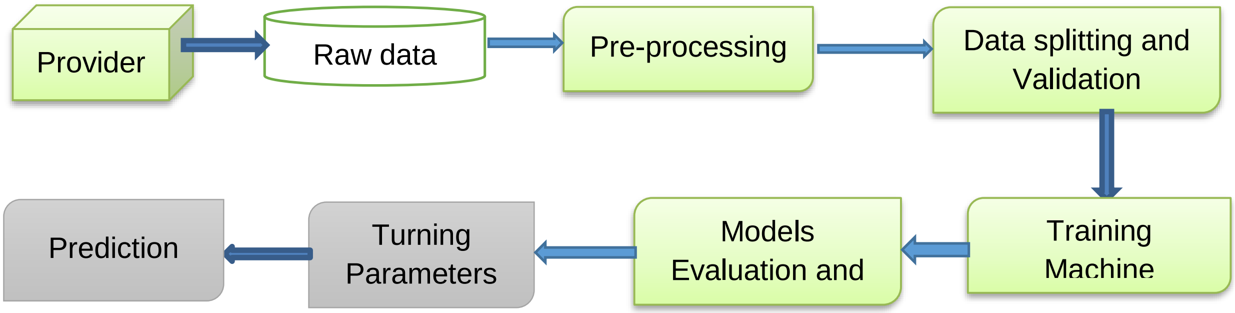


Figure 4.Machine Learning Process

Data Collection: The standard and quantity of your data dictate how accurate our model is, the result of this step is usually a representation of knowledge which we'll use for training. **Data Preparation:** Wrangle data and prepare it for training. Cleaning data includes; remove duplicates, affect missing values, correct errors, normalization, data type conversions than on. Randomize data, which erases the consequences of the actual order during which we collected and/or otherwise prepared our data. Visualize data to assist detect relevant relationships between variables or class imbalances or perform other exploratory analysis and split into training and evaluation sets **Choose a Models:** Different algorithms are for various task and you choose the proper one **Train the Model:** The goal of coaching is to answer an issue or make a prediction correctly. **Evaluate the Model:** Uses some metric or combination of metrics to "measure" objective performance of the model. Test the model against previously unseen data. This unseen data is supposed to be somewhat representative of model performance within the world but still helps tune the model.

Parameter Tuning: This step refers to hyper parameter tuning, which is an "art form" as against a science. Tune model parameters for improved performance. Simple model hyper parameters may include a number of coaching steps, learning rate, initialization values and distribution, etc. **Make Prediction:** Using further (test set) data which have, until now, been withheld from the model (and that class labels are known), are wont to test the model; a far better approximation of how the model will perform within the world.

Chapter IV: Data Analysis

Different classification learning methods are heavily dependent on quantity and the quality of the data provided for training of the model. In this chapter, we will have an overview of the loan repayment data set from Bank of Kigali and perform exploratory data analysis in order to preprocess the data and improve the prediction results. The data will also be split into test sets (30%) and the training sets (70%).

Training set is a data initial dataset that helps the program to understand how to learn and apply sophisticated technology. Also, training models determine the good values for all weights and bias from the labeled example. For the train dataset, the Machine learning algorithm builds models that examine many examples and attempts to find the model that minimizes the loss. The test data set is independent of the training data dataset, but it has the same probability distribution as the training dataset, and it is used to measure the performance.

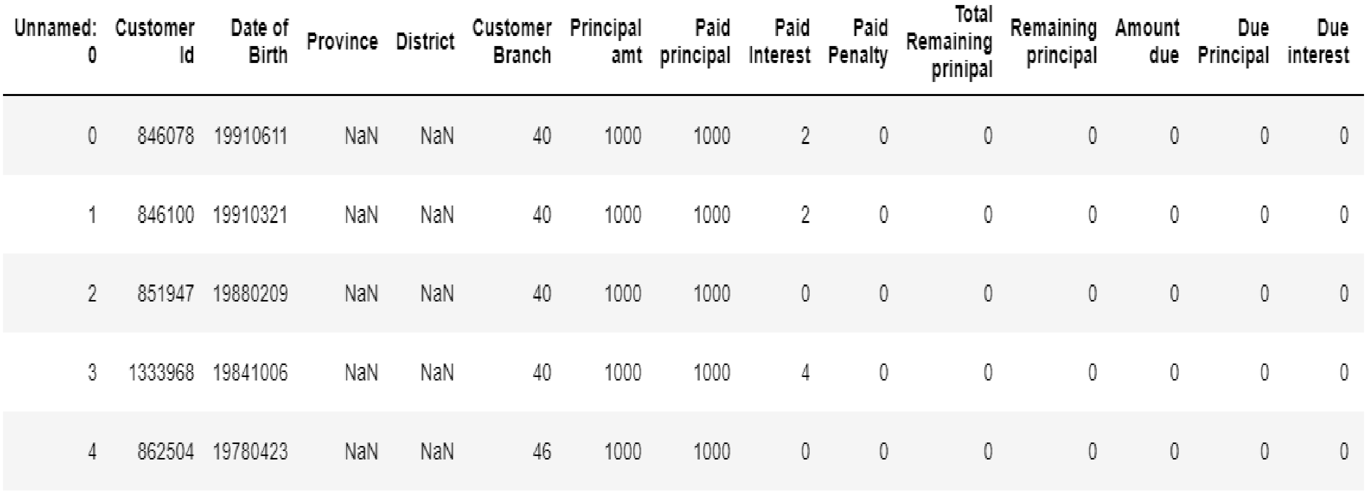
Training dataset will be used to fit the model, and the test dataset will be to evaluate the best model to get an estimation of generalization error.

4.1 Features analysis

In this study, the data used was from Bank of Kigali, the biggest financial institutions in terms of assets in Rwanda with more than 350,000 customers (Mpaka, 2019). Loan repayment data from Bank of Kigali contains more than 58096 observations and 29 variables.

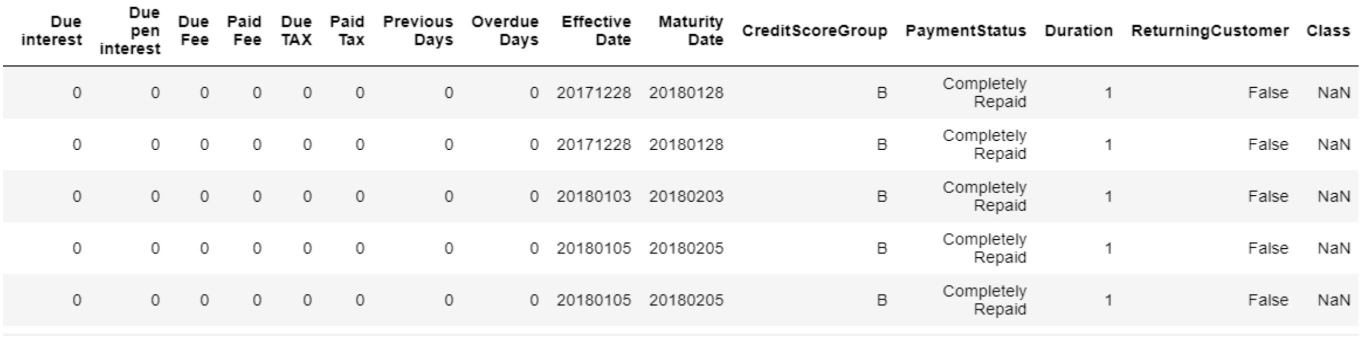
The tables below are the first five observations from the dataset before we create dummy variables.

Table 1 *The first five columns of dataset (left side)*



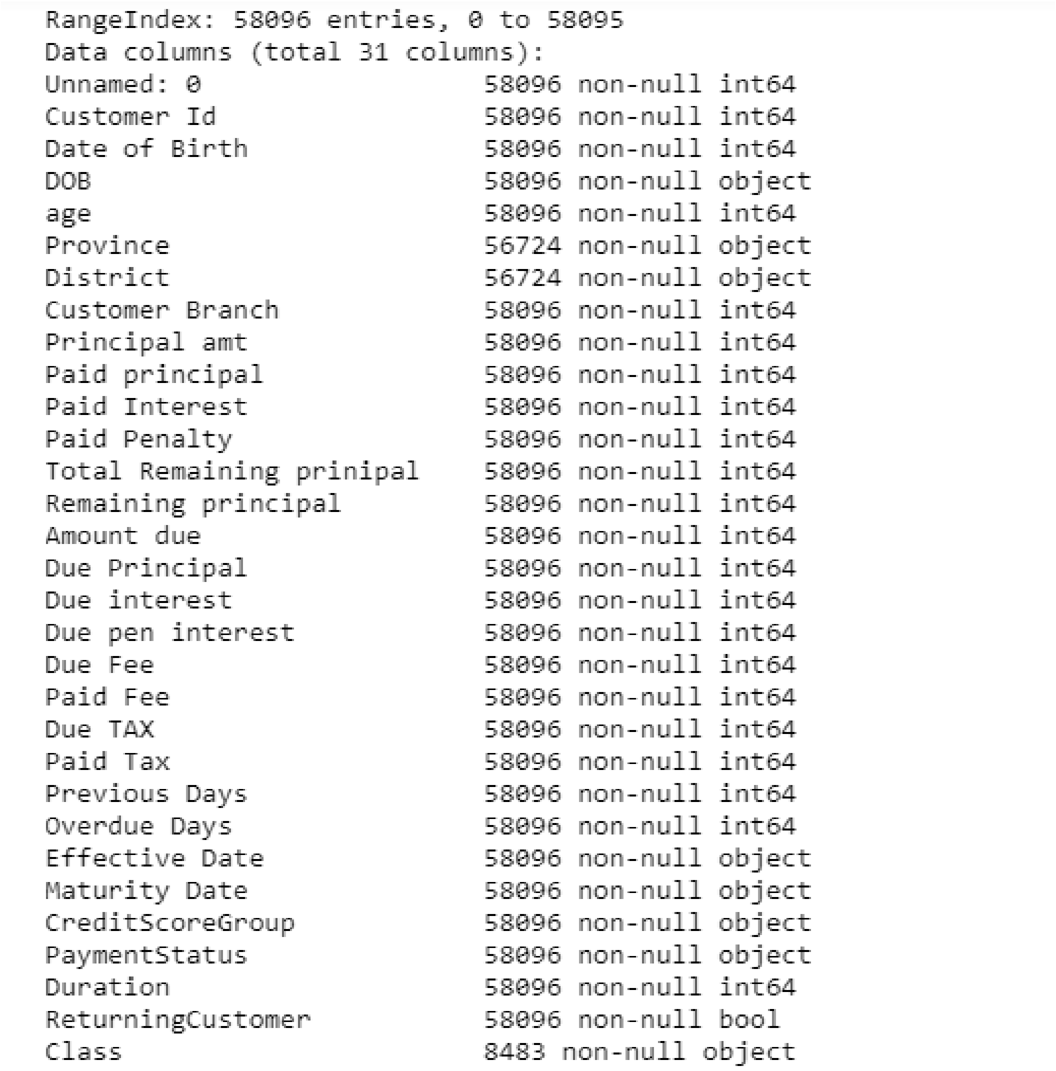
Unnamed: 0	Customer Id	Date of Birth	Province	District	Customer Branch	Principal amt	Paid principal	Paid Interest	Paid Penalty	Total Remaining prinipal	Remaining principal	Amount due	Due Principal	Due interest
0	846078	19910611	NaN	NaN	40	1000	1000	2	0	0	0	0	0	0
1	846100	19910321	NaN	NaN	40	1000	1000	2	0	0	0	0	0	0
2	851947	19880209	NaN	NaN	40	1000	1000	0	0	0	0	0	0	0
3	1333968	19841006	NaN	NaN	40	1000	1000	4	0	0	0	0	0	0
4	862504	19780423	NaN	NaN	46	1000	1000	0	0	0	0	0	0	0

Table 2 The first five columns of dataset (left side)



Due interest	Due pen interest	Due Fee	Paid Fee	Due TAX	Paid Tax	Previous Days	Overdue Days	Effective Date	Maturity Date	CreditScoreGroup	PaymentStatus	Duration	ReturningCustomer	Class
0	0	0	0	0	0	0	0	20171228	20180128	B	Completely Repaid	1	False	NaN
0	0	0	0	0	0	0	0	20171228	20180128	B	Completely Repaid	1	False	NaN
0	0	0	0	0	0	0	0	20180103	20180203	B	Completely Repaid	1	False	NaN
0	0	0	0	0	0	0	0	20180105	20180205	B	Completely Repaid	1	False	NaN
0	0	0	0	0	0	0	0	20180105	20180205	B	Completely Repaid	1	False	NaN

Table 3 *Type of data in Bank of Kigali dataset*



```
RangeIndex: 58096 entries, 0 to 58095
Data columns (total 31 columns):
Unnamed: 0                  58096 non-null int64
Customer Id                 58096 non-null int64
Date of Birth               58096 non-null int64
DOB                         58096 non-null object
age                         58096 non-null int64
Province                    56724 non-null object
District                    56724 non-null object
Customer Branch             58096 non-null int64
Principal amt               58096 non-null int64
Paid principal              58096 non-null int64
Paid Interest               58096 non-null int64
Paid Penalty                58096 non-null int64
Total Remaining prinipal    58096 non-null int64
Remaining principal         58096 non-null int64
Amount due                  58096 non-null int64
Due Principal               58096 non-null int64
Due interest                58096 non-null int64
Due pen interest            58096 non-null int64
Due Fee                     58096 non-null int64
Paid Fee                    58096 non-null int64
Due TAX                     58096 non-null int64
Paid Tax                    58096 non-null int64
Previous Days               58096 non-null int64
Overdue Days                58096 non-null int64
Effective Date              58096 non-null object
Maturity Date               58096 non-null object
CreditScoreGroup            58096 non-null object
PaymentStatus               58096 non-null object
Duration                    58096 non-null int64
ReturningCustomer           58096 non-null bool
Class                       8483 non-null object
```

The list above shows types, number of entries of each variable. In our dataset we have 3 types: int63[integer], object [string] and Bool [Yes, No].

Regarding the number of entries, class, province and district have few numbers of entries compared to others and this means that there are missing data in those 3 variables, where missing variables correspond to 85.3% for class, and 2.4% for province and district.

4.2 Dictionary for BK digital loan data

Customer Id: Customer's identification number
Date of Birth: Customer's date of birth
Province: Province in which the national Id was issued
District: District in which the national Id was issued
Customer Branch: Branch in which the account was opened
Principal amt: Disbursed loan amount
Paid principal: Amount paid of the disbursed amount
Paid Interest: Interest paid on the loan
Paid Penalty: Penalty paid due to late payment
Total Remaining principal: Total outstanding amount of the loan
Remaining principal: Outstanding amount minus amount for the next installment
Amount due: Total amount due for the current installment (due principal plus interest)
Due Principal: Due amount of the principal for the current installment (does not include interest)
Due interest: Interest amount due for the current installment
Due pen interest: Due penalty interest for the current installment
Due Fee: Processing fee paid on disbursement (1% of the principal amount on loans disbursed after 26th April 2019)
Paid Fee: Same as Due fee
Overdue Days: Days passed without paying due amount for each installment
Effective Date: Loan value date
Maturity Date: Date after which the loan expires if not paid
CreditScoreGroup: Credit score to which a given customer belongs to.
PaymentStatus: The status of the customer's current loan
Duration: Loan duration in months
ReturningCustomer: Indicates whether a customer is returning or new
Class: Customer's class depending on the number of overdue days
overdue days they have

Overdue Days >= 1 & Overdue Days < 30: Acceptable Risk
Overdue Days >= 30 & Overdue Days < 90: Special Mention
Overdue Days >= 90 & Overdue Days < 180: Substandard
Overdue Days >= 180 & Overdue Days < 360: Doubtful
Overdue Days >= 360: Loss while Overdue Days = 0: Normal

3.3 Correlation analysis

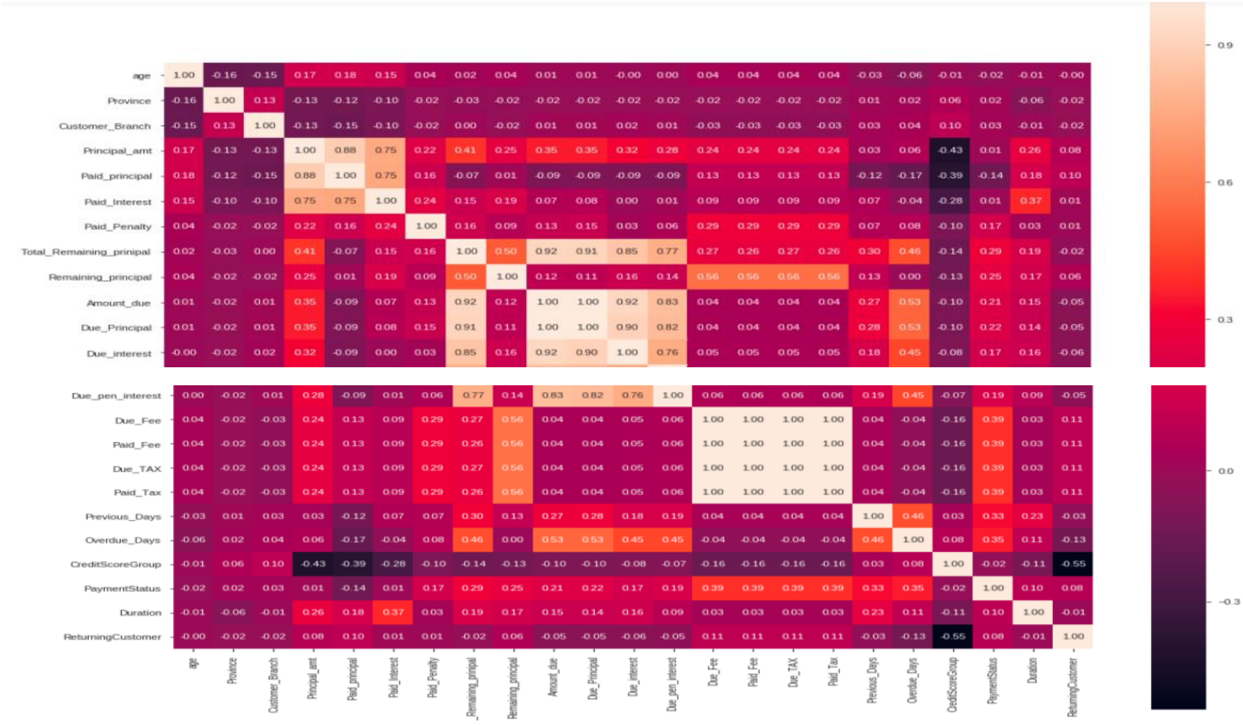
The correlation is the relationships between two or more variables, when the correlation between two or more variables is zero, this means that one variable can’t help to predict the other one and have no relationships between them, while correlation is equal to one, means that they are perfect correlated and increase of one variable cause of the increase of another with the same value.

Correlation is used in prediction from one variable to another and is used as a basic quantity of modelling techniques.

From the observation below, all variables are perfectly corrected to themselves, while others are correlated to certain extent. Here are the variables which are perfectly correlated; Due_fee, Paid fee, Due tax and Paid tax. Also, amount due and Due_principal are perfectly correlated.

Perfect correlation means that the value of one variable predicts exactly the value of another variable(s).

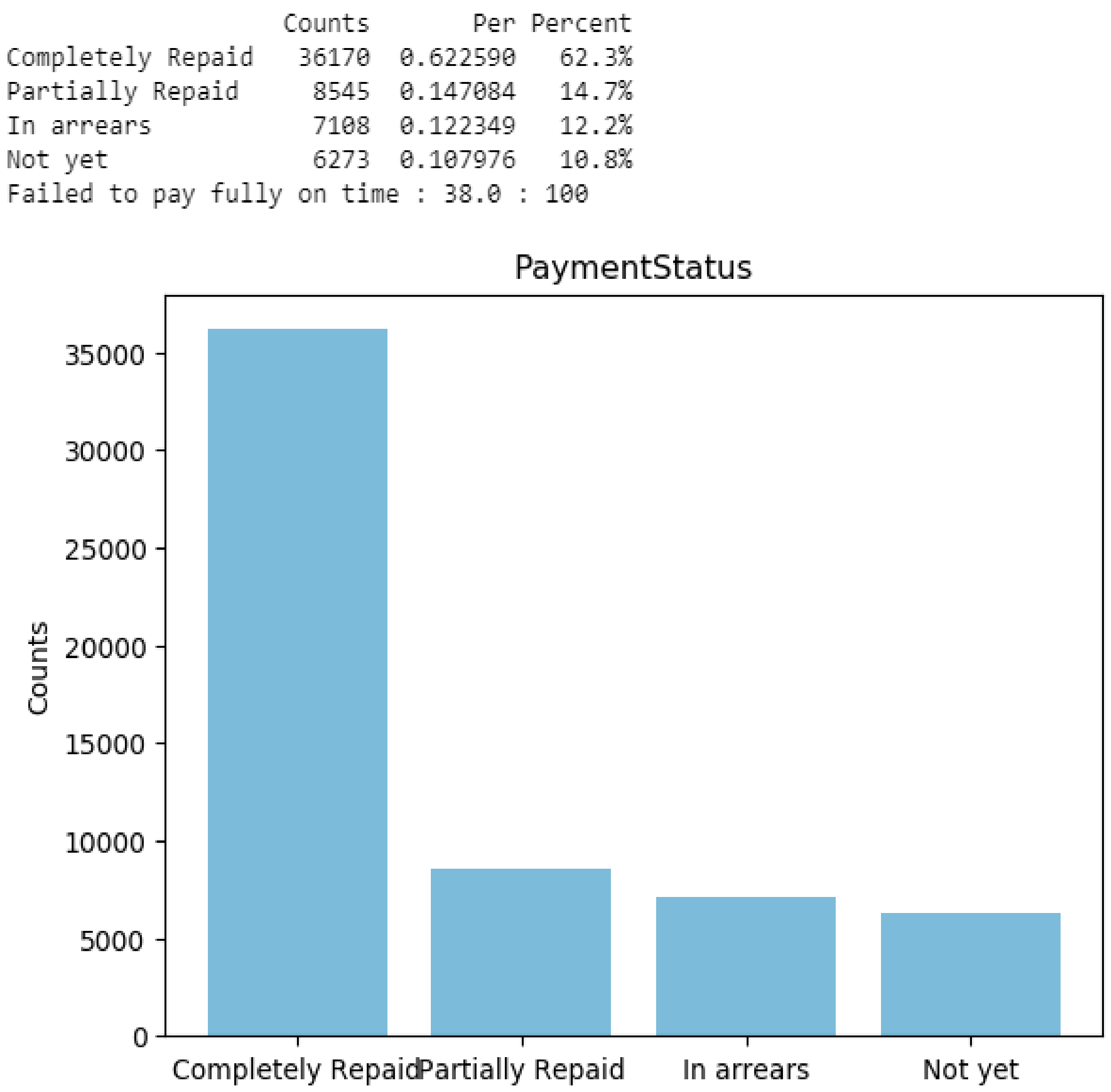
Table 4 Correlations of variables



4.4 Techniques of dealing with Payment Status imbalanced data

Most of the time the classification problems face the problem of an imbalanced dataset where there is a huge difference between the minority and the majority of the data.

When there is a big difference between the minority and majority of data, it results with the problems of accuracy. Therefore, it is good practice to use the binary featuring to check missing values in certain columns.



```
                   Counts       Per Percent
Completely Repaid   36170  0.622590   62.3%
Partially Repaid     8545  0.147084   14.7%
In arrears           7108  0.122349   12.2%
Not yet              6273  0.107976   10.8%
Failed to pay fully on time : 38.0 : 100
```

Figure 5.Distributions of payment status

From the above PaymentStatus imbalance check we can see that the 62.3 % Completely Repaid, 14.7% Partially Repaid, 12.2% in arrears and 10.8% not yet, which sum up to 38 % who failed to pay the bank on time.

Payment status distribution

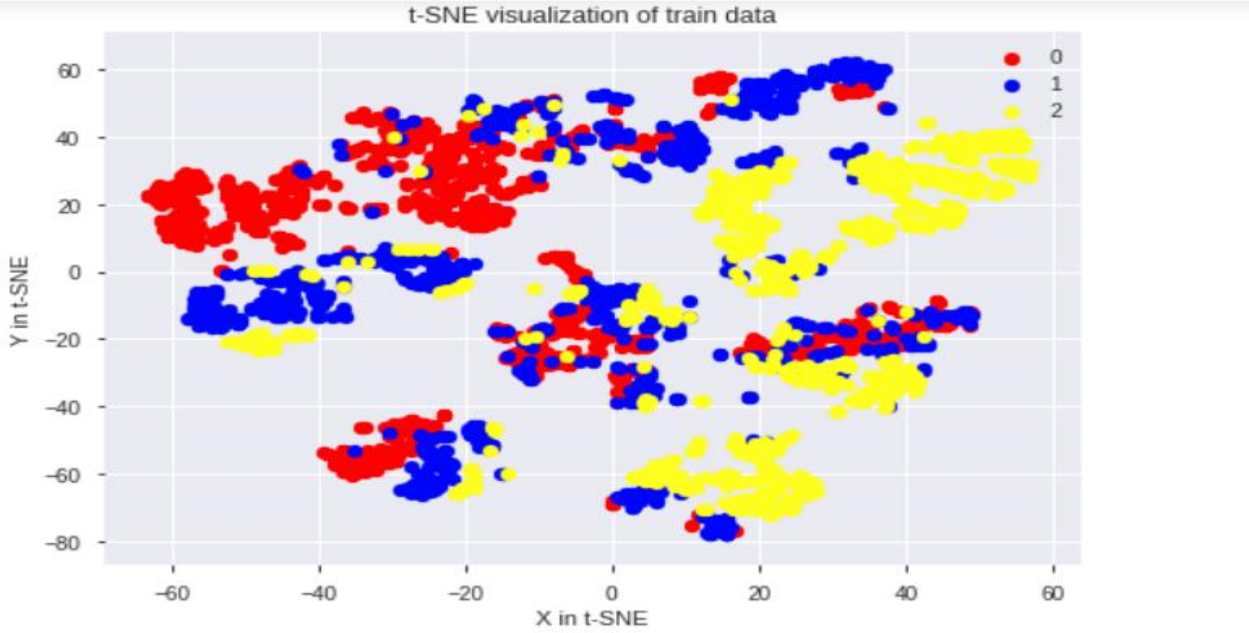
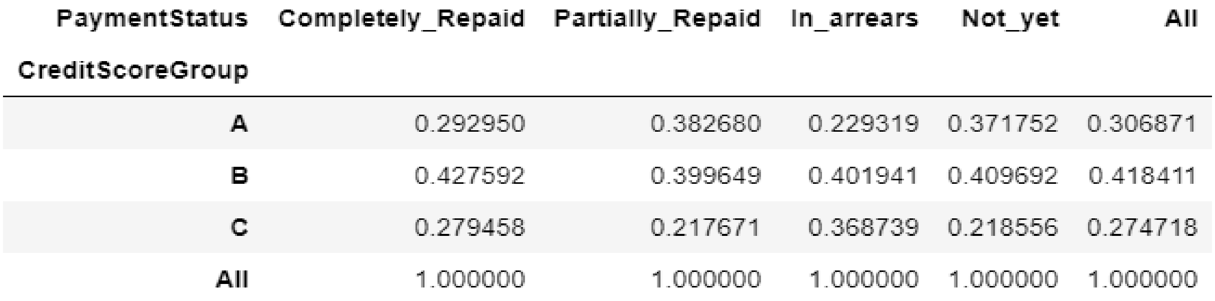


Figure 6. Payment status distribution

4.5 Relationship between Credit Group Score and Payment status

Table 4 *Payments status and Credit Score group*



PaymentStatus	Completely_Repaid	Partially_Repaid	In_arrears	Not_yet	All
CreditScoreGroup					
A	0.292950	0.382680	0.229319	0.371752	0.306871
B	0.427592	0.399649	0.401941	0.409692	0.418411
C	0.279458	0.217671	0.368739	0.218556	0.274718
All	1.000000	1.000000	1.000000	1.000000	1.000000

The figure above shows the relationship between the Credit Score Group and Payments status.

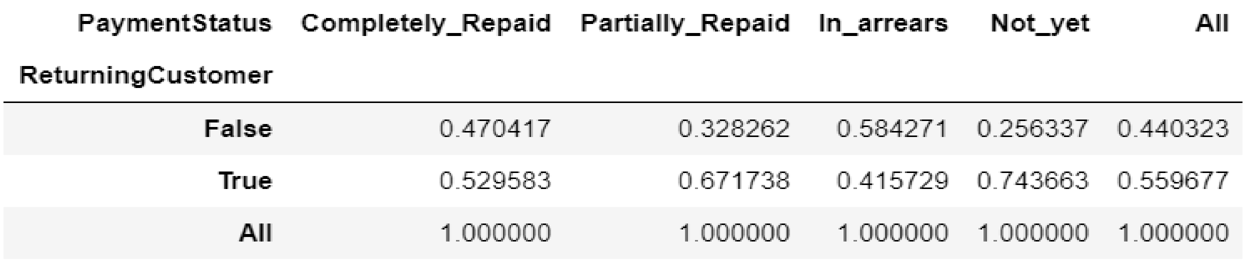
We can see customers in A whose payment is in arrears have the lowest credit score while customers who have partially paid have the highest credit score.

We can see customers in B whose payment status is Completely Repaid have a higher credit score while customers who have partially paid have the lowest credit score.

We can see customers in C whose payment status is In arrears have higher credit score while customers who have partially paid have lower credit score.

4.6 Relationship between Return Customers and PaymentStatus

Table 5 *Payment Status*

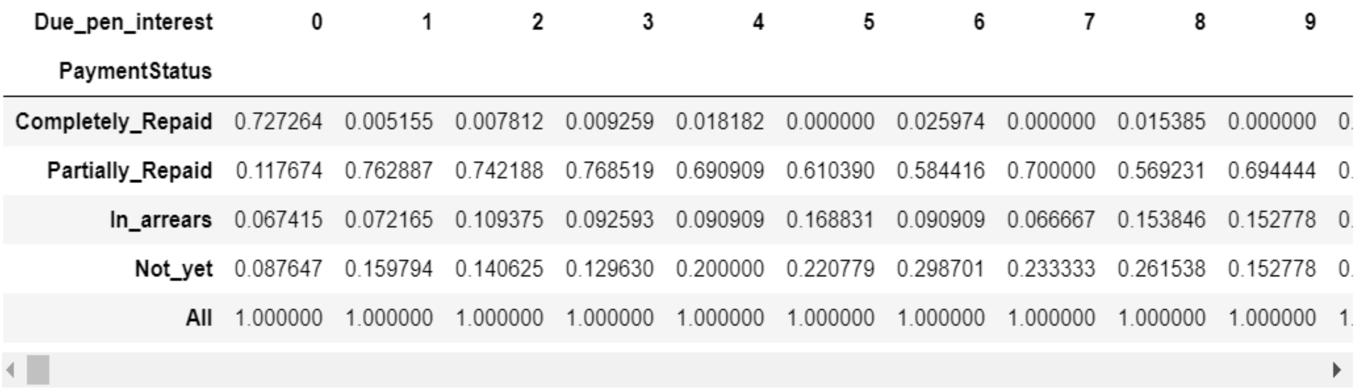


PaymentStatus	Completely_Repaid	Partially_Repaid	In_arrears	Not_yet	All
ReturningCustomer					
False	0.470417	0.328262	0.584271	0.256337	0.440323
True	0.529583	0.671738	0.415729	0.743663	0.559677
All	1.000000	1.000000	1.000000	1.000000	1.000000

The table above shows the percentage of customers who brought customers. From different PaymentStatus, we can see that 52.9% of 36170 completely repaid customers returned new customers.
While 58.42% of 7108 In arrears customers did not return new customers

4.7 Relationship between the Due_pen_interest and Credit Score Group

Table 6 *Due penalty and payments status*

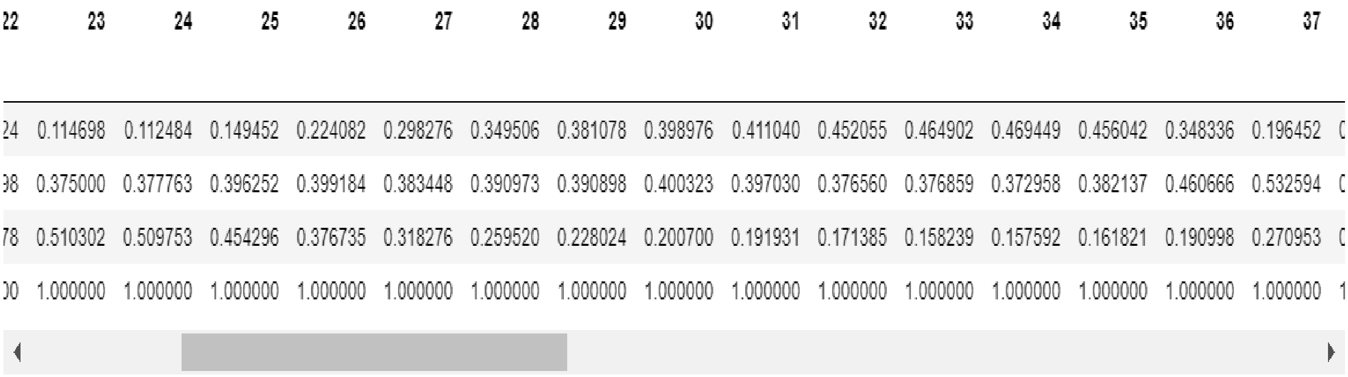


Due_pen_interest	0	1	2	3	4	5	6	7	8	9	
PaymentStatus											
Completely_Repaid	0.727264	0.005155	0.007812	0.009259	0.018182	0.000000	0.025974	0.000000	0.015385	0.000000	0.
Partially_Repaid	0.117674	0.762887	0.742188	0.768519	0.690909	0.610390	0.584416	0.700000	0.569231	0.694444	0.
In_arrears	0.067415	0.072165	0.109375	0.092593	0.090909	0.168831	0.090909	0.066667	0.153846	0.152778	0.
Not_yet	0.087647	0.159794	0.140625	0.129630	0.200000	0.220779	0.298701	0.233333	0.261538	0.152778	0.
All	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.

Due Pen interest is the amount of money you pay to the bank because you did not pay the installment on time. As the table above shows us, the customers in Completely paid with 0 Due_pen_interest have a high probability of paying back the given loan.

4.8 Relationship of Credit Score Group and Age

Table 7. *Credit score group and Age*

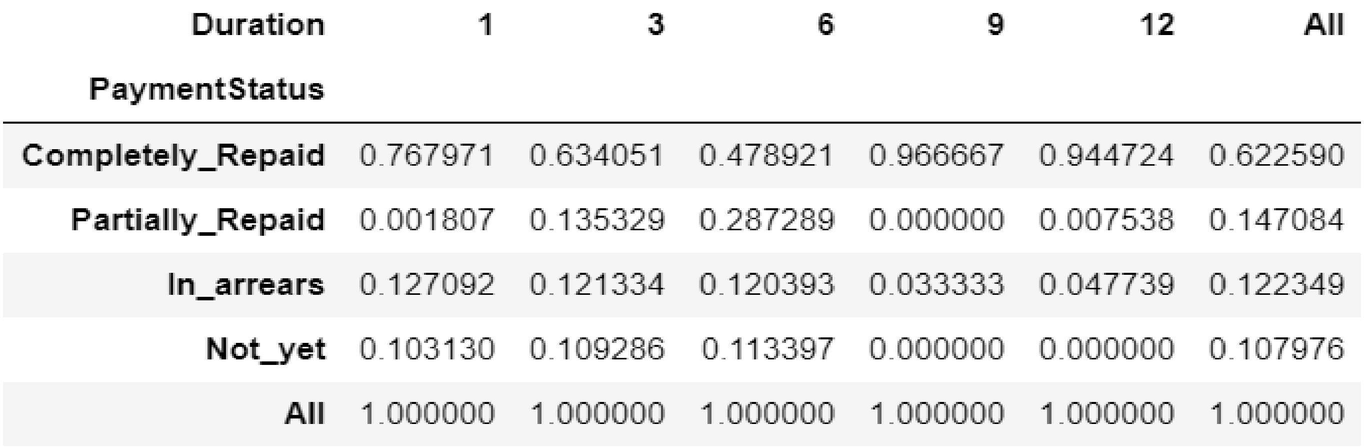


22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	
24	0.114698	0.112484	0.149452	0.224082	0.298276	0.349506	0.381078	0.398976	0.411040	0.452055	0.464902	0.469449	0.456042	0.348336	0.196452	C
38	0.375000	0.377763	0.396252	0.399184	0.383448	0.390973	0.390898	0.400323	0.397030	0.376560	0.376859	0.372958	0.382137	0.460666	0.532594	C
78	0.510302	0.509753	0.454296	0.376735	0.318276	0.259520	0.228024	0.200700	0.191931	0.171385	0.158239	0.157592	0.161821	0.190998	0.270953	C
00	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1

We can see customers in A whose age is between 28 to 36 have a high credit score. We can see customers in B whose age is between 19 to 62 have a high credit score. In C customers at age of 18 have a higher credit score than the rest, also credit score reduces as age increases.

4.9 Relationship between Loan Duration and Credit Score Group

Table 8 *Loan Duration and Credit score group*



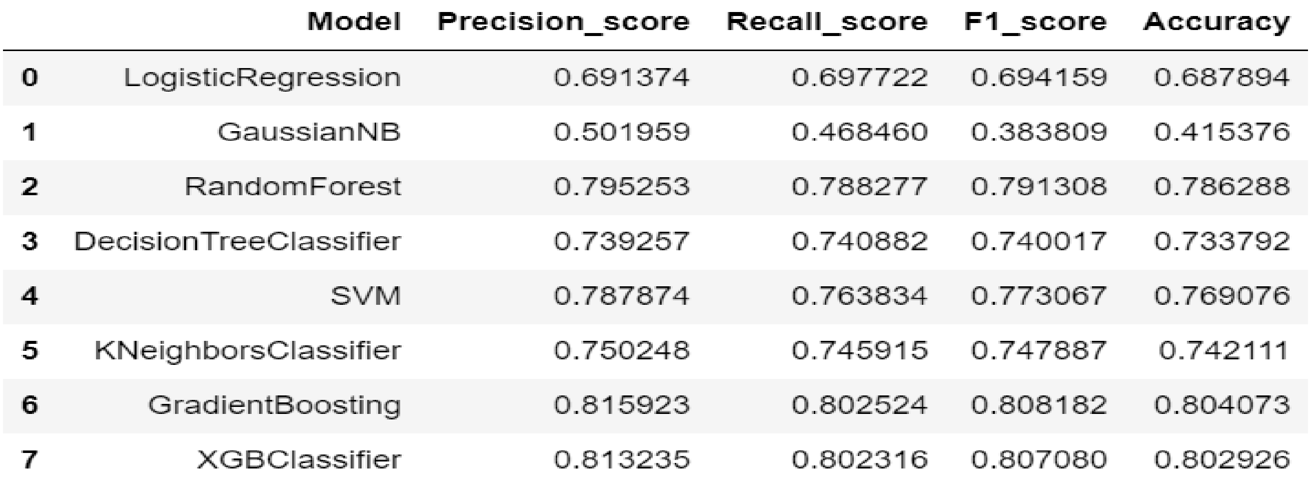
Duration	1	3	6	9	12	All
PaymentStatus						
Completely_Repaid	0.767971	0.634051	0.478921	0.966667	0.944724	0.622590
Partially_Repaid	0.001807	0.135329	0.287289	0.000000	0.007538	0.147084
In_arrears	0.127092	0.121334	0.120393	0.033333	0.047739	0.122349
Not_yet	0.103130	0.109286	0.113397	0.000000	0.000000	0.107976
All	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Duration represents the duration of loan in months, this table shows the customers with loan duration which equal to nine months have the and completely repaid have highest probability compared the remaining ones.

4.10 Discussion

Ensemble learning is a machine learning paradigm in which a number of learners are trained to solve the same problem with the goal of obtaining better predictive accuracy that could have been achieved from any of the constituent learning models alone [Zhou, 2015]. It is a well-established and widely employed methodology designed to enhance the generalizable signal by averaging out noise from a diverse set of models

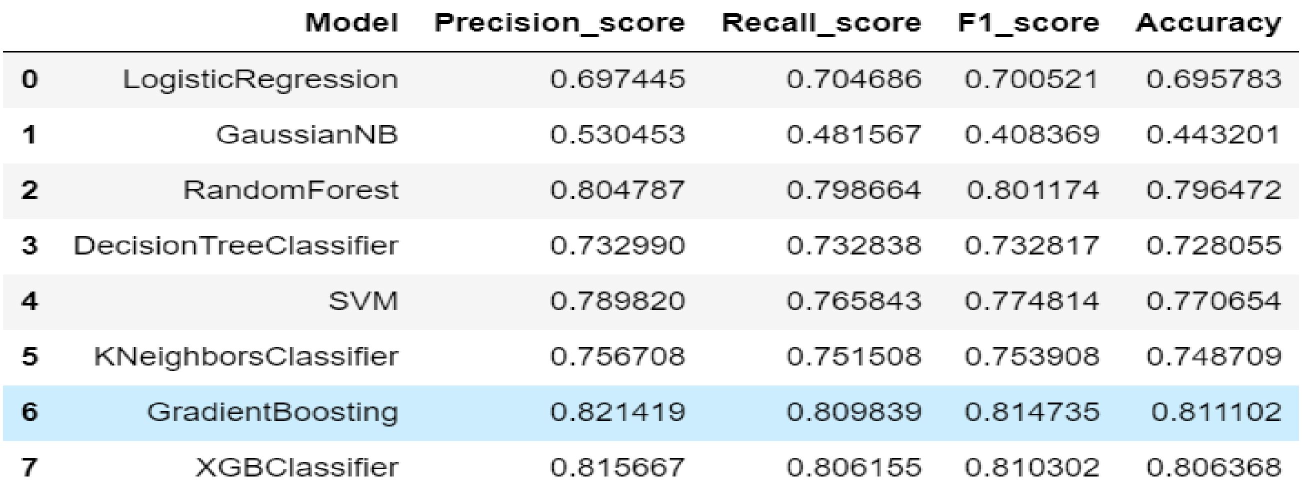
Table 9 *Machine Learning Models*



	Model	Precision_score	Recall_score	F1_score	Accuracy
0	LogisticRegression	0.691374	0.697722	0.694159	0.687894
1	GaussianNB	0.501959	0.468460	0.383809	0.415376
2	RandomForest	0.795253	0.788277	0.791308	0.786288
3	DecisionTreeClassifier	0.739257	0.740882	0.740017	0.733792
4	SVM	0.787874	0.763834	0.773067	0.769076
5	KNeighborsClassifier	0.750248	0.745915	0.747887	0.742111
6	GradientBoosting	0.815923	0.802524	0.808182	0.804073
7	XGBClassifier	0.813235	0.802316	0.807080	0.802926

From the finding we can see that Gradient Boosting is the best model to be used there, but we can improve accuracy in selected models through algorithm tuning knowing that machine learning algorithms are driven by parameters. These parameters majorly influence the outcome of the learning process. Since algorithm tuning finds the optimum value for each parameter to improve the accuracy of the model.

Table 10 *Revised Machine Learning Models*



	Model	Precision_score	Recall_score	F1_score	Accuracy
0	LogisticRegression	0.697445	0.704686	0.700521	0.695783
1	GaussianNB	0.530453	0.481567	0.408369	0.443201
2	RandomForest	0.804787	0.798664	0.801174	0.796472
3	DecisionTreeClassifier	0.732990	0.732838	0.732817	0.728055
4	SVM	0.789820	0.765843	0.774814	0.770654
5	KNeighborsClassifier	0.756708	0.751508	0.753908	0.748709
6	GradientBoosting	0.821419	0.809839	0.814735	0.811102
7	XGBClassifier	0.815667	0.806155	0.810302	0.806368

Logistic Regression produced results with a lower accuracy but overall performance is lower than other models. Decision Tree models dominated over Logistic Regression in all cases

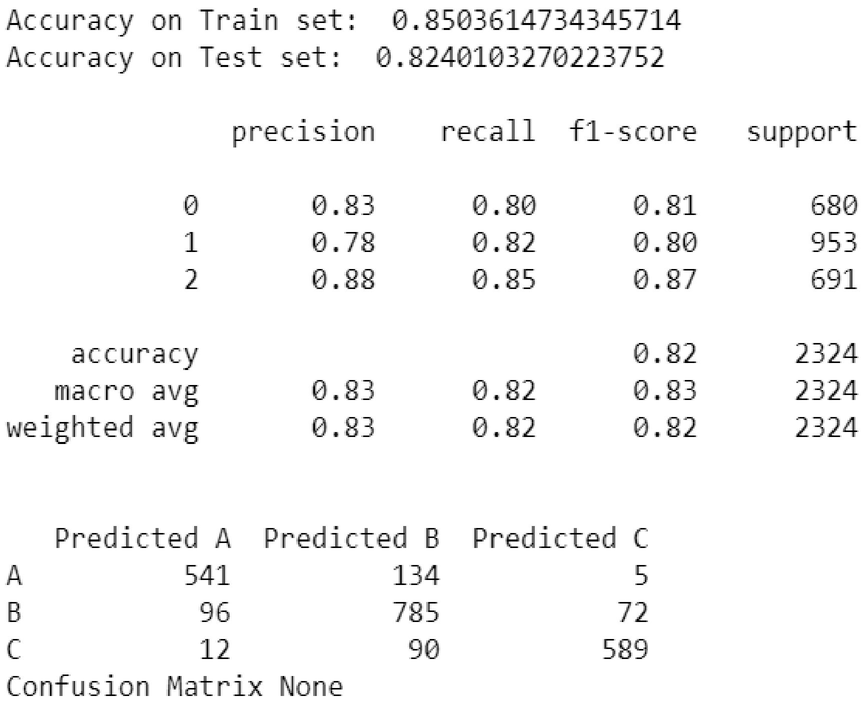
Gradient Boosting model with feature selector f_class would be the best method to use because it has the Highest recall value and highest accuracy.

The XGBClassifier model with feature selector f_classf would be the second-best method to use because it has the second Highest recall value and second highest accuracy.

As we have seen in the problem statement, this study serves the purpose of comparing the models and comes up with the best models which can be used to predict the probability of defaulting when the customers apply for a bank loan.

In my Ensemble, I used different machine learning algorithms on one dataset. Result above shows the names of models and their performance on Precision_Score, Recall_score, F1_score and Accuracy.
When we take a look at the accuracy of our models, we can conclude that the Gradient Boosting is the best performer among the remaining models and followed by the XGBoosting Classifier.

Table 11 *True Value table*



```
Accuracy on Train set:  0.8503614734345714
Accuracy on Test set:  0.8240103270223752

              precision    recall  f1-score   support

           0       0.83      0.80      0.81       680
           1       0.78      0.82      0.80       953
           2       0.88      0.85      0.87       691

    accuracy                           0.82      2324
   macro avg       0.83      0.82      0.83      2324
weighted avg       0.83      0.82      0.82      2324


   Predicted A  Predicted B  Predicted C
A          541          134            5
B           96          785           72
C           12           90          589
Confusion Matrix None
```

$$The\ classification\ metrics\ of\ i\ f\ this\ fairly\ imbalanced\ dataset\ are:$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F_{1=}\ 2\ (Precision)\ (Recall)/(Precision\ +\ Recall)$$

Depending upon bank operational costs & ideology a large bank may follow the principle that fewer False Positives are preferable over a few more False Negatives to be able to lend more & spend less on investigations on the contrary a conservative approach would go with the opposite i.e. more accuracy.

Chapter V: Conclusion and Recommendations

5.1 Conclusion

Since decades ago, banks loan have been playing a big role in the economies of countries and has biggest assets on financial market. Also, financial institutions make money through lending, and without critical analysis of credit worthiness, they may loss the capital and interest due to the inability of the customers to back the given amount of loan with agreed interest. With this, loan default prediction is very essential tool for financial institutions, due to the rapid speed of technology in data collection, big data era, competitive financial markets, development of machine learning and artificial intelligent. Financial institutions should leverage the technology to collect enough data for future use especially for loan prediction, customers ’segmentation and products recommendation.

This study aims to help the banks to predict the probability of defaulting from their customers. And this will help banks to lend their money with the low risk of losing. In this study we have used data from Bank of Kigali, the biggest financial institution in Rwanda. In this study the data was cleaned first, the features with a high percentage of missing data were removed. Also, Exploratory data analysis was performed, imbalanced data and data outliers were performed in order to remain with well cleaned data for machine learning techniques. Different machine learning techniques like Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, Decision Tree Classifier, K Nearest Neighbors, and Ensemble learning-VotingClassifier, Ensemble learning-AdaBoostClassifier and Ensemble learning-XGBClassifier were performed to evaluate the best model to predict the Bank loan default based of different variables.

In this study the data was cleaned first, the features with a high percentage of missing data were removed. Also, Exploratory data analysis was performed, imbalanced data and data outliers were performed in order to remain with well cleaned data for machine learning techniques.

5.2 Recommendation and further study

I would recommend financial institutions to use machine learning techniques because it saves money and time for both sides. Moreover, Findings show that the customers with Credit Score B will have low probability of defaulting. In this work we used data from one financial institution and we would recommend anyone who might want to further this study to consider using data from different financials institution across the region to capture the insight

REFERENCES

Abdou, H. & Pointon, J. (2011). *Credit scoring, statistical techniques and evaluation criteria*: a review of the literature, *Intelligent Systems in Accounting, Finance & Management, 18* (2-3): 59-88.

Addo, M. P., *Guegan*, D., *Hassani*, B. *(2018). Credit Risk Analysis Using Machine and Deep Learning Models*.

Alshouiliy, K., Alghamdi, A., and Agrawal, D. P. (2020). *Azure ML based analysis and prediction loan borrowers creditworthy The 3rd Int*. Conf. on Information and Computer Technologies (ICICT) 1 pp 302–6

Basel, (2021). Machine Learning Approach for Micro-Credit Scoring, Risks; Basel Vol. 9, Iss. 3, (2021): 50. DOI:10.3390/risks9030050

Calcagnini, G., Cole, R., Giombini, G. and Grandicelli, G. (2018). *Hierarchy of bank loan approval and loan performance*," Economia Politica: Journal of Analytical and Institutional Economics, Springer; Fondazione Edison, vol. 35(3): 935-954, December

Hand, D.J. and Henley, W.E. (1997). *Statistical Classification Methods in Consumer Credit Scoring:* A Review. Journal of Royal Statistical Society, 160, 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Kuizinienė, D., Krilavičius, T. (2019). *Deep learning for credit scoring", International Journal of Design, Analysis and Tools for Integrated Circuits and Systems*. Hong Kong 8(1): 66-71, (Oct 2019).

Kurapati, N., *Bhansali*, P.K. (2018). *Predicting the credit defaulters using machine learning techniques*. Int. J. Manag. Technol. Eng. 8(11), 6 (2018)

Machine Learning in Credit Risk Management", Computational Economics; Dordrecht Vol. 57, Iss. 1, (Jan 2021): 203-216. DOI:10.1007/s10614-020-10042-0.

Mpaka, N. (2019). https://www.ktpress.rw/2019/08/bank-of-kigali-eyes-2bn-value-in-the-next-six-years/

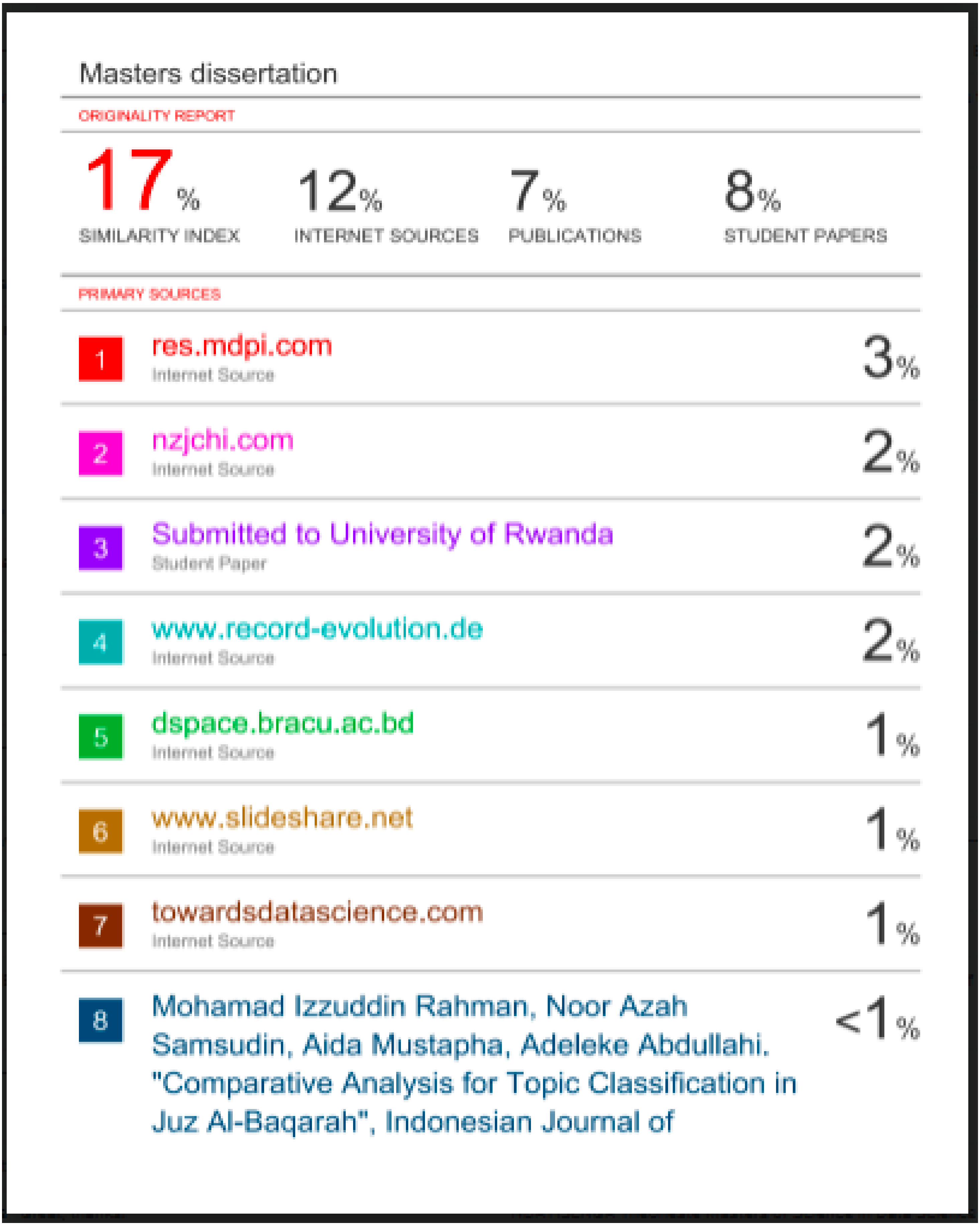
Network Algorithm", International Journal of Modern Education and Computer Science; Hong Kong Vol. 10, Iss. 5, (May 2018): 9. DOI:10.5815/ijmecs.2018.05.02

Salame (2011). *Applying Data Mining Techniques to Evaluate Applications for Agricultural Loans*. PhD dissertation, University of Nebraska, Agricultural Economics Department, Aug. 2011.

Sarma, K.S. (2013). *Predictive Modelling with SAS Enterprise Miner: Practical Solutions for Business Applications*. SAS Institute

Sawant and Chawan (2013). *Comparison of Data Mining Techniques used for Financial Data Analysis*. International Journal of Emerging Technology and Advanced Engineering.

Stuart, R,. & Peter, N. (2010). *Artificial Intelligence a Modern Approach Third Edition*. Library of Congress Cataloging- in - Publication Data on File.



Masters dissertation

ORIGINALITY REPORT

17%	12%	7%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	res.mdpi.com Internet Source	3%
2	nzjchi.com Internet Source	2%
3	Submitted to University of Rwanda Student Paper	2%
4	www.record-evolution.de Internet Source	2%
5	dspace.bracu.ac.bd Internet Source	1%
6	www.slideshare.net Internet Source	1%
7	towardsdatascience.com Internet Source	1%
8	Mohamad Izzuddin Rahman, Noor Azah Samsudin, Aida Mustapha, Adeleke Abdullahi. "Comparative Analysis for Topic Classification in Juz Al-Baqarah", Indonesian Journal of	<1%