AFRICAN CENTRE OF EXCELLENCE IN DATA SCIENCE

SUPERVISED MACHINE LEARNING MODELING FOR CREDIT RISK PREDICTION AND EVALUATION IN MALAWI

By

Alexander Peter Chidothi

Registration Number: 221000146

A Dissertation submitted in partial fulfillment of the requirements of the degree of Master of Science in Data Science in Actuarial Science in the African Center of Excellence in Data Science, College of Business and Economics at the University of Rwanda

Supervisor: Associate Professor Charles Ruranga

March, 2023

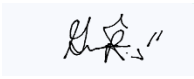
Declaration

I, **Alexander Peter Chidothi (Reg. No 221000146)** declare that the dissertation entitled "**Supervised machine learning modeling for credit risk prediction and evaluation in Malawi"** is the result of my own work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Name: Peter Alexander Chidothi

Student Reg. No: 221000146

Signature:



Date: 27th July, 2023.

Approval sheet

This dissertation entitled **supervised machine learning modeling for credit risk prediction and evaluation in Malawi** written and submitted by Peter Alexander Chidothi in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in Actuarial Science is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 18 % which is less than 20% accepted by the African Centre of Excellence in Data Science (ACE-DS).

Associate Prof. Charles Ruranga

Supervisor

Dr. Ignace Kabano

Head of Training

Dedication

This dissertation is most definitely dedicated to my mother for all the support she has been giving me throughout my academic journey. You are heavenly sent and I adore you.

Acknowledgements

I am indebted to Prof. Charles Ruranga for his tireless efforts in the process of supervising this work. The guidance and support he rendered are greatly appreciated.

I am deeply thankful to the World Bank through the African Center of Excellence in Data Science at the University of Rwanda for providing me with a scholarship, which made my studies possible.

Finally, I wish to thank God for his amazing grace which he abundantly poured on me the past two years I was doing my studies. Glory be to Him always.

Abstract

Credit institutions in Malawi are facing difficulties in their operations due to their unreliable methods of assessing credit risk. They tend to group subjects to predict their credit risk, which results in biased predictions. Some use qualitative methods to determine credit risk, and their data is limited to their own portfolio. This study aimed to overcome these challenges by creating a robust and reliable machine learning model to predict credit risk as probability of default and credit score sequentially, based on demographic and loan details. Big data was sourced from a credit reference bureau that collects data from all credit institutions in the country. The research design was quantitative where data was quantitatively analyzed using Knowledge Discovery in Databases process. The process involved problem formulation, data collection, pre-processing and cleaning, transformation, mining tasks and methods, result evaluation and visualization and knowledge discovery. Upon cleaning, transforming and splitting the data into training, validation and testing datasets, six classifier models were trained namely; KNearest Neighbors, Naïve Bayes, Logistic Regression, Support Vector Machine, Decision Tree and Random Forest. The target variable was credit status which was binary with values either 'defaulted' or 'paid up' status. The features were sex, marital status, age, district, no of dependents, principal amount, payment terms, source of income, collateral type, interest type, interest calculation method, participation code, liability type, loan duration which was a combination of numerical and categorical datatypes. A decision tree with maximum depth 10 was preferred to the other based on performance evaluation using accuracy score, ROC curve, precision score, recall score and confusion matrices. This model was used to predict probabilities of default using the testing data. Credit score was predicted based on the probabilities of default by the analogy of equating maximum FICO score to the minimum probability and minimum FICO score to maximum probability. Recommendations to implement either unsupervised or deep learning models for the same exercise subsequently followed.

Keywords Credit risk · Machine learning · Probability of default · Credit score · Knowledge Discovery in Databases · FICO scores

Table of Contents

List of tables

List of figures

List of Abbreviations

AUC	Area Under the Curve
CM	Confusion Matrix
CS	Credit Score
DM	Data Mining
ED	Exposure at Default
FN	False Negatives
FP	False Positives
IBM	International Business Machines
KDD	Knowledge Discovery in Databases
KNN	KNearest Neighbors
LGD	Loss Given Default
MFI	Microfinance Institution
ML	Machine Learning
PD	Probability of default
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
RWA	Risk Weighed Asset
SQL	Structured Query Language
SVM	Support Vector Machine
TN	True Negatives
TP	True Positive

CHAPTER 1: GENERAL INTRODUCTION

This chapter aimed at introducing the key terminologies of the study including credit risk and machine learning. More light had to be shed for the picture to be cleared and a cornerstone of the study to be laid.

The background of credit risk and machine learning and other associated terms and concepts covered their respective definitions and other relevant information based on different write ups reviewed. On the other hand, the problem statement and relevance of the study sections focused on the importance of the paper while objectives were about the targeted achievements.

1.1 Background

Credit Risk

Credit risk is generally defined as the likelihood or ambiguity of a borrower defaulting on their debt obligations (Mosch, 2013). In light of this, financial institutions such as banks, microfinance organizations, and insurance firms analyze the risk of default among borrowers in an effort to minimize losses and other fraudulent activities. According to Vaidya (2020), lenders typically take the chance that they won't get paid for the principal and interest portion of the debt. An interrupted cash flow and higher collection costs may result from this. It may also be linked to other risks, such as the failure of a bond issuer to make timely payments or the failure of an insurer to reimburse a claim.

According to Spuchákova (2015), credit or default risk refers to a customer's or counterparty's inability or unwillingness to fulfill obligations related to loans, transactions, hedging, settlements, and other financial transactions. Operating, default, and portfolio risk make up the majority of credit risk. According to Fidler's (2008) theory, credit risk is the likelihood that a loan will not be repaid in accordance with contractual obligations and terms that were previously predetermined. It is connected to problems with credit collection, which are essentially defaults, failures, etc. which have been encountered. On

the other hand, credit risk is a concept that looks to the future and focuses on the likelihood of future credit difficulties.

In banking, credit risk is defined as the potential of a bank borrower or counterparty not being able to repay a loan as per agreed terms. It is mainly determined by whether the borrower can repay the credit loan on time (Bank for International Settlements, 2020). It is the possibility of a loss yielded from a borrower's failure to return the money. In another aspect, it is more of a risk of the lending party not being able to acquire back the money from the borrower, thus causing interrupted cash flows (Jameson, 2018).

When a bank receives money as a deposit from a customer, it tends to lend it to others to make a profit. There is always a risk when lending money to a second party because there is always a possibility or chance of defaulting. This is what banks refer to as credit risk. (Ohman et al., 2019).

Microfinance is one of the industries closely related to credit risk. They are engaged in providing financial services to low-income people who lack bank loans. It is part of a financial inclusion strategy that identifies opportunities for individuals and businesses to obtain a range of useful and customized financial products and services at low cost to meet their needs. In this context, microfinance provides low-cost, high-quality financial services to low-income people. (Benouna, 2019).

Spuchľáková (2015) argued that credit risk needs to be measured and quantified. The purpose of credit risk measurement is to quantify potential losses from operations of credit. The amount of loss is not known with certainty and must be estimated.

Technically, credit risk is quantified by credit score, Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (ED). Credit score is basically a number, mostly 3-digits, that lenders use to determine the likelihood of an obligor paying them back. Based on factors like previous loan late payments, amount owed, credit history, types of credits in use, the credit score is computed to measure how likely the borrowers are to repay the money. The borrower's likelihood of defaulting on the loan is indicated by the second indicator, known as PD, while the third indicator is the ratio of exposure loss as a result of

counterparty default to outstanding principal. The outstanding balance is known as Exposure at Default. (Chee, 2021).

Machine Learning

On the other hand, machine learning is a branch of computer science that has developed from the study of artificial intelligence pattern recognition and computer learning theory. Learning and building algorithms that can learn and make predictions based on datasets, rather than following strict static program guidelines, they work by building models from sample inputs to make predictions or selections based on data. (Simon, 2015). Urbanowicz (2017) defined machine learning as the subset of artificial intelligence in computer science that uses statistical methods to enable computers to learn from data without being explicitly programmed. Technically, the ability to learn is the gradual improvement in the performance of a related task.

Machine learning involves the study of computer algorithms for learning behavior. For example, researchers may be interested in learning how to complete a task, make an accurate prediction, or act intelligently. The training provided is always observational or data-based. For example, direct experience or instruction. So, in general, machine learning is about learning how to improve future performance based on what was learned in the past. The focus of machine learning is on automated processes. The goal is to develop learning algorithms that perform learning automatically without human intervention or support. (Schapire, 2014).

Prediction and evaluation of credit risk can be implemented on traditional or machine learning modern models. The financial industry has undergone significant changes as a result of the deepening integration of technology and finance brought about by the internet's ongoing development. Traditional modeling is not trustworthy but easy to implement while the machine learning approach although being complicated yields truthful results (Zimmerman, 2021). Bazdok (2018) claimed that analytical results yielded from machine learning models enhance well informed decisions in business and other related institutions.

Types of Machine Learning Algorithms

Based on whether you know the labels of your data, machine learning (ML) algorithms can be divided into three main categories: supervised learning, unsupervised learning, and reinforcement learning. (Luckert, 2015).

Supervised learning algorithms are used when the labels of the data to be predicted are known. Email spam filters are a good example of how supervised learning can be used for classification. Here one knows that the received email is either "spam" or "non-spam" and can be used as a label for the sample population and the learning algorithm can classify these two categories. Another application of supervised learning is using regression to predict company profits. Consider the problem of predicting the monthly earnings value of a company. In this case, the input parameters are years of business activity, investment amount, type of business, location, etc., and the output is the company's monthly income. One can create a training set from known company monthly revenues and train a learning algorithm to predict other inputs. Generalizing, the goal of supervised learning is to learn a mapping of inputs to outputs for which the correct values are provided by the supervisor.

Technically, data is split into three distinct groups namely; training set, validation set and test sets for supervised learning models. The first one is used to train the model for perfect prediction. The second set is used to confirm the perfection by evaluating perform while the last one is used for deployment. Figure 1.1 shows a typical supervised learning algorithm with respect to a theory by (IBM Cloud Education, 2020).

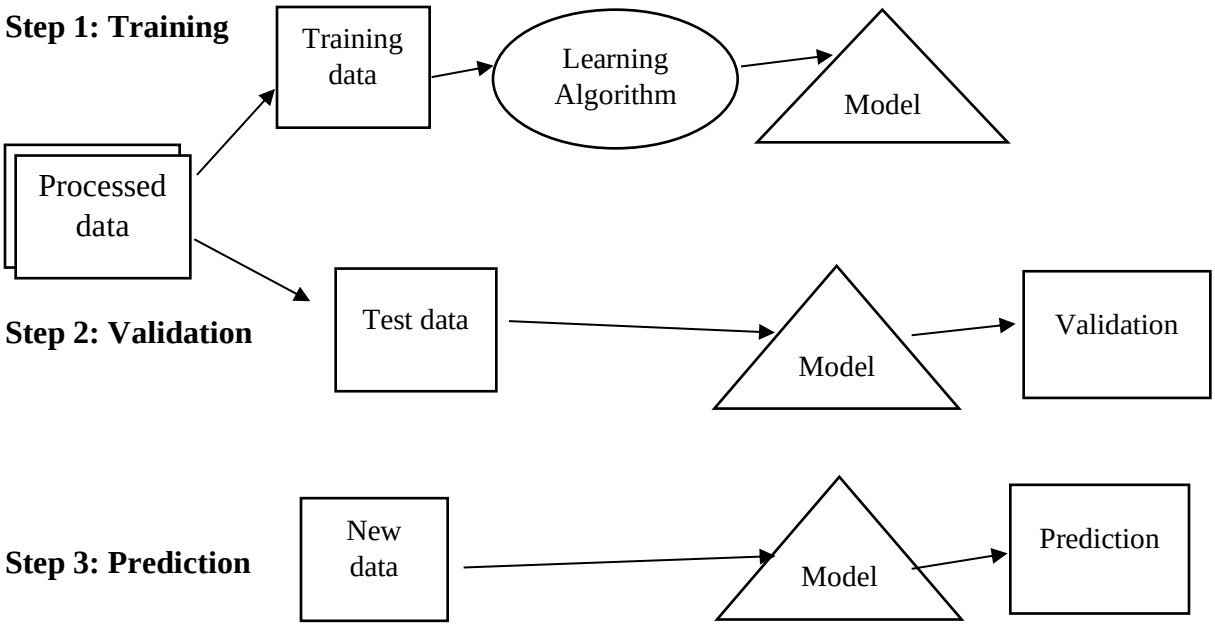


Figure 1. 1: Typical supervised learning algorithm

However, unsupervised learning is used in classification problems where the labels of the data are not known. An example of such a problem is collaborative filtering, where items are grouped based on items purchased together. For example, given a database of user preferences, the model should predict the preferences of new users. Another good example of collaborative filtering is predicting which new movies you like based on past tastes, past likes who were similar in the past, or new movie tastes. In this case, the number of categories and labels are undefined. A machine learning application needs to group articles based on some common words and provide supervisor data that can be used to label the grouped groups. The goal of unsupervised learning is to find regularities or correlations in input data without the explicit need of a supervisor.

On the other hand, according to Le et al. (2019), in Reinforcement Learning, the rules of the game are unknown. No supervisor, just late feedback reward signal. Agent's actions affect data that is received.

Figure 1.2 sums up all the types of machine learning algorithms with a single and vivid illustration.

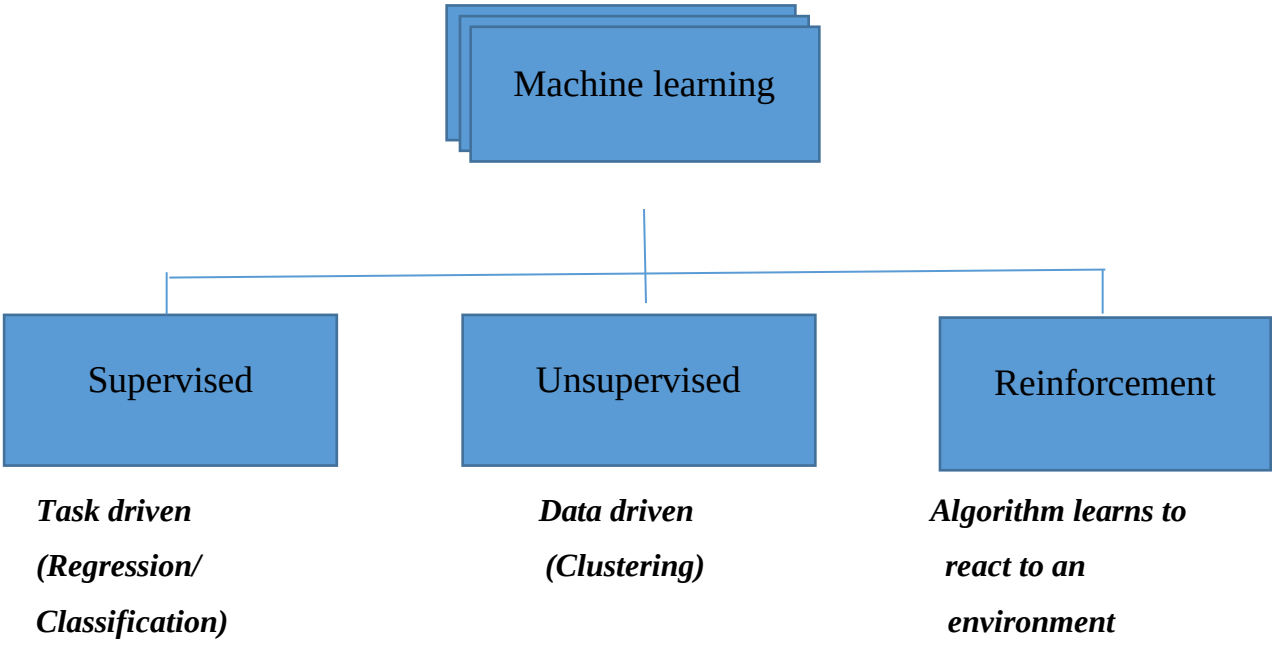


Figure 1. 2: Types of Machine Learning

Knowledge Discovery in Databases(KDD)

This is an iterative and interactive process which, according to Brownlee and James (2014), it is used alternatively with the term data mining. Apparently, machine learning modeling is a tool or stage of data mining which is never skipped. KDD process or data mining is defined as the process of discovering patterns in data where machine learning is involved to train models like human beings for perfection of determination and prediction without explicitly programming it to do so.

According to Schapire (2014), KDD refers to the broad process of learning from data and emphasizes the advanced uses of particular data mining techniques. Researchers in a number of fields, including artificial intelligence, machine learning, pattern recognition, databases, statistics, knowledge acquisition for expert systems, and data visualization are interested in this area. In the context of huge databases, the primary goal of the KDD process is to extract information from data. Data mining algorithms are used to identify what is considered knowledge in order to accomplish this.

Database knowledge discovery is regarded as programmed exploratory modeling and analysis of sizable datasets. KDD is a structured method for identifying practical and simple patterns in huge and complex datasets. The KDD process is based on data mining, which involves analyzing data, creating models, and deriving algorithms to discover previously unidentified patterns. With the help of this model, data can be analyzed, predicted, and knowledge can be drawn from them. Nowadays, knowledge discovery and data mining have become crucial and essential issues due to the availability and abundance of data. It is understandable that experts and professionals now have access to a range of techniques given recent developments in this field. (Purohiti et al. , 2015).

Nine steps are included in the knowledge discovery process. Since each stage of the process is iterative, going back to the earlier steps might be necessary. Because there is no single formula or complete scientific classification of the appropriate choices for each step and application type, the process requires a great deal of imagination. The process, as well as the various needs and opportunities at each stage, must therefore be understood.

Determining the KDD objectives is the first step in the process, which concludes with the application of the learned knowledge. The loop is then closed, and active data mining commences. Thereafter, modifications would have to be made to the application domain. Giving cell phone users a variety of features, for instance, can help to lower churn. This completes the cycle, and the effects are then assessed once more on the new data repositories and the KDD procedure. The nine-step KDD process is summarized as follows by Watson (2019), starting with problem comprehension:

i. Problem formulation and understanding the application domain

This is the initial preliminary step. It involves identifying the goal of the process and gathering prior needed knowledge about the application domain. It develops scenes to understand how to handle various decisions such as transformations, algorithms, and representations. The one responsible for the KDD project needs to understand and characterize the end-user goals and the environment in which the knowledge discovery process takes place, including relevant prior knowledge.

ii. Data Collection or creation

Once the goals are defined, one needs to determine the data that will be used in the knowledge discovery process. This is simply about choosing an appropriate dataset to extract knowledge from. This involves identifying accessible data, procuring key data, and consolidating all knowledge discovery data into a set that includes the quality considered in the process. Data mining is an important process because it learns and discovers from the available data. This is the evidence base for building the model. If some important attributes are missing at this point, the more attributes considered, the more likely the entire investigation will fail in this regard. On the other hand, organizing, collecting, and operating advanced data archives is expensive and gives you the opportunity to best understand the phenomenon. This placement is related to the interactive and repetitive aspects of KDD. It starts with the best dataset available and is later expanded to observe the impact in terms of knowledge discovery and modeling.

iii. Data preprocessing and cleansing

This step improves the reliability of the data. This includes data cleaning to handle missing values and remove noise and outliers. This may involve complex statistical techniques or use data mining algorithms in this context. For example, if one suspects that a particular attribute is unreliable or has a lot of missing data, that attribute may be subject to algorithms monitored by data mining at this point. Predictive models for these attributes are built to predict missing data. How much one sees this level depends on many factors. Regardless, studying aspects of enterprise data frameworks is important and often enlightening in itself.

iv Data Transformation

In this phase, suitable data are developed and prepared for data mining. Techniques used in this context include attribute transformation (such as discretization of numerical attributes and functional transformation) as well as dimension reduction (for example, feature selection, extraction, and record sampling). Here is where data reduction to a representable format is necessary. For instance, it may be necessary to remove variables or parameters that are ineffective for achieving the task's goal. This step, which is typically very project-specific, can be crucial for the overall KDD project's success. For instance,

the ratio of attributes, rather than each one individually, may frequently be the most important factor in medical assessments. In business, it's important to consider efforts, short-term problems, and impacts that are out of your control. The transformation needed in the following iteration can be acquired in an amazing way if the correct transformation at the beginning is not well utilized. As a result, the KDD process reinforces itself and encourages comprehension of the necessary transformation.

v. Prediction and description

This is where the choice of the type of data mining to use, such as classification, regression, clustering, etc., is made. is created. In addition to the prior steps, this is largely dependent on the KDD goals. Two key objectives of data mining. The first is a prediction, and the second is an explanation. While the term "monitored data mining" is often used to refer to predictions, descriptive data mining also includes the unmanned and visual aspects of data mining. Inductive learning is the foundation of most data mining techniques. By generalizing from a sufficient number of initial models, inductive learning creates a model either explicitly or implicitly. The prepared model is valid for upcoming cases is the fundamental premise of the inductive approach. For a specific set of readily available data, this method also considers the degree of meta-learning.

vi. Selecting the Machine Learning or Data Mining algorithm

Having the technique, the next step is to decide on a strategy. At this stage, a specific technique to use to find the pattern that contains multiple inductors gets selected. For example, considering accuracy and comprehension, the former is better at neural networks and the latter is better at decision trees. There are several ways to succeed in any meta-learning system. Meta-learning focuses on clarifying whether data mining algorithms are successful for a particular problem. Therefore, this methodology seeks to understand the situations in which data mining algorithms are best suited. Each algorithm has parameters and strategies of learning, such as ten fold cross-validation or another division for training and testing.

vii. Utilizing the Data Mining algorithm

Finally, the implementation of data mining algorithms is carried out. At this stage, you may need to use the algorithm several times until you get satisfactory results. Switch the control parameters of the algorithm for example, the minimum number of instances per leaf in the decision tree.

viii. Evaluation

Finally, the data mining algorithm has been put into practice. To get a satisfactory result at this point, the algorithm may need to be used more than once. For instance, the algorithms can be turned to control variables like the minimum number of instances in a single decision tree leaf.

ix. Using the discovered knowledge

Finally, after everything has been prepared, this stage involves incorporating the information into another system for further use. Insofar as system changes and impact assessments are possible, the knowledge becomes useful. The success of this step determines how well the KDD process as a whole works. This step presents many difficulties, such as losing the "data processing conditions" under which they have been established. For instance, if knowledge was discovered from a particular static representation, the data that was static at the time is now dynamic. Data structures may alter certain quantities that cease to exist and the data domain may change, for example, an attribute that may now have a value that was not anticipated in the past.

1.2 Problem statement

Credit risk is the most prominent cause of banking problems and is being managed as the most material risk by the Standard Bank of Malawi (Standard Bank PLC, 2021). Cunningham, Kiley, & Mihov (2017) argued that the rise in American mortgage defaults during the global economic crisis was a loss associated with credit risk. According to Ohman et al. (2019), banks face the risk of loan default by borrowers and must maintain a minimum level of capital as stated by the Basel Accords to protect against financial crises. The risk-weighted assets (RWA) calculation, based on probability of default and loss given default, is used by banks to determine the minimum capital needed. Banks are seeking ways to more accurately predict the likelihood of default to reduce capital requirements.

The idea of credit risk has been present since ancient times and is an integral part of borrowing and lending. Although the concept remains unchanged, the methods for predicting and assessing credit risk have changed over time. Banks and other lending organizations have moved away from relying solely on traditional analysis techniques and have adopted statistical methods that consider grouped risk. Today, banks use a combination of statistical techniques and human intuition to measure credit risk. Credit scores are calculated using either statistical credibility theory based on grouped risks or human intuition(Coutte, 1998; Fair Isaac Corporation, n.d). Ghosh (2019) pointed out that intuition-based prediction of credit risk lacks a data-driven approach, leading to uninformed decisions. Jia (2017) also stated that credibility theory has limited predictive power for individual borrowers because it is based on grouped risk. This highlights the need for a more reliable and accurate approach for predicting credit risk for better-informed decisions in the industry.

Machine learning algorithms, on the other hand, are trained to make accurate predictions by continually updating and refining models based on new data. During training, algorithms identify patterns and relationships in input data to make predictions which are then evaluated. Key factors affecting accuracy include training data size and quality, algorithm choice, number of training iterations and overfitting or underfitting (James, Witten, Hastie, & Tibshirani, 2013). According to Bishop (2006), techniques such as cross-

validation, regularization, and ensemble methods are employed to improve accuracy and prevent overfitting, which occurs when the model becomes too tailored to training data and loses its ability to generalize.

Reflecting on all that, there is a clear need to incorporate machine learning in credit risk prediction in the Malawian finance industry. The current methods of credit risk prediction, including human reasoning and statistical methods, are subject to bias and are not reliable. As the data on credit obligors continues to grow at an exponential rate, traditional techniques are unable to keep up. Machine learning models, when trained and tuned effectively, can provide more accurate and trustworthy results for credit risk prediction on any data, regardless of its size or complexity. The credit risk of each individual obligor can be determined based on their unique personal and loan information, not as a group. With the use of machine learning models, creditors in Malawi can make informed decisions on an increasing number of obligors. This research aims to fill the gap in the field by exploring the potential of machine learning in credit risk prediction.

Furthermore, the significance of this study is heightened by the numerous risks associated with bank or microfinance loans. Banks and MFIs are particularly vulnerable to capital loss, making risk analysis and baseline assessments imperative. Despite holding vast amounts of data on consumer behavior, these institutions are unable to accurately predict the likelihood of loan defaults. Machine learning, a branch of data mining, has emerged as a powerful tool for data analysis, capable of extracting valuable insights from large and complex datasets and making robust predictions. With its flexibility and accuracy, machine learning modeling is well suited for use in credit portfolio analysis and can help reduce problems in the banking and finance industries.

1.3 Relevance of the study

Currently, there are many risks associated with bank or microfinance loans, especially banks and MFIs that are exposed to the risk of capital loss. Then risk analysis and baseline assessments become important. These institutions hold vast amounts of data about consumer behavior and cannot determine whether applicants are likely to default. Machine learning, a subfield of data mining, is a promising area of data analysis focused on

extracting useful knowledge from large and complex datasets and making robust and reliable predictions. Machine learning modeling is flexible as far as data analysis is concerned. Statistical modeling is part of it. Such malleability and accuracy of results is surely required in credit portfolio analysis for reduction of banking and finance problems.

1.4 Objectives of the study

1.4.1 General objective

To model the prediction and evaluation of credit risk in Malawi using machine learning methods.

1.4.2 Specific Objectives

1. To model credit risk as probability of default
2. To model credit risk as credit score
3. To compare the performance of different machine learning models on credit risk prediction.
4. To determine relevant predicting factors for credit risk modeling.

1.5 Scope of the Study

Generally, this research focused on modeling credit risk in Malawi using supervised machine learning models. Specifically, Credit risk was modeled as Probability of Default and Credit Scores using classifier models. The study focused on these two measures of credit risk, disregarding Loss Given Default and Exposure at Default.

Recent studies and researches were used as reference in finding out the background of credit risk modeling and prediction both in Malawi and other countries.

1.6 Structure of study

The research paper was outlined by the pattern; introduction and motivation, literature review, mythology, results and conclusion. Figure 1.3 depicts the pattern of the structure vividly.

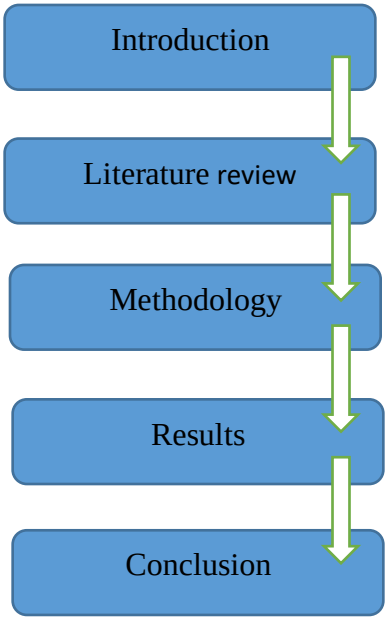


Figure 1. 3: Structure of the study

CHAPTER 2: LITERATURE REVIEW

This chapter aims at the analysis and evaluation of literature of the past researches about credit risk modeling in crediting institutions and outlining theoretical and conceptual framework of the study. The review runs from papers about evaluating the risk based on traditional methods to machine learning itself. The theoretical framework runs from reviewed theoretical background about machine learning for credit risk to the theories which were implemented in this study.

2.1 Reviewed literature

In the aspect of traditional financial analytics methods, (Chatterjee, 2015) used Vacisek model under Basel II accord based on International Financial Reporting Standard 7. By incorporating an economic capital allocation, the research aimed at meeting the need to develop quantitative estimates of the amount of economic capital needed to support a bank's risk-taking activities. Probability of default was the main measure of credit risk in the paper and it was modeled by structural models which the author claimed to be part of the Vacisek model.

Frydman et al., (2008) used the Markov mixture model to analyze the data. In their research, the original Markov chain model was expanded to include two Markov chains that were mixed together based on how quickly credit ratings changed. The primary distinction between the two studies is that the latter made use of Bayesian estimation while the former relied on the maximum likelihood method.

Fast forward to machine learning Sun et al. (2006) studied credit risk in commercial banks using support vector machines (SVM), and the experiments revealed that the binary model has a high level of classification accuracy. Huang et al. In order to improve the accuracy of credit scoring, (2007) proposed a hybrid SVM-based credit scoring model that looked for the best model parameters and feature subset. For the train and test data, respectively, the best model with accuracy of 92 and 93 percent was chosen and fitted.

In a study by Li, J., Liu, D., & Zhu, H. (2020), the authors compared the performance of different machine learning algorithms, including decision trees, logistic regression, k-nearest neighbors, support vector machines, and random forests, for credit risk prediction.

They evaluated the models on a real-world dataset and compared the results based on different performance metrics. On the other hand, a survey paper by Park, J. H., Kim, H. J., & Lee, H. K. (2019) provided a comprehensive overview of the recent research in credit risk prediction using machine learning techniques. The authors discussed the different machine learning algorithms used for credit risk prediction and the datasets used in the studies. They also provided insights into the challenges and future directions of research in this area.

In another study, authors Guo, X., Li, Y., & Wu, C. (2021) proposed a deep learning neural network model for credit risk prediction. They evaluated the model on a real-world dataset and compared the results with other machine learning models. They demonstrated that their proposed model outperforms other models in terms of accuracy and performance.

Classifiers and regression trees (CART) and multivariate adaptive regression splines (MARS) were used as feature selection methods in Yao's 2009 comparison of seven feature selection methods for measuring credit risk applied to public data sets from Australia and Germany. made that clear. While preserving the best predictions, CART can clean the tree and speed up execution. Birla et al. have produced comparable work. (2016), in which the authors examined credit risk on the basis of disproportionate data and discovered that logistic regression, classification and regression trees (CART), and random forests all performed admirably on the aforementioned disproportionate credit risk data.

Authors Purohiti et al. (2015) compared the best predictive models for credit scoring with SVMs, decision trees, and logistic regression and found that SVMs, decision trees, and logistic regression were the most accurate. Loan application classification. Turkson et al. (2016) conducted a thorough study in which the authors assessed 15 machine learning algorithms for binary classification. All algorithms, regardless of near-central Bayes and naive Gaussian, performed well with accuracy rates of 76–80 percent. Furthermore, they discovered that the prediction accuracy and other metrics did not significantly differ even when there was a total of 23 three signs. Hamid and Ahmed (2016) calculated credit risk using naive Bayes, neural networks, and decision trees. Decision trees are the best algorithms for accuracy, according to the findings of this study. Age, gender, type of

employment, place of residence, type of checking and savings account, loan size, term, and purpose were used to model credit risk.

The best algorithms for categorizing risk credits, according to Gahlaut and Singh's 2017 comparison of decision trees, support vector machines, adaptive reinforcement models, linear regression, random forests, and neural networks, are random forest algorithms. Age, duration, and frequency were also discovered to be the most significant factors. In the meantime, Khrishmar (2018) investigated 25 classification algorithms for binary credit risk prediction and discovered that neural networks perform classification more accurately and random clustering is superior in ensemble learners. Age, income, ownership, tenure of employment, willingness to borrow, credit score, and loan amount were contributing factors. The process of Knowledge Discovery in Databases was followed throughout.

In order to improve random forest algorithms, Zhang et al. (2018) proposed a reliable credit scoring model called NCSM based on feature selection and grid search. This model outperformed other linear models in terms of predictive accuracy because irrelevant features had a smaller impact. Khemchem and others (2018), estimating credit risk with the help of SVMs, neural networks, and linear regression. Their research compares the effectiveness of forecasting techniques both before and after data balancing. The findings demonstrate that, when compared to unbalanced data, the synthetic minority oversampling method (SMOTE), a sampling strategy, enhances the performance of the predictive model. When analyzing models for a task, keep the SMOTE sampling strategy in mind. In a research paper, Onay and ztürk (2018) examined 258 articles on credit scores and conducted a very thorough analysis of credit scores. After the majority of studies in 2010 used a single statistical method, this paper summarizes studies that used multiple statistical methods on the same data set. The most popular technique seems to be logistic regression.

In conclusion, credit risk prediction using machine learning techniques has gained considerable attention in recent years. The studies discussed above demonstrate the effectiveness of machine learning algorithms for credit risk prediction. However, there is still room for improvement in terms of accuracy, interpretability, and scalability of the models. Thus, the need for this study.

2.2 Theoretical Framework

In finance, credit risk prediction is a crucial issue, and the utilization of machine learning models has demonstrated remarkable capability in tackling this challenge. This section will delve into the theoretical foundation of machine learning models for credit risk prediction and the concepts used in this study based on the same.

Feature selection

In the context of credit risk prediction with machine learning models, feature selection plays a crucial role. The primary objective is to determine the most pertinent features that can precisely forecast credit risk. This can be accomplished through various feature selection methods, including but not limited to correlation analysis, principal component analysis, and mutual information-based feature selection. (Alimadadi et al., 2018)

Class imbalance

According to Tahir et al., (2021), imbalanced data is a common occurrence in credit risk prediction, where the number of non-default loans (good loans) far outweighs the number of default loans (bad loans). This imbalance can lead to inadequate model performance, as the model may exhibit a bias towards the majority class. To overcome this challenge, several techniques have been introduced, including but not limited to oversampling, undersampling, and cost-sensitive learning. These methods aim to mitigate the effects of imbalanced data and improve the performance of credit risk prediction models.

Model Selection and Evaluation

Tahir et al., (2021) also argued that selecting an appropriate machine learning model and evaluating its performance is crucial for credit risk prediction. Various models such as decision trees, logistic regression, support vector machines, and neural networks have been used for credit risk prediction. The performance of these models can be evaluated using various metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC).

Explainability and Interpretability

Further to that, in the realm of credit risk prediction, explainability and interpretability of machine learning models are crucial aspects to consider, given the potential financial impact of model decisions. Explainable models facilitate understanding of the factors that affect the model's decisions and provide insights into its behavior. Several techniques have been proposed to achieve explainability in credit risk prediction, including feature importance, decision rules, and partial dependence plots, among others. These techniques allow stakeholders to understand how the model works and to validate its decisions, ultimately increasing trust in the model's predictions. (Wang et al., 2019)

Data Preprocessing

Lastly, Data preprocessing is also an essential step in machine learning models for credit risk prediction. The data may contain missing values, outliers, and other anomalies that can affect the model's performance. Data preprocessing techniques such as imputation, outlier detection, and feature scaling are used to ensure that the data is clean and ready for analysis. (Alimadadi et al., 2018)

Reflecting on the reviewed literature and the theoretical background of machine learning modelling for credit risk prediction, credit risk is vividly and conveniently modeled with supervised machine learning models using factors like feature selection, imbalanced data, model selection and evaluation, explainability and interpretability, and data preprocessing. The same was implemented in this study. The type of models were supervised because data had a labelled target variable. Apart from the traditional and financial models reviewed, the other researchers modeled credit risk with labeled models like Support Vector Machines (SVM), Naïve Bayes etc.

That being so, this research was based on the aforementioned Knowledge Discovery in Database (KDD) process, focusing on using labeled dataset to train algorithms. On the modelling stage of the KDD process, supervised machine learning models for classifying ‘defaulted’ status from ‘paid up’ ones and predicting Probability of Default were incorporated. Like (Purohiti et al., 2015), the classifier models included Logistic Regression, Support Vector Machine, Decision Tree and Random Forest Classifier, Naïve

Bayes and KNearest Neighbor Classifier. These models were grouped into probabilistic and non-probabilistic models based on how the decisions, reflecting on the decision boundary, were made.

Class imbalance was handled, and feature selection models were also implemented. Reflecting on the reviewed argument by Wang et. al, (2019) , the model was also hinged on being explainable interpretable.

Credit score was modeled based on the prediction of probability of default where the least probability,0, was corresponding to the least credit score and the highest probability, 1, corresponded to the greatest credit score. The credit score ranges were based on the base FICO credit score scale which ranges from 300 to 850.

The features or independent variables in the models based on (Gahlaut and Singh, 2017) and (Purohiti et al., 2015) were age, source of income, loan intent, loan or liability type, loan amount or principal amount, sex, district, no of dependents, loan duration, payment terms, collateral type, interest type and calculation method and purpose.

2.3 Conceptual Framework

Figure 2.1 illustrates the conceptual framework that was used in this study.

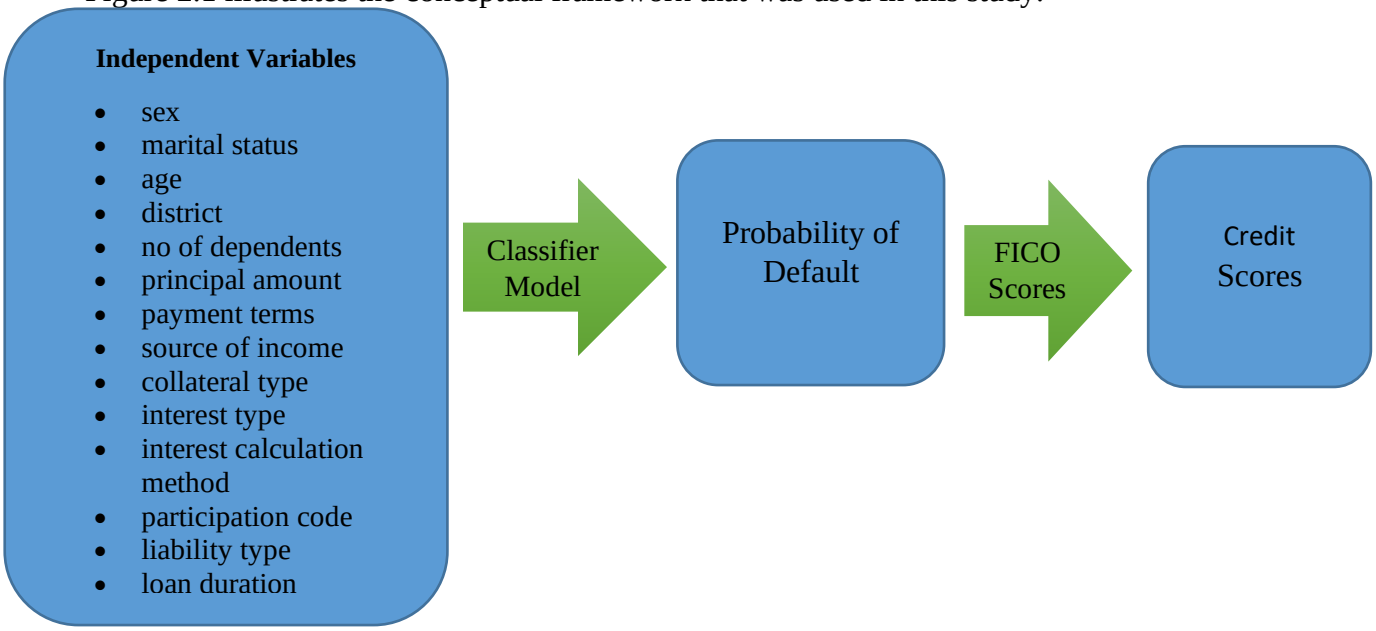


Figure 2. 1: Conceptual models of the study

CHAPTER 3: METHODOLOGY

This chapter aims at explaining the research type of this project and the Knowledge Discovery in Databases(KDD) process used. KDD includes explanations of the data used and its sources. The explanation of the data includes the type and expected target and independent variables. The process is basically the analytics methodology used to meet the research objectives and has been clearly outlined to stretch a clear picture.

3.1 Research design

The research was strictly quantitative. The justification simply being the fact that decisions were made on quantitatively collected and analyzed data. The results were based on evidence from the correspondent fitted models and other machine learning analytical techniques. Unlike qualitative research, this project was not based on empirical evidence.

Data collection and preprocessing were performed on PostgreSQL and analysis and modeling on python programming platform. That is because the data that was used was collected from PostgreSQL database and python is both flexible and performs well on big data, such that machine learning models were conveniently fitted and learnt.

3.2 Knowledge Discovery in Databases (KDD)

Based on the KDD process stages outlined by Watson (2019), this study was hinged on six summarized stages. The stages were namely; Problem formulation, data collection, pre-processing and cleaning, transformation, choosing mining tasks and methods and result evaluation and visualization. Figure 3.1 illustrates the 6 stage process.

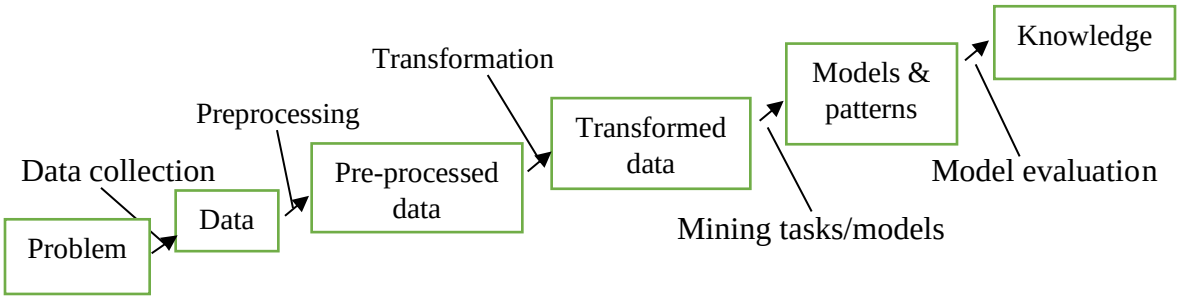


Figure 3. 1: 6 stages KDD process

i. Problem formulation

This was basically about the problem statement and research objectives. The whole issue of modeling credit risk using machine learning techniques on big data of Malawian subjects was the main and prominent problem identified and tackled in this study. As the reliability and accuracy of credit reports in Malawi is still questionable and a problem, the incorporation of the study in the banking and finance industry was seemingly a very good idea.

It was literally the first step of the process. The whole study was based on the concept of trying to model credit risk using machine learning models with Malawian credit data. The idea was based on the fact that crediting institutions in Malawi, including commercial banks and microfinance institutions, use traditional financial methods to determine risk associated with a specific or prospective obligor, which is not that reliable. As it stands, the crediting institutions come up with credit risk based on either intuition, empirical analysis of Markov chains, which are based on groups.

The study aimed at overcoming that by modeling credit risk with machine learning which produced reliable and sensible results. With the machine learning models which were fit and trained in this study, the prediction of credit risk of a specific individual is explicitly based on the explicit details of that person. Unlike the other methods which are being practiced on the ground, where the risk is predicted based on a group.

To solve the problem, the main hypothesis of this study was stated as:

Null hypothesis: Credit risk in Malawi cannot be modeled using machine learning algorithms.

Alternative hypothesis: Credit risk in Malawi can be modeled using machine learning algorithms.

While solving the problem, the task was branched into different types of classification machine learning models used. The idea was to compare probabilistic models to non-probabilistic models, to see which ones were to perform better.

On top of that, there was also a problem of determining which features are the best for credit risk modeling for machine learning. It is always a challenge to identify the right independent variables for prediction of a target variable. This study sorted out that problem using machine learning techniques.

ii. Data collection

The type of data that was used in this research project was relational with both numerical and categorical data types. That literally means that the data had a schema with numerical variables like 'principal amount' and categorical like 'marital status'. The target variable was binary, with the outcomes encoded by 1 and 0 for credit status 'default' or 'paid up' respectively. The data, theoretically, had features which were presumed correlated since most factors associated with credit risk are usually related. The target variable as expected, was dominated by individuals which did not default since the obligors which are victimized by that are practically less.

The data was sourced from CreditData Reference Bureau, a company which collects, processes and stores credit data of Malawian banks and microfinance institutions. It was licensed by Reserve Bank of Malawi to impose submission of credit data of creditors to such bureaus, to process, store and produce credit reports.

iii. Data pre-processing and cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. (Soofi, 2017) used the phrase 'garbage in garbage out' referring to messy results of the right model fitted with messy data. With unclean data it does not matter how good the model is. The results are always bad.

There are numerous ways which statisticians and data scientists use to clean their datasets ahead of analysis upon exploring it. Big datasets usually have inconsistences, missing values and noise. The inconsistences literally mean that there are discrepancies of codes, names and values in the data. For example, the data might indicate that after-tax net salary

is MWK 500,000,000, while the gross salary is MWK320,000,000. That is obviously an inconsistence since gross salary is supposed to be more than net salary. The noise is about variability and the presence of outliers in the data while missing values are just unrecorded records in some cells when other cells have records.

Based on such substantial motivation, although there is no standard with regards to data cleaning and preprocessing, the following steps were basically the methodology followed while cleaning the data

Step 1: Remove duplicate or irrelevant observations

This involved removing unwanted observations from the dataset, including duplicate observations or irrelevant observations and get more records with less missing values. Duplicate records were expected to be prevalent in the dataset due to the possibility of people or companies having gotten multiple loans from multiple creditors. When one combine datasets from multiple sources, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication was one of the largest areas that was considered in this process.

Irrelevant observations on the other hand were also handled. These are records that do not fit into the specific problem one is trying to analyze. For example, relating to the credit data on top of loan details, the dataset included records of insurance policies and utilities, which were removed as irrelevant. This would more likely make analysis more efficient and minimize distraction from the primary target, as well as creating a more manageable and more performant dataset.

Step 2: Fix structural errors

Structural errors occur when one measures or transfers data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can lead to mislabeling of categories or classes. For example, one may notice that both "N / A" and "Not applicable" are displayed, but they should be analyzed as the same category.

These kinds of structural errors were monitored in the dataset and fixed accordingly.

Step 3: Filter unwanted outliers

Often, there is usually one-off observations where, at a glance, they do not appear to fit within the data one is analyzing. If one has a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data one is working with.

However, sometimes it is the appearance of an outlier that will prove a theory one is working on. Just because an outlier exists, does not mean it is incorrect. This step was needed to determine the validity of such records. If an outlier found in the data proved to be irrelevant for analysis or was a mistake, removing it was considered.

iv. Data transformation

Before models were yet to be fit on the cleaned data, a number of transformations were implemented. The most prominent kinds of transformations which were implemented in this project were data dimensionality and numerosity reduction, normalization and recoding. The data as expected was large with many cases and features and the concept of dimensionality and numerosity reduction was implemented.

Numerosity reduction techniques like sampling and parameterizing the data was also implemented. If similar small sized datasets were found, they would have possibly been combined since machine learning models learn better when data is enough.

Data normalization of numerical variables was also incorporated to deal with differences in the scales of the features. Some features as anticipated were for example scaled in Malawi Kwacha currency and some in days or months which were all normalized in between 0 and 1. This was done to do away with concept of different scales of values in numerical factors which leads to biased predictions.

On the other hand, on top of all that, to enhance models' performance and accuracy, a number transformations were done to deal with the issue of class imbalance in the target variable. Obviously, as expected, the class of default obligors were found to be a minority which would have affected the reliability of models since that is the class of interest. If the class imbalance was not handled, the accuracy of the model would be based on the majority class of companies or individuals which did not default. The following ways opted for and

the resultant data yielded from the best technique was fitted on different models to assess performance:

1. Under-sampling

On this class balancing technique, the majority class got under-sampled. That is to say, the values in the class was randomly sampled down to the minority class, thus balancing them.

2. Over-sampling

On this one, the minority class got over-sampled. The values in the class were duplicated until the classes balanced. Apparently, this was the method that was opted for abd implemented for modelling data.

3. Reweighing

This is arguably the most practiced. technique of dealing with class imbalance in machine learning (Pan, 2009). None of the classes gets under-sampled or over-sampled on this one. The weights of the class get adjusted using machine learning techniques such that they both have approximately equal weights. Alternatively, the reweights could also be done in favor of the minority class.

v. Mining tasks and methods

To achieve the objective of modeling the prediction of credit status or credit score, a number of machine learning models were fitted and opted for the one which performed better. All the models were supervised and followed the general steps illustrated by figure 3.2

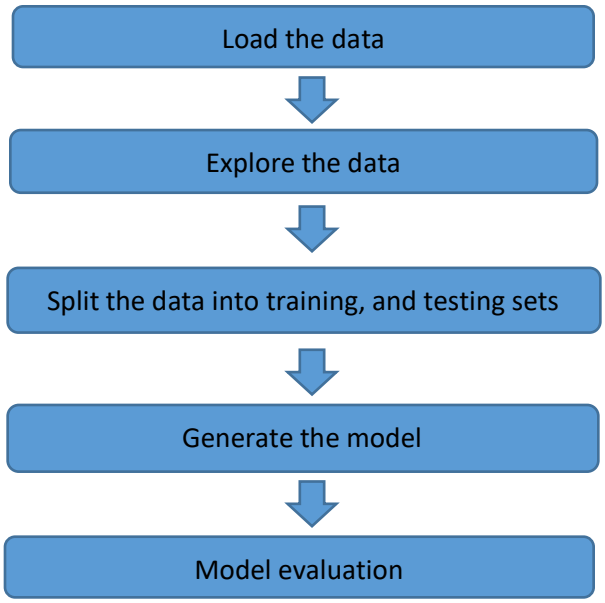


Figure 3. 2: Steps of a supervised learning model

where training data is what was given to the classifier during the training process. This is basically the paired data for learning f(x) = y, where x was the input (a vector of features) and y was the output (a label with classification). During training, the aim was to try to find a good model of f(x) that replicated the input-output relationship of the data. On the other hand, test data was used with a trained classifier to assess it.

With a testing dataset, there is always an x, but no y. Usually, not having y in research is a pretense, but in the real world, one truly won't have it. The trained system predicted y the one with the testing input, and assessment was conducted on how well the system works.

Reflecting back on this project, because the target variable was binary, only classification models were implemented with the data split into training and testing sets. However, the models were classified into two namely; probabilistic and non-probabilistic models. These models in general included, Logistic Regression, KNearest Neighbor Classifier, Decision Tree Classifier and Random Forest Classifier.

1. Probabilistic models

In the theory of machine learning, classifier models are predictive models that learns on data and predict a class label based on an input. But some classifier machine learning models do not directly predict a class for an example of the given input but instead report a probability. This type of classification model is referred to as probabilistic classification model (Bazdok, 2018).

Just as an example, the model might predict that the chance for the observation or value to be positive is 75%. Naturally, the choice should be assigning the predicted value as positive literally because the probability of prediction is more than 0.5. Nonetheless, adjustments on the threshold can be made such that one does not have to be fixed or stick to 50%. The threshold can be raised such that the model should only classify observations as positive if the models predict probability which is more than 0.9. The increased threshold of probability enhances the model to only make positive predictions whenever confidence is high to do so. On the contrary, if the threshold dropped, the model will be at liberty of making and assigning positive labels recklessly. Adjusting the threshold affects the model's precision and recall heavily.

If the classes are clearly separated, then one side of the threshold or decision boundary is clearly classified as one class and everything on the other side as the other class. However, when the classes overlap and cannot be clearly separated, then the points on each side of the decision boundary get classified with some probability. This decision is called a soft decision and that is what probabilistic models are based on.

Theoretically, these soft decisions are made by modeling the probability of a point being labeled class a function of its distance to the decision boundary. The further from the boundary, the more certain one is of the label. The closer to the boundary the more uncertain one is of the label.

The following are the probabilistic models fitted and trained in this project:

i. Naïve Bayes

Based on a paper written by (Luckert, 2015), Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

It is one of the simplest algorithms for learning, more interestingly in some cases it may outperform most of the sophisticated learning algorithms. The naive Bayes uses maximum-likelihood estimation to classify new examples. It is based on the Bayes' theorem which states that,

$$\mathrm{P(A|B)} = \frac{P(B|A) \times P(A)}{P(B)}$$

Where P(A), P(B) being the probabilities of A and B, and P(A|B) and P(B|A) are conditional probabilities of A, given B and B, given A respectively and P(B) is greater than 0.

During the training of the Naive Bayes, the probability of finding an example of each class in the sample population is calculated and stored as the prior probability for that class. It also calculates probability for instances x given its class c. Under the assumption that attributes in x are independent, it simply becomes a product of probabilities of each single attribute 23 3.3 Machine Learning Algorithms for Classification (Gosavi, 2014). Hence Bayesian theorem, when applied to classification problem using Naive Bayes, becomes,

$$\mathrm{P}(A_i \,|\, B) = \frac{P(B|A_i) \times P(A)}{P(B)}$$

The probability $\mathrm{P}(A_i|B)$ is referred to as posterior probability i.e. probability of event after evidence is seen and P(A) the prior probability i.e. probability of event before evidence is seen. A class A_i is chosen if $\mathrm{P}(A_i|B) = max_k\{P(A_k|B)\}$ i.e. the class with highest posterior probability.

In the aspect of this project, the posterior probability was the probability of either predicting credit status 'default' or 'paid up' as A_1 and A_2 respectively where B is the set of features. If $P(A_1|B)$ was found to be less than 0.5, then the prediction would be 0 corresponding to 'default' and 'paid up' otherwise. Unlike the other models, it was the only model whose parameters were not tuned.

The fundamental Naive Bayes assumption is that each feature makes an independent and equal contribution to the outcome. With relation to the dataset used in this project, this concept can be understood as: An assumption that no pair of features in the dataset are dependent. For example, principal amount being a big sum had nothing to do with the Sex being 'male' or age being high. Hence, the features were assumed to be independent. That is why the model is called 'Naïve'. Secondly, each feature is given the same weight (or importance). For example, knowing only sex and age alone can't predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing equally to the outcome.

A clear advantage of using the naive Bayes is that it is fast to train and fast to classify the data. This is because it needs to scan the database to compute probabilities and store it in a table during training and use this table to classify future examples. Also, the naive Bayes is inherently more robust against irrelevant features (J. Kim, 2008), as the likelihood of a class is product of probabilities of each single attribute. On the other hand, for prior probabilities to be realistic, the sample population needs to be truly representative of the actual data. Another major disadvantage is that the classifier assumes features to be independent of each other. However, in many cases Naive Bayes classifier performs reasonably well even in cases where features are dependent on each other.

This was literally the only model that was fitted without tuning parameters. The reason was simply because the only parameter that can be changed for a Naïve Bayes model is the prior probability. This means that if the prior probability was to be specified, the priors would not get adjusted according to the data which is a loophole to the objective of the study. Priors were set to 'None' so that the model could learn based on the natural distribution and structure of the data.

ii. Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables, Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. (Bazdok, 2018).

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

In the case of the data that was targeted in this project, the logistic regression model was perfectly and conveniently fit for credit status since the outcomes of the target variable will either be defaulted or not, encoded by 0 and 1. The model was used for modelling prediction of credit risk and determination of the influence of each of the features on the respective dependent variables.

Mathematically, as a more advanced analysis of the relationship between the target and independent variables, the logistic regression function was incorporated as:

$$Logit(y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

and

$$P(y = 1) = \frac{e^{(Logit(y)}}{1 + e^{(Logit(y)}}$$

where y represented the target variable credit status encoded by 1 for credit default and 0 for no. On the other hand, $P(y = 1)$ was the probability of an obligor going default.

X represented values of independent variables whereas β_i expressed coefficients of the independent variables and measured the effect of X_i on the odds (the chance-the probability) that an obligor has defaulted. The basic formula for the analysis with logistic is $P/(1-P = e^{\beta_0}.e^{\beta_1}$ where $P/(1-P)$ expressed the probability of success of an event (the probability of an obligor going default) divided by the probability of no event happening (probability of an obligor not going default.). e^{β_i} values, i was the position of the independent variable, represented the odds ratio and measured the degree of change in odds for every unit increase in the independent variables.

This model was fitted several times, such that its performance was monitored each time with the performance metrics; accuracy score, confusion matrix, precision, recall and ROC curve. The tuning for the betterment of the performance of the model was based on changing figures of three parameters namely, solvers, maximum number of iterations and number of features. The solvers were basically algorithms used to optimize the problem while maximum number of iterations were the number of iterations taken for the solvers to converge. On the other hand, the number of features was the count of the independent variables, amongst the 10, which were considered for the modeling credit status. All these are inbuilt parameters that come with the logistic regression function in python.

2. **Non-probabilistic models**

If classes are clearly separated, then classification is made on one side of the decision boundary as one class and everything on the other side as the other class. This decision has no uncertainty and it is called a hard decision. Non-probabilistic machine learning models are based on that simple concept. These models are also known as non-parametric models where machine learning algorithms do not make assumptions about the arrangement of the mapping function.

Given observed patterns, non-parametric machine learning algorithms tend to make assumptions about the data of interest. A popular example of non-parametric machine learning algorithm is KNN which matches observations by training patterns for new and unseen values. The theory behind this algorithm is based on an assumption that the patterns trained which are the most alike are highly likely to have a similar classification result. The

advantage of non-parametric machine learning algorithms is that they are flexible despite being slow and requiring big data to train properly.

The following are the non-parametric models used in this project:

i. KNearest Neighbor Classifier

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection (Clark J. D., 2015). Based on (Patel, 2017) , A KNearest Neighbor classifier classifies an object by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.

This KNearest Neighbor model was fitted as a classifier where the target variable was classified by the plurality vote of its neighbors, with the variable being assigned to the class which is most common among its k nearest neighbors. It was fitted to predict credit status of an obligor through determining probability of default. The model is arguably the simplest machine learning model, let alone among the classifiers.

The data was fitted with a convenient value of k while being cautious of overfitting and underfitting. Underfitting and overfitting occur when the models is setup to underperform and over perform with wrong hyper parameters respectively. In the case of this classifier, the model would underfit if an extremely small value of k is implemented and overfit if an extremely large value was opted. The challenge to find the proper one would be determined by tuning and k-fold cross validation.

A series of KNearest Neighbor (KNN) machine learning models were created and trained. The performance of each model was compared and monitored to determine the best one. The difference between the models was based on the number of nearest neighbors (k) set as a parameter in the model. To find the best value of k, the models were fit starting with small values of k and increasing in increments of 1.

ii. Decision Tree and Random Forest

Decision trees machine learning models are hierarchical, where each step is involves splitting values of a variable based on a test against a preset threshold value. Decision trees are comprised of decision nodes and branches where the former is where the product of the splitting or outcomes is indicated while latter are literally leaves where splitting is performed. For all the desired outcomes to be achieved, the recursively nodes apply tests on values of a variable of the data interest, until it exhausts the whole tree and hits a leaf node, which represents the output (Gosavi, 2014).

A typical decision tree is illustrated in figure 3.3. Training involves choosing features which provide the most information about the training set. It then subsequently develops a tree using a top-down approach. Decision trees use a "white-box" approach, wherein the internal decision making and structure of the tree is visible to the user. Arguably, this contributes heavily to the easy visualization and interpretation of decision trees. Decision trees are disadvantageous because sometimes they can create complex trees that do not generalize the data well by increasing generalization error or overfitting to the data.

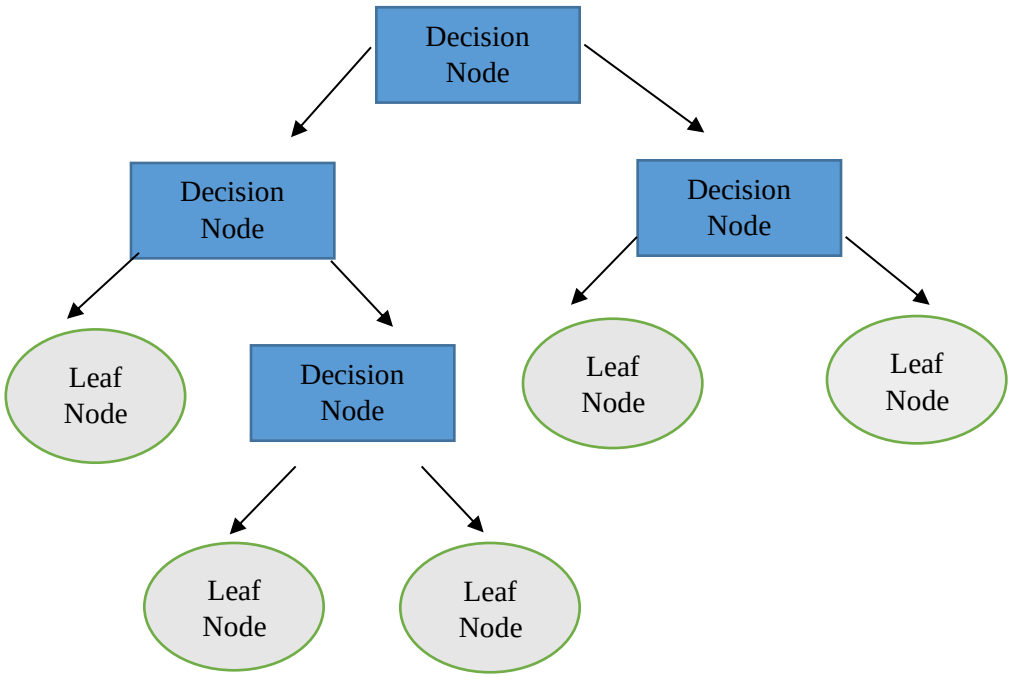


Figure 3. 3: Typical Decision Tree Classifier Model

In this model, the only key parameter that was adjusted for improved learning was the maximum depth of the tree. Decision trees have a depth that can be adjusted to shape the way the model learns. Sometimes, a small maximum depth value can result in underfitting, and a large value can lead to overfitting. The goal was to find a specific value of the maximum depth parameter that would result in a well-learned decision tree classifier model with neither overfitting nor underfitting.

On the other hand, as per an article by (Yiu, 2019), random forest, like its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes or better performance becomes the model's prediction (see figure 3.4)

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. The reason that the random forest model works so well is based on the fact that a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. They are basically rooted on the quote "None of us is as smart as all of us" by Jean-Francois Cope.

Random forests are based on the following prerequisites for them to perform better:

1. There needs to be some actual signal in the features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

In the case of this project, both decision tree and random forest classifiers were fit and the leaf nodes were credit default or not, while the decision nodes were where the data is split. Literally, the leaf nodes were the outcomes of the modelling and predictions. Each branch was correspondent to each independent variable, thus the number of features which were be fitted determined the depth of the tree or random forest.

On the other hand, the Random Forest is just a combination of decision trees. Numerous decision trees were fitted using multiple splits of data and Random Forest of Decision Tree Classification and Regression models will be generated. The best performing model out of

the ensemble will automatically be chosen by the virtual analyzer and thus being the performance of the Random Forest.

The training set was used to fit and learn the model just like the others. The validation set was also used to tune parameters and perfect its performance. The parameter that was tuned for enhancement of the performance of the Random Forest model and better learning to the data was forest depth. Unlike, in the decision tree model, this depth was for all the trees ensembled such that each tree would succumb to that.

iii. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression. The goal of the SVM algorithm is to use a training set of objects (samples) separated into classes to find a hyperplane in the data space that produces the largest minimum distance, called margin between the objects or samples that belong to different classes. Technically, the hyperplane is known as the maximum margin hyperplane. The model only uses the objects (samples) on the edges of the margin called support vectors to separate the objects rather than using the differences in class means. The name of the model is known as Support Vector Machine because the separating hyperplane is supported or defined by the vectors or data points nearest the margin (Montilla et al., 2012).

According to (Xia, 2020), SVM has been shown to perform well in many real learning problems with a variety of settings and is often considered one of the best "out-of-the-box" classifiers. SVM has the advantages of increasing class separation and reducing expected prediction error. Additionally, SVM is flexible for both linear and nonlinear-based discriminatory analyses, generally referred to as 'kernel' functions and it is suitable for analysis of high-dimensionality datasets with small sample size when combining with feature selection approaches. The best linear hyperplane is determined by comparing marginal distances. The one with the highest marginal distance is considered as the best hyperplane for the modeling. Figure 3.4 and 3.5 illustrate the concept SVM model with linear and non-linear hyperplanes respectively.

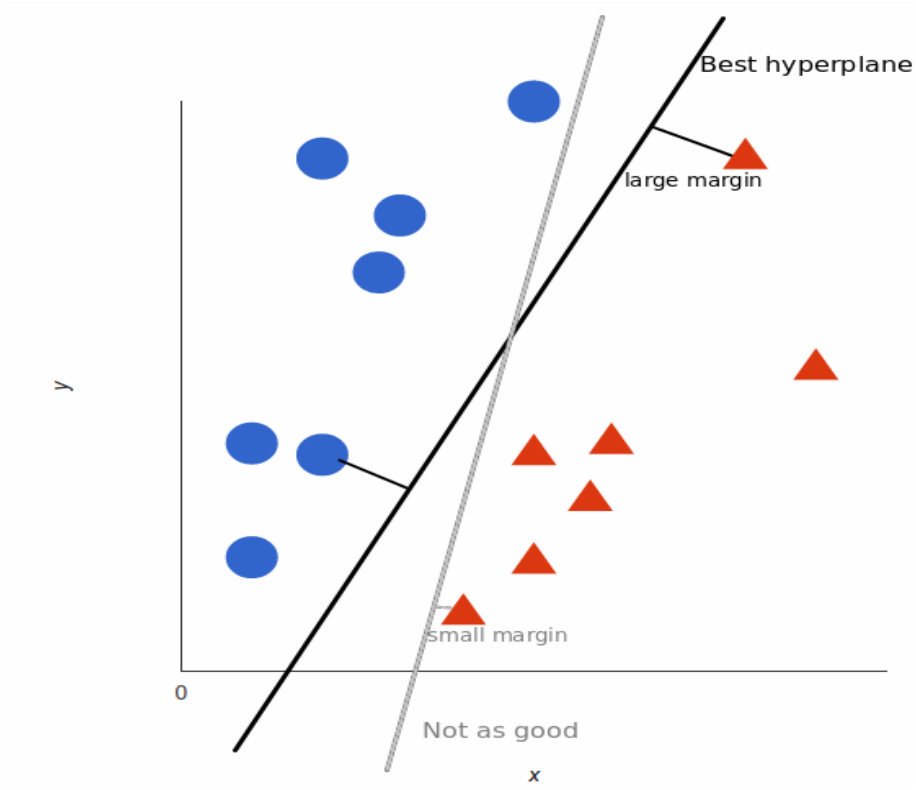


Figure 3. 4:SVM with linear hyperplane

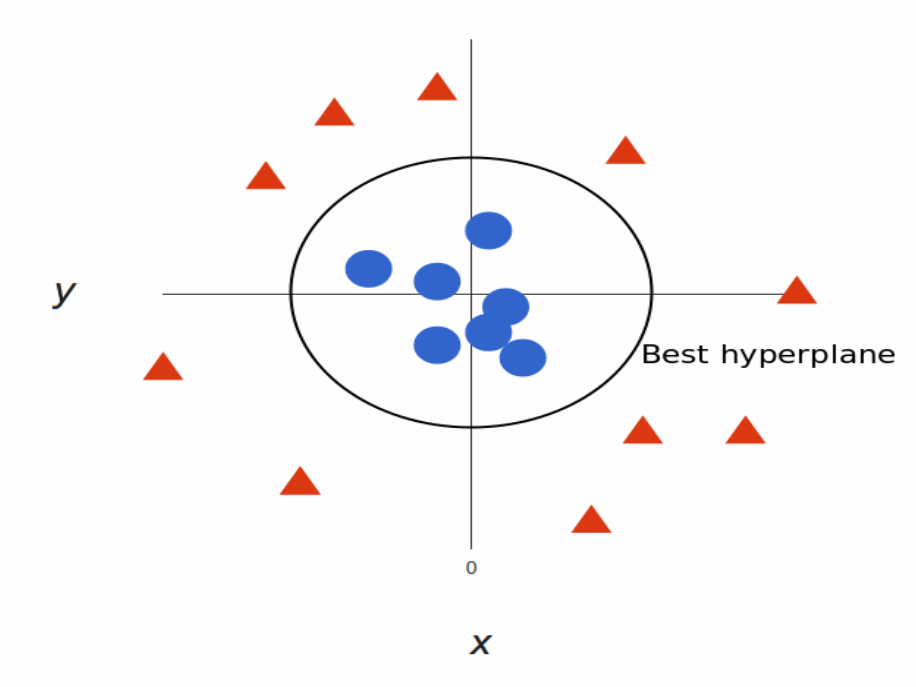


Figure 3. 5: SVM with non-linear hyperplane

Different Support Vector Machines models were fitted with different kernel functions namely, polynomial, linear and radial basis function (RBF). Simply put, a kernel function is a technique for taking data as input and transforming it into the format needed for processing data.

The performance of each model was monitored by different metrics, and they were all compared and evaluated.

vi. Results evaluation and visualization

This is where the performances of the models were assessed and evaluated. The question that was basically answered here was "how well did the models work after training and testing?". A model's performance can be evaluated using training dataset. However, the performance does not really tell how well the model can generalize to other data by lowering generalization error. In other terms, the model just memorizes the training data but tends to be clueless when new data is involved. Therefore, test dataset which is different from the training dataset to get a better insight of how well the model performs. So model performance was evaluated both on training and testing stages.

There are a number of evaluation measures for machine learning classifiers, but however, only five metrics were used in this research as follows:

1. Confusion Matrix

A confusion matrix is basically a table in form of a matrix that allows visualization of the performance of a machine learning algorithm, specifically a supervised one. Each column of the matrix represents the count of values predicted while each row represents the frequency of actual values. It depicts a story of the classes which are either most confusing or accurate. High diagonal numbers imply good performance. For a normalized confusion matrix, the sum of the values in the rows amount to 1.

For a binary response variable, the confusion matrix is composed of four entities called True Positives(TP), True Negatives(TN), False Negatives(FN) and False Positives (FP). As for the evaluation by confusion matrix of this project, the TP were counts of obligors which had a 'Paid up' credit status and were predicted as such, while TN were the ones

which were predicted to have defaulted on a loan when they truly defaulted. On the contrary, FN and FP are basically the vice versa of TP and TN where their respective values correspond to wrongly predicted statuses.

2. Accuracy score

This is the fraction of points correctly classified. It is basically a metric used in classification problems used to tell the percentage of accurate predictions. The accuracy can be calculated to be the sum of True Positives(TP) and True Negatives(TN), divided by all data points. In other terms, the calculation is by dividing the number of correct predictions by the total number of predictions. Reflecting on the evaluation of the credit risk models fitted in this project, the accuracy score was a product of total number true predictions of either defaulted credit statuses or paid up statuses divided by total number of predictions. Nonetheless, when the classes of the target variable are imbalanced, the overall accuracy can be biased and misleading. Mostly, it is important to predict the minority class correctly compared to the majority class.

3. Precision and Recall score

Precision is calculated by dividing the total number of examples that are classified as positive by the number of examples that were correctly classified as positive. Recall is calculated as the proportion of positive examples in the test set that were correctly classified out of the total number of positive examples. In object detection, precision-recall is a key metric that is particularly helpful when categories are unbalanced. High recall denotes that an algorithm returned the majority of the relevant results, whereas high precision denotes that an algorithm returned significantly more relevant results than irrelevant ones.

Considering the fact that the credit statuses in the dataset were also unbalanced, precision and recall scores were also incorporated as evaluation metrics of the classification machine learning models which were fit.

4. ROC Curve

The performance of binary machine learning classifiers can be illustrated by a graphical plot called a Receiver Operator Characteristic (ROC) curve. Although it was first employed in the theory of signal detection, it is now utilized in a wide range of other fields, including artificial intelligence, machine learning, agriculture, and health care. (Chan, 2020).

Plotting the True Positive Rate (TPR), also known as sensitivity, against the False Positive Rate (FPR), also known as 100 percent less specificity, where specificity is the True Negative Rate (TNR), results in the construction of a ROC curve. The percentage of observations out of all positive observations that were correctly predicted to be positive is known as the true positive rate (i). e. (TP/(TP FN)) Similar to the false negative rate, the false positive rate is the percentage of observations out of all negative observations that are incorrectly predicted to be positive. e. (FP/ (TN) FP) For instance, in the context of this study, the true positive rate represented the proportion of obligors who were correctly identified as having made their loan payments.

An individual point on the ROC space is produced by a discrete classifier that only outputs the predicted class. For probabilistic classifiers, however, a curve can be made by adjusting the cutoff for the score, which provides a probability or score that indicates how strongly an instance belongs to one class as opposed to another. It should be noted that by "looking inside" their instance statistics, many discrete classifiers can be transformed into scoring classifiers. A decision tree, for instance, uses the percentage of instances at a leaf node to infer the class of the node. (Zhou X.H, 2007).

The trade-off between sensitivity and 100% specificity is technically represented by an ROC curve. Classifiers perform better when their curves are closer to the top-left corner. Points along the diagonal where FPR = TPR are anticipated as a baseline from a random classifier. The test becomes less accurate as the curve approaches the ROC space's 45° diagonal. The Area Under the Curve (AUC), which measures performance, is another factor. The closer to 1 the AUC is, the better the model.

The fact that the ROC is independent of the class distribution must be pointed out with care. This makes it useful for assessing classifiers that forecast uncommon events like

defaulted loans. However, using accuracy scores calculated by (TP + TN)/(TP + TN + FN +FP) to measure performance would favor classifiers that consistently foresee a bad outcome for rare events.

Credit score evaluation

Subsequently, since the prediction was first done for probability of default, FICO score ranges were used to predict credit scores proportionally on the best model chosen. This was done by equating the highest achievable probability of default to the lowest FICO score and vice versa.

CHAPTER 4: DATA COLLECTION, CLEANING AND ANALYSIS

This chapter basically hinges on outlining the Knowledge Discovery of Databases(KDD) process that was conducted in this study. Explicitly, it runs from data collection subsequent to the predefined problem, up to data preprocessing which includes data cleaning and transformation, then analysis and machine learning modeling.

4.1 Data collected

The data that was used in this study was sourced from CreditData Reference Bureau. CreditData is a Malawian company with a strong understanding of the local market in the aspect of credit information referencing. The company's business entails collection of data, which assist it to give out credit stories. This data is collected from lending institutions licensed by the Reserve Bank of Malawi, a central bank in the country. They also collect credit data from any institution that transact on credit as they carry their businesses.

CreditData collects data for both individual and company obligors. The data that was collected for this analysis was only for individuals. The one for companies was avoided because they have different features as compared to individuals. Getting both datasets would mean a model for each, thus imposing too much work. The database source was in PostgreSQL and a query was used to pull and export it to csv.

The data was collected with 1,000,000 rows in a comma-separated values(csv) file. It had 21 columns of which 20 were features and 1 was the target variable. The target variable was credit status id which had binary values. The values were either 0 for 'paid up' status or 1 for 'defaulted' status. 'paid up' status means the correspondent obligor paid back all his or her debts while 'defaulted' means he or she never managed to pay back until the agreed date. On the other hand, the collected features were:

> 'sex', 'marital status', 'date of birth', 'home address', 'home village', 'residence plot no', 'district' 'loan date','no of dependents', 'principal amount', 'principal amount local', 'opening balance local', 'clearing date', 'payment terms', 'source of income',

’loan intent’, 'collateral type', 'interest type', ‘interest calculation method’, 'participation code' and ‘liability type’

Sex was basically a demographic feature which gives details of the obligor’s gender. As per common sense, the values of the sex feature were either female or male. Marital status is another demographic factor. It entailed values including ‘single’. ‘married’, ’divorced’ and ‘widowed’ where all those in a relationship or engaged fell under ‘single’. Date of birth was simply the day the obligor was born which was later transformed to age while district is the location where he or she stayed at the time of loan. Home address and home village entailed where the individual currently stays and where he or she originates from respectively. Residence plot number was basically the house number of the correspondent individual. Number of dependents were the number of human beings the obligor is responsible for at the time of the loan.

On top of the demographic details, details about the loan taken were also incorporated beginning from when the loan was taking, all the way through the intention of the loan to how many got that loan. ‘loan date’ was about the day the loan was taken by the obligor and given by the creditor while ‘principal amount’ entailed the total amount of money borrowed. The currency of ‘principal amount’ was general while that of ‘principal amount local’ was specifically Malawi kwacha such that the former contained amounts either in the local currency or foreign currency. Clearing date was when the loan was officiated. On the other hand, ‘payment terms’ was about the agreed intervals of paying back the loan. Some maybe agree to be paying on weekly basis which some may agree on monthly or quarterly basis. 'source of income' was a feature about how the one who got the loan earns income. The values of this feature were simply three namely: business, white collar job or blue-collar job where white collar jobs are usually highly paid which blue collar are involved with peanuts pays. A good example of a white-collar job is a bank manager while that of a blue collar job is a security guard.

The intention of the loans was recorded under the feature ‘loan intent’ while the type of collateral used by ‘collateral type’. The types of collateral collected in the dataset were cash, shares and property. ‘Interest type’ and ‘Interest calculation method’ entailed about the interest correspondent to the loan. Each loan is associated to an interest such that one

does not give back the exact amount of money borrowed. ‘participation code’ was about the number of people participating in the loan. Some get a loan independently while some get it as a group of company. Lastly, liability type was a feature which shed light on the literal kind of liability or account one has. For example, some have a loan which they intentionally got from a bank or a micro finance company while some have overdrafts and insurance policies.

4.2 Data preprocessing and cleaning

The raw data with all 1,000,000 cases which was sourced from CreditData Reference Bureau was preprocessed and cleaned as the third stage of the KDD process. Subsequent to getting the data from PostgreSQL, it was then loaded into python where everything else in the aspect of analysis was performed.

Data cleaning was performed in two stages. First it was done to the whole dataset in general and then each variable, be it target or independent, were cleansed if there was a need. The general data cleaning was performed as follows:

- Duplicates

Based on the fact that the database of CreditData Reference Bureau has duplicated cases, cleaning of such had to be the first step. Having worked at the company for about a couple of months, it was an apparent thing to note the duplicates issue in the database.

Upon dropping the duplicates, the total number of cases reduced to 989,818 from 1,000,000. Only one out of the duplicated records was left behind without dropping it. In order to maintain accuracy and prevent false positives in statistics, it is crucial to remove duplicate rows of values and columns from datasets when collecting data for machine learning models.

- Feature reduction

This was the next step. Feature reduction is an example of dimensionality reduction which involves cutting down the number of independent variables before fitting machine learning models. With too many input variables, machine learning algorithms' performance can suffer. It is frequently helpful to reduce the dimensionality when working with high

dimensional data by projecting the data to a lower dimensional subspace that perfectly encapsulates the data.

The dimensionality reduction in this study was done in two phases. The first stage involved removing features which seemed unnecessary. The second one was based on multi-collinearity of the features such that the features which were found to have a high correlation coefficient be it positive or negative, one of them was dropped.

The irrelevant features which were removed based on intuition were 'home address', 'home village' and 'clearing date'. The rationale was basically the fact that they would clearly not have any impact on credit risk prediction.

On the other hand, other features with numerical data type were reduced by multicollinearity method based on the findings on figure 4.1.

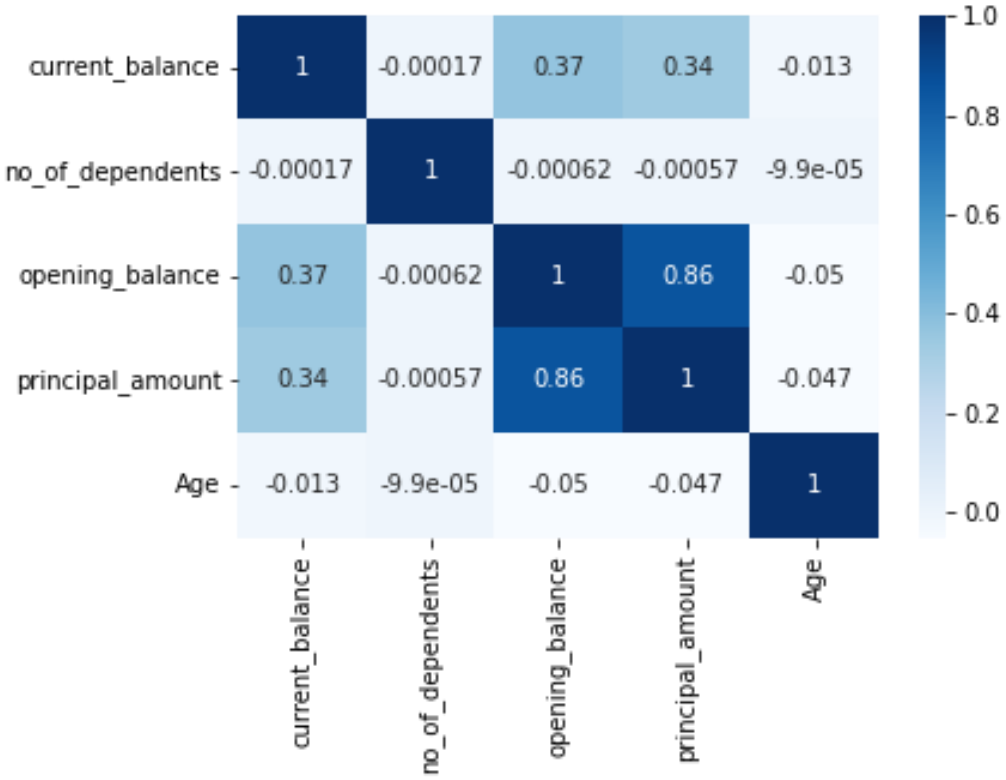


Figure 4. 1: Heat map of multicollinearity test of numerical variables

Multi-collinearity is the term used to describe the linear relationship between independent features in a dataset. It is founded on the relationship between two or more dataset variables. While working on a regression-based problem, it is beneficial to have a strong correlation between independent and dependent features. However, if the independent features are also correlated with the dependent features, it may have an impact on how well a machine learning model performs. For this reason, feature selection based on the multicollinearity theory was put into practice.

According to the theory of correlation, a correlation coefficient is a number between -1 and 1, which indicates the strength and direction of a relationship between variables i. e. opening balance and current balance were removed from the dataset because they were discovered to have a high positive value, as shown in figure 4.1; this measure reflects how similar the measurements of two or more variables are across a dataset. It is referred to as a perfect positive correlation when one variable change and the other variables change in the same direction. Zero correlation indicates that there is no relationship between the variables, while perfect negative correlation describes a situation in which one variable change and the other variables change in the opposite direction.

In the case of this study's data, there are exactly four perfect positive correlations with a correlation coefficient of 1. This means the variables being compared in that case are identical, which makes total sense reflecting on figure 4.1. features 'principal amount', 'opening balance' and 'current balance' were also found to have a perfect positive and strong correlation with coefficients 0.37,0.34 and 0.86 for pairs opening balance and current balance, principal amount and current balance, opening balance and principal amount respectively. That literally means all of them cannot be used as features for predicting an outcome since they are more likely to affect performance of a model no matter how good it is. Only 'principal amount' was opted for while the other two were dropped.

On the other hand, only variables 'age' and 'no of dependents' were found to have no correlation with any other numerical feature. The correspondent correlation coefficient with any other feature was negligibly small, be it positive or negative. That literally means that the two features were found with zero correlation with other variables because no

perfect positive correlation or negative perfect correlation were observed. Based on that, none of these two variables were dropped as part of a feature selection task that was performed on this stage.

- Missing values

Another crucial step in the preprocessing and cleansing of data was dealing with missing values. When using the dataset for any machine learning algorithm, missing values are a big problem. Missing values are still undesirable even if we ignore the algorithmic perspective. They impair data visualization and analysis. A machine learning project pipeline includes both of these as crucial steps. The project's mandate was to eliminate the prevalence of missing values in the credit risk knowledge discovery process.

Monitoring of missing values for each variable including the target variable was conducted. The following were the counts of each missing data corresponding to each variable, the target variable first:

Variable	Number of nulls
Credit status	0
Sex	0
Marital status	84
Date of birth	0
District	0
Opening balance	0
Number of dependents	0
Number of participants	241
Liability type	0
Interest type	0
Interest calculation method	188
Principal amount	42
Payment term	0
Days in arrears	56
Opening balance	0
Current balance	0
Profession	52423
Collateral type	125
Participation code	19289

Further to that, to have a clear understanding of the gravity and severity of the missing data, a heat map of the same was plotted. Sometimes the seriousness of the missing values can vividly be determined by a visualization. The way forward of handling missing data

depends on how much of a proportion it is to the whole dataset. Figure 4.2 vividly illustrates that.

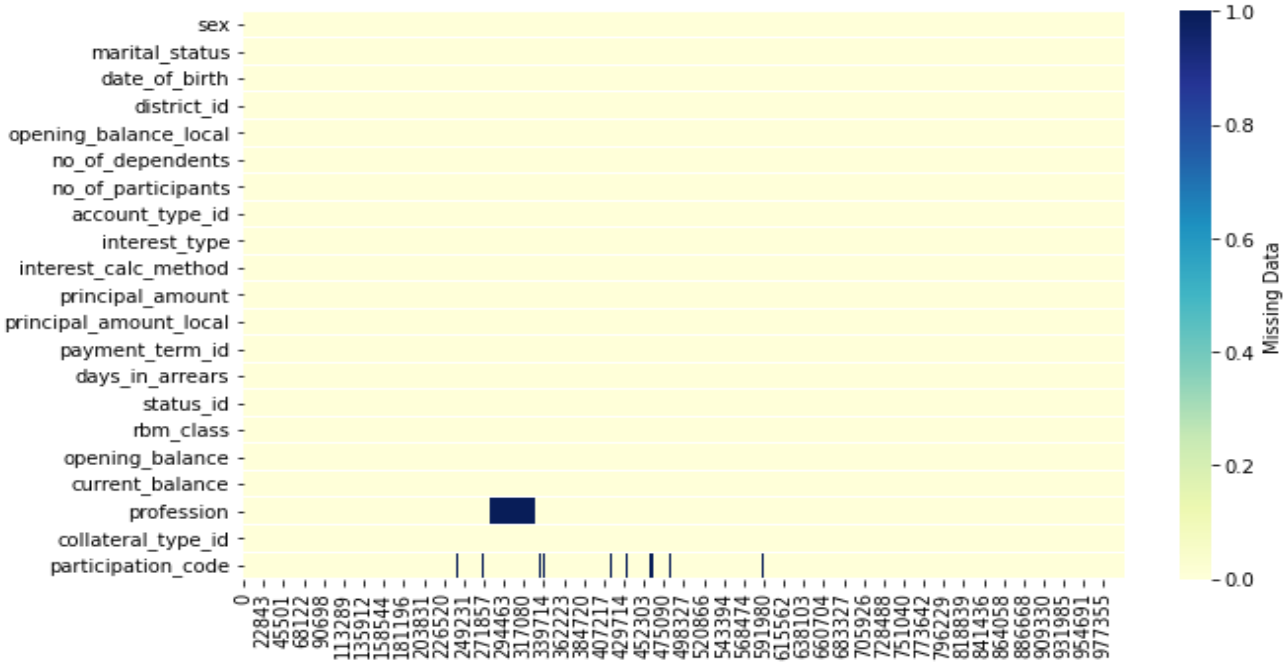


Figure 4. 2: Heat map of missing data

Apparently, the missing values found in features 'profession/source of income', 'interest type', 'collateral type' and 'participation code' were not that severe. Figure 4.2 clearly shows that the vast majority of the data has no null values thus concluding to drop all the missing data since the remaining data is big enough for analysis. If the missing values were so prevalent, other techniques like imputing values using models would have been implemented to handle the missing data.

Having dropped all the rows with the null values, the count of subjects decreased to 918,019 from 989,818. This was the data that went through to the next phase of data pre-processing and cleaning.

- Recoding and general cleaning

Subsequent to cleaning the missing values, the 918,019 rows of data were involved in recording of categorical variables and general cleaning of other things. The categorical

variables which were recorded included the target variable. The variables were ‘sex’, ’marital status’, ’district’, ’payment terms’, ‘source of income’, ‘collateral type’ and ‘participation code’.

On top of that, as it has already been mentioned, general cleaning of the data as a whole was also conducted. There were some categorical variables with category ‘UNKNOWN’ which obviously, would be ambiguous and confusing for decision making. For example, if a ‘sex’ category is ‘unknown’, it would be very hard to determine if the subject was a male or a female. That way, such noise would compromise the integrity and cleanliness of the data towards modeling.

The total number of records with ‘unknown’ as a value were 78,087, representing 8.5% of 918,019. Based on that, a decision to drop all rows with such a value was made since the proportionality was not substantial and the remaining data would still be enough for analysis and modeling.

The encoding or recoding of categories of the categorical variables into numbers was conducted variable by variable. The following is a detailed breakdown on how each categorical variable was recoded as part of data cleaning:

Credit status

This was basically the target variable in this project. As per expectations, it was supposed to only have two distinct categories listed under it i.e., ‘defaulted’ and ‘paid up’ status. But based on the raw data that was collected, there were additional statuses like ‘paid in full’ and ‘resumed pay after default’. All these had different codes in the raw data. The codes were 1,2,4,5 for ‘defaulted’, ‘paid up’, ’paid in full’ and ‘resumed pay after default’.

Grounded on the knowledge gathered at CRB, the extra credit statuses, ’paid in full’ and ‘resumed pay after default’ are about a subject, individual or company, having paid all the debt and restarted paying back a loan after defaulting respectively.

As such a decision was made to recode all ‘paid in full’ statuses as ‘paid up’ statuses. ‘Resumed pay after default’ status was also, on the other hand, recoded into ‘defaulted’ status considering the fact that the respective subject had defaulted before.

All ‘paid up’ statuses were recoded as 0 while ‘defaulted’ statuses were recoded as 1s. This means that for any probability predicted below 0.5 would be a prediction of a ‘paid up’ status while the vice versa would be about ‘defaulted’ status.

Sex

Fast forward to independent variables or features involved in the study, the first categorical variable that was recoded was sex of an individual. Since there are no people with transgender sex, the only recorded sex categories were male and female.

Because of data collection and processing issues at CRB, there were other anomaly values under the ‘sex’ variable namely ‘Mr.’,’Mrs.’ and ‘Ms.’ which are basically salutations. Since it is obvious for a Mr. to be a man and Mrs. and Ms. to be a woman, all the values with the former category were recoded into males and the latter two salutations recoded into females.

The recoding of the sex categories was encoded as 0 and 1 for female and male subjects respectively.

Marital status

Upon inspecting the distinguished categories of the ‘marital status’ feature, the observed groups were either single, married, widowed, and divorced. The variable had exactly four categories which had anomalies such that no questions were raised for decision making of the kind of numbers to be assigned to each category.

Numbers 0,1,2,3 were assigned to the categories in the order of single, married, widowed and divorced. Although the numbers are in increasing order, the variable marital status is nominal not ordinal i.e., the categories are not ranked. In other words, there is no category which is higher or lower than the other.

District

Malawi has 28 distinctive districts and based on that, the ‘district’ feature was encoded from 1 to 28 for each respective district in no particular order. No districts were combined for one code to reduce the categories because of the distinctiveness. The recoding was as follows:

District	Code
Blantyre	1
Balaka	2
Chikwawa	3
Chiradzulu	4
Chitipa	5
Dedza	6
Karonga	7
Kasungu	8
Lilongwe	9
Likoma	10
Machinga	11
Mangochi	12
Mchinji	13
Mtchisi	14
Mulanje	15
Mwanza	16
Mzimba	17
Neno	18
Nkhatabay	19
Nkhotakota	20
Nsanje	21
Ntcheu	22
Phalombe	23
Rumphi	24
Salima	25
Thyolo	26
Zomba	27
Dowa	28

Payment terms

Payment terms were basically about the agreed terms of an obligor paying back the loan to the creditor. The interval between the last and the next payment differs with different obligors and creditors. Thus making the variable 'payment terms' worthy of having an impact on credit risk by contributing to prediction of probability of default and credit score.

The categories under 'payment terms' were: installments payment, once off payment, Salary reduction, on demand payment, daily payment, weekly payment, monthly payment, quarterly payment, half-yearly payment, annual payment, indefinite overdraft and irregular schedule. Installments payment was about an obligor agreeing with the creditor to be paying the loan little by little while once off is one-time repayment. Salary reduction was a repayment method of deducting a proportion of salary of an obligor when it gets credited while on demand payment was based on the need of the money by the creditor. Daily,

weekly, monthly, quarterly, half-yearly, annual payments were basically a breakdown of the payments with installments. As the names speak for themselves, they were about the specified timeframes in between the last and next payment of the loan by the associated party.

The correspondent codes assigned were 0 to 12 respectively. Although the numbers are ordered, the feature 'payment terms' was nominal not ordinal.

Source of income

This was about what the means of earning a living by the obligor. The data was also recorded in categories, thus making it a categorical variable. The categories were observed as business, farming, white collar job and blue collar job.

To clarify the differences between the two categories of employment, a white-collar position is one in which a professional works in an office setting and carries out specialized tasks that call for training or education. On the other hand, blue collar jobs are basically jobs executed outside the office setting. Most blue collar jobs involve trade or manual labor.

Since they frequently work on salaries rather than receiving hourly wages, which is typical of many blue-collar jobs, white-collar employees frequently earn more than their blue-collar counterparts. Thus being a source of income, being an interesting feature for credit risk prediction since business men, white collar and blue collar job workers have observable distinctions with regards to income earned and the pattern of earning it.

As three distinct categories under source of income, business, farming, white collar job and blue collar jobs were encoded as 1, 2,3,4 respectively.

Collateral type

Types of collateral or surety of the loan were also recorded in categories as well. The observable categories in the dataset used in this study were cash, shares and property. They were encoded as 0,1 and 2 respectively.

Interest type

Based on the fact that loans or insurance policies always have interest rates involved, it was definitely a very good idea to include the feature 'interest type' for credit risk prediction.

There were basically four types of interest listed under the feature, namely: fixed, floating, reducing balance and straight. Fixed and floating interest types are static and volatile with time respectively.

However, the interest method is when one calculates the interest on the diminishing principal amount. Every month when one pays installments, the principal amount reduces. When the interest is calculated on this new amount, at the time of installments payment, it means one is using the reducing balance interest method. While straight or flat interest type works when the interest rate is calculated on the full amount of the loan. It means that the rate of interest is calculated on the principal amount of the loan and will remain the same throughout the tenure.

The encoding was 1,2,3,4 for fixed, floating, reducing balance and straight interest types correspondingly.

Participation code

Either a loan being taken as an individual or a group, were the categories under feature participation code. Some loans are given to distinguished persons while some are rendered to a number of people. These two categories accounted for that hence being vital to prediction of credit risk in the aspect of probability of default and credit score.

The variable was a nominal one and the encoding of 0 and 1 to single and group loan respectively as the categories under the feature participation was readily implemented.

Liability type

There were basically seven categories of liabilities collected and observed under 'liability type'. The types of liabilities were: general loan, business loan, government loan, student or education loan, overdraft loan, mortgage loans and utilities loan.

General loans were a combination of any unspecified and generalized loans. For example, emergency loans or just simply ‘loans’. Government loans, as the name suggests, were basically liabilities where one was obliged to pay money for activities related governmental or citizenry tasks while a business loan was about individuals who got a loan from either a bank, MFI or SACCO. Student loan was about students who got loans for school tuition fees, accommodation or upkeep. On the other hand, mortgages were about subjects who took a lot by pledging their property. (Kagan, 2022) defines a mortgage as a particular kind of loan used to buy or maintain a home, piece of land, or other piece of real estate. The borrower agrees to make periodic payments to the lender, usually in the form of a series of regular installments that are split into principal and interest. The property then acts as security for the loan. However, overdrafts were liabilities of one withdrawing more money than he or she owns in a bank account while utility loans were basically about one defaulting on postpaid house or business bills.

All the liabilities were encoded from 1 to 7 in the general loan, business loan, government loan, student or education loan, overdraft loan, mortgage loans and utilities loan.

- Outliers

Under this section, outliers were monitored and analyzed specifically for numerical variables in the dataset. This was based on the fact that most machine learning algorithms do not work well in the presence of outliers. So it was desirable to detect and remove outliers.

The prominent numerical variables in the data set of interest were age, number of dependents and principal amount. The rest of them were categorical or binary such that analyzing their respective outliers was not that much recommended or of substantial importance.

The cleaning of the outliers was basically in phases. The first phase was about plotting the distribution of the three features pairwise so that the outliers would be vividly seen and monitored. The second phase, subsequent to that, a decision was made on how to handle the outliers.

Phase 1

The three numerical features were plotted pairwise using a sea-born library in python as previously outlined. Figure 4.3 clearly showcases the distribution where an analysis was made to decide if outliers were available and prevalent.

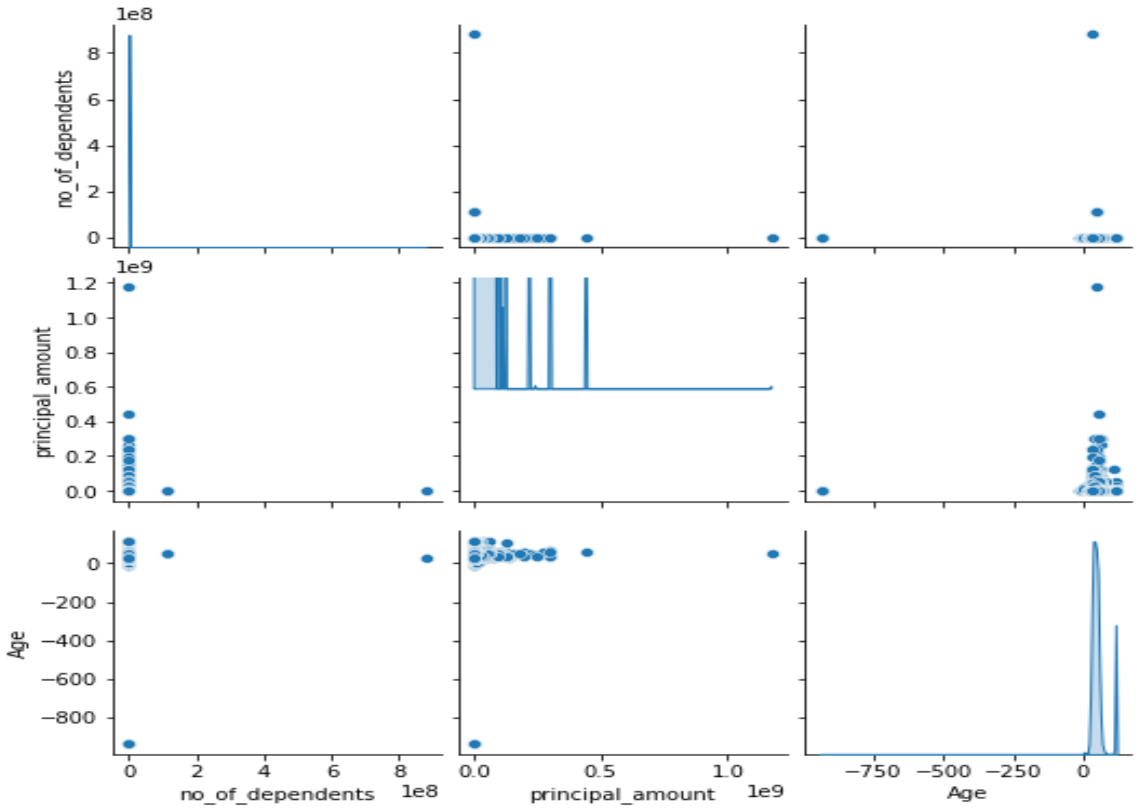


Figure 4. 3: illustration of the distribution of numerous features

Apparently, according to figure 4.3, all the 3 variables had a number of outliers. The plot shows that most of the data for variable 'number of dependents' were distributed between 0 and 2. For any data point outside that, both technically and theoretically, were outliers. The data point above 8, which represents a number of values in the variable, was obviously an outlier.

A similar observation was observed in the feature 'principal amount' of which the numbers were recorded and measured in millions. The majority of principal amounts for the distinctive subjects were clustered between 0 and 0.5 million such that any amount outside that was considered an outlier.

On the other hand, age also had vivid outliers. There were age points observed below 0 which is an inconsistency. This definitely was calling for attention and cleaning so that age should be above 0.

Phase 2

Reflecting on what was observed in phase 1, a decision to drop all rows associated with the outliers was made. Upon executing that, the total number of rows left were 640,149 with a 23.8% decrease from 839,932.

Using the cleaned data, free of outliers, a similar plot to figure 4.3 was made to showcase the impact. Figure 4.4 illustrates the development.

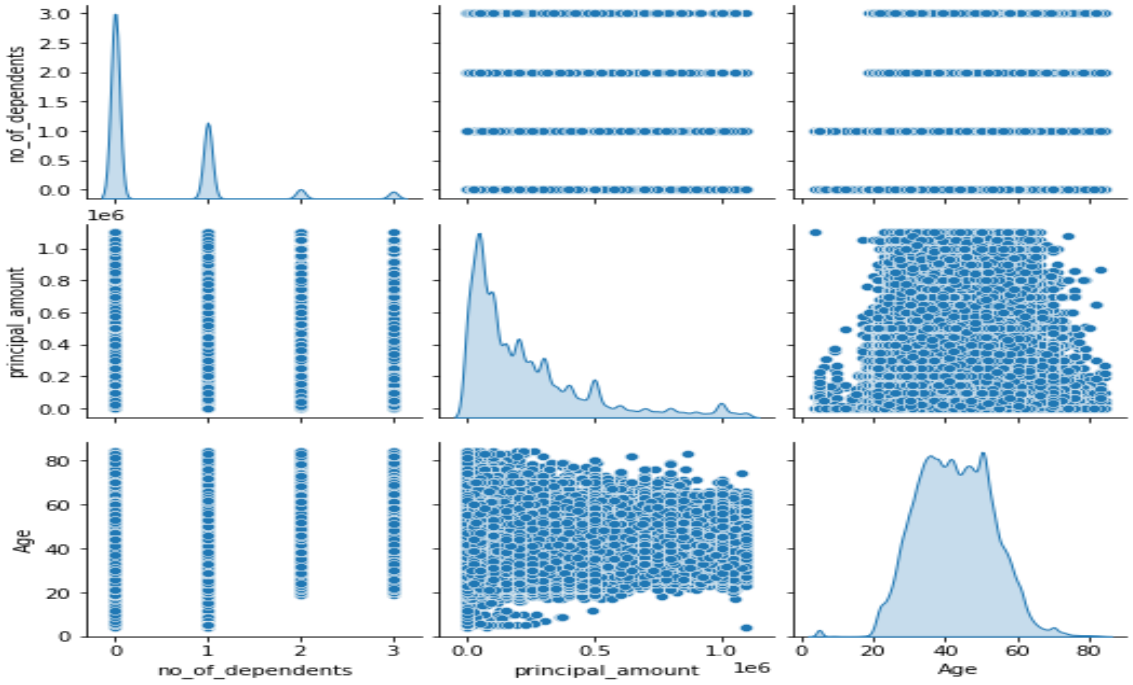


Figure 4. 4: Illustration of the numerical features after removing outliers

As desired, all the three features were cleaned off outliers. Variable 'age' no longer had negative ages and both principal amount and number of dependents were distributed with a regular shape clustered with a reasonable distance.

4.3 Data transformation

After cleaning and preprocessing the data, the next step as per the Knowledge Discovery in Databases process, was data transformation. A number of changes to the structure of the data had to be made in preparation for efficient and well performing machine learning models.

The distinguished data transformation tasks performed in this project were; new variable creation, normalization, handling class imbalance and one hot encoding. All the other tasks, after the new variable creation, were performed in no particular order.

Variable creation

Based on two columns 'date of birth' and 'loan date', a feature 'age' was created. The feature was about the age of an individual at the time the loan was given to him or her. The transformation was executed successfully by subtracting the loan date from date of birth and get the corresponding year for the value of interest.

Another variable that was created from two different features was loan duration. Since the difference between the clearing date of a loan and the date the loan record was created is the time required for the loan to be fully paid back, an algorithm to compute the duration of a loan was based on that. Unlike the other numerical features, 'loan duration', rounded up to years, had no observable outliers. The values ranged from 0 to 6, which were reasonably and almost evenly distributed

This one was the first step of data transformation as date of birth, clearing and loan date, recorded with timestamp data types, were transformed to integers containing years of age subject and the allowable years of the loan.

Normalization

There are a number of normalization methods but the one that was opted for in this study was min-max normalization where, for every feature, the minimum value got transformed to 0, the maximum to 1 and the values in between to a decimal number between 0 and 1. This was only performed to the continuous variables which had different scales and measurements of values. It was necessary because the difference in the scales lead to bias

in the prediction of the target variable since other features definitely would have bigger numbers than the other.

The numerical features, age, no of dependents and principal amount were normalized with min-max normalization algorithm as desired such that the maximum values in each feature were 1s and minimums were 0s and the rest in between.

Handling class imbalance

Since class imbalance is usually an issue when modeling using machine learning classifiers, the number of classes in the target variable, credit status, was monitored. It was observed that there was a class imbalance between the statuses 'paid up' and 'defaulted', where the former had 488,481 and the latter 151,668. That literally meant that the model would perform well for predicting more 'paid up' statuses than 'defaulted' statuses. That kind of biased prediction would only benefit a prospective obligor, but the creditor would be at more risk of borrowing money from someone who never really is likely to pay back a loan.

Based on that, a class imbalance technique of oversampling the minority class was opted for and implemented. Synthetic Minority Oversampling Technique (SMOTE), a specific model, was applied in this case. SMOTE functions by randomly selecting a point from the minority class and calculating its k-nearest neighbors. Between the selected point and its neighbors, the artificial points are added. A minority class is chosen as the input vector for the algorithm, and then one of its k closest neighbors is selected. A synthetic point is then placed anywhere along the line connecting the point under consideration and its selected neighbor. Until the data is balanced, these steps are repeated (Shah, 2022).

An over-sampling technique was preferred to an under-sampling technique because under sampling would imply losing data while oversampling yields more data. More data was required for the models to fit well with no bias thus avoiding any more tasks of decreasing the subjects.

The 640,149 rows of data were fit into the SMOTE model and the minority class 'defaulted' was over sampled from 151,668 to 488,481, leveling up with class 'paid up'. Because of this, the total number of subjects in the dataset increased from 640,149 to 976,962.

One hot encoding

With the interest of enhancing performance of the machine learning models to be fitted, one hot encoding for all the categorical features was performed. Technically, it was about encoding categorical values using a one-hot scheme. One hot encoding is a process that transforms categorical data variables so they can be supplied to machine learning algorithms to improve predictions, making it a crucial component of feature engineering in the field.

The categorical features which were one-hot encoded were: 'sex', 'marital status', 'district' 'payment terms', 'source of income', 'collateral type', 'interest type', 'interest calculation method', 'participation code' and 'liability type'.

4.4 Clean data

Having preprocessed and transformed the raw data as outlined, expounded and illustrated, the frequencies of classes of each categorical variable including the target, was as table 4.1 indicates.

Table 4. 1: Descriptive statistics of the target variable and features

Variable	**Frequency(Percentage)**
Credit status	
Paid up	488,481 (50%)
Defaulted	488,481 (50%)
Sex	
Female	450,152 (46.08%)
Male	526,810 (53.92%)
Marital	
Single	232,836 (23.83%)
Married	721,317 (73.83%)
Divorced	15,998 (1.64%)
Widowed	6,802 (0.70%)
District	
Blantyre	403,473 (43.39%)

Balaka	73,348 (7.89%)
Chikwawa	32,006 (3.44%)
Chiradzulu	30,583 (3.29%)
Chitipa	5,122 (0.55%)
Dedza	2,272 (0.24%)
Karonga	14,586 (1.57%)
Kasungu	4,718 (0.51%)
Lilongwe	17,635 (1.90%)
Dowa	100,871 (10.85%)
Likoma	12,355 (1.33%)
Machinga	4,338 (0.47%)
Mangochi	20,163 (2.17%)
Mchinji	15,747 (1.69%)
Mtchisi	21,140 (2.27%)
Mulanje	292 (0.03%)
Mwanza	25,892 (2.78%)
Mzimba	10,140 (1.09%)
Neno	34,211 (3.68%)
Nkhatabay	1,517 (0.16%)
Nkhotakota	5,223 (0.56%)
Nsanje	7,745 (0.83%)
Ntcheu	17,982 (1.93%)
Phalombe	16,028 (1.72%)
Rumphi	18,319 (1.97%)
Salima	2,231 (0.24%)
Thyolo	12,717 (1.37%)
Zomba	19,223 (2.07%)
Liability type	
General loan	457,567 (46.84%)
Business loans	38,366 (3.93%)
Government loans	56,181 (5.75%)
Student/education loans	304,064 (31.12%)
Overdraft	108,627 (11.12%)
Mortgage	9,343 (0.96%)
Utilities	2,814 (0.29%)
Interest type	
Fixed	274,822 (28.13%)
Floating	303,827 (31.10%)
Reducing balance	232,170 (23.76%)
Straight	166,143 (17.01%

Payment terms	
Installments	3,321 (0.35%)
Salary deduction	12,686 (1.33%)
On demand	97,046 (10.16%)
Daily	205,468 (21.51%)
Weekly	667 (0.07%)
Fortnightly	33 (0.003%)
Monthly	636,185 (66.59%)
Quarterly	161 (0.02%)
Half-yearly	21 (0.0001%)
Annually	7 (0.000005%)
Bullet	1,590 (0.17%)
Revolving	7,868 (0.82%)
Irregular Schedule	11,909 (1.25%)
Source of income	
Business	409,334 (41.90%)
Farming	184,519(18.89%)
Blue collar job	298,311 (30.53%)
White collar job	84,798 (8.68%)
Collateral type	
House	392,372 (40.16%)
Land/Estate	194,948 (19.95%)
Business	194,555 (19.91%)
Other	195,087 (19.97%)
Participation code	
Single	517,699 (52.99%)
Group	459,263 (47.01%)

4.5 Models

This was the stage where data mining was performed. A number of classifier machine learning models were fitted and learned with the preprocessed, cleaned and transformed data. As it has already been outlined in methodology, the models were namely; Logistic Regression, Naïve Bayes, KNearest Neighbor, Decision Tree, Random forest and Support Vector Machine.

The 976,962 rows of data with 12 variables were split into training, validation and testing sets for the machine learning models. The target variable was first detached from the features so that they would be split separately. Subsequent to that, both the features and target variable data frames were split into training and testing sets in proportions of 0.9 and 0.1 respectively. The testing data, with a pair of features and the dependent binary variable, that was observed here, was the final data that was used to examine the true performance of the models at deployment level.

The target variable was credit status of a subject. It was a binary variable with classes ‘paid up’ and ‘defaulted’ encoded as 0 and 1 respectively. The features were a combination of categorical and numerical datatypes namely; 'sex', 'marital status', ’district’ 'payment terms', 'source of income', 'collateral type', 'interest type', ‘interest calculation method’, 'participation code' and ‘liability type’, ’age’, ’no of dependents’, ’principal amount’ and ‘loan duration’.

The 0.9 proportion of the training set was then subsequently split into a final training set and a validation set. The proportions were 0.75 and 0.25 out of the 0.9 for the former and latter respectively. The training set was used to train and learn the models while the validation set was used to assess the performance of the models for adjustments and modifications towards accuracy and better learning.

All this splitting was doable because the data was big enough, such that K-Fold validation algorithm (algorithm used as a tool to evaluate machine learning models, by training a number of models on different subsets of the input data whenever number of subjects are limited) was uncalled for.

Using these three sets of data were used to model the classifiers and the results of the modeling were grouped as probabilistic and non-probabilistic based on the classes of the models as follows:

i. Probabilistic models

These are models which yielded a probability distribution over a set of classes for each input of the features. These models were basically Logistic Regression and Naïve Bayes.

1. Logistic Regression

This was the first model to learn, and it was fitted several times, such that its performance was monitored each time with the performance metrics; accuracy score, confusion matrix, precision, recall and ROC curve as outlined in methodology.

Along the way of tuning the three parameters (solvers, maximum number of iterations and number of features) of the model for a good fit, the values of solvers, maximum number of iterations and number of features which were settled on were 'saga', 150 and 9 respectively. 'saga' is basically a solver which is usually fast for large datasets as the one used in this project. The other values literally meant that the maximum number of iterations required for the model to fit well was 150 while the number of features is 9.

Along the way of tuning the three parameters for a good fit of the logistic regression model, the values of solvers, maximum number of iterations and number of features which were settled on were 'saga', 150 and 9 respectively. 'saga' is basically a solver which is usually fast for large datasets as the one used in this project. The other values literally meant that the maximum number of iterations required for the model to fit well was 150 while the number of features is 9.

At this point, reflecting on accuracy score of 0.73 and 0.7 for training and testing sets respectively and a confusion matrix with a total of 423,901 true positives and negatives, the model was declared to have a better fit as compared to other metrics when the parameters were set otherwise.

When the model was set with other parameters like maximum number of iterations greater than 150, it over-fitted yielding a greater training accuracy score than that of the testing set. The vice versa, underfitting, occurred when the parameters were set otherwise. So although, the model with the 'saga' solver and 150 maximum number of iterations metrics was determined to be the best fit of the logistic regression model, it was not convincing and satisfying enough for credit risk prediction.

The roc curve for the model was also plotted to monitor its performance using sensitivity and specificity. Figure 4.5 shows the ROC curve of the logistic regression model with parameters solver and maximum number of iterations set to 'saga' and 150 respectively.

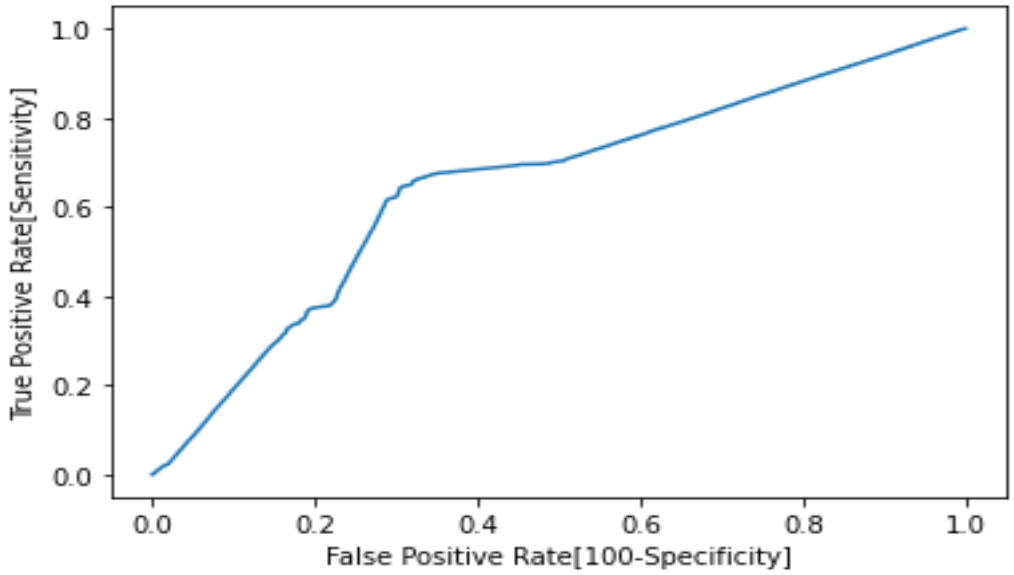


Figure 4. 5: ROC curve of the best logistic regression model

The ROC curve literally backs up the idea that the best logistic regression model did not fit well to the data for credit status prediction. Technically, using this model to predict the probability of defaulting would be unreliable and misleading. This is because the area under the curve is minimal due to a higher rate of false positive rate as compared to True positive rate. In other words, the specificity and sensitivity rates were relatively small such that the sharp curve was far away from the top left corner, thus leading to a small area under the curve. The target is to maximize the area under the ROC curve for a better fit to the data.

2. Naïve Bayes

Just like in Logistic regression, the model was trained with the training set and its performance was evaluated using the validation set. Performance measurement metrics were computed and evaluated. The accuracy score for the training and testing sets were 0.66 and 0.79 respectively. On the other hand, the precision scores were 0.67 and 0.63 for training and testing sets respectively and recall scores were 0.65 for both.

An ROC curve for the training set was also plotted. Just like in the preceding model, the area under the curve 0.65 was relatively small for the model to be considered good. A combined sum of 223,123 for false negatives and false positives was substantially big while

the training and testing accuracy scores were small although they were relatively the same, indicating no over-fitting and under-fitting.

ii. Non-probabilistic models

Also known as deterministic models, these models never modeled the probability distribution of classes but rather modeled credit status by separating the features space and returning the class associated with the space where a sample originated from. These models were namely; KNearest Neighbor, Decision Tree, Random Forest Support Vector Machine.

1. KNearest Neighbor

Fitting the models started with small values of k, going up high on intervals of 1. Fifteen KNN models were fit with a value of nearest neighbors starting from 5 to 20. For all values of k between 5 and 14, the values of accuracy scores of the training set were technically less than those of the testing set. This literally means that the models were underfitting to the data since it was performing better, by far, on the testing set. However, fitting with k values above 15, the models seemed to be overfitting. The accuracy scores of the training sets turned out greater than those of the testing sets. This was not good because it leads to generalization error, thus making the model not good enough.

All in all, a good fit was determined on 15 as a value of nearest neighbors. At this value, the model yielded an accuracy score of 0.86 for the training set and 0.84 for the testing set. The confusion matrices were as follows:

$$\begin{bmatrix} 270182 & 59421 \\ 34626 & 295219 \end{bmatrix} \qquad \begin{bmatrix} 88105 & 21709 \\ 13383 & 96620 \end{bmatrix}$$

CM for training set *CM for testing set*

The counts of true positives and true negatives ,270182 and 295219 respectively for the training set were big enough proportionally as compared to the false positives and negatives, 34626 and 59421 respectively. The interpretation applies to the testing set as it has vividly been shown. Basically, the idea was to have less number of subjects predicted wrongly as compared to the ones predicted right. In this case, that goal was achieved almost the wrong predictions were also part of it. This model would not be considered if another one was found with far less wrong predictions and higher accuracy scores. However,

arguably, this was considered to be a good model for prediction probability of defaulting for credit risk modelling.

On top of this, to cement the goodness of fit of this model an ROC curve was also fitted. Figure 4.6 shows the distribution of the true positives and negative rates.

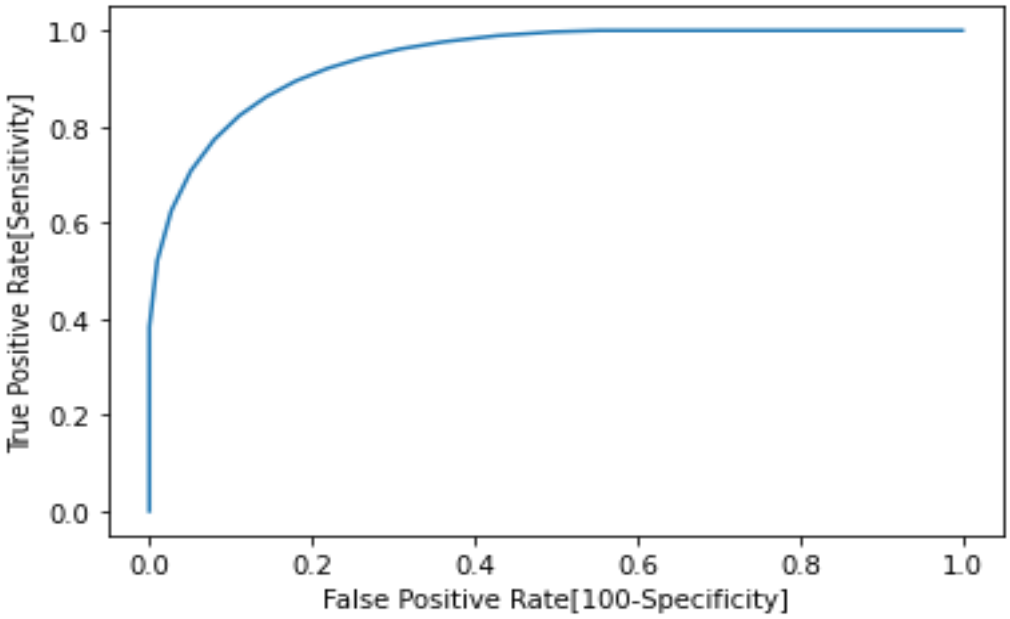


Figure 4. 6: ROC curve of KNN model with k=15

Although the curve did not have a sharp curve towards the top left corner, the curve was more curved than the preceding two models. The area under the curve computed was 0.94 which is closer to the maximum 1.

2. Decision Tree

This model was fit numerous times while adjusting depth of the tree. Initially, the model was fit without specifying any parameters and performed poorly, with an accuracy score of only 0.67 for both the training and testing sets. After finding a good fit, the value of the maximum depth of the model was determined to be 10, which produced accuracy scores of 0.96 for each dataset.

The Area under the Receiver Operating Characteristic (ROC) curve was approximately 0.99, which is very close to 1. Based on these metrics, it was concluded that the model was good enough for credit risk prediction.

3. Random Forest

This model, besides the Decision Tree Classifier, was arguably a best fit to the data .For all depths of the forests below 11, the model was underfitting with accuracy scores of the training sets being less than those of the testing set. While on the other hand, with a depth above 12, the model would be overfit with testing sets models having accuracy score less than that of training sets. A better model, with balanced and high accuracy scores was settled at the depth of the forest being set to 12. The accuracy score, recall score and precision score for the training set were 0.96,0.95 and 0.97 respectively. On the other hand, the accuracy score, recall score and precision score for the testing set were 0.96,0.96 and 0.94 respectively.

To shed more light on how the two different classes were correctly and wrongly predicted, confusion matrices were computed. The accuracy and trustworthiness of the model was concreted by Receiver Operating Characteristic curve shown by figure 4.7. The Area Under the Curve(AUC) was computed and recorded as 0.99 which is as good as 1.

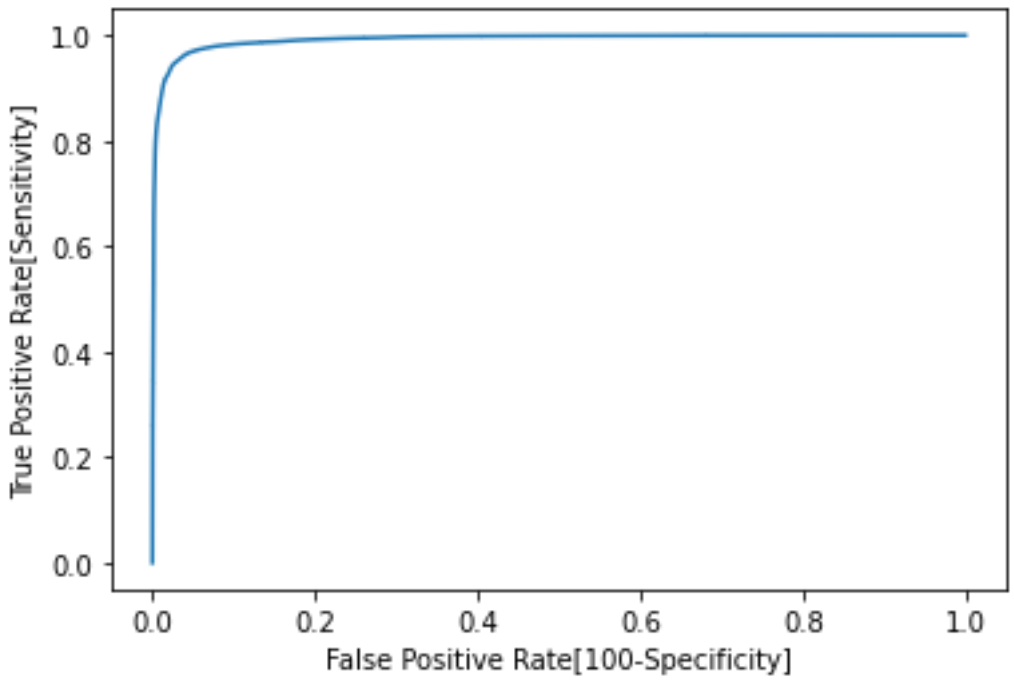


Figure 4. 7: ROC curve of Random Forest classifier with depth 12

Vividly, this is an ROC curve of the model which has fitted well to its input data. The sharp curve intimated to the top left corner indicates the truthfulness of high rate of sensitivity as compared to 100 minus specificity rate. Obviously, the rate at which defaulted credit statuses were being correctly predicted was exponentially higher than that of predicting paid up status as defaulted.

Immediately after 0.0, false positive rate negligibly increased, arguably static, while the sensitivity rate explosively hiked up. The moment the true positive rate almost hit 1, it asymptotically maintained until the false positive rate almost got to 1. That is exactly what is required of a model which has fit well to data for prediction of a binary target variable.

3. Support Vector Machine

The model only showed to have performed well was determined to be an SVM model with kernel function RBF. Although the measurement metrics of this model's performance were recorded low, relatively, as compared to the models set with the other functions, they were high. For example, the accuracy scores of training and testing sets for the RBF model turned out 0.78 and 0.71 while those of linear were 0.65 for both sets.

The area under the curve for the SVM model with kernel function RBF was 0.77 while the recall and precision scores for the training data set were 0.64 and 0.66 respectively. Although, this was the best SVM model fitted to the data, arguably and reasonably, it was not a good fit to the data.

4.6 Knowledge discovery (Results)

Subsequent to fitting and training the six machine learning classification models, comparisons had to be made carefully to select the best model. The metrics used, on top of confusion matrices, were accuracy scores, precision scores, recall scores and area under ROC curves for both training and testing datasets.

Reflecting on the values of the metrics, performance of each model was evaluated and the best model was found. This model was the one that was used for predicting probability of default and credit score using the final testing dataset which was completely new to the model.

4.6.1 Comparison of models performances

Tables 4.2 and 4.3 show how each model performed in the aspects of accuracy scores, precision scores, recall scores and area under ROC curves for training and testing datasets respectively.

Table 4. 2: *Performance evaluation of each model based on training dataset*

Model	**Accuracy score**	**Precision score**	**Recall score**	**Area under the ROC curve**
Logistic Regression	0.73	0.61	0.01	0.77
Naïve Bayes	0.68	0.7	0.63	0.74
KNN	0.86	0.83	0.9	0.94
Decision Tree	0.95	0.96	0.95	0.99
Random Forest	0.96	0.97	0.95	0.99
SVM	0.81	0.68	0.6	0.83

Table 4. 3: Performance evaluation of each model based on testing dataset

Model	**Accuracy score**	**Precision score**	**Recall score**	**Area under the ROC curve**
Logistic Regression	0.7	0.6	0.01	0.73
Naïve Bayes	0.69	0.74	0.63	0.76
KNN	0.84	0.82	0.88	0.91
Decision Tree	0.96	0.95	0.94	0.99
Random Forest	0.97	0.97	0.95	0.99
SVM	0.8	0.68	0.6	0.84

Clearly, random forest and decision tree classifier models were standouts in all the aspects followed by KNearest Neighbors and then Support Vector Machine. The other two models;

Naïve Bayes and Logistic Regression model were the worst with accuracy scores less than 75% and AUC less than 80%. The top performing models, random forest and decision tree had accuracy scores greater than 95% with Area Under the curve of 0.99, almost 1, each. Decision tree model had 0.95 and 0.96 for accuracy scores of training and testing sets respectively. On the other hand, random forest classifiers had 0.96 and 0.97 accuracy scores for training and testing sets respectively. Both the former and the latter had an area under the curve 0.99 apiece. Both training and testing precision scores for random forest classifier model were slightly greater than those of the decision tree model. Only the testing recall score of the decision tree model was found to be 0.01 less than that of the random forest.

For careful comparison of the two models and reasonable choice of the best model, confusion matrices for both training and testing sets fitted to both models were incorporated. Since the performance based on the other metrics which have already been outlined were too close, analysis of the total number of true positives, true negatives, false negatives and false positives had to be brought into the picture.

The following were the respective confusion matrices for each of the two models, classified by the data set used.

a. Decision tree confusion matrices

$$\begin{bmatrix} 317197 & 12406 \\ 12599 & 317246 \end{bmatrix} \qquad \begin{bmatrix} 105512 & 4302 \\ 4447 & 105556 \end{bmatrix}$$

CM for training set *CM for testing set*

b. Random forest confusion matrices

$$\begin{bmatrix} 319420 & 10183 \\ 15914 & 313931 \end{bmatrix} \qquad \begin{bmatrix} 106279 & 3535 \\ 5499 & 104504 \end{bmatrix}$$

CM for training set *CM for testing set*

The decision tree model had 12406 false negatives and 12599 false positives while random forest had 10183 false negatives and 15914 false positives both for training dataset. This literally means that the former had less false positives than the latter while the latter had less false negatives than the former. The scenario was the same with the testing data set. The decision tree model was placed at 4302 for false negatives and 4447 for false positives

while random forest produced 3535 and 5499 false negatives and false positives respectively.

The true positives for the decision tree model had 2223 less true positives than random forest but 3315 more true negatives than random forest. The numbers of true positives and true negatives for the decision tree model were 317197 and 317246 respectively while true positives and true negatives for the decision tree model were 318729 and 313931 respectively.

Since high number of false negatives imply high number of subjects who really defaulted on their loan but were predicted with a 'paid up' status and high number of false positives imply high number of subjects who really paid back all their loan but were wrongly classified as 'defaulted', careful scrutiny had to be made. Based on the fact that both models were close, in terms of performance, the model with less number of false positives had to be chosen over the other to avoid defaming subjects who never defaulted on a loan.

People who have never defaulted would fail to get a loan from a bank or a microfinance company simply because they have been reported wrongly as if they did. Although the trade would be made with false negatives, where a subject who defaulted would be given a loan due to a report which would indicate a clean record, the risk is more severe on the former. Economy or business of the crediting institution would not get affected that much with high numbers of false negatives. But however, an individual with a clean record who would like to get some loan for progressive use, would be much affected if he or she is rejected of a loan before a report is wrong about him about his last defaulted loan.

Based on this reasoning, the decision tree classifier model was chosen over the random forest model. It had a lesser number of false positives and greater number of false negatives as compared to the other. With the decision tree model with depth 10, there is high confidence of predicting more subjects who do not default on their loans, unlike the random forest classifier model.

4.6.2 Prediction of probability of default

The decision tree model with the specific depth of 10 was used to predict credit risk using the final testing data that was initially separated from the two datasets which were used to train and tune the model based on performances monitored.

The predictions literally had either 0 or 1 values for 'paid up' and 'defaulted' statuses. To get the probability of default for each correspondent prediction, a command for specifically computing probability associated to each value of the binary target variable was implemented. All 1s for 'defaulted' status turned out with probabilities greater than 0.5 while the 0s turned out with probabilities less than 0.5. This means that all subjects who were predicted 'defaulted' had a probability of default above 0.5, while the 'paid up' statuses less than 0.5.

With these results, the probabilities were considered as credit risk, specifically as probabilities of default such that for a new data, with all the features involved in this research provided, the correspondent credit risk can be computed that way.

4.6.3 Credit scores

Credit risk was modeled as credit scores based on the probabilities of default. The analogy was simply based on equating the least credit score to the highest probability and the highest credit score to the least probability. This is because a high credit score implies the subject is less likely to default on a loan such that giving him the loan would be less risky as compared to one with a much less score.

It is of common knowledge that the least and highest probabilities are 0 and 1 respectively. However, there are different agencies of credit scores. For example, TransUnion credit scores range from 300 to 850 while Experian ranges from 330 to 830. On the other hand, FICO ranges from 300 to 850 while VantageScore ranges from 501 to 990. Appendix A vividly lists the ranges for each type.

Out of all these different score ranges, Fair Isaac Corporation(FICO) was the preference. This means that for all probabilities of default 0, the credit scores were the maximum 850 while probabilities 1 were the minimum credit scores 300. For probability 0.5, the credit

score was the half of the subtract of 300 from 850 added to 300 which is 575. For all the probabilities between 0 and 0.5, the credit scores were between 575 and 850 while for probabilities between 0.5 and 1, the scores ranged from 300 to 575.

Further that, the results of the credit scores were categorized as very poor, fair, good, very good and excellent based on FICO ranges. The categories were classified based on ranges 300 to 579, 580 to 669, 670 to 739, 740 to 799 and 800 to 850 respectively. This means that all subjects with credit score less than 580 were categorized as poor and those with credit score above 800 were classified as excellent. Refer to Appendix B for a visualization of this categorization of credit scores. On a credit score report, this is what would appear besides the truthful figure.

4.7 Discussion

The Probability of Default (PD) is one of the important quantities to quantify credit risk. Together with Loss Given Default (LGD), the PD will lead to Expected Loss (EL) calculation. These quantities enable the estimation of regulatory capital and risk-weighted assets, which are required by financial regulators.

The methodologies for PD estimation are quite well established. According to (Chee, 2022), they generally involve grouping the obligors into rating grades and estimating the likelihood of the specific rating going into default. Making it more complex, the PD estimates are extended to multiple years instead of one year. In addition, rather than estimating defaults from each rating, transition matrices of probabilities are computed, representing each rating's probability to transition to other ratings, inclusive of the probability of remaining in the same rating. These models have served the banks and regulators well and can nonetheless be widely used these days.

The key difference between the findings of this study and others, is that probabilities were predicted at obligor level. In traditional methodologies, probabilities are determined by groups only which is arguably questionable and unreliable. The analysis of subsequent analysis diverges from traditional methods after having the probabilities at obligor level, because with these probabilities, all sorts of possibilities get opened up which are not based on assumptions and intuition.

With Machine Learning coming into the picture, the tendency of one avoiding predicting and evaluating credit risk based on scores of groups associated with a number of assumptions which are subject to substantial errors, should be of high priority. Credit scores and PDs predicted by the Decision Tree Classifier, are solely dependent on the values of the features of each subject independently. This was definitely a more meaningful and trustable way to go with, unlike predicting credit risk based on a whole bunch of subjects' risk.

On the other hand, FICO credit scores, traditionally, are calculate based on weights of factors namely; a person's payment history, amounts owed, length of credit history, new credit accounts and types of credit used with contributory proportions of 35%,30%,15%,10% and 10% respectively (Arya et al., 2013). These are not reliable because the proportions are based on assumptions and empirical evidence unlike the credit scores evaluated in this study by using a machine learning model. The Decision Tree automatically predicts the credit scores based on probabilities of defaults without explicitly defining weights.

The performance of the Machine learning model which has been settled for in this study has been evaluated with higher metrics which makes it more trustworthy. Each time a prediction is made, one would be about 96% sure of the accuracy of the credit risk prediction. Further that, the model handles processing big data as per the anatomy of machine learning models, thus making the results of this study more relevant.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1 Results summarization

To achieve the main objective of the study, where the task was to model credit risk using machine learning in Malawi, different classifier models were fitted and trained where the targeted variable was credit status. This major objective was achieved by modeling credit risk with a robust and highly performing decision tree classifier model with maximum depth of 10 with an average accuracy score of 95.5% and Area under the ROC Curve of 0.99. The total number of false negatives and false positives as per the training set was 12406 and 12599 respectively. These measurements of performance of the model were good enough for the main objective to be declared achievable.

As per the specific objectives of modeling credit risk as probability of default and credit score, the decision tree model was used to predict probabilities associated with the credit statuses predicted. These were basically probabilities of default where values in between 0 and 0.5 indicated a low chance of a subject to default on a loan. The vice versa, as in the chance of a subject to default on a loan were given by probabilities between 0.5 and 1. As for the credit scores, credit score ranges by FICO were used to compute the correspondent credit scores proportional to the predicted probability of default.

An objective to come up with the best model from a number of classifiers based on performance and generalization to new data was also achieved. Six distinguished models; KNearest Neighbor, Decision Tree, Random Forest, Support Vector Machine, Naïve Bayes and Logistic Regression classifiers were fitted and learned to the same data such that only two, random forest and decision tree, stood out. Decision tree was preferred to the random forest because of lower rates of false negatives and false positives.

The relevant features towards prediction of credit risk as per the best model were exactly the initial input, without adding or removing any. The features were The features were a combination of categorical and numerical datatypes namely; 'sex', 'marital status', ’district’ 'payment terms', 'source of income', 'collateral type', 'interest type', ‘interest calculation method’, 'participation code' and ‘liability type’, ’age’, ’no of dependents’, ’principal

amount' and 'loan duration'. These were a combination of numerical and categorical data types.

All these results based on the objectives were achievable using a data mining strategy called the Knowledge Discovery in Databases process. It involved understanding the problem, data collection, data preprocessing, transformation, modeling and knowledge discovery. The process was very instrumental to the accuracy and reliability of the results.

5.2 Research and policy recommendations

Although this study was aimed at solving the problem of trustworthy credit risk determination using classification supervised learning models, there is more to this as far as machine learning is concerned.

As it has already been outlined in chapter 1, where different types of machine learning algorithms were illustrated and explained, unsupervised models for predicting either a categorical or numerical target variable can also be implemented. Using clustering algorithms to cluster subjects into distinguished ranges of credit scores as per Appendix B is very much implementable. This approach would not require labels of the credit statuses as they are not required in unsupervised learning.

Based on that, it is recommended of other researchers to model credit risk in Malawi using unsupervised learning. K-means clustering, DBSCAN clustering, Gaussian Mixture Model, BIRCH, Affinity Propagation clustering, Mean-Shift clustering algorithm, OPTICS and Agglomerative Hierarchy clustering algorithms are some of the prominent clustering models which other researchers can implement to take in from here. That way, the results of credit scores will not be based on probabilities but percentages associated with attributes of credit histories and loan characteristics.

Further to this, deep learning, another aspect of supervised learning which is prominent was not implemented in this research. It involves fitting and training a neural networks model which is an analogy of the human nervous system. Because of neural networks ability of fault tolerance, generalization and learning to non-linear and complex relationships, this model is wholeheartedly being recommended for credit risk prediction in Malawi.

All in all, the decision tree model determined for credit risk prediction in this research is highly recommended for all crediting institutions to engage whenever credit referencing is involved. Since the methods used at hand are not trustworthy and reliable enough. The predictions by machine learning are substantially trustable because they are based on a model which was tunefully trained for high performance and accurate results. With a machine learning model that has fit well to the data, one is always sure, by some high proportion, that the results being predicted are true.

The predictions of probability of default and credit score by the decision tree model were based on data inputs of each subject such that the results were not based on groups, assumptions and probabilities. This is not the case with the traditional methods at hand, which crediting institutions, like banks and microfinance companies use.

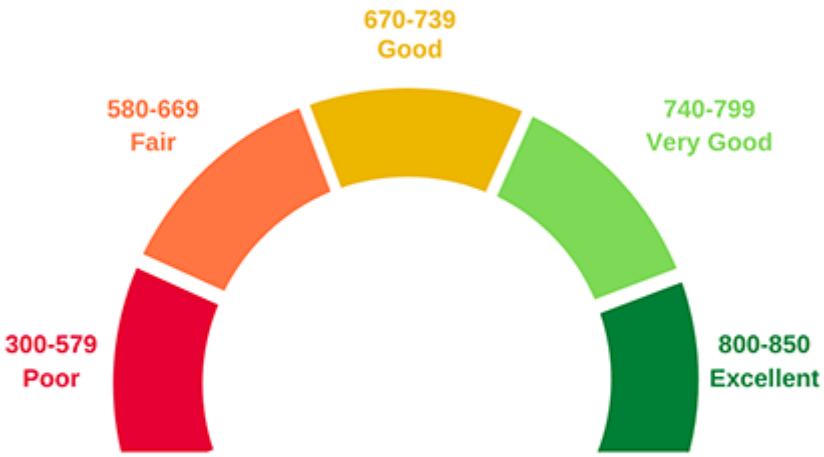
Because of this, it is recommended to the Reserve Bank of Malawi to enforce credit risk determination using this model. With all the data provided for the specific subject whose credit score or probability of default to be determined, credit risk prediction and evaluation should no longer be a big problem.

Appendices

Appendix A: List of different credit scoring ranges

SCORING AGENCY	RANGE
FICO	300-850
TransUnion	300-850
Experian	330-830
Equifax	300-850
VantageScore	501-990

Appendix B: Visualization of FICO credit score ranges and categories



References

Alimadadi, S., Aghabozorgi, S., & Pourhoseingholi, M. A. (2018). *Feature selection methods in credit risk prediction: A survey.* Expert Systems with Applications, 94, 205-224.

Arya et al., S. (2013). *Anatomy of the credit score. Journal of Economic Behavior & Organization Vol 95, 175-185.*

Bank for International Settlements, T. (2021). *Principles for Management of Credit Risk.* Basel: Bank for International Settlements.

Bazdok, D. (2018). *Statistics versus Machine Learning.* Researchgate.

Bekker, A. (2017). *Spark vs Hadoop MapReduce.* Chicago: ScienceSoft.

Benouna, G. (2019). Scoring in microfinance. *Procedia Computer Science*, 522-531.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (1st ed.).* Springer.

Brownlee, J. (2014). *What is Data Mining and KDD.* Machine Learning Mastery.

Cathcart, B. (2009). *Investment portifolio optimization for economic development.* New Jersey: GetInfo Press.

Chan, C. (2020). What is a ROC Curve and How to Interpret It. *Displayr*.

Chatterjee, S. (2015). *Modelling credit risk.* London: Bank of England.

Chee, C. (2021). *Probability of Default using Machine Learning techniques.* Hong Kong: Clinton Chee.

Chee, C. (2022). *Turning data into machine learning predictions and live insights.* Sydney: University of Sydney.

Clark, G. F. (2011). Financial risk modelling of New York commercial banks. *DataFi*, 1-9.

Clark, J. D. (2015). *Basics of classification models:Machine learning.* Ohio: Ohio University.

Coutte, A. (1998). The evolution of credit risk management, 16(7). *International Journal of Bank Marketing*, 242-249.

Cunningham, L. M., Kiley, M. T., & Mihov, I. (2017). The role of credit risk in the financial crisis of 2007-2009. *Journal of Economic Perspectives, 31(2)*, 3-26.

Dr. Benett et al., M. (2019). *Similarities and Differences between Machine Learning and Traditional Advanced Statistical Modeling in Healthcare Analytics.*

Fair Isaac Corporation. ((n.d)). *Understanding FICO Scores*. Retrieved from https://www.myfico.com/credit-education/whats-in-your-credit-score/

Fidler, E. R. (2008). *Measures of Credit Risk and Experience.* Helsinki: NBER.

Frydman et al, N. (2008). *Structured Models for Credit Risk Evaluation.*

G.Clark, J. (2014). *What Every Leader Shoud Know About Employee Performance.* New York: Quantum Workplace.

Gahlaut and Singh. (2017). *Credit risk modelling: Supervised learning.*

Gosavi, S. S. (2014). *Machine Learning Methods for Fault Detection.* Stuttgart: University of Stuttgart.

Guo, X., Li, Y., & Wu, C. (2021). Credit risk prediction using machine learning: A systematic review and future research directions. Applied Sciences, 11(6), 2896.

Hu, Y. (2022). Research on Credit Risk Evaluation of Commercial Banks Based on Artificial Neural Network Model. *Science Direct*, 1169-1176.

Huang et al. (2007). *Credit scoring with Support Vector Machine.*

IBM Cloud Education, .. (2020, August 19). *IBM Cloud Learn Hub*. Retrieved from https://www.ibm.com: https://www.ibm.com/cloud/learn/supervised-learning

J. Kim, D. X. (2008). *Naive Bayes Classifier for Extracting Bibliographic.* Biomedical Online Articles.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

Jameson, J. (2018). Credit Risk Evaluation . *Elsevier*.

Johnson, R. (2017). *Writing an effective research problem.* Ohio: Ohio University.

Kadam et al., M. (2008). *Stochastic Credit Risk Models.* Cambridge: Cambridge University Press.

Kadam et al., M. (2008). *Stochastic Models for Credit Risk.*

Kagan, J. (2022). *Mortgages expounded.* New York: Investopedia.

Khemakhem et al., K. (2018). *Linear Regression for Credit Risk.*

King, T. (2019). *Big Data in Five Years.* New York: Solutions Review.

Kundar, K. (2013). *Introduction to Deep Learning 3rd Edition.* New Delhi.

Le, J. (2019). Machine Learning Classifier: Basics and Evaluation. In M. Daniels, *The Theory Behind Machine Learning Algorithms* (pp. 68-78). Data Notes.

Li, J., Liu, D., & Zhu, H. (2020). Credit risk prediction using machine learning: A survey. Journal of Financial Data Science, 2(3), 47-65.

Luckert, M. (2015). *Using Machine Learning Methods.* London: University of Oxford.

Maayan, G. D. (2020). *Eight Big Data Analytics Options on Microsoft Azure.* Jerusalem: Dataversity.

Microsoft. (2012). *Solution Brief.* Los Angeles: Microsoft.

Mohiuddin, I. A. (2019). Apache Spark: A Big Bata Processing Engine. *Apache Spark: A Big Bata Processing Engine* (p. 7). Khobar: Prince Muhamad University.

Montilla et al., J. (2012). 3D Modeling with Support Vector Machines. *Journal of Information Technology*, 1-11.

Mosch, V. J. (2013). What determinesmicrocredit interest rates? *Appl. Financ. Econ.*, 1579–1597.

Ohman, P., Erlandsson, S., & Enberg, S. (2019). *Banks and the management of credit risk. In Credit Risk Management (pp. 1-20).* Springer.

O'Reilly, M. (2021, January 1). *O'Reilly*. Retrieved from O'Reilly: http://www.oreilly.com

Ostrowski, R. (2015). *Introduction to Apache Spark.* New York: Developers.

Pan, J. (2009). Class imbalance and classification. In J. Pan, *Classification machine learning models* (pp. 213-227). New York.

Panneerselvam, R. (2004). *Research Methodlogy.* New Delhi: Prentice Hall of India.

Park, J. H., Kim, H. J., & Lee, H. K. (2019). Credit risk prediction using deep learning neural networks. Applied Sciences, 9(13), 2683.

Patel, S. (2017). K Nearest Neighbors Classifier. In J. Standev, *Intro to Machine Learning Algorithms.* Machine Learning 101.

Pointer, I. (2020). *Apache Spark.* New York: InfoWorld.

Purohiti et al., M. (2015). *Advanced machine learning in banking.*

Roos, T. (2011). *Introduction to Data Science.* Melbourne: Monash University.

Schapire, R. (2014). *Basics of Machine learning.* Princeton: Princeton University.

Shah, R. (2022). *10 Techniques to deal with Imbalanced Classes in Machine Learning.* New Delhi: Analytics Vidhya.

Simon, A. e. (2015). An Overview of Machine learning and its applications. *International Journal of Electrical Sciences & Engineering (IJESE)*, 22-24.

Soofi, A. (2017). Classification Techniques in Machine learning. *ResearchGate*, 459-465.

Spuchľáková, E. (2015). The Credit Risk and its Measurement, Hedging and Monitoring. *Procedia Economic and Finance*, 675 – 681 .

Standard Bank PLC, M. (2021). *Risk and Capital Management Report.* Blantyre: Standar Bank PLC.

Sun et al, ,. (2006). *Machine learning in banking.*

Tahir, M., Sadiq, S., Mehmood, A., Rehman, M. A., & Khan, F. (2021). *Credit risk prediction using machine learning algorithms:* A review. IEEE Access, 9, 55110-55130.

Thomas, L. (2020). Research designs explored. *Sribbr*, 23-67.

Tian, W. (2015). *Optimized Cloud Resource Management and Scheduling.* Beijing: ScienceDirect.

Urbanowicz, R. (2017). *The theory behind machine learning.* Pretoria: University of Pretoria.

Vaidya, D. (2020). *Credit Risk.* New Delhi: WallStreetMojo.

Ventura, L. (2020). The Worlds Biggest Bankruptcies 2020. *Globlal Finance*, 2.

Wang, F., Zhang, Y., Hu, J., & He, W. (2019). *Credit risk prediction based on feature selection and explainable deep neural network.* Expert Systems with Applications, 135, 78-91.

Watson, A. J. (2019). *Understanding Knowledge Discovery in Databases and Data Mining.* Java T point.

Wikipedia. (2021, November 8). *DJIA Wikipedia*. Retrieved from Wikipedia: www.wikipedia.com/DJIA

Xia, Y. (2020). The Microbiome in Health and Disease. *Progress in Molecular Biology and Translational Science*.

Yiu, T. (2019). Understanding Random Forest. *Towards Data Science*, 12-17.

Young, D. P. (2008). Research Methodology . In L. Redman, *Social Surveys and Research* (pp. 45-59). Bournemouth: Bournemouth University.

Zhang et al., Z. (2018). *Random Forest for NCSM Credit Score.*

Zhou X.H, e. a. (2007). Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in Medicine*, 2170-2183.

Zimmerman, C. (2021). *Traditional and machine learning models for Credit Risk prediction.* New York: Elsevier.

SUPERVISED MACHINE LEARNING MODELING FOR CREDIT RISK PREDICTION AND EVALUATION IN MALAWI

ORIGINALITY REPORT

18%	18%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.javatpoint.com Internet Source	4%
2	elib.uni-stuttgart.de Internet Source	3%
3	www.mdpi.com Internet Source	3%
4	www.coursehero.com Internet Source	2%
5	experfy.com Internet Source	1%
6	www.displayr.com Internet Source	1%
7	huggingface.co Internet Source	1%
8	personal.article.dyndns.info Internet Source	1%
9	www.intechopen.com Internet Source	1%

10	www.diva-portal.se Internet Source	1 %
11	Submitted to University of Rwanda Student Paper	1 %
12	www.ismiletechnologies.com Internet Source	1 %
13	dr.ur.ac.rw Internet Source	1 %

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	Off		