INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Machine Learning Approach to Credit Scoring for Fintech Start-Ups Using Micro Finance Banks in Nigeria

Akinwunmi, Adeboye A.
Associate Professor, Department of Banking and Finance,
Achievers' University, Owo, Ondo State, Nigeria
Dare Festus Oluwafemi
Post Graduate Student, Department of Business Administration
Achievers' University, Owo, Ondo State, Nigeria

Abstract:
In the Nigeria FINTECH markets, lack of recorded credit history is a significant impediment to assessing individual borrowers' creditworthiness, deciding fair interest rates, and exposing this company to humongous risk. Thus, this research compares various machine learning algorithms (logistic regression model alongside Decision tree, Naive Bayes, Adaptive Boosting (Adaboost), Random Forests, K-Nearest Neighbour, and Gradient Boosting methods) on real micro-lending data of LAPO microfinance bank domiciled in Nigeria to test their efficacy at classifying borrowers into various credit categories. The results were validated with test metrics such as confusion matrix, accuracy, recall, precision, and area under the curve, which revealed that machine learning algorithms, can be used successfully to classify new customers into various risk classes, while some machine learning algorithm outperforms the others. Also, risk can be mitigated by ascertaining the creditworthiness of an individual applying for a loan. Finally, customers with no credit history should not be classified as high-risk customers. Generally, Random forest, Decision tree Classifier, and KN Neigbhour machine learning algorithms showed better performance with real-life data than others. The study also demonstrated that classifiers such as random forest algorithms can perform this task very well, using readily available data about customers (such as age, occupation, and gender). This presents an inexpensive and reliable means for FINTECH institutions around the developing world, especially Nigeria, to assess creditworthiness without credit history or central credit databases.

Keywords: *Machine learning, credit scoring, Fintech start-ups, microfinance, machine learning algorithm*

1. Introduction

One of the challenges of low-income and emerging economies like Nigeria is the high cost of credit. The high cost of borrowing and credit rationing gives rise to the financial exclusion of small borrowers such as Small- and Medium-sized Enterprises (SMEs) and households that play a significant role in the macroeconomy (Sahay et al., 2015). As a promising solution, modern technological advances have enabled new business models to employ modern data analysis techniques on big data and automate tasks to make credit decisions more efficiently. FINTECH credit promises to offer loans at a higher speed and lower cost. Therefore, granting loans to a fraction of the population results in elevated financial inclusion.

Credit scoring is a system used by creditors (banks, insurance companies, and FINTECH companies) to assign credit applicants to either a good credit group (the one that is most likely to repay the debt, or a bad credit group (the one that has a high possibility of defaulting on debt or any financial obligation, i.e., not paying within the given deadline).

Financial technology (FINTECH) is used to describe new technology that seeks to improve and automate the delivery and use of financial services. At its core, FINTECH is used to help companies, business owners, and consumers better manage their financial operations, processes, and lives by utilising specialised software and algorithms used on computers and smartphones. FINTECH is a combination of Financial Technology.

The financial markets in Nigeria are dynamic and spontaneous, which calls for constant monitoring and perpetual change of the firms' credit policies. Lending money to a bad client is not only costly to the firm but also a loss of equity for the stakeholders (Hooman et al., 2013). The loss has always been the failure to predict payment defaults before the event. Wehinger (2012) assumed that the financial crisis had brought other financial woes, such as fraud and scandals, to FINTECH start-ups in Nigeria. Such maladies have brought low confidence in the financial industry and raised anxiety about the structural flaws in the methods used by start-ups to function.

Credit risk assessment aids in objective decision making, deciding whether to lend or not and how much to charge for the loan. The construction and implementation of predictive models are powerful strategy tools (Moin and Ahmed, 2012). At the heart of modern predictive analytics are various machine learning algorithms that extract hidden insights from masses of data. The data may be multimedia data, text data, web data, time-series data, or spatial data (Moin & Ahmed, 2012). Harnessing this data at that scale helps the start-up make profitable decisions daily (Sudhakar et al., 2016).

In Nigeria, it has become rampant for people to receive calls from unknown operatives of FINTECH companies stating how someone collected a loan from their organisation and how they have to reach out to the person or else the individual and the client will be black-listed. Also, there have been cases where people who require money from FINTECH firms are denied loans or given a meagre amount of money because of a lack of credit history. Machine learning algorithms can recognise patterns in the data of existing borrowers, which can be used to predict the credit behaviour of a new customer with some level of accuracy. Several studies have been conducted on FINTECH companies, the creditworthiness of clients, and conventional mode credit worthiness ascertainment. This study attempts to fill a noticeable gap in adopting machine learning.

2. Review of Literature

2.1. Conceptual Clarifications

The methods used in credit scoring are evolving from traditional statistical techniques to more innovative approaches like artificial intelligence, which includes machine learning such as random forests, gradient boosting, and deep neural networks. The core philosophy of machine learning (ML) is to apply potentially complicated algorithms running on machines to learn patterns in data with the primary aim of making predictions. ML models are designed to analyse large amounts of information contained in data from various sources. These models can identify patterns in the data that standard econometric models cannot.

There has been a surge in recent years in the use of ML tools for estimating credit risk, especially since the establishment of Basel II, which called for the development of internal credit rating models by banks, and since the global financial crisis. However, internal credit rating models based on the standard linear econometric approach have been generally shown to exhibit poor performance in forecasting losses given default (Altman & Hotchkiss, 2010). Studies of credit risk show that while ML models outperform traditional models, their performance depends on the specific ML model and the sample data used in the analysis (Žliobaitė, 2017).

2.1.1. Traditional Credit Scoring Methods

Statistical discrimination and classification methods are the most prominent techniques used to develop credit scorecards (Hand & Henley, 1997). These include linear regression models, discriminant analysis, logit and probit models, and expert judgment-based models.

2.1.1.1. Linear Regression

Regression analysis is particularly useful in credit scoring because the statistical approach is easy to explain and predicts risk parameters, such as the probability of default. In linear regression, the label (dependent variable or target outcome) is projected onto a set of features (covariates or independent variables). Parameters that minimise the sum of squared residuals are chosen.

2.1.1.2. Discriminant Analysis

Discriminant analysis is a variation of regression analysis used for classification. The label is based on categorical data. The simplest variation is a label with two categories (for example, 'default' versus 'non-default'). The dichotomous linear discriminant analysis was originally developed by Sir Ronald Fisher in 1936 (Fisher, 1936). In default prediction, linear discriminant analysis was the first statistical method applied to systematically explain which firms entered bankruptcy based on accounting ratios and other financial variables. Altman's 1968 model is still a leading model in practical applications (Altman, 1968). The original Altman Z-score model, developed using data of publicly held manufacturers, was as follows:

$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$

Where:

- X_1 = working capital / total assets
- X_2 = retained earnings / total assets
- X_3 = earnings before interest and taxes / total assets
- X_4 = market value of equity / book value of total liabilities
- X_5 = sales / total assets

2.1.1.3. Probit Analysis and Logistic Regression

For the dichotomous label in credit scoring, there have been several efforts to adapt linear regression methods to domains where the output is a probability value instead of any infinite real number. Many efforts focused on mapping the binary range to an infinite scale and applying linear regression to these transformed values. In the probit model, an abbreviation for 'probability unit,' the inverse standard normal distribution of the probability is then modelled as a linear combination of the features (Salisu, 2016).

The logit function uses the log of odds, which is an abbreviation for 'logistic unit,' following the analogy for probit. In the logit model, the log of the odds ratio of the label is modelled as a linear combination of the features. The logit model is a popular model for estimating the probability of default because it is easy to develop, validate, calibrate, and interpret.

Rather than choosing parameters that minimise the sum of squared errors (as in ordinary regression), estimation in logistic regression chooses parameters that maximise the likelihood of observing the sample values.

2.1.1.4. Judgment-Based Models

Multiple methods may be employed to derive expert judgment-based models. One such is called the Analytic Hierarchy Process (AHP), a structured process for organising and analysing complex decisions. The AHP model is based on the principle that when a decision is required on a given matter, consideration is given to information and factors, which can be represented as an information hierarchy. The decision-makers decompose their decision problem into a hierarchical structure of more easily comprehended sub-problems, each of which can then be independently analysed. The prominent element of the AHP is that human judgments, not only the underlying information, be used to perform the evaluations. Human judgment is particularly critical in evaluating exceptions and instances that do not have precedence or are significantly underrepresented in the data. Bana et al. (2002) developed a categorical credit scoring model for business loans based on concepts of the AHP.

2.1.2. Artificial Intelligence (AI) and Machine Learning in Credit Scoring

The adoption of the term AI in modern times is attributed to John McCarthy, who is widely recognised as the father of Artificial Intelligence (AI). In 1956 during an academic conference on Artificial Intelligence in Dartmouth, McCarthy defined AI as 'the science and engineering of making intelligent machines'. In depicting AI, Allan Turin proposed the limitation game (Turing Test). Any computer that passes the Turing Test is therefore said to be intelligent (Kolade – Faseyi, 2021)

Artificial intelligence (AI) is an application of computational tools to address tasks traditionally requiring human sophistication (SAS, 2019). AI enables machines to learn from experience, adjust to new inputs, and perform human-like tasks (FSB, 2017). Most AI examples that are popular today— from self-driven machines to superhuman doctors— rely heavily on deep learning and natural language processing. These techniques leverage the ability of computers to perform tasks, such as computer vision and chatbots, by learning from experience. Today's evolving AI is made possible by rapid development in foundational technologies such as computing power, big data, and innovative algorithms.

By using these technologies, computers can be trained to accomplish specific tasks by processing and recognising patterns in the data, while the data could be of different types and from different sources. AI is a broad field, and machine learning is a subcategory. Machine learning can also be defined as a method of designing a sequence of actions to solve a problem, known as an algorithm, which optimizes automatically through experience with limited human intervention (SAS, 2019). These techniques can be used to find complex patterns in large amounts of data from increasingly diverse and innovative sources (SAS, 2019).

Deep learning is a form of machine learning that uses algorithms that work in layers inspired by the structure and function of the human brain (SAS, 2019). Deep learning algorithms can be used for supervised, unsupervised, or reinforcement learning. Recently, deep learning has led to remarkable results in fields such as image recognition and natural language processing. The deep learning approach was designed to mitigate the weakness of other machine learning algorithms (Onova & Omotehinova, 2021). The most prominent weakness is the popular 'CURSE OF DIMENSIONALITY', where the algorithm becomes less effective as the number of features it has to analyse becomes very large. However, deep learning provides a better option when it comes to working with data of complex features. For example, deep learning may be used to classify images, recognise speech, detect objects, and describe the content. Voice recognition systems are powered, in part, by deep learning.

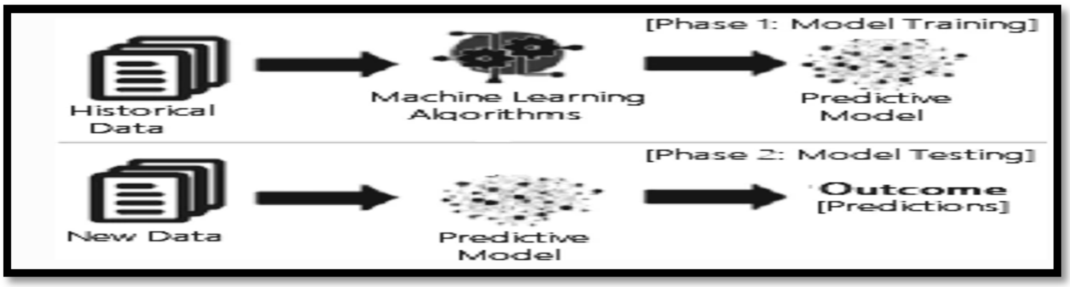

Figure 1: Overview of Machine Learning Process for Credit Scoring (Sarker, 2021)

At a high level, the application of machine learning algorithms for credit scoring involves the following high-level process shown in figure 1.

Innovative credit scoring methods include supervised, unsupervised, and reinforcement learning.

2.1.2.1. Supervised Learning

In supervised learning, the algorithm is developed using data that contain a label (dependent variable or event) and independent features (variables). The algorithm then predicts future or unknown values of the labels of interest, using features (independent variables). For instance, a data set of counterparties may contain labels on some data points identifying those that are default and those that are not in default. The algorithm will learn a general rule of classification that it will use to predict the labels for other observations in the data set. Some of the supervised techniques include regression, decision trees, random forests, gradient boosting, and deep neural networks.

2.1.2.2. Unsupervised Learning Techniques

Unsupervised learning refers to methods where the data provided to the algorithm do not contain labels (events). The algorithm is required to detect patterns in the data by identifying clusters of observations that demonstrate similar underlying characteristics, for example. In other words, rather than predict new or unknown data, these algorithms explore the properties of the data examined. Unsupervised techniques include clustering, K-means clustering, and hierarchical clustering (Hand & Henley, 1997).

2.2. Theoretical Framework

2.2.1. Supervised Learning

The algorithm is developed using data that contain a label (dependent variable or event) and independent features (variables). The algorithm then predicts future or unknown values of the labels of interest, using features (independent variables).

2.2.2. Decision Trees

Decision trees are typically schematic while showcasing a tree-shaped diagram used to show a statistical probability. Classification and regression trees (CART) are the most well-established supervised learning techniques. CART works by repeatedly finding the best feature to split the data into subsets. The partition improves the isolation of the label with each split. Decision trees can be used for either classification, for example, to determine the category of observation (that is, default or no default), or for prediction, for example, to estimate a numeric value (that is, the loss-given default).

2.2.3. Random Forests

Random forests are a combination of tree predictors such that each tree depends on a sample (or subset) of the model development data (or training data) selected at random (Breiman, 2001). Working with multiple different sub-datasets can help reduce the risk of overfitting. Random forests or random decision forests are ensemble methods for regression and classification problems. It relies on constructing a multitude of decision trees and outputting the class that may be the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

2.2.4. Gradient Boosting

Gradient boosting is an ensemble method for regression and classification problems. Gradient boosting uses regression trees for prediction purposes and builds the model iteratively by fitting a model on the residuals. It generalises by allowing optimisation of an objective function.

2.2.5. AdaBoost

Adaptive boosting, in short AdaBoost, is an ensemble algorithm incorporated by (Freund & Schapire, 1997). It is a model that trains and deploys trees in time series. Since then, it has evolved as a popular boosting technique introduced in various research disciplines. It merges a set of weak classifiers to build and boost a robust classifier that will improve the decision tree's performance and improve accuracy (Schapire, 2013)

2.2.6. Naive Bayes

It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple. Naive Bayesian model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

2.2.7. kNN (k- Nearest Neighbours)

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case assigned to the class is most common amongst its K nearest neighbors measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski, and Hamming distances. The first three functions are used for continuous functions and the fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing kNN modelling.

2.3. Empirical Review

According to Khandani et al. (2010), machine learning technique was employed for forecasting approaches that build the nonlinear nonparametric measure of consumer credit risk. Credit office data sets and commercial bank customer transactions were used to establish a forecast estimation that identifies credit cardholders' defaults.

There was a cost saving from 6% to 25% of total losses when machine learning forecasting techniques were used to estimate the delinquency rates. Besides, the research led to further questions of whether there may be an improvement if systematic risk estimation of aggregated customer credit risk is analysed.

Yap et al. (2011) identified potential club member subscription defaulters by using historical payment data from a recreational club and established credit scoring techniques too. From the study, no model outperforms the others among a credit scorecard model, logistic regression, and a decision tree model all generated almost identical accuracy figures.

Zhao et al. (2017) used a German credit dataset to train and estimate the multi-layer perceptron (MLP) neural network's accuracy of the credit scores efficiently. Despite containing nine hidden units, the results indicated an MLP model achieving a classification accuracy of 87%, higher than other similar experiments. Their study results proved the trend of MLP models' scoring accuracy by increasing the number of hidden units.

In Addo et al. (2018), Machine learning and deep learning techniques were employed to examine the credit risk scoring by incorporating ten key features, and the stability of the classifiers was tested through the evaluation of separate data performance. The findings revealed that logistic regression, random forest, and gradient boosting modelling were more accurate than neural network-based models incorporating various technicalities.

Wijewardhana et al. (2018) used historical data from a US-based collection agency to attempt debt repayment behaviour prediction of customers. It was pointed out that data sets taken for the research study are directly related to prediction accuracy. Thus, the centrality of discussions related to prediction study while assessing credit scores revolves around accessibility to proper relevant and adequate data. Key banking operations like fraud detection, credit assessment, customer churn prediction, etc., are the areas where banking institutions have experienced difficulty in coming up with an excellent ML-based algorithm. Data mining techniques even offer little help.

Boughaci and Alkhawaldeh (2018) evaluated German and Australian credit data sets and compared this with well-known classifier benchmarks. They used the local search method (LS), the stochastic local search method (SLS), and the variable neighborhood search (VNS) method combined with the support vector machine (SVM) model for the credit score assessment. The result obtained using the method was promising, showing an accuracy of 85 percent on the German dataset and 87 on the Australian data set.

Petropoulos et al. (2019) studied a dataset of loan-level data of the Greek economy to examine credit quality performance and quantification of the probability of default for an evaluating period of 10 years. The authors used an extended example of classifications of the incorporated machine learning models against traditional methods, such as logistic regression. Their results identified that machine learning models had demonstrated superior performance and forecasting accuracy through the financial credit rating cycle.

Assef and Steiner (2020) classified borrowers' adequacy. Researchers analysed the adequacy of the borrowers by using Brazilian Bank's loan database and explored various ML methods. Data sets are mainly comprised of low-income borrowers from large financial institutions in Brazil. The default rate of the portfolio was almost 48%. They developed an ML-based model based on real data and showed that RF and AdaBoost performed better than other models. A few authors recommended a decision tree model to classify the lender as a performing or non-performing loan risk. This researcher used the C5.0 algorithm (decision tree model) and recommended that if Indonesia's rural banks (Bank Perkreditan Rakyat) could have adopted this method, they could reduce their non-performing loan risk to a considerable extent.

Ozgur et al. (2021) have shown the impact of 19 bank-specific, macroeconomic, and global variables on bank loans for the period between 2002 Q4 and 2019 Q2 in Turkey. They compared the regression model with the ML-based methods to assess the impact of these factors. Authors further observed that the standard linear regression methods could not handle large dimensional datasets as compared to ML-based algorithms, and ML-based models have the flexibility to accommodate the complex nature of variables. Banking institutions mainly depend on third-party sources for their debt recovery management, which incur higher costs and market risks.

Hence, it is always recommended to have a strong debt repayment prediction method in place before disbursing any credit to the borrowers. Mathematically, data mining and statistical models are used to assess the debt repayment behavior of a customer with considerable accuracy. However, sample selection bias is the commonest issue for most research authors in consumer credit literature. Researchers have attempted to identify various factors to be considered by the rural bank for assessing credit applications. They used a decision tree model using a data mining methodology for a credit assessment to minimise non-performing loans.

They identified five discrete or non-continuous variables, gender, type of collateral, type of business activities, source of funding, credit status, and use of the loan, and eight continuous variables, age, monthly income, credit amount, and expenses per month, current payment per month, savings, collateral values, and loan period for the modeling phase of the data mining process. It has been found that collateral value is one of the most important factors to be considered by the rural banks for credit assessment for rural borrowers, specifically to minimise non-performing loans. Most of the researchers stressed that credit scoring is a classification problem.

3. Methodology

3.1. Data Collection

The data used in this research were obtained from LAPO Microfinance Bank (MFB), a micro-lending institution in Nigeria. It started operating in 1987. The data are an extract of information that LAPOMFB could make available on micro-loans from January 2016 to December 2020. A total sample size of 4000 customers was extracted, but 30 rows were entirely deleted due to many missing values in those rows.

3.2. Variable Identification

Variables or indicators that are typical in this type of credit scoring research are shown in table 1. The variables can be divided into four main categories: Demographic, Financial, employment, and behavioural.

Demographic Indicators	Financial Indicators	Employment Indicators	Behavioural Indicators
Age Sex	Housing Savings Account	Job	Credit Amount Purpose of the Loan Duration of the Loan

Table 1: Indicators Employed in Training the Algorithms

The demographic profile of the datasets consists of elements such as gender and age. In figure 2, the highest number of borrowers falls between the ages of 25-30. There are no borrowers below the age of 19, and there is a steady decrease in the number of borrowers as the age increases, with the only sharp fall occurring from the late thirties to early forties. There are no borrowers above the age of 75.

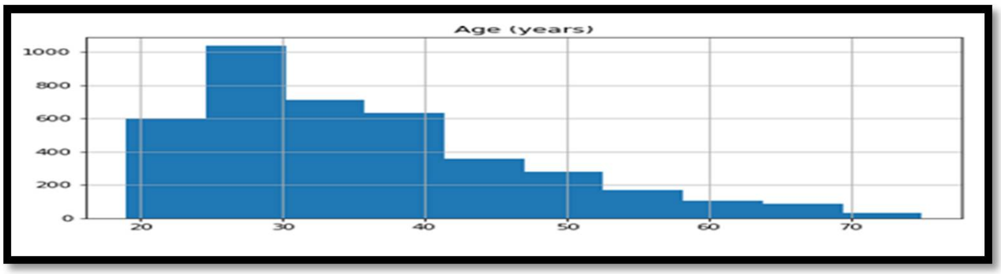

Figure 2: Distributions of the Ages

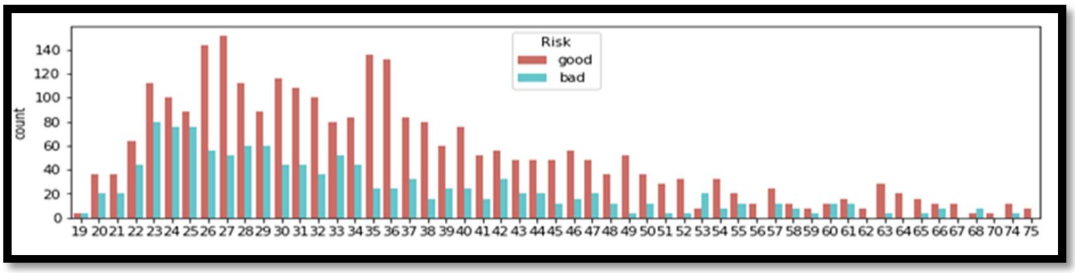

Figure 3: Age Count by Risk

The plot in figure 3 depicts the age count by risk. From the graph, it is more likely for borrowers between the ages of 24-25 to default. It was also observed that borrowers between the ages of 34-35 are good with repayment. In figure 4, it was seen that more men are inclined to loan as they outnumbered their female counterparts.

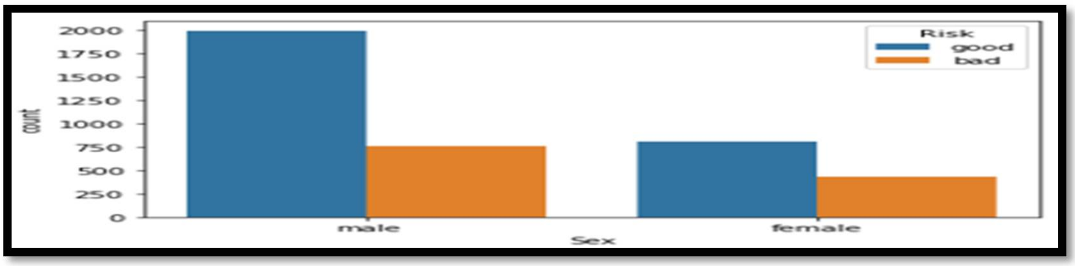

Figure 4: Gender Risk Distributions

Apart from the behavioural indicator of the datasets, the financial indicator comprises the savings of an individual and the kind of housing the borrower can afford presently. The saving account is categorised into little, moderate, quite rich, and rich. These depend on the preference of the Fintech Company. Figure 4 shows the distribution of the saving account of borrowers in the dataset. It is observed that people with little in their accounts borrow more than people with so much in their accounts. Moreover, people with little in their savings accounts are more prone to default.

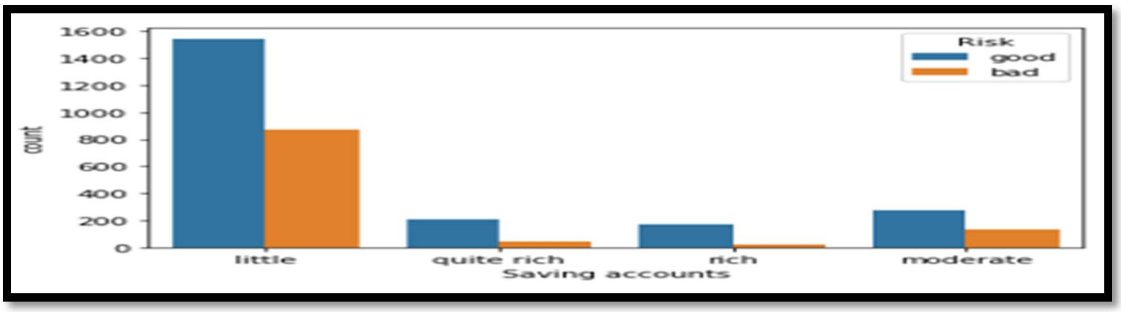

Figure 5: Saving Accounts Risk Distribution

Another Crucial Contributor to the financial indicator is the type of housing. It tells us where the borrower is domiciled, whether it is rented, or owed by the client. Figure 6 shows that people who own their apartment request for a loan the most. They account for more than half of the population and are followed by those in rented apartments.

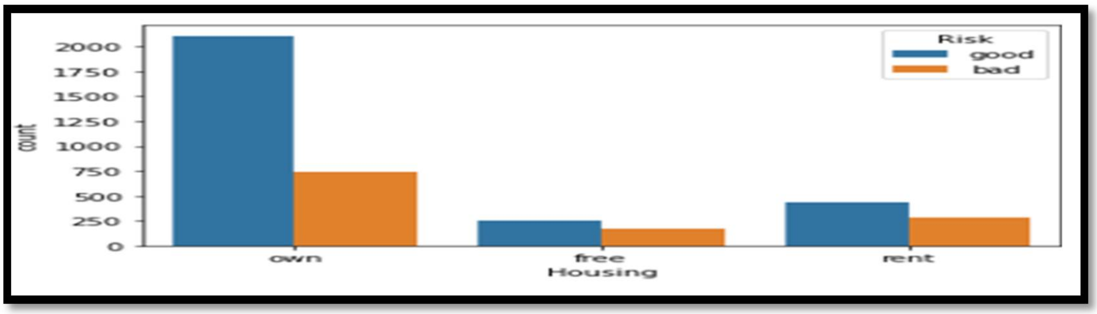

Figure 6: Housing Risk Distributions

The dataset also contained data about the job of the clients. The job is divided into unskilled, unskilled resident, skilled, and highly skilled. Figure 6 shows that most of the borrowers are skilled.

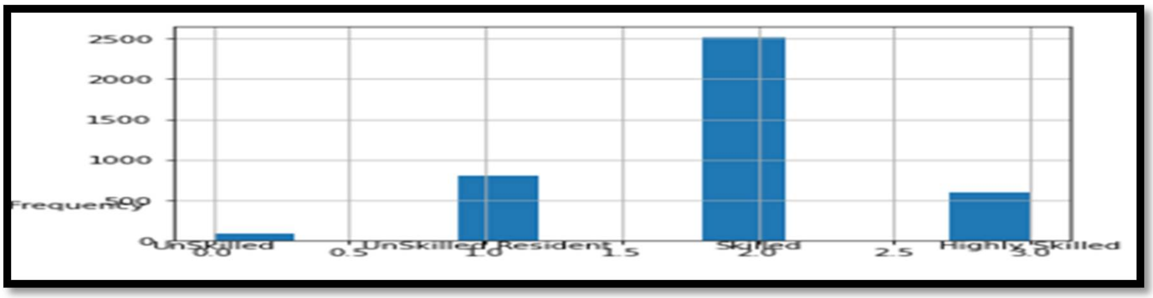

Figure 7: Job Distribution

The purpose of the loan, the amount of the loan, and the duration of the loan are all classified as a set of behavioural indicators. There are several reasons for applying for a loan. Most of the borrowers in this data are seen to be interested in acquiring a car, with more than half of that group of people returning the loan. Borrowing was also seen in figure 7 to satisfy the following needs:

- Purchasing a radio or TV,
- Education,
- Purchasing furniture or equipment,
- purchasing car,
- Business,
- Repair of domestic appliances or vacation

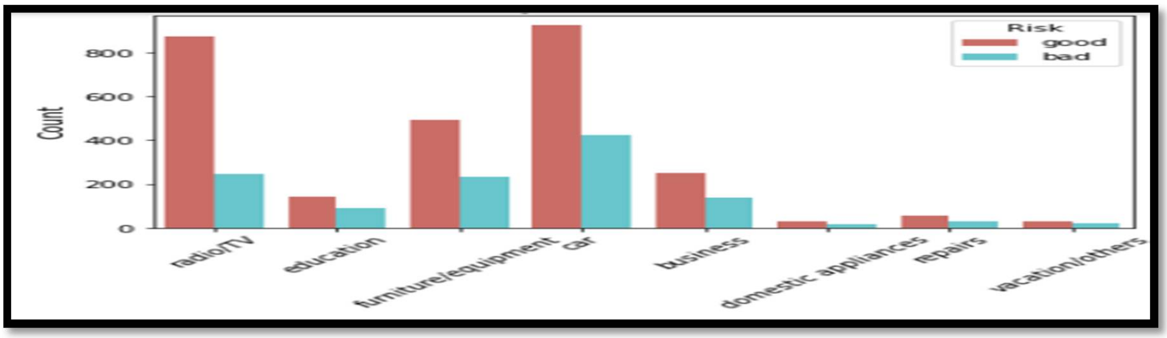

Figure 8: Job Risk Distributions

The duration of the loan is the length of time it takes for a loan to be completely paid off. In the dataset, the minimum number of months is 4, while the maximum is 72. More people complete payment within 12 or 24 months. High default rates were experienced at 12, 18, 24, and 36 months. Another important factor that was considered in the dataset is the credit amount. The amount ranges from twenty thousand naira to one million and six hundred thousand naira. Figure 9 shows a trend line depicting the amount of the loan disbursed with respect to the duration of the loan repayment. The trend line suggests that those with higher loan amounts are prone to default.

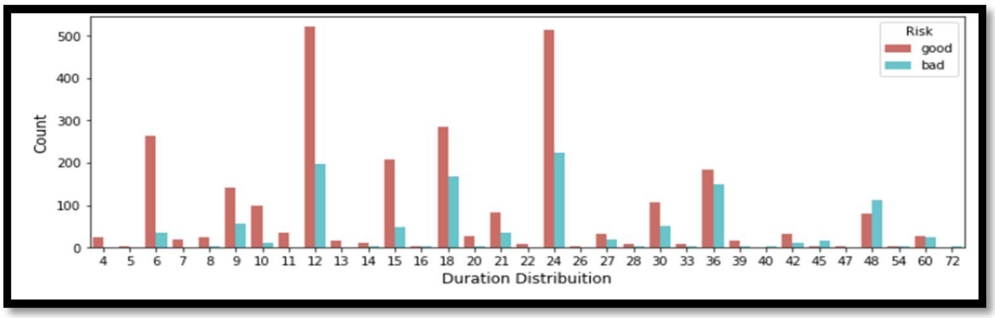

Figure 9: Age Risk Distribution

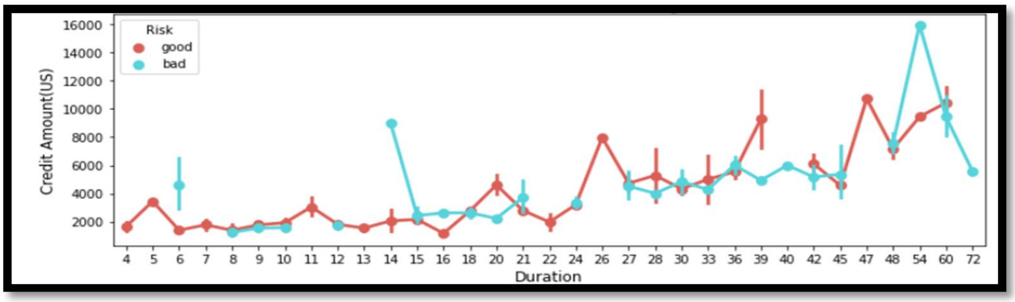

Figure 10: Credit Amount Distributions by Risk Duration

3.3. Training–Test Set Split

The data is partitioned into features and target (The target variable is generally an 'output' of the model. It contains the information on the available data that is to be predicted in future data. In credit scoring, it is commonly called good or bad). The data shows 2800 good credit target classes and 1200 bad credit target classes. In the case of this work, the 4000 dataset is splinted into training and validation sets. The training set has a feature set that is presented by the X_train and the target represented by y_train, while the validation set that is used for validating the model is divided into X_test and y_test.

3.4. Data Balancing

The dataset has 2800 training target values (y train) that have 1973 Non-defaults and 873 bad defaults. It implies an unbalance dataset that requires the application of a balancing technique known as SMOTE (synthetic minority oversampling technique) balancing technique which increases the number of samples of the smallest class up to the size of the biggest class. Figure 11 shows the nature via a scatter plot of the dataset before and after applying the SMOTE balancing technique.

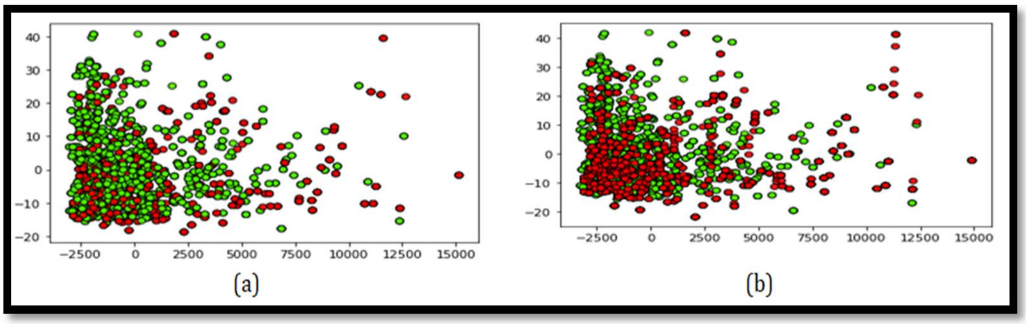

Figure 11: Scatterplot of Credibility Distribution (A) Before (B) After SMOTE

In figure 11 (a), we observed an unbalanced distribution with one instance dominating the other. However, once SMOTE was introduced, we experienced an even distribution of the class in figure 11 (b).

4. Results

This section presents the analytical results of the various machine learning models adopted in this paper. In this paper, all model diagnostic metrics are based on the validation/test set.

4.1. Prediction Accuracy

The idea here is also to determine which model performs best with our data, and as a first step, we considered each model's overall out-of-sample prediction accuracy on the test set. Note that as a rule of thumb, it is advisable to use the global f1-scores for model comparison instead of the accuracy metric. However, recall and precision metrics were also used for all the classifiers in our case. The results are shown in the table below:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Decision Tree Classifier	98.3333	99.0279	98.5489	98.7878
Logistic Regression Classifier	70.5000	83.6415	71.1003	76.8627
Random Forest Classifier	99.0000	99.5139	99.0326	99.2727
AdaBoost Classifier	73.9166	99.5139	72.3095	79.2577
Xtreme Gradient Boosting Classifier	81.5833	90.7258	81.6203	85.9325
KNNeighbours Classifier	85.1667	94.7586	83.0713	88.5309
Naive Bayes Classifier	70.3333	84.9925	69.1656	76.2666

Table 2: Test Set Prediction Accuracy, Precision, Recall, and F1-Score

While selecting a metric, it is essential to have the machine learning application in mind. In practice, we are usually interested not just in making accurate predictions but in using this prediction as a larger decision-making process which is to disburse the loan to people who will return the money. The consequence of choosing a particular algorithm for a machine learning application is called the business impact. From table 2, the least performing model in terms of prediction accuracy was the Naive Bayes. Despite its not-so-good accuracy, it still produced a precision value of approximately 85 percent, showing that a large amount is correct out of the predicted value. However, note that the best performing models were the machine learning ensemble classifiers (Random forest and decision tree). The random forest slightly outperforms the decision tree based on the value of the precision, recall, and the f1 score.

4.2. Confusion Matrix

One of the most comprehensive ways to represent the result of binary classification is the confusion metrics. The dataset was divided into 70 percent training data and 30 percent testing or validation set. A total of 1200 dataset was used to validate the model, and the validation test was further divided into (1 – Non-default and 0 – default) the target classes. Table 3 shows the statistical distribution of the test set.

Credibility	1 –Non-default	0 –Default
Count	827	373

Table 3: Distribution of the Credibility of the Test Set

The output of the confusion matrix is a two-by-two array, where the rows correspond to the true classes, and the columns correspond to the predicted classes. Each entry counts how often a sample belongs to the class corresponding to the row ('1' - Non-default, '0'- default). For an ideal confusion matrix, we expect to get values only on the leading/principal diagonal since they represent correct classification. Values off-diagonal are those that were misclassified. Hence, figures 12, 13, and 14 illustrate the confusion matrix for each of the three top-performing models with respect to the test sets. From the confusion matrix, the decision tree classifier could predict 815 non-defaults correctly, and 4 of the non-defaults were predicted as defaults.

In contrast, just 12 of the defaults were classified as non-defaults, and 369 of the customers that defaulted were correctly classified as defaults. Meanwhile, a slight improvement in the random forest classifier was experienced as 819 non-defaults were correctly classified, while 4 of that class were wrongly predicted. The number of correctly classified defaulters remained at 369 while the wrongly classified defaulters dropped to 4. A wide gap is seen in the confusion matrix of the KNNeighbours Classifier as the values of the leading diagonals showed the correct classification of row ('1' - Non-default, and '0'- default) as 625 and 327, respectively. All the performance metrics that have been used so far showed that the Random Forest Classifier outperforms the other entire model.

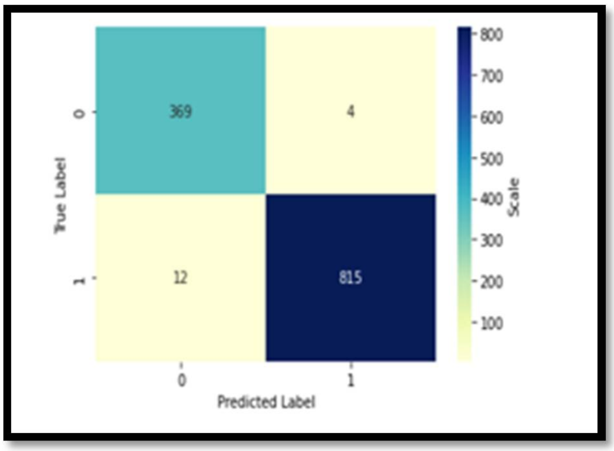

Figure 12: Confusion Matrix of the Decision Tree Classifier

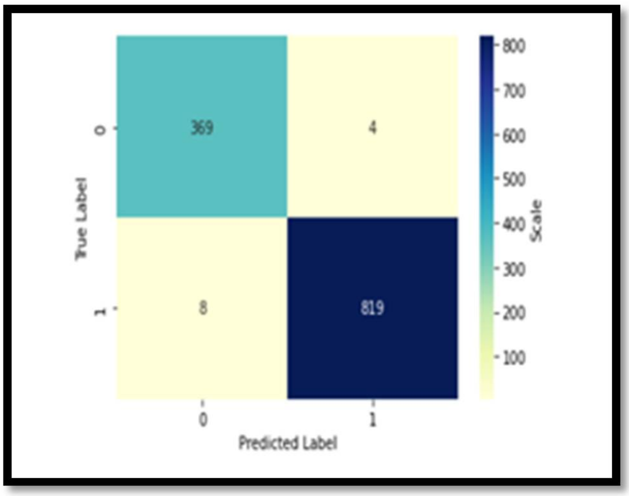

Figure 13: Confusion Matrix of the Random Forest Classifier

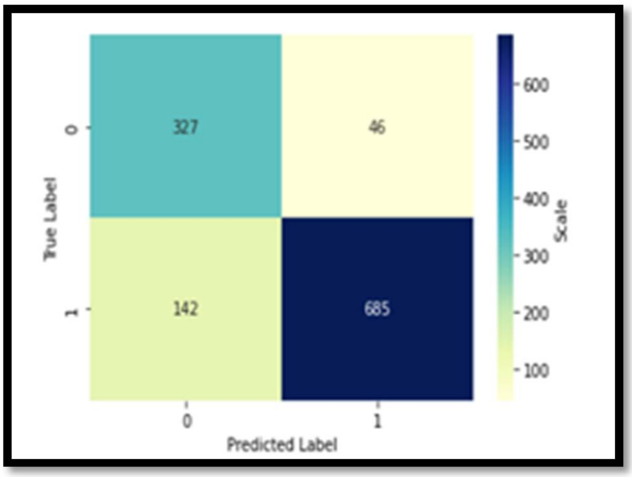

Figure 14: Confusion Matrix of the KNNeighbours Classifier

4.3. Sensitivity Analysis

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. A probability curve plots the (True Positive rate) TPR against (False Positive rate) FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model in distinguishing between the positive and negative classes. Since the data set is balanced due to the application of SMOTE, it is essential to use the ROC and AUC performance metrics. Figures 15, 16, and 17 show the ROC and AUC of the top three models. The random forest classifier has an AUC of 1, which means it can correctly classify both default and non-default. The decision tree classifier follows closely with an AUC of 0.98. The KNNeighbours Classifier also gives a value of 0.94.

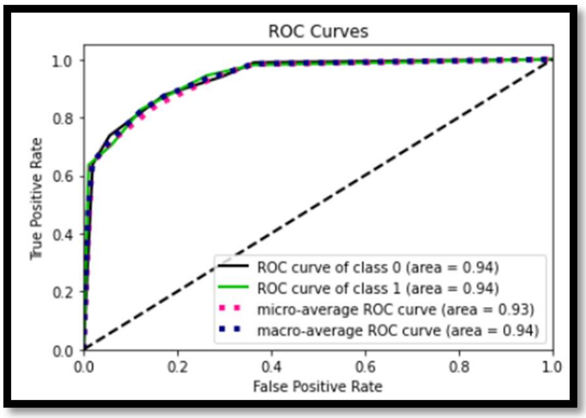

Figure 15: Receiver Operating Characteristic (ROC) Curve and Area under the Curve (AUC) for Random Forest Classifier

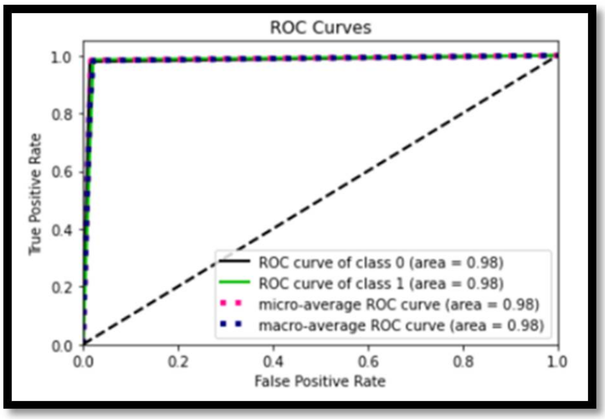

Figure 16: Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) for Decision Tree Classifier

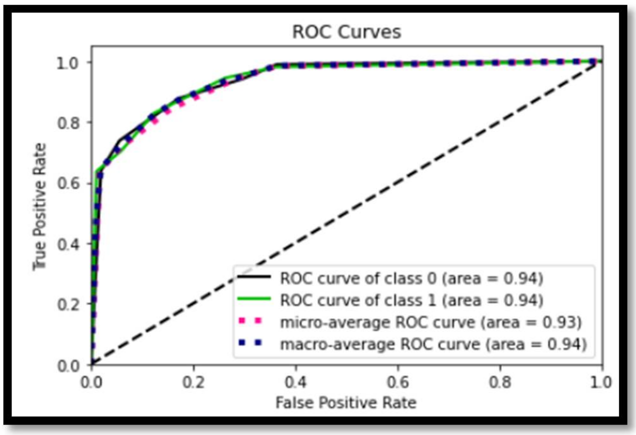

Figure 17: Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) for KNNeighbours Classifier

5. Discussion

This work has evaluated the usefulness of machine learning models in assessing credibility in a micro-credit environment. In micro-credit, there is usually no central credit database of customers and very little to no information at all on a customer's credit history. This situation is predominant in Africa, especially Nigeria, where the data was obtained from. This makes it hard for FINTECH institutions to determine whom to deny or not deny micro-loans. To overcome the drawback, it has been demonstrated that machine learning algorithms are powerful in extracting hidden information in the data set, which helps assess micro-credit defaults. All performance metrics adopted were those based on the validation/test set.

The data imbalance situation in the original data set was solved using the SMOTE algorithm. Several machine learning models were fitted to the data set. All the models recorded overall accuracy of 70% or higher on the validation set. Among the models reported in this paper, the top three best-performing classifiers (random forest, decision Tree, and KNNeigbhour) these classifiers reported an overall accuracy of more than 80% on the validation set. Other performance measures adopted also revealed that these three classifiers have good predictive power in assessing defaults in micro-credit (as shown in the confusion matrix).

The result obtained revealed lots of information, such as:

- Machine learning algorithms can be used to classify new customers into various risk classes successfully,
- Some machine learning algorithm outperforms the others, and
- Risk can be mitigated by ascertaining the creditworthiness of an individual applying for a loan

Finally, customers with no credit history should not be classified as high-risk customers since the behavior of the machine learning algorithm is independent of the data profile of the new clients.

6. Conclusions

The study has shown that modelling credit risk is crucial for the image and mitigation of risk of FINTECH companies in Nigeria. In conducting binary classification, a balance has to be struck between the number of defaulters and non-defaulters in the dataset task, and the findings were aided in this. An imbalanced data set will only make things much more difficult for the algorithms to learn, which is a common problem with imbalanced datasets, as shown by studies such as Khandani et al. (2010) and Cambria et al. (2013). For the data used in this study, it was established that the Random forest classifier had a better performance compared to the other models, and among the models, the Decision tree classifier also showed high performance at modeling credit risk. This was followed by the KNNeighbours Classifier, while the Naive Bayes Classifier performed the worst. Efficiency, in this study, was taken to be the ability of the models to optimise the company's profitability by:

- Minimising the cost of false negative (i.e., predicting non-default as default),
- Maximising the revenue through minimising the opportunity cost of false positives negative (i.e., predicting default as non-default), and
- Denying customers with no credit history loans

The measure that was used to capture this was the F1-score which was a harmonic mean between Precision and Recall.

However, the Random forest classifier still performed well even when it came to the Accuracy measure, which is the ratio of the correctly classified defaults to the total observation, with an accuracy of 99 %. However, the accuracy measure is not a good measure of performance as it does not consider the cost of misclassification, which is captured in the false positives and negatives. Therefore, the accuracy score was not used generally as the only performance metric, and this has also been found in other studies, such as the study by Butaru et al. (2016). The study concludes that machine learning models perform better in modelling credit risk while dealing with balanced datasets for real-life credit data sets. Thus, a sophisticated sampling technique such as SMOTE was introduced to help improve the imbalanced data set and improve performance.

Conclusively, the analytic results revealed that machine learning algorithms are capable of being employed to model credit risk for FINTECH start-up environments even in the absence of a central credit database and/or credit history. Generally, Random forest, Decision tree Classifier, and KNNeigbhour machine learning algorithms showed better performance with our real-life data than others. The most performing model is the Random forest classifiers (Bajari et al., 2015). Fernandez Delgado et al. (2014) found that the Random Forest classifier generated the most accurate prediction, which is the case in this study as the Random Forest classifier bettered the Decision tree by 0.67 percent. The study on a specific data set demonstrates that logistic regression, Naïve Bayes, and AdaBoost classifiers perform poorly with roughly the same prediction accuracy (within 70%).

7. **References**

i. Altman, E. (1968). Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. *Journal of Finance,* 23(4), 4589–609

ii. Altman, E. I., & Hotchkiss, E., (2010). Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyse and invest in distressed debt, *John Wiley & Sons.* 289 (10), 12.

iii. Assef, F.M., & Steiner, M.T., (2020). Machine Learning Techniques in Bank Credit Analysis. *Journal of Economics,* 2(1), 16-22

iv. Bajari, P., Denis, N., Stephen, P., & Miaoyu, Y., (2015). Machine learning methods for demand estimation. *American Economic Review,* 105(1), 481–85.

v. Bana, E. C., Carlos, A. E., Luis, A. &Joao, O.S. (2002) – Qualitative Modelling of Credit Scoring: A Case Study in Banking, *European Research Studies 5(1), 37-51.*

vi. Boughaci, D., & Alkhawaldeh, A.A., (2018). Three local search-based methods for feature selection in credit scoring. *Vietnam Journal of Computer Science* 5: 107–21

vii. Breiman, L. (2001). Random Forests Random Forests. Machine Learning.' *Statistics Department, University of California–Berkeley*, 45(3), 5-32.

viii. Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72(1), 218–239

ix. Cambria, E., Liu, Q., Li, K., Leung, V. C., Feng, L., and Ong, Y. (2013). Extreme learning machines. *IEEE Intelligent Systems* 7(6), 30–59.

x. Fernández, D., Manuel, E., Senén, B., & Dinani, A., (2014). Do we need hundreds of classifiers to solve real-world classification problems? *The Journal of Machine Learning Research,* 15(1), 3133–81.

xi. Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7(1), 179–88.

xii. Freund, Y. & Schapire, R.E., (1997).A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *Journal of Computer and System Sciences*, 55(1), 119-139.

xiii. FSB (Financial Stability Board). (2017). Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications.' FSB, Basel, Switzerland. 2(2), 22-23

xiv. Hand, D. J., & Henley, W. E., (1997). Statistical Classification Methods in Consumer Credit Scoring: *A Review. Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 160(3), 523–541.

xv. Hooman, A. Marthandan, G. & Karamizadeh, S., (2013). *Statistical and Data Mining Methods The Journal of Developing Areas*. 41(1), 2-27.

xvi. Khandani, A. E., Adlar J. K. & Andrew W. L., (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34(11), 2767–87

xvii. Kolade - Fadeyi, I. (2021) – Artificial Intelligence and the Nigerian Legal Profession, *Achievers University Law Journal (AULJ) Vol.1Issue 1 September 2021, pp 161-176.*

xviii. Moin, K. I., & Ahmed, Q. B., (2012). Use of Data Mining in Banking, *International Journal of Engineering Research and Applications*. 2(1), 738–742.

xix. Onova, C.U., & Omotehinwa., T.O., (2021). Development of a machine learning model for image-based E-mail spam detection. *FUOYE Journal of Engineering and Technology*, 6(4), 336-40.

xx. Ozgur, Ö., Erdal, K., & Fatih, C. O., (2021). Machine learning approach to drivers of bank lending: Evidence from an emerging economy. *Financial Innovation* 7(2), 20-22

xxi. Petropoulos, A., Vasilis, S., Evaggelos, S., & Aristotelis, K., (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. In Bank Are Post-Crisis Statistical Initiatives Completed *National Institute of Economic and Social Research 49(1), 21.*

xxii. Sahay, R. M., Cihak, P., N'Diaye, A., Barajas, S., Mitra, A., Kyobe, Y.N., Mooi, I., & Yousefi, S.R., (2015). 'Financial Inclusion: Can It Meet Multiple Macroeconomic Goals?' *International Monetary Fund Staff Discussion* 15(2), 17.

xxiii. Salisu, A. (2016) – Analysis of logit and probit models, *Journal of Data Science 11(2), 200-204*

xxiv. SAS. (2019). 'Artificial Intelligence: What It Is and Why It Matters.' https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.

xxv. Sudhakar, M. Reddy, C. V. K. & Pradesh, A. (2016).Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique Research Scholar, *Department of Computer Science and Technology, Rayalaseema University Kurnool, Andhra*, 5(3), 3-9.

xxvi. Wehinger, G. (2012). Banking in a challenging environment: Business models, ethics and approaches towards risks *Journal Financial Market Trends,* 2(2), 79–88.

xxvii. Wijewardhana, U., Chinthaka, B. & Thesath, N. (2018) – A Mathematical Model for Predicting Debt Repayment: A Technical Note, *Australasian Accounting, Business, and Finance Journal 12, 127-35.*

xxviii. Yap, B. Wah, Seng H. O. & Nor, H.M. (2011). Using data mining to improve the assessment of creditworthiness via credit scoring models. *Expert Systems with Applications*.38 (1), 13274–83.

xxix. Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4), 1060-1089.

xxx. Zhao, Y., Jianping L., & Lean Y., (2017). A deep learning ensemble approach for crude oil price forecasting. *Energy Economics* 66 (1), 9–16.