National College of Ireland

A Study Using Machine Learning to Predict Loan Default in Nigerian Microfinance Banks

MSc Research Project
Data Analytics

Akintomiwa Tomisin Akinyemi
Student ID: x22137149

School of Computing
National College of Ireland

Supervisor: Dr. Anu Sahni

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Akintomiwa Tomisin Akinyemi
Student ID:	x22137149
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Anu Sahni
Submission Due Date:	14/12/2023
Project Title:	A Study Using Machine Learning to Predict Loan Default in Nigerian Microfinance Banks
Word Count:	6031
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	27th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Study Using Machine Learning to Predict Loan Default in Nigerian Microfinance Banks

Akintomiwa Tomisin Akinyemi
x22137149

Abstract

The Nigerian microfinance banking industry is active and ever growing. However, a big problem for many of the banks that work in this sector is dealing with people defaulting on loans. This research project sets out to identify loan defaulters using machine learning. After implementing seven classifiers and evaluating each result using accuracy, precision, and other metrics, the boosting classifiers, Random Forest and XGBoost came out on top as the best in detecting loan defaults. Both techniques had an accuracy of 80.10% and 82.06% respectively, and a precision rate greater than 75%.

***Keywords*- Nigeria, Microfinance banks, loan, default, machine learning, Random Forest, XGBoost.**

1 Introduction

There are a total of 64 financial institutions in Nigeria, including five discount houses, twenty-one commercial banks, and five bank that specialise in development finance. Over eight hundred percent of the nation's banking institutions are microfinance banks now operating[1]. The fact that the Central Bank of Nigeria is in charge of supervising all of these different financial organisations is just another indication of the tremendous development that has taken place in Nigeria's banking system. There have been recent indications that the market for credit and loans in Nigeria is showing signs of significant increase. One of the primary reasons for this is that these financial organisations are primarily concerned with catering to the diverse requirements of the general people in terms of loans and credit. In particular, mortgages and "buy-now-pay-later" (BNPL) financing are now widely recognised, which makes it possible for a bigger number of people from a variety of economic strata to have simple access to financial products. The growing percentage of loans that are defaulted on is, nevertheless, one of the most serious and ongoing issues that the majority of financial institutions are confronted with during their existence. These financial institutions are tasked with precisely determining the loan amount for each individual consumer while simultaneously reducing the likelihood of the customer defaulting on their loan.

Despite the fact that the majority of customers of microfinance banks are members of the lower economic class or have modest incomes, these institutions continue to provide loans of up to four million Naira to assist struggling homes and companies in the hopes of improving their financial conditions. When a bank participates in the loan market, the

[1] `https://www.cbn.gov.ng/supervision/inst-mf.asp`

overall operating risk of the institution is increased because the majority of its customers have a credit history that is either nonexistent or extremely imperfect.

Kuda MFB, Moniepoint MFB, FinaTrust MFB, and Renmoney are examples of financial institutions holding microfinance banking licenses, primarily providing loans for personal or commercial purposes. Kuda Bank incurred a loss of over 14 million dollars in 2021, as reported by TechCabal [2], with a significant portion of the loss attributed to the type of loan services offered by the bank.

The aim of this study is to ascertain the most optimal classification model out of the seven implemented, such as the random forest classifier, logistic regression, and decision trees, for accurately forecasting loan defaults based on different variables, including personal, socioeconomic aspects, and geographical location. These characteristics include customer income, loan type, age, and marital status, amongst others.

(Tumuluru et al.; 2022) applied Various machine learning algorithms such as Random Forest, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression to address the challenge faced by financial institutions in estimating the risk involved in providing a loan, proposing the use of machine learning methods to extract patterns from loan-approved datasets for predicting future loan defaulters. The random forest algorithm achieved a higher accuracy than the rest.

Therefore, the objective of this study is to utilise several machine learning techniques to determine the essential characteristics required for decision-making in Nigerian microfinance banks. The objective is to evaluate loan authorization and the approved loan amount, significantly mitigating the risk of loan defaults. The research will be carried out in Nigeria, taking into account the computational expenses associated with the models in order to choose the most efficient model.

1.1 Research Question and Objectives

This research seeks to answer the question:
"How good are machine learning models at identifying individuals who fail to repay loans in Nigeria?"

The research objectives include:

1. Evaluate and compare the effectiveness of several models, and subsequently select the most efficient one.

2. Tune the model's hyperparameters to observe its effect on the results.

3. Employ multiple performance measures in assessing the results of the model.

Section 2 provides a comprehensive literature overview of relevant studies. This evaluation offers a thorough examination of prior research and scholarly contributions that are relevant to the topic. The objective of this section is to provide a foundation for understanding the existing information on the issue and identifying any research issues that exist. Section 3 examines the research methods employed in the study. This includes the strategies, procedures, and analytical tools employed to investigate the study questions or objectives. The design specifications are provided in Section 4. These specifications provide an explanation of the specific objectives and requirements that will constitute the next phase of the research, which is the implementation phase. Section 5 speaks to

[2] `https://shorturl.at/rBHLW`

how the final phase of the project was implemented. This includes the models used and the splitting of the dataset, among other processes. The next two sections then address the evaluation of the implemented models, encompassing an analysis of the findings, deductions, and consequences of the investigation. The evaluation assesses the research's effectiveness and its impact by deriving conclusions and acquiring insights from the collected and analysed data throughout the investigation. The conclusion encompasses a summary of the research findings, addresses the research questions and their resolution, and explores the study's potential for future works.

2 Related Work

Lending operations play a crucial role in the financial ecosystem as they facilitate the provision of funds to individuals, organisations, and governments for various purposes, hence promoting economic growth and stability. These operations as stated by Mishkin and Eakins (2019) comprise the methods by which financial institutions extend credit or loans to borrowers, thereby facilitating the flow of capital and investment. Lenders are able to make decisions on loan terms and interest rates when they have a thorough awareness of these aspects, which in turn reduces the likelihood of default and financial loss.

There is a wide range of loan options available, including those that are tailored to the needs of individuals, businesses, and municipalities. Unlike personal loans and credit cards, consumer loans are designed to meet the specific requirements of the borrower. On the other hand, commercial loans are designed to provide funding for organisations' day-to-day operations as well as their capital expenditures Saunders and Cornett (2008). The emergence of fintech platforms has completely transformed the lending industry with the introduction of peer-to-peer lending and online lending. Lending operations have a significant impact on the economy. In addition, well-regulated lending operations contribute to financial stability by ensuring that both lenders and borrowers adhere to prudent risk management practises Casu and Gall (2016). Effective regulation and prudent lending practices are crucial for ensuring that lending operations have a beneficial impact on both borrowers and the overall economy.

The growing demand for effective regulation that safeguards the interests of both borrowers and lenders has led to a rise in the adoption of machine learning in traditional lending practices.

2.1 Most Common Algorithm Used in Past Works

A very common machine learning algorithm used in previous research projects is Logistic Regression. According to Sheikh et al. (2020) Logistic Regression is a well-known machine learning algorithm for predictive analysis because it describes the data correlations between independent binary variables and independent nominal, ordinal, and ration level variables. Logistic regression is a common algorithm for predictive analysis. When it comes to classification issues, it is the most useful.

The table 1 shows the breakdown of 4 related papers where Logistic Regression was used as the major machine learning technique. As one of the recommendations from Sheikh et al. (2020) which spoke about adding more personal features of the borrower, like his/her gender and marital status to enhance the model, authors Gupta et al. (2020) and Sujatha et al. (2021) have included similar features to be used for their research

project and some of these features have also been identified in a report by Manglani and Bokhare (2021). Inclusion of these features saw an increase in the accuracy of the model from 81% to an average of 82%.

Model Used	Features	Results	Recommendation	Authors
Logistic Regression	Age, Purpose, CreditHistory, Amount	81%	Recommended that the borrowers gender and marital status should be included as one of the features	Sheikh et al. (2020)
Logistic Regression	Income, Loan Amount, Loan term, Credit History	97%	Combine multiple ML algorithms	Manglani and Bokhare (2021)
Logistic Regression	Gender, Marital Status, Dependents, Education, employment status, Income, loan amount	84.50%	Further research should take into consideration the country parameters	Sujatha et al. (2021)
Logistic Regression	Gender, Marital Status, Education Level, Income, loan amount, credit history	80.50%		Gupta et al. (2020)

Table 1: Table to show related works that used Logistic Regression algorithm

Several articles have examined the performance of logistic regression and other machine learning methods in predicting loan defaults. A comparative study was conducted to assess the effectiveness of logistic regression, decision tree, and K-nearest neighbour models in loan forecasting. The results indicated that the K-nearest neighbour model outperformed the decision tree model (Nagashree; 2023). A further investigation using logistic regression models to forecast loan defaults and discovered that incorporating personal characteristics of clients, alongside checking account data, enhanced the precision of the model Diwate (2023). (Zhou; 2023) also conducted a research to assess the predictive accuracy of logistic regression and random forest models in forecasting loan defaults. The results revealed that the random forest model outperformed the logistic regression model in terms of accuracy. These findings indicate that various machine learning models may exhibit variable degrees of accuracy in predicting loan defaults depending on the attributes of the dataset, with certain models outperforming others.

2.2 Other Algorithms Identified From Previous Studies

Other machine learning techniques, such as the Random Forest classifier, the Naive Bayes algorithm, and the Decision Tree algorithm in particular. A number of researchers have also investigated the use of artificial neural networks for the purpose of classifying loan defaults.

In their comprehensive study, Madaan et al. (2021) compared the Random Forest algorithm with the Decision Tree algorithm for the purpose of evaluating loan applications based on the characteristics of the applicants. According to the findings of the study, loan applicants without homeownership who were seeking to fund small companies or weddings displayed a significant chance of defaulting on their loans. Bhardwaj (2020), on the other hand, centred his research on the prediction of charged-off loans in online P2P banking by utilising borrower and loan attributes. GridSearchCV, logistic regression, a random forest classifier, k-nearest neighbour, and an artificial neural network (ANN) were all utilised in this investigation. Notably, the length of the loan, the total debt-to-income ratio, the interest rate, and the FICO score all revealed as important factors impacting the outcomes of loans.

Research done that considered the credit history of the borrower proved to better the model and yield higher accuracy results. Using features like gender, employment status, credit history, and loan amount, Rajesh et al. (2023) applied Random Forest classifier to get a model with 98.40% accuracy, whereas research with similar features was done by Tumuluru et al. (2022) without considering the borrower's credit history and the accuracy value dropped to 81%. Meanwhile Kadam et al. (2021) got an accuracy of over 80% using Naive Bayes machine learning algorithm. A conclusion was made that Random forest classifier is adaptable to future trends because it can be trained on a large array of features.

Naive Bayes has a good performance when implemented for loan prediction tasks, attaining a high level of accuracy in identifying eligible loan applicants. (Kavitha et al.; 2023a) employed it within a hybrid framework with other methods like Decision Tree and Support Vector Machine to enhance the model's accuracy. (Eweoya et al.; 2019) applied Naive Bayes in the forecasting of fraudulent activities in loan management which resulted in an accuracy rate greater than 75%. A further study conducted by (Riyadi et al.; 2022) shows a comparison between Naive Bayes and Support Vector Machine in forecasting non-performing loans, and Naive Bayes achieved the greater accuracy rate. This method also achieved a 95% accuracy rate in identifying eligible consumers for credit (Vedala and Kumar; 2012). In general, Naive Bayes has been shown to be good at loan prediction tasks, giving accurate results and making decision-making easier.

2.3 Research Niche

From the review done, it can be seen that most of the researchers do not take location of the applicant into consideration. Hence, this project will be gathering data from top microfinance banks in different parts of Nigeria. It only came up as a recommendation for future works from Ma et al. (2018) to future researchers as a feature to consider for further experiments. A comparative analysis will be done to determine if this factor affect the repayment of the loans offered by microfinance banks to low-income borrowers.

3 Methodology

This section covers important topics such as the selection of data, the pre-processing of data, the research procedure, the techniques that were employed, and the limits that were encountered during the course of this study.

3.1 Choice of Data

Loan data usually includes customer information and sometimes their transaction history could be included. Due to the sensitivity and data protection laws surrounding this type of information, most companies do well to safeguard this information. In general, there is a restricted amount of data that is available to the public in Nigeria.

To facilitate this project, data has been collected from **QORE** [3], a top financial technology firm in Nigeria that have a premium core banking automation that aid financial institutions in delivering products that take care of the day-to-day management and operations of banking.

The dataset consists of loan applications from various microfinance banks spread across the different geopolitical locations in Nigeria. The initial dataset consist of data gathered over 5 years from about 200 microfinance banks in Nigeria, however after careful consideration and analysis, the final dataset put together comprises 230,000 applications with 17 distinct attributes including personal, transactional, and geographical details of the applicants. The data is yet to be transformed using PCA as that will be done during the pre-processing stage to protect customer's information and identity. The different attributes of the dataset have been listed within Table 2 below.

S/N	Parameter	Description
1	CustomerID	Unique id to identify bank customer
2	FirstName	Bank customer first name
3	LastName	Bank customer last name
4	PhoneNumber	Bank customer mobile contact number
5	CustomerAge	Bank customer age
6	Gender	Bank customer gender
7	DOB	Bank customer date of birth
8	Address	Bank customer home address
9	City	Customer's resident location
10	MarriageStatus	Bank customer marriage status
11	EmploymentStatus	Bank customer employment status
12	BankBalance	Bank customer account balance
13	LoanID	Unique id to identify loan application
14	LoanAmount	Amount requested for loan
15	LoanTerm (Days)	Duration of loan in days
16	LoanRequestTime	DateTime the loan application was made
17	LoanStatus	Loan approval status (Repaid or Defaulted)

Table 2: Dataset Variable Description

[3] `https://qore.inc/`

3.2 Data Preprocessing

When working with loan datasets, it is very important to do data preparation to make sure that machine learning models are accurate and reliable. There are a few things that need to be done to the data in this step before the machine learning models can be built on it. This stage is highly essential since the quality of the information being used is a significant factor that has a direct impact on the ability of the model. The model can make better predictions when it cleans and changes raw loan data using pre-processing methods like handling missing values, scaling features, and storing categorical variables. Correct pre-processing operation reduces errors, makes models easier to understand, and stops them from overfitting. In addition, it helps find and fix outliers, which makes sure the model is trained on a good dataset. Financial institutions can make better choices, lower risks, and make their loan prediction models more efficient and fair by investing in strong data pre-processing.

For this project, the original dataset was large and would have required high computational power while also considering time constraint, hence the records with missing values were removed as opposed to using machine learning methods to fill them. This was done using Python programming language on Jupyter Notebook. To protect customer information, attributes like *'Firstname'* and *'Lastname'* containing personal information about the loan applicant were removed, hence leaving just the *'CustomerID'* for identifying the customer applying for the loan.

The final dataset containing 64,651 records was then saved to a csv file and read into Google Colab, a cloud based Integrated Development Environment (IDE) which provides better computational resources than the regular local system.

3.3 Data Transformation

Data transformation is an important part of data preparation process because it converts raw data into one that can be used for analysis, visualisation, and modelling. It makes the data better in terms of quality, structure, and readability.

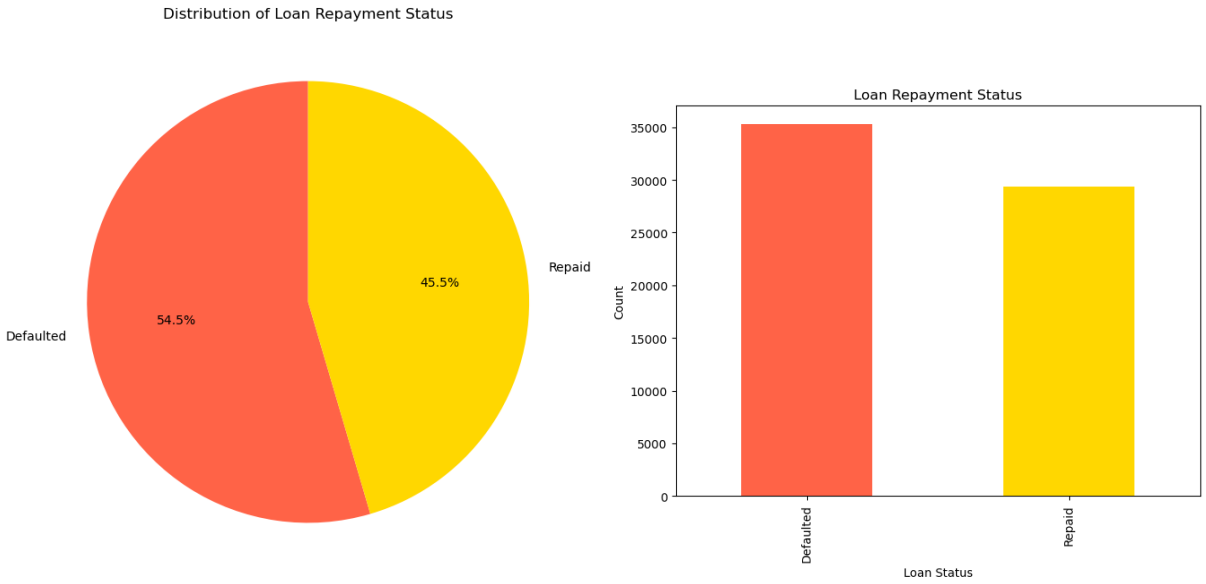
The first step taken was to encode all categorical attributes using LabelEncoder class from the sklearn.preprocessing Python package. This resulted to having numerical values for *LoanApproved*, *EmploymentStatus*, *MarriageStatus*, and *Gender*. Next, the *CustomerAge* column was converted from type float to integer to truncate the decimal parts. After the conversion, the original columns were dropped from the dataframe to avoid redundancy. Scaling the features was another method that was explored for a few of the variables, but because there was no significant effect, it was reversed.

3.4 Data Exploration and Visualisation

During the exploratory data analysis (EDA) stage, an investigation was conducted to analyse the correlation between the dependent variable and the independent variables. The purpose of this investigation was to uncover the relevant characteristics and offer a deeper insight of hidden trends within the dataset.

Figure 1 shows the analysis of class distribution which reveals a minor imbalance in the loan application dataset, with almost 55% of loan defaults, and 45.5% classified as repaid. Seeing as the difference is not much, the imbalance can be ignored.

While having a default rate of 60%, the relationship of the target variable with other variables was investigated in Figure 2 and Figure 3. The charts show the Male gender



(a) Target Variable Pie Chart (b) Target Variable Bar Chart

Figure 1: Class Distribution

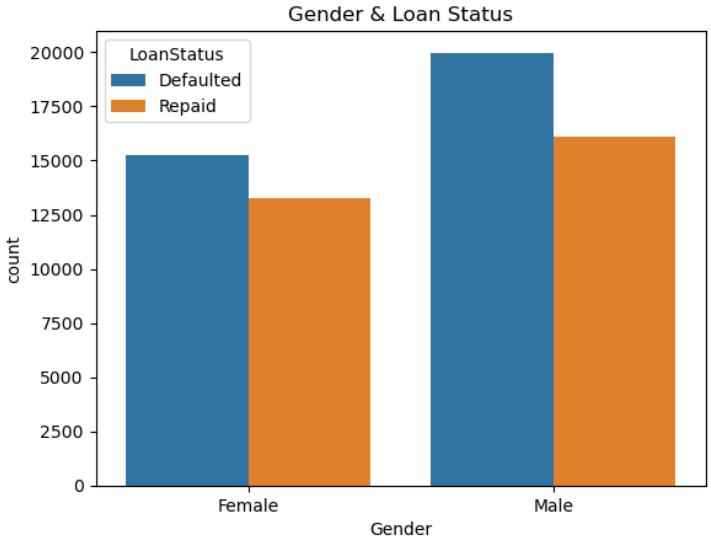


Figure 2: Distribution of loan status by Gender

had more requests so it's only relatable to have more repayments of almost 17,000 as seen in Figure 2. The SelfEmployed category has the least approved applications as seen in Figure 3 which shows the relationship between the applicant's employment status and loan status.

Also, the marriage status of the applicants were examined in Figure 4 and shows there were 6000 loans repaid by single applicants while applicants who are married had a repayment rate about 3.5 times more than that.

Finally, a correlation heatmap as seen in Figure 4 was plotted to show a visual representation of the correlation coefficients between the variables in the dataset. The map

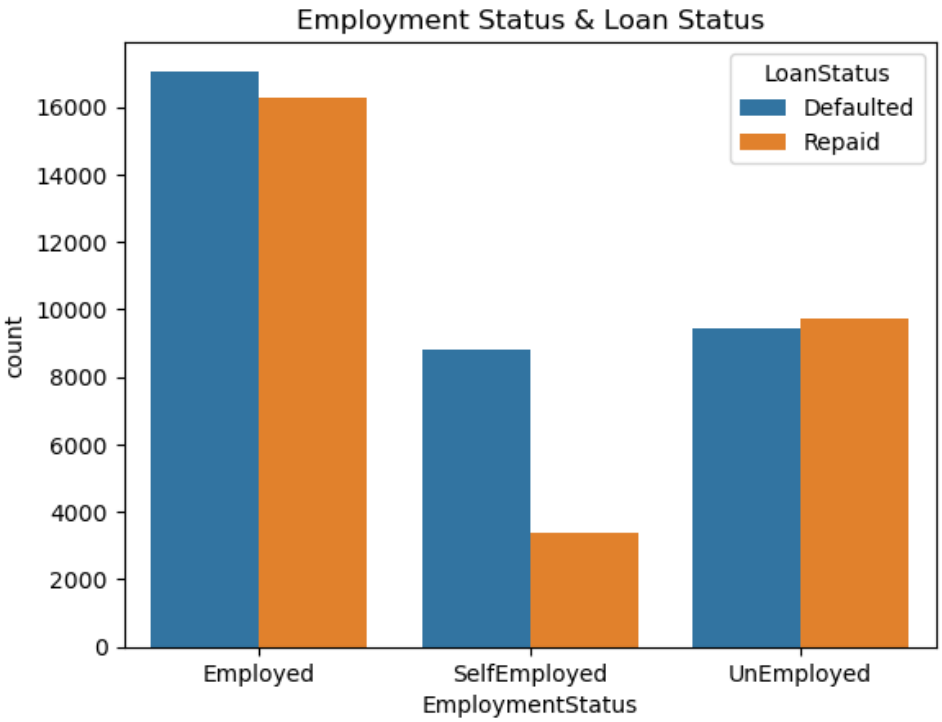


Figure 3: Distribution of loan status by Employment Status

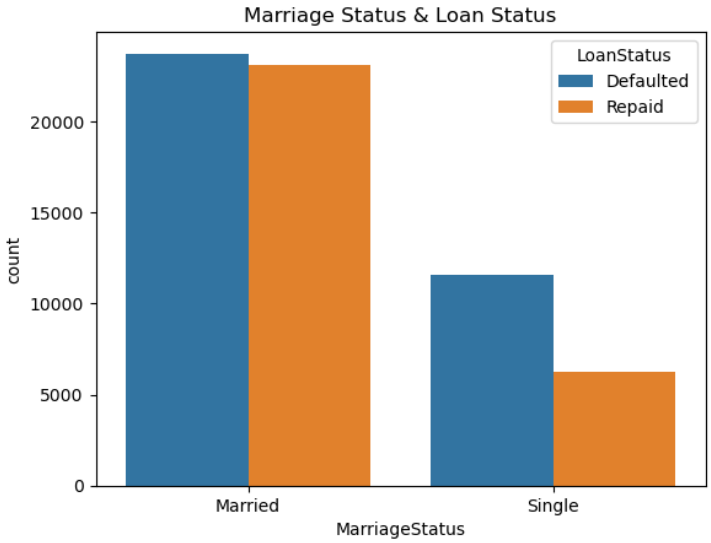


Figure 4: Distribution of loan status by Marriage Status

shows a moderate positive correlation for *CityCount* and *BankBalance*, with weak negative correlations for *CustomerAge*, while *LoanAmount*, *LoanTerm (Days)*, and *EmploymentStatus* are observed to have weak positive correlations. These correlations will be investigated later on to find their impact on the decision of loan approval.

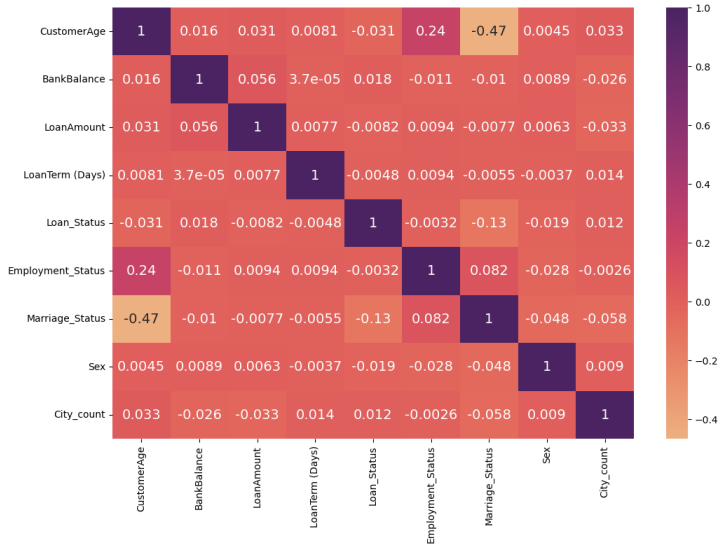


Figure 5: Correlation Heatmap

3.5 Data Modelling

In a data analysis project, data modelling is an important step that involves converting the relationship within a dataset into a mathematical model that can be used to make decisions or predictions. The fundamentals of machine learning involve a range of algorithms that facilitate the training of data and provide results that offer valuable insights for data analysis purposes. In this study, five machine learning algorithms are utilised to analyse the loan data.

3.5.1 Logistic Regression

Logistic regression is one of the most frequently used supervised machine learning algorithms. In addition to this, it forecasts numerous parameters by predicting the relationship that will exist between one or more independent variables. This algorithm will produce either a single value or a set of categorical values as its output (N and D; 2022). This estimation aids decision-making processes. With the probability estimate and the logit function, you can make an s-shaped curve that looks like a step function, resulting in a single numerical value Prakash et al. (2023).

3.5.2 Decision Tree

Decision tree algorithm in machine learning is a highly accurate and straightforward statistical model used in the banking industry for classification and regression tasks, with nodes representing diagnostic criteria, edges representing decisions, and leaves representing outcomes Gomathy et al. (2021). An increase in the quantity of sub nodes corresponds to an increase in the homogeneity and purity of the nodes that lead to the outcome of the target variable.

3.5.3 Random Forest

It is a member of the family of tree-based models and is well-known for the robustness as well as the high prediction accuracy that it possesses. Random Forest is an effective method for learning in ensembles, and it may be applied to issues involving classification as well as regression. The inherent limitations of decision trees, such as their susceptibility to overfitting and their sensitivity to particular data patterns, are amenable to being efficiently addressed through the utilisation of ensemble approaches. These methods aggregate the predictions made by numerous trees, which, in comparison to using just one tree, results in more accurate information. After training numerous decision trees on various subsets of data gathered by bootstrapping, the Random Forest model then proceeds to aggregate the learned results Zhang (2023). Each tree in the forest uses randomised data to evaluate case proximity, and this variation gives each tree in the forest its own distinct identity while yet ensuring that there is a consistent distribution of resources. The increase in the accuracy of the model can be attributed to the tree-based architecture of the classifier. The final prediction class is either the median of all the predictors or the standard of all of them Sheikh et al. (2020).

3.5.4 K-Nearest Neighbor Algorithm

In the field of machine learning, the K-Nearest Neighbours (KNN) algorithm is a flexible and straightforward technique employed for both classification and regression tasks. It belongs to the family of instance-based or lazy learning algorithms, which are characterised by the fact that the model is trained on the complete dataset, and predictions are generated based on the proximity of a new data point to points that already exist in the feature space. It is possible to swiftly sort data into a meaningful segment using the KNN algorithm, independent of the origin of the data to be sorted. In spite of the fact that the K-NN technique is usually utilised for editing and undoing, its primary purpose is to resolve issues that arise throughout the editing process. Since this method is non-parametric, it does not make any assumptions on the data that is really being used Tumuluru et al. (2022). The weight is determined in such a way that the closer neighbour has a greater influence on the prediction decision.

3.5.5 Gaussian Naive Bayes

The Naive Bayes algorithm are linear classifiers known for its simplicity but yet very efficient application. It encompasses a collection of algorithms that adhere to the principles of the Bayes theorem. It operates under a "naive" assumption that each attribute exerts an independent and equal influence on the outcome. The algorithm is commonly referred to as Naive Bayes due to its dependence on the principles of Bayes theorem, which enables the calculation of the likelihood of an event's occurrence based on the prior probability of another event (Kavitha et al.; 2023b). Over time this model has undergone significant modifications in the fields of statistics, machine learning, and pattern recognition. The advantages of this machine learning approach spreads across several key aspects. Firstly, it offers a significant reduction in computational time, hence, making it an efficient choice. Secondly, its model development is straightforward, which makes it easier to implement. Moreover, it is particularly well-suited for handling large datasets, enabling efficient processing of hidden insights within the data. Lastly, its robustness across various applications further makes it a good choice Eweoya et al. (2019).

3.5.6 Extreme Gradient Boost Algorithm

XGBoost is a popular machine learning algorithm that improves on conventional gradient boosting techniques. It is a decentralised, adaptable deep learning architecture for analysis, diagnosis, and rating scenarios based on gradient-boosted decision trees (GBDTs). XGBoost incorporates multiple tree models and builds additive models in a stepwise manner while reducing total error Mao et al. (2022). This leads to a group of base learners that perform better as a whole than as individual classifiers because of uniform contributions, shallow tree depths, and progressive refining. To improve resistance to noise and overfitting, a stochastic sampling system is added to the gradient boosting technique R et al. (2022).

3.5.7 Deep Neural Networks

Deep Neural Networks (DNNs) are a category of artificial neural networks characterised by their multi-layered structure, which enables them to acquire complex patterns and representations from input dataset. DNNs have played an important role in the recent advancements observed in the field of machine learning. The fundamental principles underlying deep neural networks (DNNs) encompass the idea of nodes, weights/biases, and activation functions, often organised into three layers. Hidden layers serve as an intermediary processing mechanism that connects the source and destination. While the neural network (NN) comprises two key components of importance: the neuron, serving as the processing element, with one or more inputs and a singular output; and the interconnections between these processing elements, which are denoted by weights Eletter et al. (2010). One of the drawbacks associated with Deep Neural Networks (DNNs), particularly when dealing with datasets with a large number of parameters, is the issue of overfitting the training data. In order to address the issue of overfitting, various regularisation strategies such as dropout and early stoppage are frequently utilised. In addition, a significant amount of computational resource is necessary to effectively process the dataset and generate a model. Bayraci and Susuz (2019) concluded that the accuracy of the model increased with the size and complexity of the dataset.

4 Design Specification

The data analytics project aims to analyze a complex dataset using machine learning techniques for prediction and classification, utilizing various technologies, computer programming languages, data visualization tools, and statistical software. The various steps involved in achieving this has been visualized in Figure 6 using a flowchart.

Some of the tools that were used to run the Python code were Jupyter Notebook and Google Colab [4], an online Integrated Development Environment (IDE).

Initially, the dataset was obtained from QORE Technologies after getting temporary access to their data warehouse. The data was extracted using SQL, a programming language for designing and managing relational databases, and exported in CSV format. However, after careful consideration of processing capabilities and time constraints, only a section of the data was used.

Subsequently, the CSV file, denoted as the "Initial Dataset," was imported into a Pandas dataframe within the Jupyter Notebook environment to facilitate further analysis

[4] `https://colab.google/`

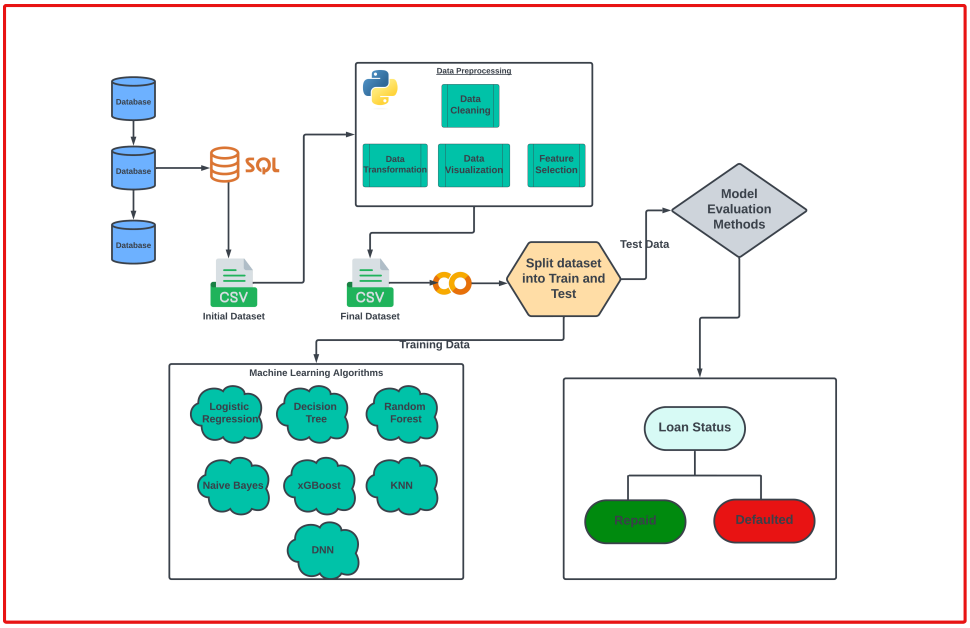


Figure 6: Design Flow Diagram

and preprocessing tasks. Furthermore, this facilitated the process of examining and visualising the dataset, thereby offering valuable insights.

Following the completion of the pre-processing stage, the resultant data was exported in CSV format and then imported into the Google Colab Integrated Development Environment (IDE) for the machine learning phase of the project. Loan default prediction was conducted using machine learning methods with the utilisation of the scikit-learn and keras libraries. Also, as part of this stage, the models were assessed by employing various evaluation metrics that were created from the classification reports. These metrics were utilised to evaluate the predictive capacities of the models and their ability in differentiating between a defaulted loan and a repaid loan. The outcomes of the models are subsequently provided using visual representations to allow accurate observations and facilitate better understanding of the outcomes.

5 Implementation

This section discusses the final implementation of the machine learning algorithms that were highlighted to be used for this project as seen in subsection 3.5.

The dataset used for this project consists of loan applications received from customers of multiple Microfinance banks in Nigeria, and parts of the dataset were randomly selected to be cleaned and prepared to train the models. This activity brought the final dataset to 64651 records and 9 attributes to indicate the personal and financial status of the applicants. The online IDE, Google Colab, was used for this step.

To initiate the process, many Python packages, including pandas, numpy, io, sklearn, and matplotlib, were loaded to facilitate data processing and model execution. One of the essential libraries utilised in this study was the *train_test_split* library from the *sklearn.model_selection* package. This library facilitated the division of the dataset into separate train and test sets, with a partition ratio of 80% for the former and 20% for

the latter. In addition, the *metrics* library was incorporated into the project to facilitate model evaluation. This library enables the visualisation of evaluation measures, such as the confusion matrix, accuracy, recall, and precision.

After partitioning the dataset, the target variable was extracted from each set and assigned to variables *y_test* and *y_train*, respectively. The training dataset consisted of 54.7% defaulted loans and 45.22% loans that were repaid.

An iterative process then began which included importing the requisite library for the model. Subsequently, the model was fitted using the training set, followed by making predictions using the trained model. The model's performance was then evaluated, and the results were visually represented through the utilisation of a confusion matrix and a ROC-AUC score graph.

To examine the impact of hyperparameter tuning on the performance of Random Forest, XGBoost, and DNN models, several iterations were conducted. The hyperparameters employed for each model have been outlined within Table 3.

S/N	**Model**	**HyperParameters**
1	Logistic Regression	Default Parameters
2	Random Forest	n_estimators = 100; n_estimators = 200
3	Decision Tree	criterion = "entropy", random_state = 100, max_depth=3, min_samples_leaf=5; criterion = "entropy", random_state = 300, max_depth=10, min_samples_leaf=5; criterion = "entropy", random_state = 300, max_depth=10, min_samples_leaf=10
4	KNN	Default Parameters
5	Naive Bayes	Default Parameters
6	XGBoost	objective="binary:logistic", random_state=42; learning_rate =0.1, n_estimators=200, max_depth=4, min_child_weight=6, gamma=0.3, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.005, objective= 'binary:logistic', nthread=4, scale_pos_weight=1, seed=27; learning_rate =0.1, n_estimators=1000, max_depth=4, min_child_weight=6, gamma=0, subsample=0.8, colsample_bytree=0.8, objective= 'binary:logistic', nthread=4, scale_pos_weight=1, seed=27
7	DNN	sequential, 3 layers, activation='relu', output: sigmoid; sequential, 5 layers, activation='relu', output: sigmoid

Table 3: Hyperparameters used for Classifiers

6 Evaluation

In order to determine which machine learning algorithms are the most effective in forecasting defaulted loans inside Nigerian microfinance banks, the purpose of this study was to assess and compare the effectiveness of these algorithms. Seven classifiers were trained and evaluated in order to accomplish this, and their performance evaluated using four metrics (accuracy, precision, recall, and roc auc score).

6.1 Experiment 1: Logistic Regression

The logistic regression model demonstrated an accuracy rate of 0.5363, indicating the percentage of correctly identified cases within the entire dataset. The model's precision, denoting its capacity to minimise false positives, was measured at 0.51, indicating that the accurately categorised repaid loans accounted for 51.12% of the total. Nevertheless, the model's recall rate of 1.52% indicates that it was only able to correctly identify a minute portion of the True Positives inside the dataset. The roc_auc score, which quantifies the balance between true positives and false positives, was shown to be 0.51, suggesting a moderate level of discerning ability exhibited by the model.

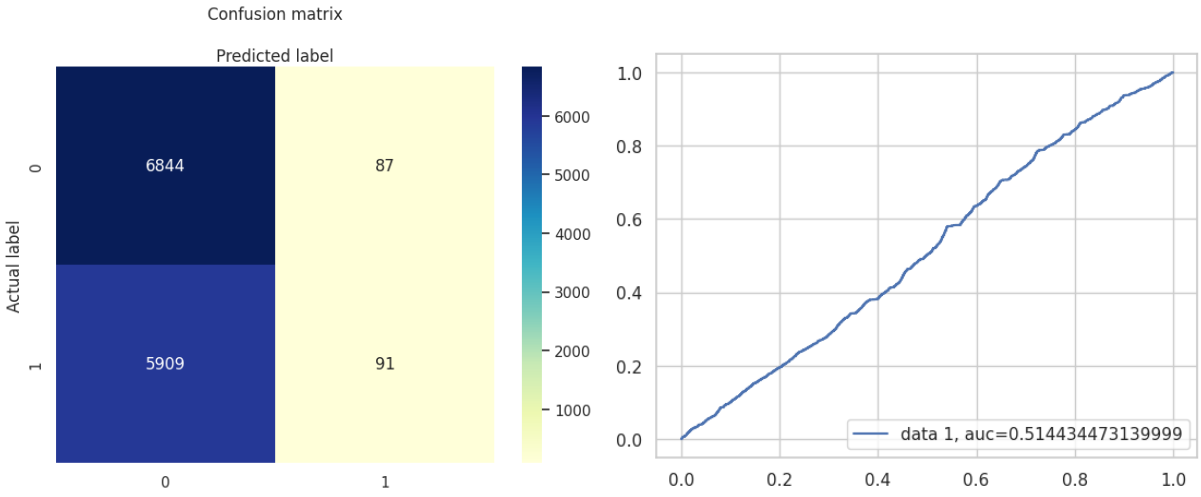


Figure 7: Logistic Regression Model Evaluation

Figure 7 illustrates a visual depiction of the performance of the model, encompassing two components: (i) the Confusion Matrix and (ii) the ROC Curve.

6.2 Experiment 2: Random Forest

The Random Forest model demonstrated an accuracy rate of 0.8010, a precision of 77.9%, a recall rate of 79.75%, and a roc score of 0.8793 as illustrated in the visualisations in Figure 8. This model performed better than the last one and shows the ability of the model to discern between True Positives and True Negatives.

6.3 Experiment 3: Decision Tree

As seen in the confusion matrix in Figure 9, the decision tree model was able to correctly identify 5576 instances of loans that were repaid, and 4754 of defaults from the test dataset. However, there were 1355 instances of False Positives, and 1246 of False Negatives,

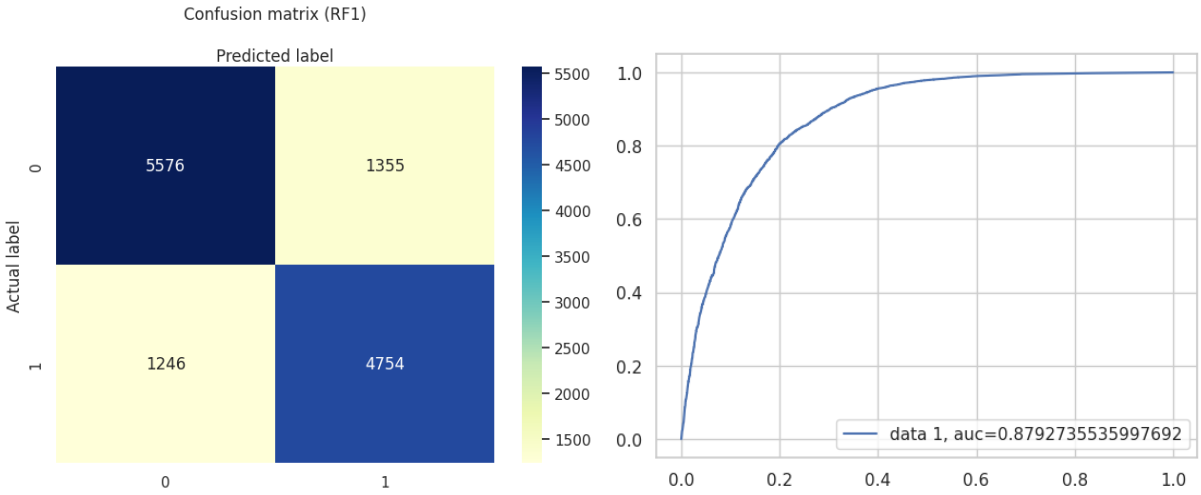


Figure 8: Random Forest Model Evaluation

hence resulting to an accuracy of 79.89%, precision of 77.82%, and a recall of 79.23%, which makes the model fair. However, there's still room for possible improvement.

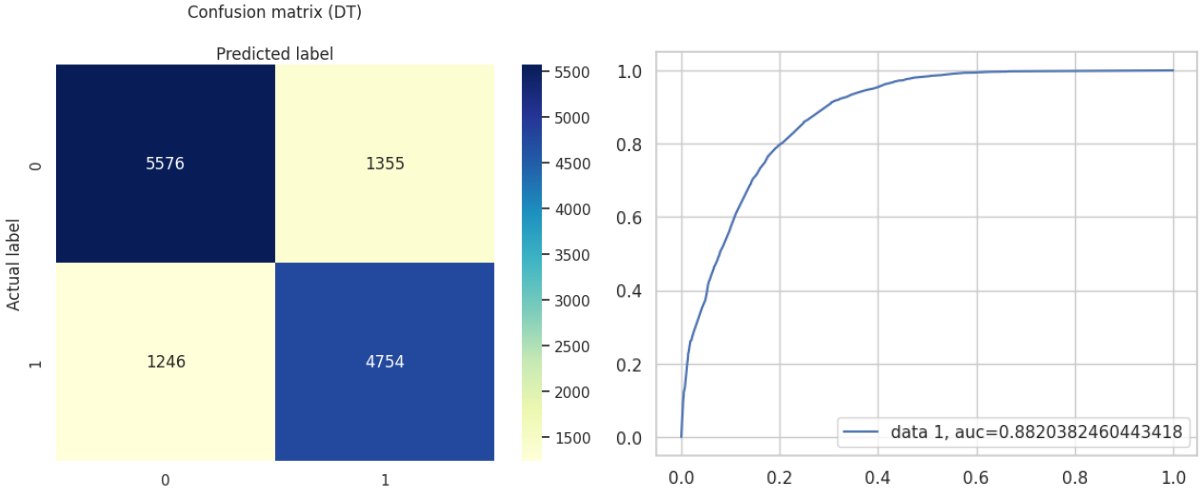


Figure 9: Decision Tree Model Evaluation

6.4 Experiment 4: K-Nearest Neighbor

The KNN model has exhibited a relatively low recall of 60.65% and precision of 61.59% which indicates an average likelihood of the model making false positive predictions. With an accuracy of 0.6419, it indicates the model was able to correctly identify loan defaults 64.19% within the dataset. The model also achieved a ROC AUC score of 0.6954, suggesting that the model is capable of distinguishing between defaulted loans and repaid loans to a moderate degree, but not exceptional.

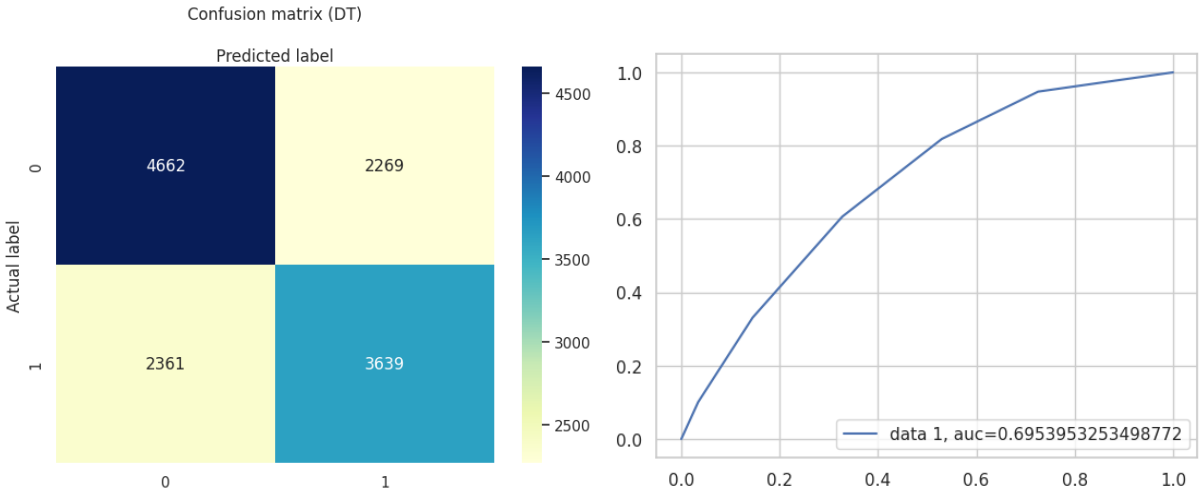


Figure 10: KNN Model Evaluation

6.5 Experiment 5: Naive Bayes

In this instance, the model attained an accuracy of 53.99%, showing the comprehensive accuracy of its prediction, however, with a recall rate of 4.53%, the model has a limited ability to correctly identify instances of loan defaults, highlighting the need for enhancements in detecting positive situations. The Naive Bayes model does a good job overall; its accuracy and precision show that its estimates are mostly right. It is recommended, however, that efforts be devoted on enhancing recall in order to guarantee that the model accurately identifies a greater number of actual loan defaults.

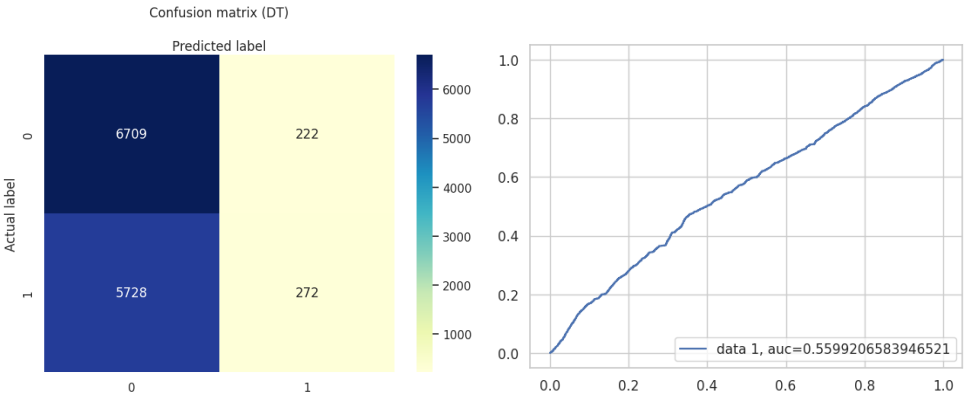


Figure 11: Naive Bayes Model Evaluation

6.6 Experiment 6: Extreme Gradient Boosting

This model's confusion matrix as seen in Figure 12 shows it correctly predicted defaulted loans 5045 times. Across a number of evaluation measures, the XGBoost model shows strong and even performance. Its high accuracy shows that it is generally right, and its precision and recall measures show how good it is at making correct predictions. With a precision of 78.72%, the model is correct approximately 79% of the time in predicting a loan default. Also the calculated F1 score of 81.29% reinforces the model's effectiveness, providing a balance between both precision and recall.

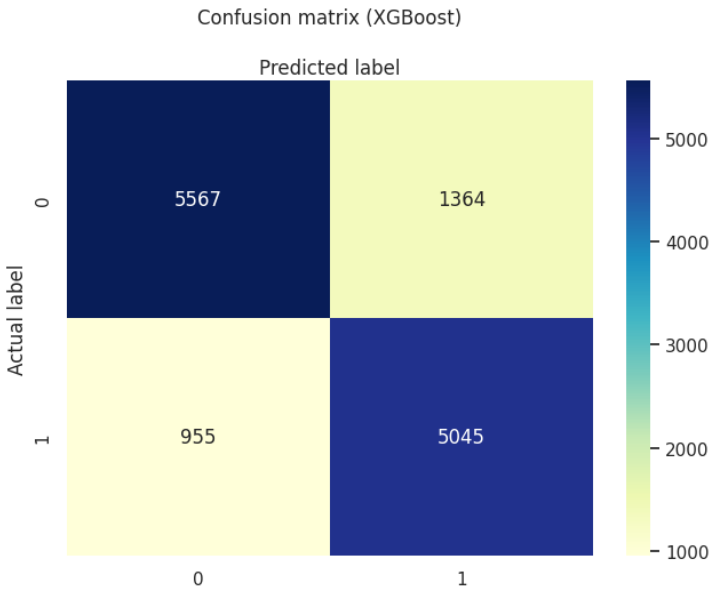


Figure 12: XGBoost Model Evaluation

The XGBoost model demonstrates exceptional performance in forecasting loan defaults, exhibiting high levels of accuracy, precision, recall, and a well-balanced F1 score. The tool's strong performance establishes it as a dependable instrument for spotting possible instances of loan default.

6.7 Experiment 7: Deep Neural Network

The Deep Neural Network of multiple layers achieved an accuracy of 53.60% suggests that the model is closely resembling guesswork or chance when making predictions about loan defaults, which is poor. Further processes can be explored to increase this value, either by performing more hyperparameter tuning, or including additional features.

6.8 Discussion

The outcome of these evaluations are summarized in Table 4 below.

Analyzing the results presented in the performance metrics of various classifiers provides valuable insights into their effectiveness in detecting and predicting loan defaults. Among the seven classifiers that were evaluated, the boosting methods, such as Random Forest, and XGBoost, performed better than the rest, hence, showing their ability to effectively identify defaulted loans. The recall rates, 0.7975 and 0.8408 respectively, suggests their ability to accurately categorise a significant proportion of actual loan defaults.

This is a critical aspect, as overlooking some defaults could lead to significant loss of funds amongst other negative implications for the microfinance bank.

While the measurement of recall holds significant importance, it is equally crucial to take into account other metrics, such as precision. The measure of precision relates to the level of accuracy in positive predictions. In this particular scenario, the XGBoost model demonstrated the highest precision, approximately 0.79. This implies that in cases where XGBoost classified a loan application as defaulted, it exhibited a higher level of accuracy in comparison to alternative models.

Model Used	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**roc_auc_score (%)**
Logistic Regression	53.63	51.12	1.52	51.44
Random Forest	80.10	77.9	79.75	87.93
Decision Tree	79.89	77.82	79.23	88.20
K-Nearest Neighbor	64.19	61.59	60.65	69.54
Naive Bayes	53.99	55.06	4.53	55.99
XGBoost	82.06	78.72	84.08	N/A
DNN	53.60	N/A	N/A	N/A

Table 4: Model Evaluation Outcome

Also, using the ROC AUC score which evaluates the model's capacity to effectively differentiate between positive and negative instances, even in scenarios where there is an imbalanced class distribution within the dataset. It gives a full and detailed picture of how well a binary classification model is working. It is a useful tool for testing discrimination and decision-making skills because it looks at the true positive and false positive rates at different levels. In this scenario, the Decision Tree and Random Forest models outperformed the other models as the showed a percentage greater than 85%.

Although all seven models performed relatively well, several limitations were faced in this research. The most prominent being the inability to use the whole dataset gotten from the provider, due to its size and the limited computing power used for the research. This could limit the ability of some of the models and therefore affect their evaluation.

7 Conclusion and Future Work

In conclusion, this research set out to address the question of how effectively various machine learning models can predict loan defaults within the Nigerian Microfinance banking space. To achieve this, a number of processes were involved from selecting key predictors, transforming the dataset, to the selection of key metrics to aid in the assessment of the models' performance.

Based on the models' performance results, the boosting classifiers were evidently better than the alternatives with their remarkable ability to detect loan defaults. These models displayed impressive accuracy, with Random Forest achieving a rate of 0.80 and XGBoost achieving a rate of 0.82. Also, their recall percentages were notably high, with Random Forest achieving 79.75% and XGBoost having 84.08%. These models exhibited notable precision, with the Random Forest model getting a precision score of 0.779, while the XGBoost model achieved a slightly higher precision score of 0.787. Furthermore, calculating the F1-scores for these models was also strong, with Random Forest and XGBoost achieving a percentage of 78.64% and 81.29%, respectively.Collectively, these measures highlight the ability of Random Forest and XGBoost algorithms in attaining a favourable balance between the identification of loan defaults and successfully repaid loans.

For future works, it is of the utmost importance to strengthen the robustness and applicability of our machine learning model for the prediction of loan default. Firstly, the introduction of new diversified datasets, geographical patterns, and macroeconomic variables has the potential to enhance the forecast accuracy and general acceptability of the model. Additionally, it is recommended to perform continuous monitoring of the model and to adjust it to changing financial landscapes in order to guarantee the model's durability and continued relevance. These many suggestions are working together with the goal of refining and improving the efficiency of loan default detection tools, which will ultimately contribute to a reduction in the amount of financial loss that lending institutions experience.

References

Bayraci, S. and Susuz, O. (2019). A deep neural network (dnn) based classification model in application to loan default prediction, *Theoretical and Applied Economics* pp. 75–84.

Bhardwaj, B. (2020). *Prediction of charged-off loans for p2p online banking using classification models and deep neural network*, Master's thesis, Dublin, National College of Ireland. Submitted.
URL: *https://norma.ncirl.ie/4433/*

Casu, B. and Gall, A. (2016). *Financial Structure of the Building Society Sector*, Palgrave Macmillan UK, pp. 61–77.
URL: *https://doi.org/10.1057/978-1-137-60208-4*$_4$

Diwate (2023). Loan approval prediction using machine learning, *International Research Journal of Modernization in Engineering Technology and Science* .

Eletter, S. F., Yaseen, S. G. and Elrefae, G. A. (2010). Neuro-based artificial intelligence model for loan decisions, *American Journal of Economics and Business Administration* **2**: 27–34.
URL: *https://thescipub.com/abstract/ajebasp.2010.27.34*

Eweoya, I. O., Adebiyi, A. A., Azeta, A. A., Chidozie, F., Agono, F. O. and Guembe, B. (2019). A naive bayes approach to fraud prediction in loan default, *Journal of Physics: Conference Series* **1299**(1): 012038.
URL: *https://dx.doi.org/10.1088/1742-6596/1299/1/012038*

Gomathy, C., Charulatha, M., AAkash, M. and Sowjanya, M. (2021). The loan prediction using machine learning, *International Research Journal of Engineering and Technology* **8**(10).

Gupta, A., Pant, V., Kumar, S. and Bansal, P. K. (2020). Bank loan prediction system using machine learning, *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, pp. 423–426.

Kadam, A. S., Nikam, S. R., Aher, A. A., Shelke, G. V. and Chandgude, A. S. (2021). Prediction for loan approval using machine learning algorithm, *International Research Journal of Engineering and Technology (IRJET)* **8**(04).

Kavitha, M. N., Saranya, S. S., Dhinesh, E., Sabarish, L. and Gokulkrishnan, A. (2023a). Hybrid ML classifier for loan prediction system, *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, IEEE.

Kavitha, M. N., Saranya, S. S., Dhinesh, E., Sabarish, L. and Gokulkrishnan, A. (2023b). Hybrid ml classifier for loan prediction system, *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 1543–1548.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q. and Niu, X. (2018). Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning, Vol. 31, pp. 24–39. **URL:** *https://www.sciencedirect.com/science/article/pii/S156742231830070X*

Madaan, M., Kumar, A., Keshri, C., Jain, R. and Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study, *IOP Conference Series: Materials Science and Engineering* **1022**(1): 2–5.

Manglani, R. and Bokhare, A. (2021). Logistic regression model for loan prediction: A machine learning approach, *2021 Emerging Trends in Industry 4.0 (ETI 4.0)*, pp. 1–6.

Mao, Q., Liu, G., Chen, Z., Guo, J. and Liu, P. (2022). Loan prepayment prediction based on svm-rfe and xgboost models, EAI.

Mishkin, F. S. and Eakins, S. G. (2019). *Financial markets*, Pearson Italia.

N, S. P. and D, V. (2022). An informative solution to predict and improve accuracy for approving bank loan using novel numerical and categorical data of the customer by comparing logistic regression over decision tree algorithm, *ECS Transactions* **107**(1): 14473. **URL:** *https://dx.doi.org/10.1149/10701.14473ecst*

Nagashree (2023). Loan default prediction using machine learning techniques, *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT* **07**(07).

Prakash, S. B., Amudha, V. et al. (2023). Efficient human action recognition using novel logistic regression compared over linear regression with improved accuracy, *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, IEEE, pp. 1–6.

R, P. B., K, A., Kumar, A., Rao, B., K, P. S. and P, S. A. (2022). An approach to predict loan eligibility using machine learning, *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp. 23–28.

Rajesh, D. M. V., Lakshmanarao, A. and Gupta, D. C. (2023). An efficient machine learning classification model for credit approval, *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp. 499–503.

Riyadi, S., Siregar, M. M., Margolang, K. F. F. and Andriani, K. (2022). Analysis of svm and naive bayes algorithm in classification of nad loans in save and loan cooperatives, *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)* **8**(3): 261–270.

Saunders, A. and Cornett, M. M. (2008). *Financial institutions management: A risk management approach*, McGraw-Hill Irwin.

Sheikh, M. A., Goel, A. K. and Kumar, T. (2020). An approach for prediction of loan approval using machine learning algorithm, pp. 490–494.

Sujatha, C. N., Gudipalli, A., Pushyami, B., Karthik, N. and Sanjana, B. N. (2021). Loan prediction using machine learning and its deployement on web application, *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pp. 1–7.

Tumuluru, P., Burra, L. R., Loukya, M., Bhavana, S., CSaiBaba, H. and Sunanda, N. (2022). Comparative analysis of customer loan approval prediction using machine learning algorithms, *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp. 349–353.

Vedala, R. and Kumar, B. R. (2012). An application of naive bayes classification for credit scoring in e-lending platform, *2012 International Conference on Data Science & Engineering (ICDSE)*, IEEE.

Zhang, Q. (2023). Loan risk prediction model based on random forest, *Advances in Economics, Management and Political Sciences* **5**(1): 216–222.

Zhou, Y. (2023). Loan default prediction based on machine learning methods, *Proceedings of the 3rd International Conference on Big Data Economy and Information Management, BDEIM 2022, December 2-3, 2022, Zhengzhou, China*, EAI.