AFRICAN CENTRE OF EXCELLENCE IN DATA SCIENCE

Effect of Machine Learning in Early Prediction of Personal Loan Defaulting on Commercial Bank in Tanzania (2009-2020).

Jesca Elisante Mwende

Registration Number: 220000004

A dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Data Science. University of Rwanda, College of Business and Economics
The African Centre of Excellence in Data Science (ACE-DS)

Supervisor: Dr. Daniel Ruturwa

June 2022

Declaration

I, Jesca Elisante Mwende declare that this thesis/dissertation is the result of my work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Name: Jesca Elisante Mwende

Signature: ………………………

Date: 28th June, 2022.

Approval Sheet

This dissertation entitled Effect of Machine Learning in Early Prediction of Personal Loan Defaulting on Commercial Bank in Tanzania (2009-2020) is written and submitted by Jesca Elisante Mwende in partial fulfillment 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 10 %, which is less than 20% accepted by ACE-DS.

Dr. Ruturwa S. Daniel

Supervisor

Dr. Kabano Ignance

Head of Training

Dedication

To:

- My husband Daudi L. Sambai.
- My son Nathan.
- My parents Elisante and Catherine Mwende.
- My siblings Joshua, Esther, and Dorcas.

Acknowledgment

I thank God for everything He has blessed me with and mostly for the opportunity to pursue a Master's degree.

I also want to appreciate the support from my sponsors through the African Center of Excellence in Data Science for their financial assistance.

I have an appreciation for my supervisor Dr. Daniel Ruturwa for his time, advice, and supervision. To my honorable teachers at the University of Rwanda, African Center of Excellence in Data Science for their valuable support, constructive criticism, and inspirational attitude throughout my studies.

Lastly, I am grateful to my friends and family members, who have been my vocal supporters throughout my academic career.

Abstract

The issuing of loans is one method the bank conveys itself, which has been shown to raise credit default risk in the past. Despite many bank measures to undertake pre-assessment of loan applications, the case remains, providing a warning flag for the bank to produce a rapid, cost-effective, and optimal approach to reduce and perhaps combat credit risk on loan defaulting before banks experience large losses.

The main aim of this study is to develop machine learning models to predict personal loan default and analyze the performance of various models to identify the borrower's features in an early prediction of personal loan default.

In this study, three classical machine learning algorithms K-Nearest Neighbors, Gradient Boosting, and Random Forest were trained on a Historical credit dataset of 5012 observations obtained from a Commercial bank in Tanzania. The dataset was imbalanced and the use of data imbalance techniques namely SMOTE was applied to give us more insight into the classified datasets and reduced erroneous in the conclusion. This dataset was divided into training and test sets respectively with optimized parameters for each of the algorithms. The performance comparison for each model was done by AUC and plotting the ROC Curve, the performance evaluation of classifiers was done to find out the accuracy, recall, precision, ad F1-Score for each classifier in classifying the different types of loan defaults.

The finding showed the features for early prediction of borrowers' loan default were monthly income, total loan amount, and age. The three models RF, KNN, and GB were developed and Random Forest performed well with an accuracy of 84 percent, recall of 85 percent, the precision of 82 percent, F1 Score of 84 percent, and AUC ROC of 91 percent in predicting either loan defaulting or not.

Although, the study managed to implement the high-performance model further studies should be conducted particularly with the use of deep learning or other machine learning models, the involvement high dimensional dataset from many banks in Tanzania.

Keywords: Loan default, Commercial Bank, Machine Learning, Tanzania.

List of Tables

List of Figures

List of Acronyms

AUC	Area Under Curve
BOT	Bank of Tanzania
CART	Classification and Regression Tree
CR	Credit Risk
CRM	Credit Risk Management
EDA	Exploratory Data Analysis
KNN	K-Nearest Neighbor
ML	Machine Learning
URT	United Republic of Tanzania
RF	Random Forest
RFE	Random Forest Estimator
ROC	Receiver Operating Characteristic
SACCOS	Saving and Credit Cooperative Societies
SMOTE	Synthetic Minority Oversampling Technique
TZS	Tanzanian Shillings

Table of Contents

CHAPTER ONE

INTRODUCTION

1.1 Background

The trend of loan defaults is growing to be a problem for the nation's economy, as well as the banking industry. It negatively affects the banks' ability to finance themselves, which harms the nation's overall socio-economic development. Lending has been the bank's core business for a decade among its suite of products. (Opa & Tabe-Ebob, 2019). This development cannot always be achieved due to the loan default issue. Given the fact that loan default can be either voluntary or involuntary, the problems with default may result from both the nature of the business and the attitudes of the borrowers (Aslam et al., 2020).

Increased loan default rates potentially contribute to banks, the finance sector, and the economy as a whole. Additionally, failing to properly manage non-performing loans over time has an impact on how profitable commercial banks can be (Stephen Kingu et al., 2018). The Bank of Tanzania's laws, as outlined in the Bank of Tanzania Act of 1995 regulate commercial banks in Tanzania. Commercial banks support the economic development in the country by channeling funds from surplus areas to deficiency areas (Kaaya & Pastory, 2013). However economic downtown has led to the closure of many businesses and most clients have not been able to pay back their loans (Kaaya & Pastory, 2013)

By 2013 banks in Tanzania had begun the application of the credit reference bureau which is a specializes in gathering and selling data on how well people and businesses have managed their credit (Bank of Tanzania, 2012), it plays an important role in improving loan defaults and access to loans, in turn, facilitates growth in the economy by enabling banks to make quicker and give early warnings to lenders to make more effective lending decisions. The decision to approve a loan is based on several considerations, including the borrower's past credit history, the overall amount requested, the economic situation, and the loan's intentional use, occupation, and so on. Additionally, the borrower must be able to repay the loan in the given timeframe (Torvekar & Game, 2019).

Due to technological advancement, there is now an increased interest by the banking sector in using machine learning techniques to predict customer's loan default, partially due to evidence that conventional techniques are insufficient and prone to inaccuracy. There is evidence that by utilizing machine learning techniques, credit risk management for loan default can be greatly improved because these techniques enable meaningful interpretation of unstructured data (Aziz & Dowling, 2018).

Although it is becoming more common, commercial banks have been using machine learning approaches to model credit risk for some time. The study that has been undertaken on loan defaulting, and in Tanzania banks are non-machine learning such as (Mungure, 2015). Furthermore, Tanzanian banks rely on traditional methods and credit bureau references to decide if a customer would default or not default before the provision of a loan. Machine learning techniques that form are transforming and will revolutionize, how we approach credit risk management. Through the development of machine learning-driven solutions, everything related to understanding and managing risk is now possible, from determining how much a bank should lend to a customer to alerting businesses on the financial markets about position risk to identifying customer and insider fraud, improving compliance, and lowering model risk. (Leo et al., 2020).

1.2 Problem Statement

In traditional credit scoring models, human judgment and instinct are important factors that influence whether to accept or reject an application. The borrower's capacity, character, condition, capital, and collateral are the main considerations (ElMasry, 2019)

- Capacity: What is their ability to repay? What amount of free revenue do they have?
- The character of the person: Are the borrowers or family member’s familiar to you?
- Condition: What are the market conditions?
- Capital: How much is requested?
- Collateral: What resources is the borrower willing to contribute?

Therefore, the process of providing credit is both a science and an art. Branch offices and corporate divisions regularly monitor the activities, performance, and conduct of each borrower daily to determine whether any adjustments are necessary to protect the non-repayment of credits. (Richard et al., 2008). When a borrower is estimated to default and could have reimbursed off the loan otherwise, and the bank rejects to offer the loan, in this case, the bank might have missed the opportunity to advance out from interests that could have been paid. In reality, default loans are bad loans that banks are unable to profit from since it is difficult to predict whether the borrowers will be able to make payments on the amount borrowed or owed (Opa & Tabe-Ebob, 2019).

In the big data revolution, more advanced models that surpassed the traditional classifier model logistic regression have been built, and default prediction has emerged as the most fascinating field to be researched (ElMasry, 2019). Within financial institutions, interest in and acceptance of strong and analytical solutions like machine learning and artificial intelligence are rising. The urge to improve analytical capabilities for handling and mining the growing volumes and variety of data has led to this rise. (Van Liebergen, 2017). Machine learning relies heavily on learning from data that is already accessible, therefore it may be subject to the same biases and issues that plague traditional statistical methods. It would be helpful to assess and comprehend how issues with traditional statistical research methods perform when treated by machine learning approaches when they are compared to machine learning methods. To help banks, improve their analytical skills and develop their risk management process, it would be good to research how machine learning may be used to improve risk measurement, risk reporting, risk assessment, and risk aggregation. (Leo et al., 2020).

The study seeks to fill the gap of a few research conducted on loan defaulting in Financial Institutions, which are non-machine learning such as (Mungure, 2015), and other ML research conducted did not focus on Commercial Banks such as (Ngimbwa, 2020). Motivated by these studies, the study will cover the most common and powerful ML algorithms that have been applied for CRM to develop a Machine Learning model for loan default prediction. A predictive ML model that will classify a loan whether will default or not, evaluate which borrowers' features are important in predicting loan default, and find answers to the question; What borrower's variables predict loan defaulting, what machine learning models to be used for predicting loan defaulting and what is the performance of the proposed the model.

1.3 Objectives

- To find out which borrower's variables predict loan defaults.
- To develop a Machine Learning models that early predicts loan defaulting at the Bank.
- To compare which model performs well in the early prediction of personal loan default.

1.4 Significance of the Study

The study is significant in helping banks to reduce loan defaulting with the proposed machine learning model. Further, the study will offer the following advantages to the banks, such as increased liquidity, extended provision of loan services to customers, speeding up loan provision process, increase reputation, and reduced operational costs. To the government, the study will increase revenue, reduce bank monitoring costs, and form policies.

1.5 Scope of the Study

This study was conducted using a private credit dataset from 25th December 2009 to 28th December 2020 from a commercial bank in Dar es Salaam region, Tanzania, which has other branches in Tanzania regions, to be used for an early prediction of loan default.

1.6 Limitations of the Study

The rules and limitations of data sharing in Commercial Banks, restrict access to certain of the features. As a result, several features that were helpful in building and training the model for loan default prediction were not shared. Another is a wide range of timeframe but few numbers of the dataset.

1.7 Outline of the Thesis

The remainder of the study is divided into four chapters Chapter two describe relevant previous work and identify the relationship of works in the context of their contribution to the topic. Chapter three includes what things and methods were done to generate results. Chapter four machine learning processes conducted and interpretations of statistical results. Chapter five describes the summary findings, conclusion, and recommendations for further research.

CHAPTER TWO

LITERATURE REVIEW

21 Factors Influencing Loan Default

A study by (Mungure, 2015) in Tanzania Microfinance Finance Institution showed that the number of dependents has an impact on loan repayment since loans might be used to fulfill family responsibilities. The fact that 80% of loan officers believed that there is a default problem and loan payment delays was also made clear. The factors that led to default included insufficient loan repayment follow-up, loans without collateral requirements, high-interest rates, limited loan utilization monitoring, a need for loan usage training, multiple borrowing, unplanned loan use, and poor sales.

The study by (Aslam et al., 2020) was conducted in Bangladesh using binomial logistic regression to determine the elements that affect the probability that borrowers will notify loan default. There are fourteen explanatory variables in the study and concludes that living status, repayment amount, loan product, and interest rate may contribute to less or high loan default.

The aforementioned factors were identified as loan default causes by a study that was done in Ghana Micro-Finance Institution identified by the customers include late loan repayments, financial difficulties, unattractive terms of payment, rising interest rates, insufficient loan amounts, unanticipated situations, such as illness and death of a family member, and an absence of customer training before and following granting loans. (Addae-Korankye, 2014).

22 Related Studies on Loan Default Prediction

The study by (Zhu, 2019) used Lending club public datasets on various machine learning algorithms, including ensemble models, extreme gradient boost, neural networks, and logistic regression, to explore how they are being applied to loan default prediction. The study used class weights, SMOTE, and ADASYN to balance the data, and each model improved to a different extent. The best performing model was extreme gradient boosting which had high accuracy of 68%. The client's interest rate, annual income, and possession of property were the most significant variables in this study that clients provided that may be potential, according to an initial data interpretation.

According to (Ngimbwa, 2020) in resolving the problem of loan default, his study analyzed Saving and Credit Cooperative Societies datasets in Tanzania by using machine learning, to determine the factors affecting credit rating. The study concluded that the foremost variables found by the random forest algorithm in influencing SACCOS members' credit ratings were age, interest rate, and membership years with an accuracy of 95%, while the least variables were marital status and gender The foremost factors found by logistic regression with an accuracy of 74% include age, loan period, and interest rate while the least factors include membership years and marital status.

In a study conducted on China loan data by (Wang et al., 2020), used to determine if a person meets the requirement for lending, the naive bayesian model, logistic regression analysis, RF, DT, and K-NN classifier are used to evaluate applicant's credit information. The study shows that RF performs best in terms of AUC, precision, recall, and accuracy it had 96.53%.

(ElMasry, 2019) used a variety of single classification machine learning techniques to predict mortgage loan defaults based on the publicly available dataset, including ensemble, SVM, RF, K-NN, DT, Logistic Regression, and SVM. The output probabilities of the aforementioned techniques were combined in the study using a meta-algorithm ensemble approach stacking, which increased prediction power. The accuracy that Stacking Ensemble had was the highest at 89 percent. The top five factors used to forecast loan default include zero balance code, current loan delinquent status, loan age, credit score, and original total loan-to-value.

Further studies were done by (Abrahamsson & Granstrom, 2019) to determine how each machine learning technique, from a selected group, performs best in default prediction, taking into account the selected model assessment parameters. Kendall's Tau performed worse than RFE, which was utilized as a feature selection method. The study found that the model's accuracy was influenced by the number of features it contained. The techniques that were looked into including support vector machines, AdaBoost, XGBoost, random forests, decision trees, and logistic regression. The machine learning algorithm that produced the relative greatest performance, known as extreme gradient boost (XGBoost), got an AUC score of 89 percent for the 7 features used.

In the study by (Chang, 2019), several machine learning models are trained using data from Lending Club to predict whether the borrower can repay the loan. Randomized Search Cross-Validation and Grid Search Cross-Validation methods are applied in a different situation to tune the parameters. Also evaluated were the RF, SVM, K-NN, and logistic regression performance of the proposed models. Then the selected models were compared using the confusion matrix and ROC curve. Random forest was found as the best fit for the dataset with the highest accuracy of 70.4%. The study concluded that annual income and Fico scores are features that influence prediction.

A study by (Madaan et al., 2021) used two machine learning techniques, random forest, and decision tree, to investigate, examine, and create a machine-learning technique to accurately determine whether a person, given certain traits, has a high likelihood to default on a loan. Then the study compared the performance metrics by using the confusion matrix and F1 score to quickly understand the two models' results for accuracy and other factors. The accuracy of the Random Forest Classifier was 80%.

CHAPTER THREE

METHODOLOGY

3.1 Study Design

The design framework is summarized in figure 1 below

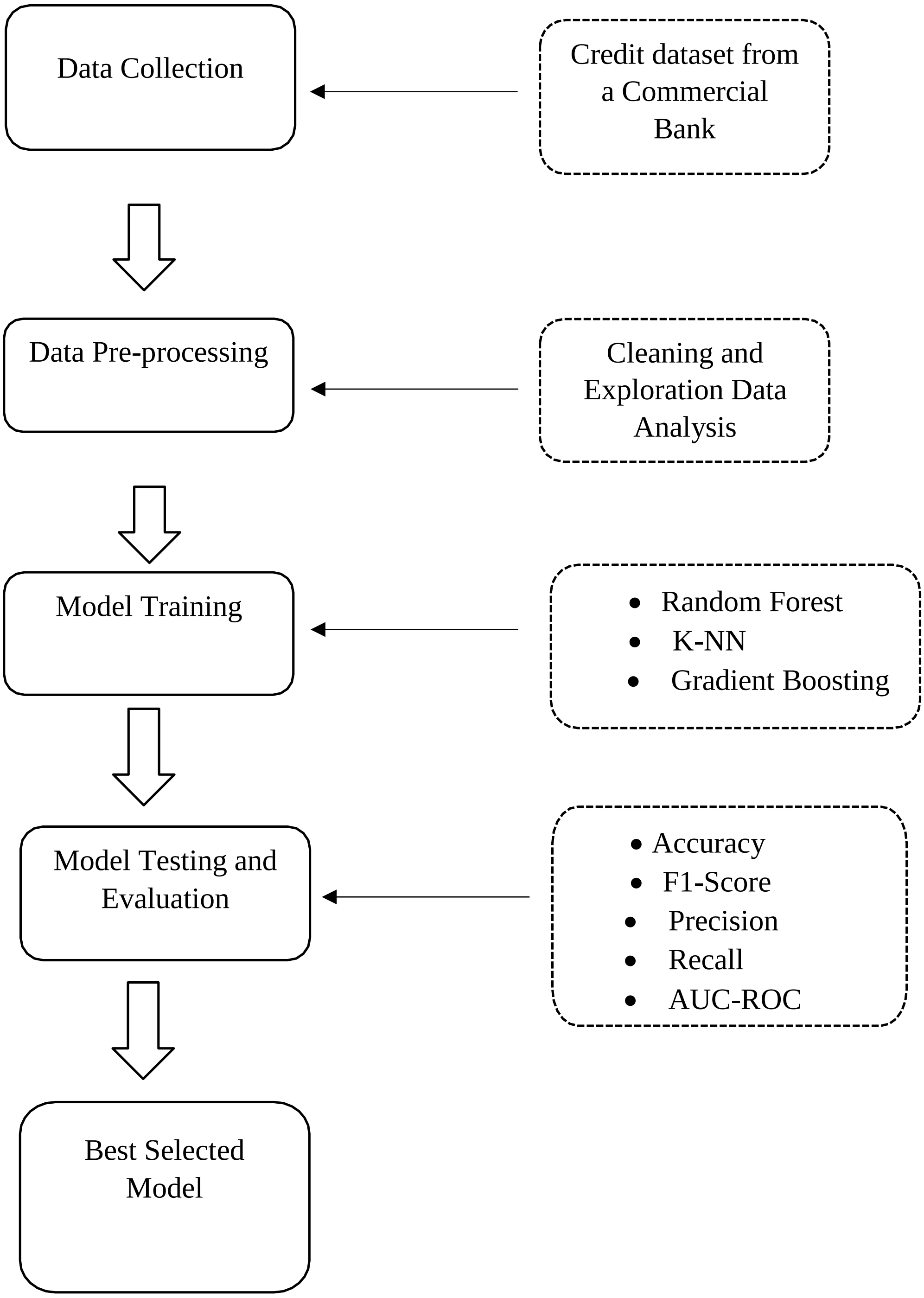


Figure 1 Machine Learning Experiment Approach

3.2 Data Collection

Secondary data, including lending information from other branches in Tanzania, was gathered from a commercial bank. The dataset, which included performance and personal data, was utilized to train and test the model. It consists of 5485 customers of which 5181 are non-defaulters and 304 are defaulters with 31 attributes.

3.3 Working Environment

For our study, all the activities including processing the data, cleaning and visualizing the data, analyzing the data, model building, and evaluation were done on the Jupyter Notebook 6.0.3 using Python programming language.

3.4 Machine Learning Methods

The study aimed at using three classification algorithms in the early prediction of loan default for an individual loan.

3.4.1 K-NN

It is a non-parametric technique for regression and classification. A data point is classified by calculating the distance to the closest nearby training case. The classification of the sample will depend on the value of that point. The sample is classified using the value of the majority and the k nearest points in the k-nearest neighbor classifier. The existence of noisy or irrelevant features, or if the feature scales are inconsistent with their significance, can significantly reduce the accuracy of the K-NN method (Alpaydın, 2010).

3.4.2 Random Forest

To obtain extremely high classification accuracy, bagging is combined with random decision trees. This method involves growing a forest of trees, then before fitting each tree, putting the training data onto a randomly selected subdomain. The concept of randomized node optimization, which substitutes a randomized technique for a deterministic optimization, is the final one.

Random forest is useful to lessen the connection between different classifiers when a variety of qualities are available. This significantly improves the performance of the final model at the cost of a slight increase in bias and some loss of interpretability (Alpaydın, 2010).

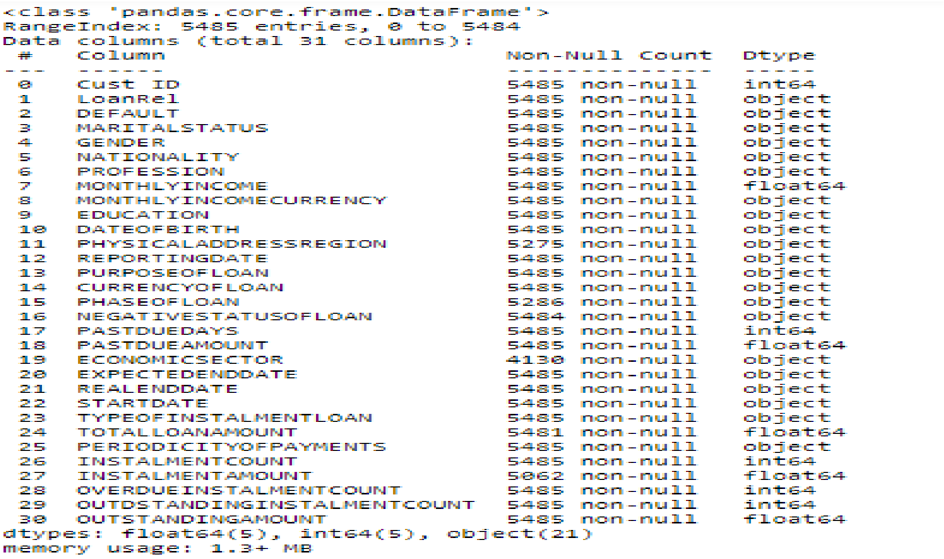
3.4.3 Gradient Boosting

Iteratively combining weak learners into a single strong learner, gradient boosting is a machine learning technique for regression and classification problems that yields a prediction model in the form of an ensemble of weak prediction models. Gradient boosting is frequently used as the base learner with decision trees, particularly CART trees, that have a fixed size (Alpaydın, 2010).

Implementations of gradient tree boosting frequently additionally use regularization by limiting the least number of observations in the tree's terminal branches. By eliminating splits that lead to branches with a low number of instances from the training set than this threshold, it is used during the tree-building procedure. This limit's imposition aids in lowering prediction variance at the leaves.

3.5 Identification and Definition of Variables

The dataset collected contained a total of 31 features (1 independent and 30 dependent) with 5485 observations. Figure 2 shows the identified features in the dataset with their attributes.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5485 entries, 0 to 5484
Data columns (total 31 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Cust ID                          5485 non-null   int64
 1   LoanRel                          5485 non-null   object
 2   DEFAULT                          5485 non-null   object
 3   MARITALSTATUS                    5485 non-null   object
 4   GENDER                           5485 non-null   object
 5   NATIONALITY                      5485 non-null   object
 6   PROFESSION                       5485 non-null   object
 7   MONTHLYINCOME                    5485 non-null   float64
 8   MONTHLYINCOMECURRENCY            5485 non-null   object
 9   EDUCATION                        5485 non-null   object
 10  DATEOFBIRTH                      5485 non-null   object
 11  PHYSICALADDRESSREGION            5275 non-null   object
 12  REPORTINGDATE                    5485 non-null   object
 13  PURPOSEOFLOAN                    5485 non-null   object
 14  CURRENCYOFLOAN                   5485 non-null   object
 15  PHASEOFLOAN                      5286 non-null   object
 16  NEGATIVESTATUSOFLOAN             5484 non-null   object
 17  PASTDUEDAYS                      5485 non-null   int64
 18  PASTDUEAMOUNT                    5485 non-null   float64
 19  ECONOMICSECTOR                   4130 non-null   object
 20  EXPECTEDENDDATE                  5485 non-null   object
 21  REALENDDATE                      5485 non-null   object
 22  STARTDATE                        5485 non-null   object
 23  TYPEOFINSTALMENTLOAN             5485 non-null   object
 24  TOTALLOANAMOUNT                  5481 non-null   float64
 25  PERIODICITYOFPAYMENTS            5485 non-null   object
 26  INSTALMENTCOUNT                  5485 non-null   int64
 27  INSTALMENTAMOUNT                 5062 non-null   float64
 28  OVERDUEINSTALMENTCOUNT           5485 non-null   int64
 29  OUTDSTANDINGINSTALMENTCOUNT      5485 non-null   int64
 30  OUTSTANDINGAMOUNT                5485 non-null   float64
dtypes: float64(5), int64(5), object(21)
memory usage: 1.3+ MB
```

Figure 2 Dataset Features

3.5.1 Independent Variables

Table 1 List of Independents

Loan Default = a0 +a1X1+a2X2+a3X3+a4X4+a5X5+a6X6+ a7X7+a8X8+……. +U

- a0=Intercept,
- a1=CustID
- a2=LoanRel
- a3=Marital Status
- a4=Negative Status of Loan
- a5=Reporting Date
- a6=Type of Installment Loan
- a7=Education
- a8=Monthly Income Currency
- a9=Monthly Income
- a10=Physical Address Region
- a11=Nationality
- a12=Overdue installment count
- a13=Gender
- a14=Economic sector
- a15=Periodicity of payments
- a16=Purpose of the loan
- a17=Start date
- a18=Profession
- a19=Date of birth
- a20=Instalment Count
- a21=Outstanding Instalment Count
- a22=Past Due Days
- a23=Past Due Amount
- a24=Instalment Amount
- a25=Outstanding Amount
- a26=Total Loan Amount
- a27=Phase of loan
- a28=Real End Date
- a29=Start Date
- a30=Expected End Date

3.6 Exploratory Data Analysis

The features were explored individually to summarize statistics, data visualization, summarizing statistics, perform a transformation, and discover the important features for loan default prediction.

The figure below visualizes the distribution of loan applicant or customer professions over the default classes, which shows high number of customers with a status of self – employment did not default as well as defaulted.

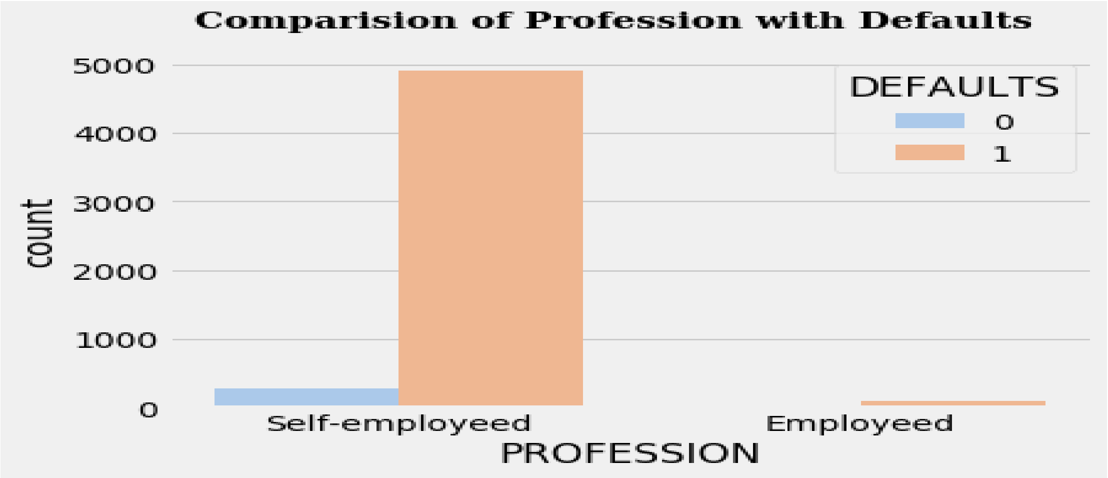


Figure 3 Comparison of Professions with Default

Figure 4 plots the monthly income over the total loan amount for the target variable classes "default". The observation shows that high number of customer who did not default have a monthly income between 1-1000000Tzs and they took a total loan amount of between 1-15000000

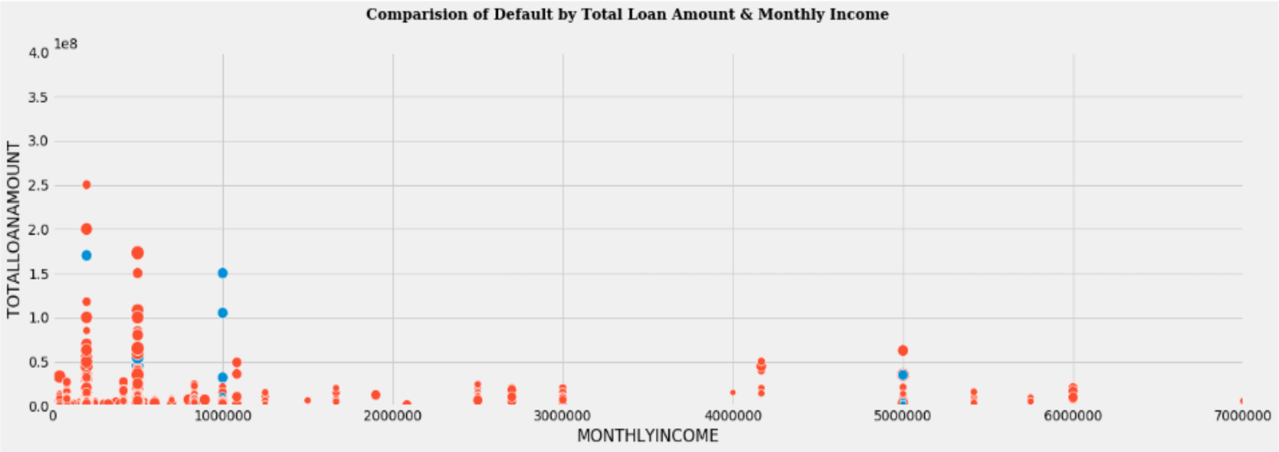


Figure 4 Comparison of Monthly Income over Total Loan Amount

3.7 Univariate Analysis

The target feature (DEFAULT) was descriptively analyzed and showed that 94.6% of the observations are of **class 1**(no-defaulters) and 5.4%of the observations are **class 0** (defaulters). Figure 5 shows the percentage of the dependent variable.

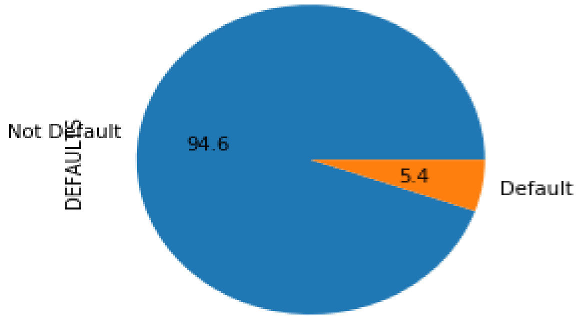
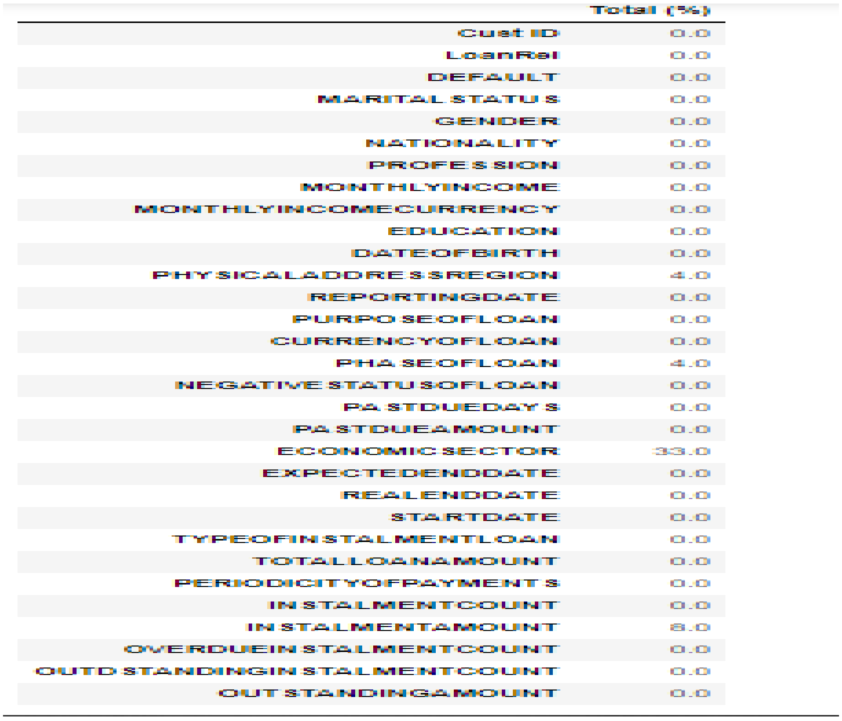


Figure 5 Percentage of Defaulted and Not Defaulted

3.8 Missing Value Imputation

Missing values include all entries in data that their measures were entered wrong, not recorded, or observed. These values have the necessary information for the efficiency of prediction. Thus, it is necessary to properly deal with missing values. The amount of missing values in the data is displayed in Figure 6.



	Total (%)
Cust ID	0.0
LoanRel	0.0
DEFAULT	0.0
MARITALSTATUS	0.0
GENDER	0.0
NATIONALITY	0.0
PROFESSION	0.0
MONTHLYINCOME	0.0
MONTHLYINCOMECURRENCY	0.0
EDUCATION	0.0
DATEOFBIRTH	0.0
PHYSICALADDRESSREGION	4.0
REPORTINGDATE	0.0
PURPOSEOFLOAN	0.0
CURRENCYOFLOAN	0.0
PHASEOFLOAN	4.0
NEGATIVESTATUSOFLOAN	0.0
PASTDUEDAYS	0.0
PASTDUEAMOUNT	0.0
ECONOMICSECTOR	33.0
EXPECTEDENDDATE	0.0
REALENDDATE	0.0
STARTDATE	0.0
TYPEOFINSTALMENTLOAN	0.0
TOTALLOANAMOUNT	0.0
PERIODICITYOFPAYMENTS	0.0
INSTALMENTCOUNT	0.0
INSTALMENTAMOUNT	8.0
OVERDUEINSTALMENTCOUNT	0.0
OUTDSTANDINGINSTALMENTCOUNT	0.0
OUTSTANDINGAMOUNT	0.0

Figure 6 Percentage of Missing Values

From figure 2 it has been observed that the dataset contained a total difference of 210 from the physical address region and 4 from the total loan amount. Missing values can be dealt with through different mechanisms such as;

3.8.1 Filling Missing Values

The missing data for the variable physical address region was filled by the most frequent value of the variable.

The numerical variables, total loan amount, monthly income and installment amount, and missing data were filled with the median values.

3.9 Outliers

3.9.1 Detecting Outliers

To identify features that significantly contain outliers z- score method was applied through all features and a new data frame of similar shape was created which contained respective z - score values only.

$$\boldsymbol{z = \frac{x - \mu}{\sigma}}$$

Assuming values follow a normal distribution, a value with a z - score of more or equal to 5 has zero probability to belong in the distribution. Therefore, a z-score of 5 was located as the optimal for data to follow the distribution of the dataset. All features that had a z-score of greater than or equal to 5 were then filtered excluding categorical variables and including total loan amount and monthly income features were then visualized with boxplots in figure 7.

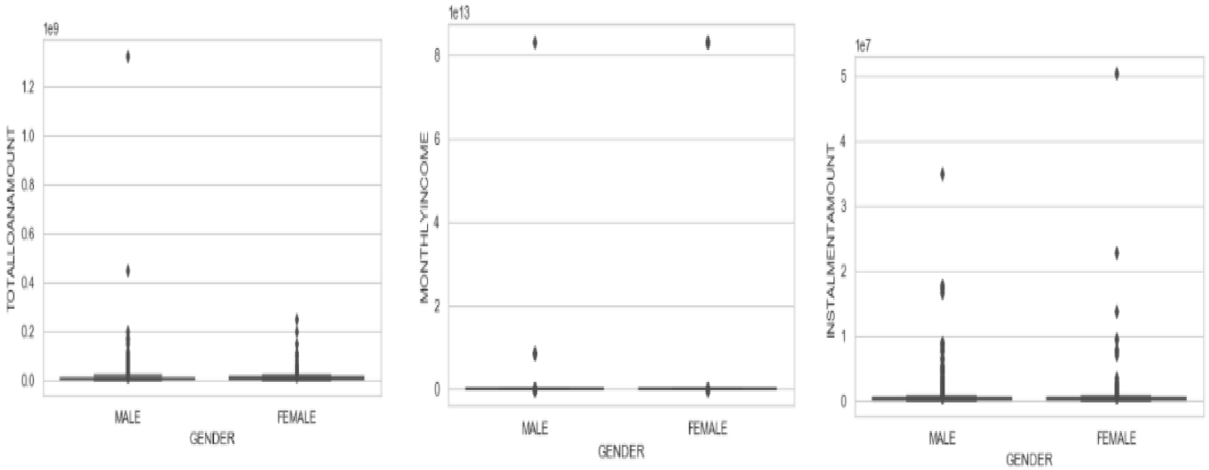


Figure 7 Box plots for features with outliers

3.9.2 Fixing Outliers

From the monthly income feature, most of the outliers (approximately 34%) were 0 (zero) values, which were considered missing values and were imputed by the median value. Median of the features was used to fill the outliers where the lower and upper limits were fixed by the relation between the quartiles and frequency distribution of the monthly income and were fixed between (2000000 and 6000000).

3.10 Correlation of Features

The correlation of 1 in the analysis explained the maximum correlation between features. The highest positive correction obtained is of tenor and installment count, also total loan amount and installment amount have attained the limit; installment amount and installment count are then being dropped due to high correlation with the total loan amount and tenor respectively.

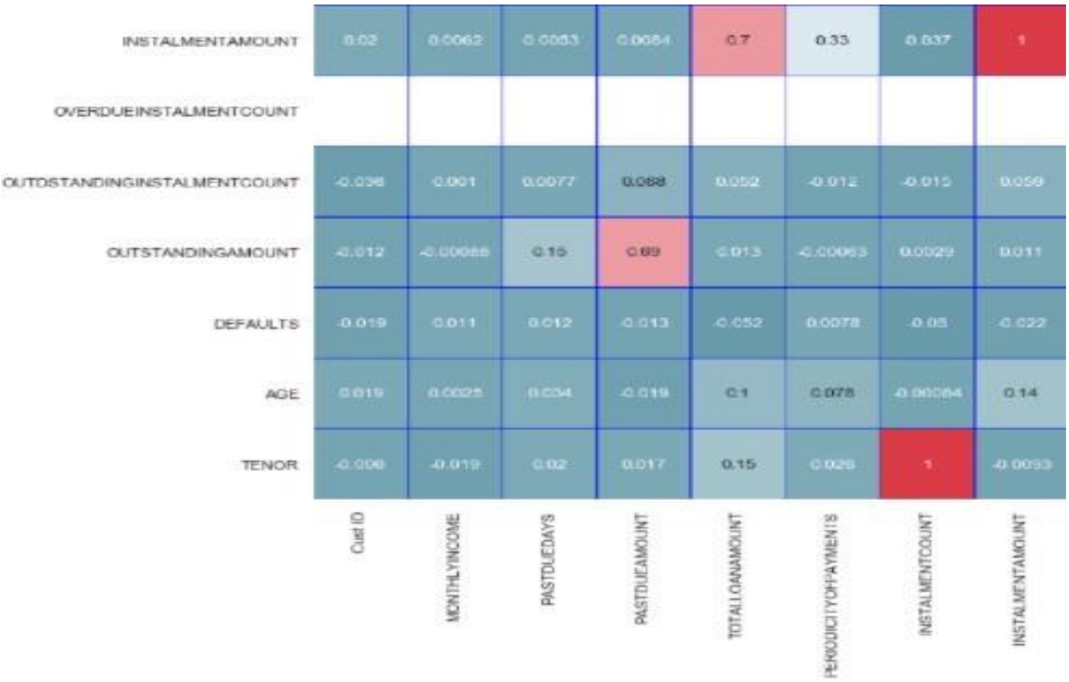


Figure 8 Correlation of Features

3.11 Features Reduction

The dataset containing features such as “Cust ID” and “LoanRel” were dropped because they are used for tracking customers and they don’t have a direct impact on the loan. The type of Instalment loan was dropped because the type was common for all datasets.

3.12 Cleaning of Incorrect Recordings

The physical address region contained mixed recordings for regions, districts, and unapplied entries. Observations were restructured to present unit regions on matching entries and hence unique observations in the feature were reduced from 58 to 18.

Furthermore, the profession feature was treated as employment status where the observation contained some of the mixed observations, and profession was treated as employment status where all professions listed were grouped as employed, whereas other and others were grouped as self-employed.

3.13 Selection of Features

The process of screening the identified features and their behaviors was further subjected to check their suitability for the study. The status of what our model wants to predict may not relate to some of the features provided but we have to get familiar with the features for the guidance of the model regularization. The criteria used involve checking:

3.13.1 Variability Filtering

Features that had low variance include past due days and past due amounts. Features such as the past due amount which cannot be interpreted when assessing new customers were dropped. Features that are policy progress observations including overdue installment count, outstanding amount, and outstanding installment count, which cannot be accessed from a new customer were also dropped.

From figure 6 above, the economic sector was dropped because it has the highest amount of missing values. The phase of loan and installment amount were kept due to their relevance and importance in the domain of loan.

3.13.2 Feature Mapping

Date of birth values was converted to year values by creating a new feature age. The ages of customers who were above 80 years were dropped. Then the date of the birth feature was dropped. Figure 9 shows the distribution of age.

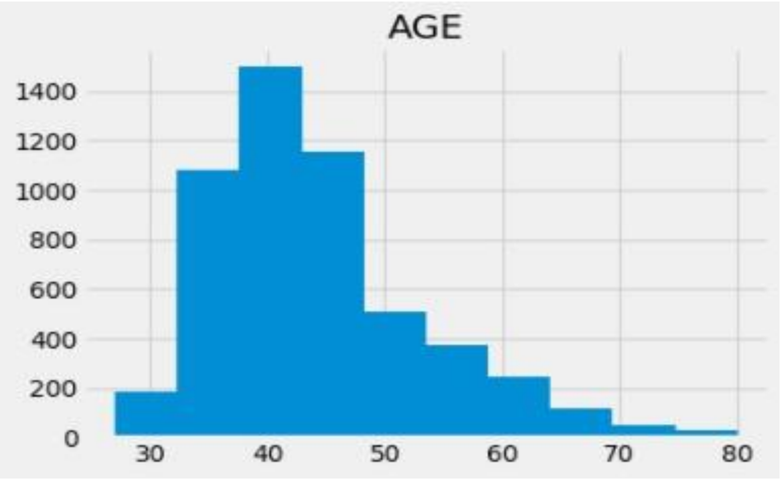


Figure 9 Distribution of Age

Default values were converted, all the values with the filled number to 0 which means default, and all empty values to 1 which means not default.

Periodicity of payments values was converted to days and then multiplied by installment count to get a new feature tenor which is the length of time that will be taken by the borrower to repay the loan along with the interest.

3.13.3 Selection of Features Importance using Random Forest Classifier

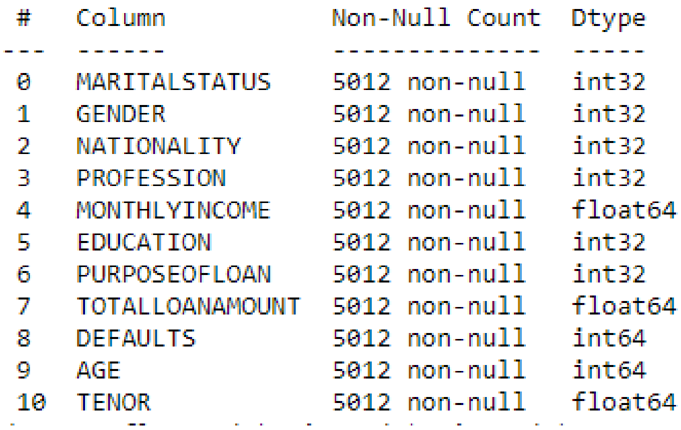
Which features are best capable of predicting the target variable are indicated by their importance. Feature importance can assist in figuring out which features are most crucial to our model and which ones we can safely disregard. We can then use this to simplify our models and make them easier to understand. In this study, feature importance was determined using the Random Forest Classifier.(Kumar, 2022). After fitting the Random Forest Classifier model, it gathers the feature importance values so that the same may be accessed via the feature importance's property. The attribute "feature importance’s" indicates the weights assigned to each feature based on how they are organized in the training dataset. Random forest can be used to naturally order the relevance of variables in a classification or regression issue. Fitting a random forest to the data is the initial stage in determining the variable importance in a data set where $Dn = f(Xi;\ Yi)gn,\ i = 1$. The standard deviation of these discrepancies is used to standardize the score. Features that provide high values for this score are given higher priority than those that produce low values (Alpaydın, 2010). The important features selected using Random Classifier in our study where age, total loan amount and monthly income.

3.14 Label Encoding/Transformation

Machine learning models require all data to be transformed into a numerical format and most of the datasets contained some non-numerical values. The encoding process used was Label Encoder from sklearn which encodes all the categorical variables into integer values. All non-numerical data except “nan” values were transformed into numerical abstract/dummy values.

3.15 Final Dataset after EDA

The final dataset after pre-processing remained with 5012 observations and 11 features, as on figure 10 below



```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   MARITALSTATUS   5012 non-null   int32
 1   GENDER          5012 non-null   int32
 2   NATIONALITY     5012 non-null   int32
 3   PROFESSION      5012 non-null   int32
 4   MONTHLYINCOME   5012 non-null   float64
 5   EDUCATION       5012 non-null   int32
 6   PURPOSEOFLOAN   5012 non-null   int32
 7   TOTALLOANAMOUNT 5012 non-null   float64
 8   DEFAULTS        5012 non-null   int64
 9   AGE             5012 non-null   int64
 10  TENOR           5012 non-null   float64
```

Figure 10 Final Dataset after EDA

3.16 Class Imbalance

The distribution of the two classes default and non-default in our dataset is unbalanced. where just 5% of the data is classified as class 0 (default) and 95% of the observations are classified as class 1 (non-default). Because of how the majority class will affect the minority class, the minority class will not be identified. When fitting the model to the training dataset and predicting classes to the testing dataset, data balancing techniques are used.

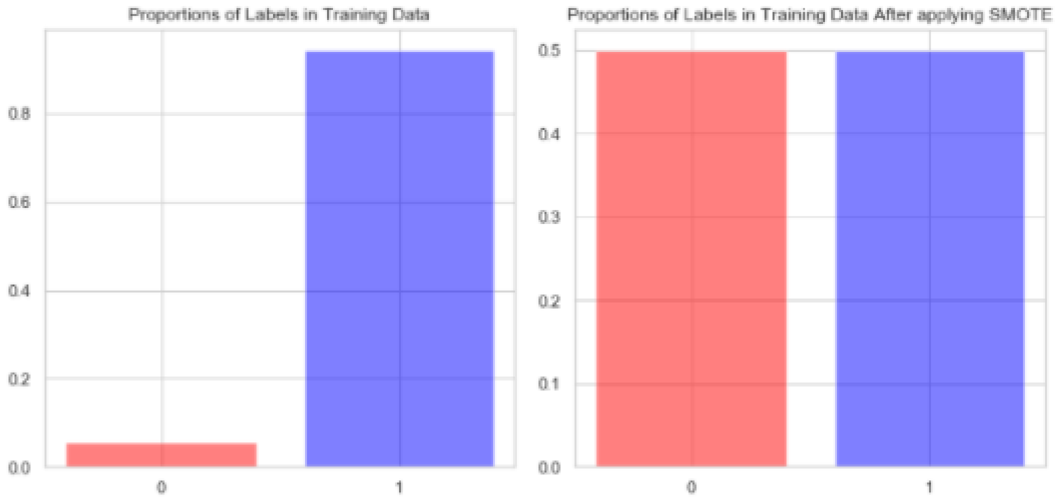


Figure 11 Proportions of Dependent Variable before SMOTE and after SMOTE

3.16.1 Synthetic Minority Oversampling Technique

To balance the data, oversampling techniques reproduce the observations from the minority class. The identical observation is however added to the original data to create overfitting, which results in high training accuracy but low accuracy over testing data. In contrast, the majority of classes are excluded using the under-sampling techniques to rebalance the data. The training data loses essential information from the majority class when observations are removed (ElMasry, 2019). SMOTE locates random points among each minor observation's closest neighbors and creates new minor observations using boosting techniques. The overfitting issue will no longer occur because the new data are not identical to the old data, and we did not want to lose the same or more information as we would with under-sampling techniques (Chawla et al., 2002).

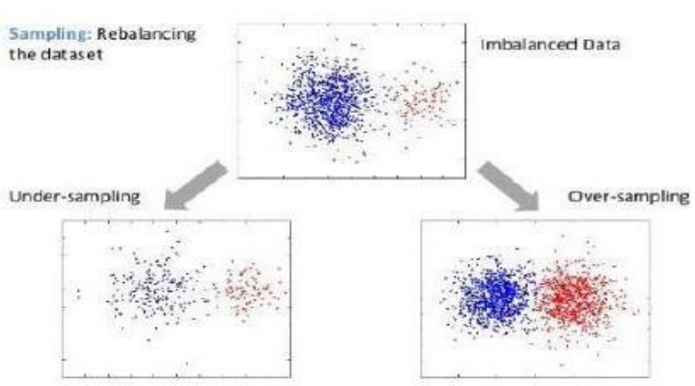


Figure 12 Under-Sampling and Over-Sampling (By Zhu, Leon)

3.17 Evaluation Metrics

Evaluation of the model is a very important step in developing a machine learning model as it allows you to evaluate, compare, and qualify how well a model performed. The evaluation metrics used in our test dataset for this study were accuracy, precision, recall, F1 score, and AUC ROC.

Accuracy

Its definition states that it is the ratio of accurately anticipated observations to all observations.

$$\mathbf{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}}$$

True Positive (TP)-Number of cases correctly identified as the default

True Negative (TN)- Number of cases correctly identified as not default

False Positive (FP)-Number of cases incorrectly identified as the default

False Negative (FN)-Number of cases incorrectly identified as not default

Precision

Defined as the measure of observations that were correctly predicted by our model over the correct and incorrect predictions.

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP + FP}}$$

Recall

Defined as the measure of observations that were correctly predicted by our model over the total amount of prediction

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP + FN}}$$

F1 Score

Is being used to combat prioritizing one score over the other, it strikes a balance between recall and precision score.

$$\mathbf{F1\ Score} = \frac{\mathbf{2 * Precision * Recall}}{\mathbf{Recall + Precision}}$$

AUC ROC

Area Under the Curve Receiver Operating Characteristics is the most significant evaluating metric which can be used to see how well a classification model is performing.

CHAPTER FOUR

RESULTS AND DISCUSSION

41 Factors Influencing Borrower's Loan Defaulting by Applying Machine Learning

The factors which were obtained from Commercial Bank include gender, marital status, total loan amount, birth date(age), default, monthly income, the purpose of the loan, physical address region, installment count (tenor), nationality, education, and profession. Random Forest Classifier was used to select the appropriate factors influencing personal loan default in the bank. The importance of features differs as shown in Figure 13

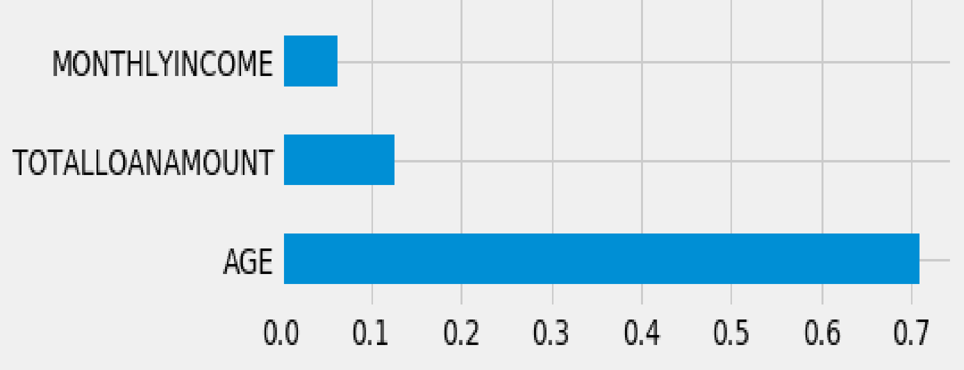


Figure 13 Three Features Selected Using Random Forest Classifier

4.1.1 Descriptive Analysis

The study included a total of 5012 borrowers involved, with 4727 (94%) of them being non-defaulters and 285 (6%) being defaulters. Borrowers who took a loan at an age between 36-64 were defaulted by 86 percent time compared to other age groups. The majority of defaulted borrowers who had a monthly income of 100,000 to 1,000,000 Tanzanian Shillings, were defaulted 94.7 percent. When compared to borrowers who took a total loan amount of below 1,000,000 TZS, borrowers above that amount were defaulted by 90.5 percent. Table 2 below summarizes the findings

Table 2 The Distribution of the Sample According to Selected Features and Default Status

	DEFAULTS	
Age-Group	**Default(%)**	**Not-Default(%)**
18 - 35	11.2	13.2
36 - 65	86.0	83.7
Above 65	2.8	3.1
Monthly Income		
Below 100000 TZS	1.8	1.7
100000 TZS - 1000000	94.7	95.3
Above 1000000 TZS	3.5	3.0

Total Loan Amount		
Below 100000 TZS	0.7	1.0
100000 TZS - 1000000	8.8	8.3
Above 1000000 TZS	90.5	90.8
Purpose of the Loan		
Purchase of Personal Consuming Products	36.5	38.1
Working Capital	27.5	25.7
Construct	26.2	27.3
Syndicated Loan	9.5	8.2
Development	0.2	0.4
Others	0.1	0.3
Gender		
Female	30.0	36.3
Male	70.0	63.7
Profession		
Employed	98.0	95.4
Self-Employed	2.0	4.6
Marital Status		
Married	31.4	37.0
Single	65.1	61.2
Divorced	3.5	1.8
Education		
Education	98.2	98.9
No Education	1.8	1.1
Tenor		
0-25	30.5	57.4
26-51	66.2	41.1
52-77	4.3	1.5

4.1.2 Multivariate Regression

Table 3 below results show that borrowers who are divorced are statistically significant, with an odds ratio of 0.11. Borrowers who are male have a statistical significance of 5%, the odds ratio is 0.01. The results on profession imply that employment is statistically significant at a 10% level of significance, and the odds ratio is 0.05. The factors monthly income, total loan amount, and age are statistically insignificant. The table below summarizes the results

Table 3 Multivariate logistic regression model of factors associated with defaulting status

Defaults	Coef.	Std. Err.	Odds Ratio	P>z	[95% Conf.	Interval]

Marital status						
Married	-.0129 09	.0073981	.7775185	0.081	-.0274125	.0015945
Divorced	-.2533185	.0733934	.1141431	0.001	-.3972018	-.1094352
Gender						
Male	.0133317	.0068957	.0133317	0.049	-.000187	.0268503
Profession						
Employed	.052433	.0222102	.052433	0.010	.0088912	.0959747
Monthly_Income						
100000 TZS - 1000000 TZS	.0027923	.0252896	1.042759	0.912	-.0467864	.0523709
Above 1000000 TZS	-.0073106	.0313492	.8776999	0.816	-.0687687	.0541475
Total_Loanamount						
100000 TZS - 1000000 TZS	-.0270317	.0351719	.5783217	0.442	-.095984	.0419207
Above 1000000 TZS	-.0215658	.0334087	.6382422	0.519	-.0870615	.0439299
Age_Group						
36 - 65	-.0018389	.0105803	.9690478	0.862	-.0225809	.018903
Above 65	.0074347	.0216409	1.15507	0.731	-.0349911	.0498604

4.2 Developing a Machine Learning Model for Prediction of Loan Default

To train the selected ML algorithms by using the dataset from the commercial bank, three features were selected using an extremely random tree classifier, then they were fitted to the three ML algorithms RF, KNN, and Gradient Boosting for predictions. Machine learning metrics were used for evaluations to find the best combination of features in determining the relationship between factors of borrowers and their associated loan default.

4.2.1 Training K-NN Algorithm Using Commercial Bank Loan Dataset

In training K-NN, the selected important features were added to the algorithm and then evaluated. Table 4 shows various evaluation metrics for the K-NN algorithm for the three features. The three features total loan amount, age, and monthly income were fitted to the K-NN model.

Table 4 Evaluation Metrics for K-NN

Model Name	Accuracy	Recall	Precision	F1-Score
K-Nearest Neighbor	71	78	68	70

The results shown in table 4 the accuracy, recall, precision, and F1-Score are presented in percentage. Figure 14 shows that the AUC- ROC is 79 percent.

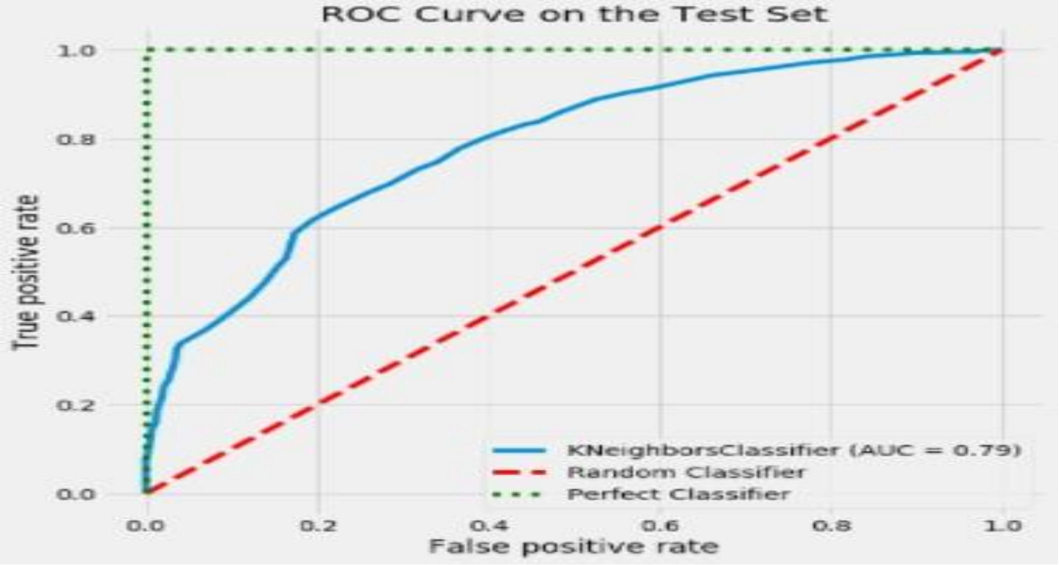


Figure 14 ROC Curve for K-NN

4.2.2 Training Gradient Boosting Using Commercial Bank Loan Dataset

In training GB algorithms, features were added according to their order of importance. The GB algorithm was fitted with three features. The first was age, monthly income, and total loan amount. After fitting GB with the features, metrics scores were found as indicated in table 5 below.

Table 5 Evaluation Metrics for Gradient Boosting

Model Name	Accuracy	Recall	Precision	F1-Score
Gradient Boosting	70	79	62	67

The results shown in table 5 the accuracy, recall, precision, and F1-Score are presented in percentage. Figure 15 below shows that the AUC- ROC is 79 percent.

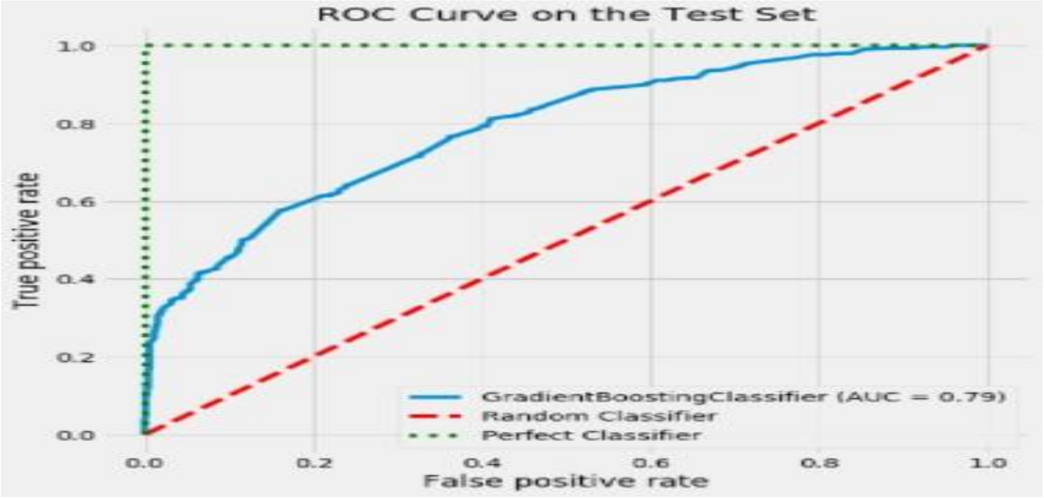


Figure 15 ROC Curve for Gradient Boosting

4.2.3 Training Random Forest Algorithm Using Commercial Bank Loan Dataset

The RF algorithm was fitted with three features during training. The first was age, monthly income, and total loan amount. After fitting RF with the features, metrics scores were found. The evaluation metrics used were Accuracy, Precision, Recall, F1-Score, and AUC-ROC as indicated in table 6. The result indicates that as features increase the score of evaluation metrics increases.

Table 6 Evaluation Metrics for Random Forest

Model Name	Accuracy	Recall	Precision	F1-Score
Random Forest	84	85	82	84

From the table above are the results of evaluation metrics for a random forest classifier in percentage. The figure 16 below, we observe that the Area Under the Receiver Operating Curve (AUC ROC) is at a score of 91%.

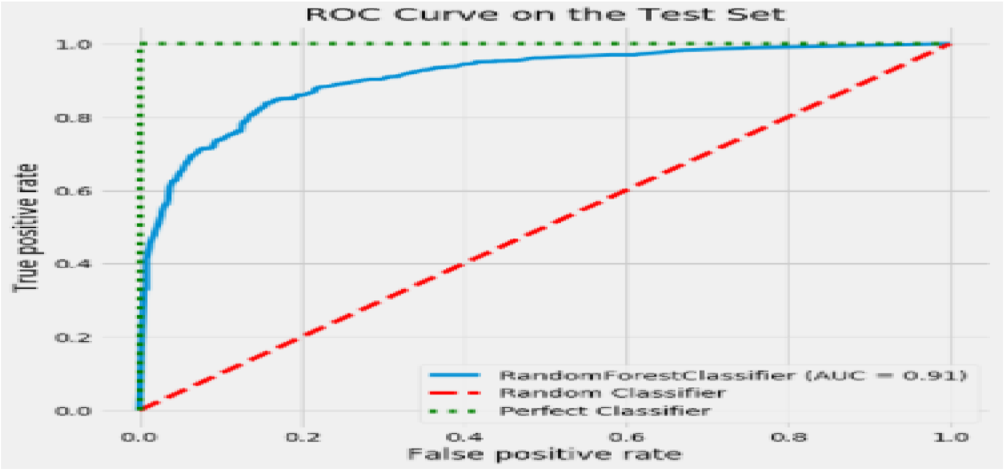


Figure 16 ROC Curve for Random Forest

4.3 Comparison of Machine Learning Model Performance

To find the best baseline model the three machine learning algorithms were compared using performance metric accuracy, recall, precision, ROC AUC and F1-Score as shown in figure 17

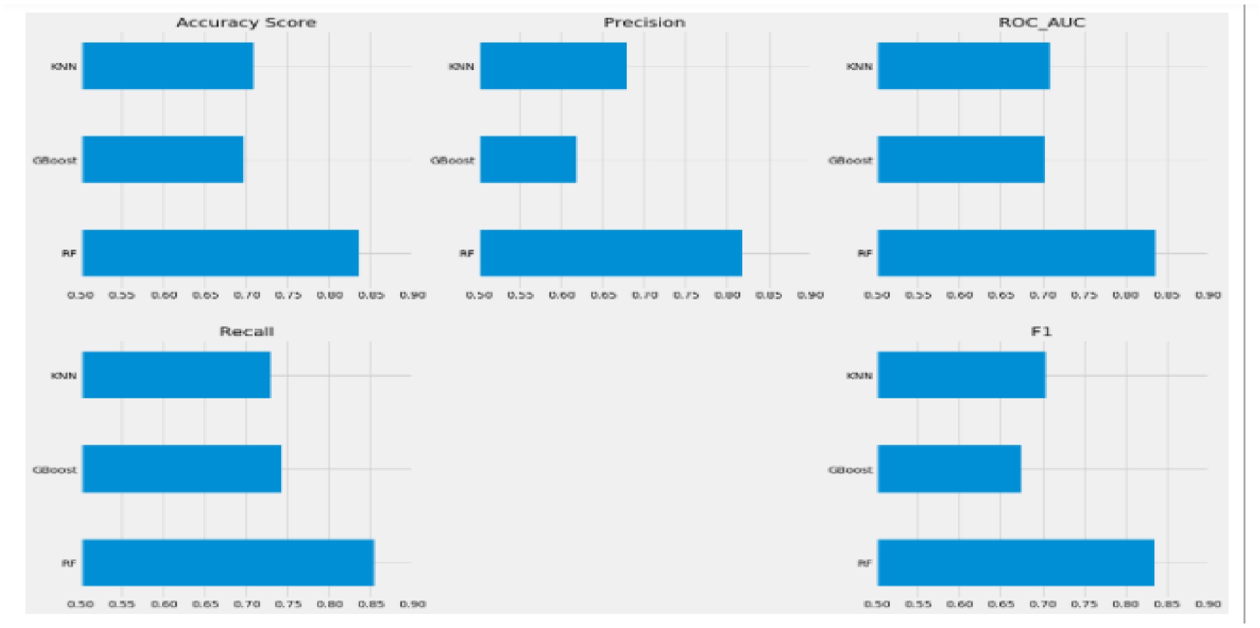


Figure 17 Comparison of all Models

CHAPTER FIVE

SUMMARY, CONCLUSION, AND RECOMMENDATION

5.1 Summary of Findings

The project aimed to learn insight into the applicability of machine learning algorithms for individual credit risk assessment on loan default on a Tanzania Commercial bank. With a broad goal of enhancing the use of machine learning algorithms, purposely for improving early loan application evaluation at the time a customer applies for a loan.

The findings are presented in an order of the study objectives which are to find borrower's variables that can predict loan default design a machine learning model that can be used to early predict loan default and to assess which model was most effective at doing so.

The findings correspond to the factors that influence loan default using Random Forest were age, monthly income, and total loan amount, the findings correspond to the study (Ngimbwa, 2020). Figure 13 shows that the importance of the features differs as age has the highest importance followed by total loan amount and monthly income respectively. The borrowers who have aged above 65 have the least chance of defaulting as most of them have settlements in business and life, compared to the borrowers who have aged below 65 as the majority of them still struggling with establishing businesses and still have less exposure to handle businesses. The findings on monthly income indicate that borrowers with a monthly income between 100,000TZS to 1,000,000TZS have the highest chance of defaulting compared to those who have above that, due to the factor that the monthly income have to have other expenses in their daily life such as rents, transportation, and other basic needs. Borrowers who took a total loan amount from 100,000 TZS and above have higher chances of defaulting this can be because some get tired of returning the loan for a higher range of time which may lead to loan default.

Findings of the factors influencing loan default using multivariate logistic regression in Table 3 were gender, marital status, and profession, it shows that males have a positive coefficient by holding all other factors constant, and the odds ratio and the significance level of 5% show that male borrowers are 0.01 times more likely to default on a loan. The borrowers who are employed have a positive coefficient of 10% level of significance and the odds ratio indicates that they are 0.05 times more likely to default on a loan. The divorced borrowers imply that borrowers are 0.11 times less likely to default and have a negative coefficient at the 1% level of significance.

The study has adapted techniques of machine learning and the model was trained using three algorithms namely K-NN, Gradient Boosting, and Random Forest. Findings corresponding to K-NN from table 4 mean that the classifier was correct in 68 percent of the cases it predicted as loan default. A recall score of 78 percent, on the other hand, represents the proportion of borrowers who were classified as default by the classifier. The harmonic mean of recall and precision is 70 percent for the F1 score. The closer the F1 score gets to 1, the better the model's performance. The 73 percent accuracy means that the model correctly predicted 73 percent of the cases as default or not default. Figure 14 shows that the AUC- ROC is 79 percent, which means that K-nearest neighbors have a 79 percent success rate.

From table 5 the findings corresponding to Gradient Boosting, show that the precision of the gradient boosting classifier is 62 percent, this implies that the classifier was correct in 62 percent of the cases it predicted as default. A recall score of 79 percent, on the other hand, represents the proportion of borrowers who were classified as default by the classifier. The harmonic mean of recall and precision is 67 percent for the F1 score. The closer the F1 score gets to 1, the better the model's performance. With a 70 percent accuracy rate, the algorithm correctly predicted 71 percent of the cases as default or not default. Figure 15 shows that the (AUC-ROC) is 79 percent, which means that the gradient boosting’s performance is at 79 percent.

In table 6 we observed the findings for the Random Forest classifier has a precision of 82%. It illustrates that 82% of the cases that the classifier predicted as defaulting were correct. A recall score of 85% on the other hand, is the proportion that the classifier has picked borrowers who were classified as default. The F1 score of 84% is the harmonic mean of recall and precision score. F1 score approaching 1 the better the performance of the model. The accuracy of 84% tells us the model accurately predicted 84% of the cases as either default or not default. We observe in figure 16 that the Area under the Receiver Operating Curve (AUC ROC) is at a score of 91%. This indicates that the performance of the random classifier is at 91%.

In figure 17, we compared the model by their evaluation metrics and performance measurement, we observed the best-performed model is the Random Forest since it out-performed K-NN and Gradient Boosting by having the highest evaluation metrics score, the scores were 84%, 82%, 85%, 84%, and 91% for accuracy, precision, recall, F1-score, and AUC ROC respectively. These results coincide with the study conducted by (Chang, 2019), who used four machine learning algorithms and RF had the highest accuracy of 70% in predicting loan default. Also, the study of

(Madaan et al., 2021) used a Decision tree and RF, where RF had a higher accuracy of 80%.

5.2 Conclusion

This study was carried out to develop a machine learning model for early prediction of personal loan defaulting before the provision of a loan to the loan applicant. The development and deployment of this technology are important in the banking sub-sector of Tanzania to enable banks to identify potential borrowers and their willingness to pay back the loan. Although loan officers can perform loan evaluation activities with great success, sometimes he/she may become biased toward loan applicants, get tired, or do it with minimum evidence and cause wrong assessment. To lessen errors and workload to loan officers, machine learning model solutions in loan application assessment are very important.

By comparing multivariate logistic regression which was used to check the factors influencing loan default, the study findings indicate that random forest achieves high performance in classifying whether a borrower will default or not before taking the loan. The study found that age, monthly income, and total loan amount can be used to predict in an early stage that the applicant will default. It enables accurate and quick results in loan assessment as compared to manual or statistical evaluation processes. The use of this model will require a little human intervention in data input as well as interpreting the results. This will save time and intensive work for loan officers in identifying the potential customer and by doing so banks will reduce operating costs and increase profits.

The study contributes a machine learning model which will be used by bank loan risk analysts such as loan officers to assess loan defaulting. Analysts will know whether a customer will be able to repay the loan or not before the provision of the loan, this will make them provide appropriate action or advice to the management. The model will enable individuals to make predictions on whether they will default before consulting banks for a loan. Therefore, conducting this work is of significance to individuals, government and banking sub-sectors.

5.3 Suggestions

The study reveals the importance of using machine learning in the financial sector, particularly in the banking sub-sector for early identification of loan applications that might default. The implementation of the model will help bankers in identifying potential customers. This is because the model identified in this study tends to concentrate on the properties of safer borrowers about his/her credit history and other borrowers, resulting in the reduction of losses due to the provision of bad debt. The reduced losses will increase profits for banks leading them to increase the provision of financial services to the society, even at a lower rate, and improve the standard of living of people in Tanzania.

Although the study managed to achieve its objectives successfully, it is far from drawing a generalization in the field (practical) manner due to some factors. For example, the use of a single banking institution and gathered dataset has the least reflection of banking sectors' behavior on credit risk management in Tanzania. In addition, the credit risk defaulting factor (dependent variable) can be affected by extraneous variables such as the survival probability of customers and financial literacy, which are not included in the dataset.

Despite the listed challenges of the study, it is evident that this study will pave the way for credit assessment of the banking financial sub-sector in Tanzania. Scholar of this study believe that the study will provide a wide view to banks on what to care about in ensuring effective provision of credit services and the smooth transition to the use of machine learning in credit risk assessment.

The study also suggests that banks should observe the quality of data being shared to help researchers build models that will perform well and should have systems that have built-in information which bankers can input customer information rather than type into the system which has proved to have erroneous in some information which were entered with a different keyword such as the Physical Address where there were more than 5 different inputs for Dar Es Salaam, features such as age there should be a built-in calendar, for numerical features such as total loan amount and monthly income a range should be provided with an exact input for the amount since there is an entry for a monthly income of a customer of more than Trillion Tshs.

In the future researchers may focus on increasing the size of the dataset, and parameters and retrain the model to attain proximal performance. This is because classical machine learning models require a high-dimensional dataset. Researchers may consider using other machine learning models such as deep learning to find out their performance with credit rating datasets in the banking sector of Tanzania. After obtaining a high-performance model the deployment could be done on mobile applications to enable society at large to conduct credit self-assessment.

References

Abrahamsson, J. &, & Granstrom, D. (2019). *Loan Default Prediction using Supervised Machine Learning Algorithms*.

Addae-Korankye, A. (2014). Causes and Control of Loan Default/Delinquency in Microfinance Institutions in Ghana. *American International Journal of Contemporary Research*, *4*(12).

Alpaydın, E. (2010). Introduction to Machine Learning, The Wikipedia Guide. *Natural Language Engineering*, *19*(2), 584.

Aslam, M., Kumar, S., & Sorooshian, S. (2020). Predicting likelihood for loan default among bank borrowers. *International Journal of Financial Research*, *11*(1), 318–328. https://doi.org/10.5430/ijfr.v11n1p318

Aziz, S., & Dowling, M. M. (2018). AI and Machine Learning for Risk Management. *SSRN Electronic Journal*, *January 2019*. https://doi.org/10.2139/ssrn.3201337

Bank of Tanzania, U. (2012). Goveerment Notice 416: The Bank of Tanzania (Credit Reference Bureau) Regulations, 2012. *The Bank of Tanzania (Credit Reference Bureau) Regulations*, 1–19.

Chang, H. (2019). *Loan Repayment Prediction Using Machine Learning Algorithms*.

Chawla, N. V, Bowyer, K. W., & Hall, L. O. (2002). *SMOTE : Synthetic Minority Over-sampling Technique*. *16*, 321–357.

ElMasry, M. H. (2019). *Machine Learning Approach For Credit Score Analysis: A Case Study Of Predicting Mortgage Loan Defaults*.

Kaaya, I., & Pastory, D. (2013). Credit Risk and Commercial Banks Performance in Tanzania: a Panel Data Analysis. *Research Journal of Finance and Accounting*, *4*(16), 2222–2847.

Kumar, A. (2022). *Feature Importance & Random Forest – Python*. Https://Vitalflux.Com/Feature-Importance-Random-Forest-Classifier-Python/. https://vitalflux.com/feature-importance-random-forest-classifier-python/

Leo, M., Sharma, S., & Maddulety, K. (2020). *Managing Operational Risk using Bayesian Networks: A practical approach for the risk manager*. *4*(6), 54–69.

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering*, *1022*(1). https://doi.org/10.1088/1757-899X/1022/1/012042

Mungure, M. E. (2015). *The Causes And Impacts Of Loan Default To Microfinance Institutions (Mfis) Activities The Case Of Pride Tanzania Ltd Pamba Branch-Mwanza*.

Ngimbwa, A. (2020). *Saccos credit rating prediction in Tanzania by using machine learning approach: A case of KKKT Arusha Road Saccos Ltd*. 33–37.

Opa, V. O. &, & Tabe-Ebob, W. T. (2019). *The Effects of Loan Default on Commercial Bank Profitability in Cameroon, Case Study BICEC Limbe The impact of devaluation on the growth of the Cameroon economy View project Loan Default View project The Effects of Loan Default on Commercial Banks Profita*.

Richard, E., Chijoriga, M., Kaijage, E., Peterson, C., & Bohman, H. (2008). Credit risk management system of a commercial bank in Tanzania. *International Journal of Emerging Markets*, *3*(3), 323–332. https://doi.org/10.1108/17468800810883729

Stephen Kingu, P., Macha, D. S., & Gwahula, D. R. (2018). Impact of Non-Performing Loans on Bank's Profitability: Empirical Evidence from Commercial Banks in Tanzania. *International Journal of Scientific Research and Management*, *6*(01). https://doi.org/10.18535/ijsrm/v6i1.em11

Torvekar, N., & Game, P. S. (2019). Predictive analysis of credit score for credit card defaulters. *International Journal of Recent Technology and Engineering*, *7*(5), 283–286.

Van Liebergen, B. (2017). Machine learning: A revolution in risk management and compliance? *Journal of Financial Transformation*, *45*, 60–67.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data. *Procedia Computer Science*, *174*, 141–149. https://doi.org/10.1016/j.procs.2020.06.069

Zhu, L. (2019). *Predictive Modelling for Loan Defaults*.

Effect of Machine Learning in Early Prediction of Personal Loan Defaulting on Commercial Bank in Tanzania (2009-2020)

ORIGINALITY REPORT

10 %	8%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	**dr.ur.ac.rw** Internet Source	1%
2	hdl.handle.net Internet Source	1%
3	link.springer.com Internet Source	1%
4	Submitted to University of East London Student Paper	1%
5	Submitted to University of Leeds Student Paper	<1%
6	www.mdpi.com Internet Source	<1%
7	Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain, Preeti Nagrath. "Loan default prediction using decision trees and random forest: A comparative study", IOP Conference Series: Materials Science and Engineering, 2021 Publication	<1%

8	**www.hund-um-schoen.de** Internet Source	<1%
9	etd.aau.edu.et Internet Source	<1%
10	Submitted to University of Rwanda Student Paper	<1%
11	Submitted to University of Central England in Birmingham Student Paper	<1%
12	repository.unam.edu.na Internet Source	<1%
13	Erepository.uonbi.ac.ke Internet Source	<1%
14	intjem.biomedcentral.com Internet Source	<1%
15	github.com Internet Source	<1%
16	ndltd.ncl.edu.tw Internet Source	<1%
17	openrepository.aut.ac.nz Internet Source	<1%
18	Submitted to Cardiff University Student Paper	<1%

19	Submitted to Higher Education Commission Pakistan Student Paper	<1%
20	Submitted to Ghana Technology University College Student Paper	<1%
21	Submitted to UNIVERSITY OF LUSAKA Student Paper	<1%
22	Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin. "A novel gene selection algorithm for cancer identification based on random forest and particle swarm optimization", 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 Publication	<1%
23	Yiqiang Chen, Meiyu Huang, Chunyu Hu, Yicheng Zhu, Fei Han, Chunyan Miao. "A coarse-to-fine feature selection method for accurate detection of cerebral small vessel disease", 2016 International Joint Conference on Neural Networks (IJCNN), 2016 Publication	<1%
24	ugspace.ug.edu.gh Internet Source	<1%
25	Ashraf Ullah, Nadeem Javaid, Muhammad Asif, Muhammad Umar Javed, Adamu Sani	<1%

	Yahaya. "AlexNet, AdaBoost and Artificial Bee Colony Based Hybrid Model for Electricity Theft Detection in Smart Grids", IEEE Access, 2022 Publication	
26	Submitted to University of Hertfordshire Student Paper	<1%
27	www.oer.unn.edu.ng Internet Source	<1%
28	www.theijbmt.com Internet Source	<1%
29	Somayeh Akhavan Darabi, Babak Teimourpour. "chapter 19 A Case-Based-Reasoning System for Feature Selection and Diagnosing Asthma", IGI Global, 2017 Publication	<1%
30	Submitted to University of Surrey Student Paper	<1%
31	essay.utwente.nl Internet Source	<1%
32	www.slideshare.net Internet Source	<1%
33	clausiuspress.com Internet Source	<1%
34	text-id.123dok.com Internet Source	<1%

35	edocs.maseno.ac.ke Internet Source	<1%
36	eprints.utar.edu.my Internet Source	<1%
37	erepository.uonbi.ac.ke Internet Source	<1%
38	etd.astu.edu.et Internet Source	<1%
39	ir.jkuat.ac.ke Internet Source	<1%
40	repository.kemu.ac.ke:8080 Internet Source	<1%
41	ujcontent.uj.ac.za Internet Source	<1%
42	Rong Jiang, H. Tagaris, A. Lachsz, M. Jeffrey. "Wavelet based feature extraction and multiple classifiers for electricity fraud detection", IEEE/PES Transmission and Distribution Conference and Exhibition, 2002 Publication	<1%
43	Saqib Aziz, Michael Dowling. "Chapter 3 Machine Learning and AI for Risk Management", Springer Science and Business Media LLC, 2019 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		