SSRG International Journal of Recent Engineering Science
ISSN: 2349 – 7157 / https://doi.org/10.14445/23497157/IJRES-V10I2P107

Volume 10 Issue 2, 46-53, Mar-Apr 2023

Original Article

Predicting Credit Risk : Using Machine Learning Algorithms to Predict the Creditworthiness of Borrowers and Determine their Likelihood of Defaulting on Loans in Nigeria

Akinmoluwa Oluseye Ayobami[1], Odeajo Israel[2], Jimoh Yusuf[3], Afolabi Moses Eniola[3]

Received: 19 February 2023 Revised: 02 April 2023 Accepted: 20 April 2023 Published: 30 April 2023

Abstract - *Due to recent advancements in the financial sector and the increasing need for obtaining loans, a significant number of people have begun to apply for bank loans. The rising rate of loan defaults is one of the most significant challenges the banking industry faces in the current economy. It is becoming increasingly challenging for banking authorities to precisely evaluate loan applications and creditworthiness, thereby mitigating the risk of individuals defaulting on loans. The present study proposed four different machine learning models that seek to predict an individual's eligibility for loan approval based on the evaluation of certain attributes, thereby aiding the banking authorities by facilitating the selection of suitable candidates from a given list of loan applicants. This paper provides a thorough comparison and analysis of four algorithms: Decision Trees, Gradient Boosting Classifiers, Random Forest, and Gaussian NB. The prediction is based on zindi Africa data. Important evaluation metrics, including Confusion Matrix, Accuracy, Recall, Precision, and F1-Score, have been calculated and presented in our result. In terms of accuracy, the Gaussian NB Algorithm outperformed the other three algorithms by a margin of 77%.*

Keywords - *Credit Risk, Credit Score, Decision Trees, Gaussian NB, Gradient Boosting, Loan Prediction, Machine Learning, Random Forest.*

1. Introduction

In terms of both macroeconomic factors and systemic risk, consumer spending is a key factor. Therefore, the analysis of consumer credit is pertinent, as individuals may obtain loans to satisfy their consumption needs [1]. According to [2], It is expected that the transaction value of the Marketplace Lending (Consumer) segment will reach $78.57 million in the year 2023 and is anticipated to grow at a 5.29% compound annual growth rate (CAGR) over the period 2023-2027, resulting in a total of US$96.57m by the year 2027. In 2023, the average transaction value per Marketplace Lending (Consumer) user is anticipated to reach $48.75m. From a global comparison standpoint, the United States will reach the maximum transaction value (US$26,180,000,000) in 2023.

The Nigerian credit market is predominantly governed by the Central Bank of Nigeria (CBN), the apex regulatory body in the banking system and is therefore responsible for the DMBs [3]. Obviously, there are credit lenders that are not governed or supervised by the CBN. These include the Primary Mortgage Institutions, which report to the Federal Mortgage Bank, and the leasing corporations that operate under the Equipment Leasing Association of Nigeria's self-regulatory body [4].

A loan is essentially an agreement between two parties, the lender and the borrower, whereby the lender grants the borrower credit in the form of cash, property, or other tangible goods if the lender believes that the individual who receives a loan is capable of repaying the borrowed funds with interest. Almost every bank's primary focus today is on approving and distributing debts. A significant component of a bank's assets comprises profits generated from disbursed loans.

Credit risk evaluation is a crucial aspect of financial risk management because banks must make critical decisions regarding whether or not to lend to a counterparty. According to [5], the most significant issue in finance is predicting bankruptcy or default. In consumer lending, the large number of potential consumers necessitates using models and algorithms that minimize or eradicate errors caused by human actions in analysing credit applications [1]. Several largest global institutions have developed advanced automated systems for modelling credit risk, providing decision-makers with crucial data. Before extending credit to borrowers, it is essential to ascertain their creditworthiness, as payment defaults can be extremely risky. Due to the dearth of appropriate data for machine learning techniques, some financial institutions continue to rely on the traditional strategy [6].

2. Literature Review

Within the framework of credit risk analysis and creditworthiness research employing machine learning techniques, there are a number of studies analyzing the adequacy of models in particular databases.

[1] employed various machine learning models, namely Support Vector Machine, Decision Trees, Bagging, AdaBoost, and Random Forest, to predict consumers' creditworthiness and compared the accuracy of their predictions utilizing a benchmark predicated on a Logistic Regression Model. They analyzed comparisons in accordance with standard classification performance metrics. According to their results, random Forest and Adaboost outperform other models. In addition, Support Vector Machine models perform poorly with both linear and nonlinear parameters.

Using Machine Learning (ML) techniques, namely k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), and Multiple Linear Regression, [7] analyzed various competencies for CR analysis (MLR). According to the findings, the MLR method performed significantly, thus improving the accuracy of credit risk prediction than both the k-NN and SVM methods.

[8] also conducted a comparative study utilizing machine learning to predict corporate credit ratings. They utilized 1881 loan applicants operating in three distinct industries. In addition, the machine learning methods and solvers LR (lassoglm), Decision Tree, and KNN were implemented using MATLAB 9.4 (R2018a). The Support Vector Machine (SVM) model was developed utilizing the LIBSVM software application. The Python programming language was employed to implement the RF (random forest classifier) and XGBoost algorithm (XGBoost classifier). Based on their findings, On all data sets, logistic regression analysis, support vector machines, random forest, and XGBoost outperforms decision tree and KNN.

[9] conducted a comparative analysis of various machine learning models, including singular classifiers such as logistic regression, decision trees, LDA, and QDA, as well as heterogeneous ensembles like AdaBoost and Random Forest and sequential neural networks. For credit scoring. Their research shows that neural networks and ensemble classifiers are better. The ML-based credit scoring model is also evaluated by two advanced post-hoc models - LIME and SHAP.

Utilizing the hierarchical clustering methodology and the k-means algorithm, [10] proposed a model for categorizing borrowers. The model employs a methodology that categorizes real borrowers into clusters that exhibit comparable levels of creditworthiness and credit risk, thereby creating homogeneous groupings. They also created a classification model for borrowers using the stochastic gradient boosting (SGB) method. The modified version of the proposed method for evaluating a person's creditworthiness cut down on credit risks and predicted the stability and profitability of financial institutions.

In a paper by [11], they utilized a variety of supervised learning algorithms, including J48 Decision Trees classification, BayesNet, NaiveBayes, and the Waikato Environment for Knowledge Analysis (WEKA) software to develop predictive models. Due to its more effective accuracy and minimal absolute mean error, the J48 algorithm proved to be the optimal algorithm for the loan risk classification, as demonstrated by their findings. Equally, the J48 algorithm has the potential to classify instances more accurately than the other techniques.

[12] evaluated the performance of different machine learning algorithms (including a logistic regression model, a decision tree, a Naive Bayes classifier, a random forest, a K-Nearest Neighbor algorithm, and a Gradient boosting algorithm) on real micro-lending data from LAPO microfinance bank, Nigeria. The Random forest classifier outperformed the other models in modelling credit risk, while the Decision tree classifier also demonstrated excellent performance. The KNNeighbors Classifier performed the next best, while the Naive Bayes Classifier performed the worst.

In work by [19], the authors devised and executed a Gradient boosting regression model to evaluate and predict the creditworthiness of borrowers. The study revealed that various parameters, including age, department of employment, position, qualification, city of residence prior to the war, gender, marital status, and availability of a guaranteed annual premium, were examined to describe an individual. The results indicated that the department and qualifications of the employee had the most significant impact on their creditworthiness.

[20] employed a methodology based on data mining and machine learning algorithms to develop a set of predictive models. The classifiers utilised in this study included LightGBM, XGBoost, Logistic Regression, and Random Forest. The optimal classifier for Random Forest yielded the best results outcomes when applied to the scenario that combined under-over sampling, resulting in a representative AUC value of 0.89. A loan default prediction model was developed by [21] utilising Lending Club's user loan data and the Random Forest algorithm. The SMOTE technique was employed to address the issue of imbalanced class distribution within the dataset. The research results indicate that the Random Forest algorithm exhibits superior performance compared to logistic regression, decision trees, and other machine learning algorithms when predicting default samples.

A predictive model was proposed by [22] for the purpose of predicting loan defaults in peer-to-peer lending communities. The study employed Logistic Regression, Random Forest, and Linear SVM to construct predictive models based on a chosen feature set. Results indicated that Random Forest exhibited superior performance, achieving an accuracy rate of 92%.

In work to estimate credit risk analysis, [23] proposed an algorithm based on XGBoost that has been optimized through the utilization of Adaptive Tree Parzen Estimators (ATPE) for the purpose of Credit Risk analysis. Upon optimization using the Adaptive TPE method, the hyperparameters of the algorithm under consideration exhibit superior performance compared to the default model.

[24] employed various classification techniques, namely Logistic Regression, Naive Bayes, Random Forest, K Nearest Neighbor, and Decision tree, to categorize loan data according to their probability of default. The algorithm's simulation outcome yields a noteworthy level of accuracy, specifically 94.6%.

[25] utilized a range of machine learning techniques, including gradient boosting, random forest, and feature selection, were employed in conjunction with decision trees. The algorithms successfully classified customers for loans as either valid or invalid. The model that has been designed is capable of classifying customers into two categories, namely, good and bad applicants. The model is trained to improve the accuracy of the customer data.

2.1. Theoretical Background

Nearly every economic sector around the globe is moving toward total automation. To accomplish this objective, ideas and methods are created daily, and many fields have been studied for a long time. Artificial Intelligence (AI) is an emerging field that has garnered the interest and dedication of scholars, researchers, and professionals. AI is the concept of creating a computer or machine with intelligence and behavior patterns resembling a human's. It dates back to the earliest days of computer construction and has since branched out into numerous disciplines, such as Machine Learning (ML), Neural Networks, Natural Language Processing (NLP), etc. [13].

Machine learning is a field of artificial intelligence (AI) whose aim is to comprehend how data is structured and model it so people can utilise it [27]. Computers can train data using machine learning algorithms and employ statistical tools to produce values contained in a specified range. Machine learning facilitates the development of models from available data that will enable decisions to be made based on these data inputs [15]. Supervised learning and unsupervised learning are the most prevalent machine learning techniques, despite the existence of reinforcement learning and semi-supervised learning methods.

Researchers and technologists have become very interested in machine learning (ML) in recent years. Various machine learning models and algorithms are being employed to facilitate significant tasks and enhance the quality of life for the general populace, especially in banking and finance. By utilizing various machine learning models, financial institutions and banks can identify patterns and make inferences regarding matters such as credit card fraud and forecasting loan defaults. It has made the process much easier now and more accurate. Literature uses many different prediction and classifier algorithms to determine if someone is creditworthy and how risky a loan is. Below, we explain the classification algorithms that were used in this paper.

2.1.1. Gaussian Naive Bayes (GNB)

Gaussian Naive Bayes (NB) refers to a set of supervised learning models used for predictive purposes. They are simple and effective models which learn the probabilities of certain features belonging to a given group [16]. The Naive Bayes classifier uses two assumptions:

- Given the class label, features are conditionally independent of each other and contribute equally to the process.
- No latent feature affects the class label prediction process.

Suppose a vector represents $(x_1, x_2, \dots x_n)$ the n features of X. Let θ represent the class label of X. The naïve theorem describes the conditional probability of observing X given the class label θ, $P(\theta|X)$ as a product of several simpler probabilities, as shown below:

$$P(\theta|f_1, f_2, \dots f_n) = \frac{P(\theta)P(f_1,f_2,\dots,f_n|\theta)}{P(f_1,f_2,\dots f_n)} \quad (1)$$

Where $(f_1, f_2, \dots f_n)$ represents the features of vector X. Following the assumption of independence, the probability can be expressed as;

$$P(\theta|f_1, f_2, \dots, f_{i-1}, f_{i+1}, \dots f_n) = P(f_i|\theta) \quad (2)$$

For all i in equation (1),

$$P(\theta|f_1, f_2, \dots f_n) = P(\theta)\frac{\prod_{i=1}^{n} P(f_i|\theta)}{P(f_1,f_2,\dots f_n)} \quad (3)$$

Assuming every feature is constant, the classification rule follows below:

$$P(\theta|f_1, f_2, \dots f_n) \alpha\, P(\theta)\prod_{i=1}^{n} P(f_i|\theta), \quad (4)$$

And,

$$\hat{\theta} = arg\max_{\theta} P(\theta)\prod_{i=1}^{n} P(f_i|\theta) \quad (5)$$

It is worth noting that for the GNB model, the probability of feature occurrence follows a Gaussian distribution.

$$P(f_i|\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta}} \exp\left(\frac{-(f_i-\mu_\theta)}{2\sigma_\theta^2}\right) \quad (6)$$

whose parameters σ_θ and μ_θ are computed using maximum likelihood.

2.1.2. Random Forest (RF)

Random forest is a supervised ensemble method that uses a collection of numerous decision trees to make predictions. The RF algorithm is very efficient, as it handles datasets that contain continuous variables, as well as categorical variables, robustly. An RF classifier contains subsets of various tree classifiers $\{h(x,\Theta_k), k = 1,2,\dots\}$ where the Θ_k are independently and identically distributed random vectors, with each tree being able to specify the modal class at input x [17]. The performance index, which solely approximates the confidence interval (CI) of the RF model, is given as

$$mg(x,y) = av_k I(h_k(x,\Theta_k) = y) - \max_{j\neq y} av_k I(h_k(x,\Theta_k) = j) \quad (7)$$

where $I(.)$ denotes an indicator function, and $av(.)$, the average value. It is observed that as the margin increases, the confidence level also increases. The generalisation error becomes

$$PE^* = P_{x,y}(mg(x,y) < 0), \quad (8)$$

where $P(.)$ denotes probability. With an increase in trees for all sequences Θ_k, PE^* converges to

$$P_{x,y}(P_\Theta(h(x,\Theta) = y) - \max_{j\neq y} P_\Theta(h(x,\Theta) = j < 0) \quad (9)$$

Convergence of this generalization error proves that the RF model does not overfit as more trees are introduced. The upper bound for the generalization error is given as

$$PE^* \leq \frac{\bar{\rho}(1-s^2)}{s^2}, \quad (10)$$

where $\bar{\rho}$ is the average correlation value, s is the strength of each tree in the model. Increased strength of individual trees and a low correlation between them produces more accurate prediction results

2.1.3. The Decision Tree Classifier

The Decision Tree Classifier model belongs to the category of supervised learning algorithms that works by partitioning the input space into a hierarchy of nested regions, each of which is assigned a class label [18]. If a target is a classification outcome taking on values 0,1, K-1, for node m,

Let

$$p_{mk} = \frac{1}{n_m}\sum_{y\in Q_m} I\,(y = k) \quad (11)$$

be the proportion of class k observations in node m. If m is a terminal node, common impurity measures are the following.

$$H(Q_m) = \sum_{y\in Q_m} p_{mk}(1 - p_{mk}) \quad (12)$$

2.1.4. GBMs Algorithm

The Gradient Boosting Machine (GBM) is a well-known ensemble method that is commonly utilized for forward learning in the field of machine learning. It is an effective method for constructing predictive models for regression and classification tasks. GBM assists us in obtaining a predictive model in the form of a collection of feeble prediction models, such as decision trees. Whenever a decision tree operates as a weak learner, gradient-boosted trees are the resulting algorithm.

The Light GBM is an improved version of the Gradient boosting machine due to its increased efficacy and speed. In contrast to GBM and XGBM, it can manage massive amounts of data without requiring complexity. In addition, in light of GBM, the primary node is divided into two secondary nodes, and then one of the secondary nodes is subdivided, as depicted in the diagram below.

3. Materials and Methods

3.1. Data

Based on data from [26], Using machine learning algorithms, we built a model to predict credit risk, the creditworthiness of borrowers, and their likelihood of defaulting on loans.

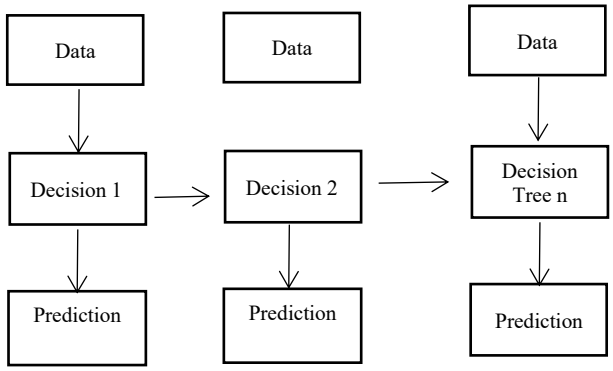

Fig. 1 GBM Algorithm

Source: Javapoint.com

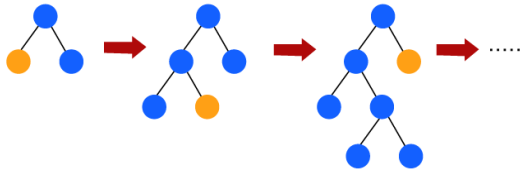

Fig. 2 Light GBM Algorithm

Source: Javapoint.com

3.2. Data Cleaning

The dataset Zindi data is divided into three (3); demographic data, performance data and previous loan data. The columns in the datasets containing null values were purged. To remove columns that do not meet a minimum threshold percentage, it is essential to determine the percentage of invalid values in each column.

Also, as shown in Figure 3, we plotted loan amount distribution, loan amount by good or bad and educational level by good/bad for better understanding and analysis of the data. This shows that in terms of loan amount and educational level, the good outperformed the bad.

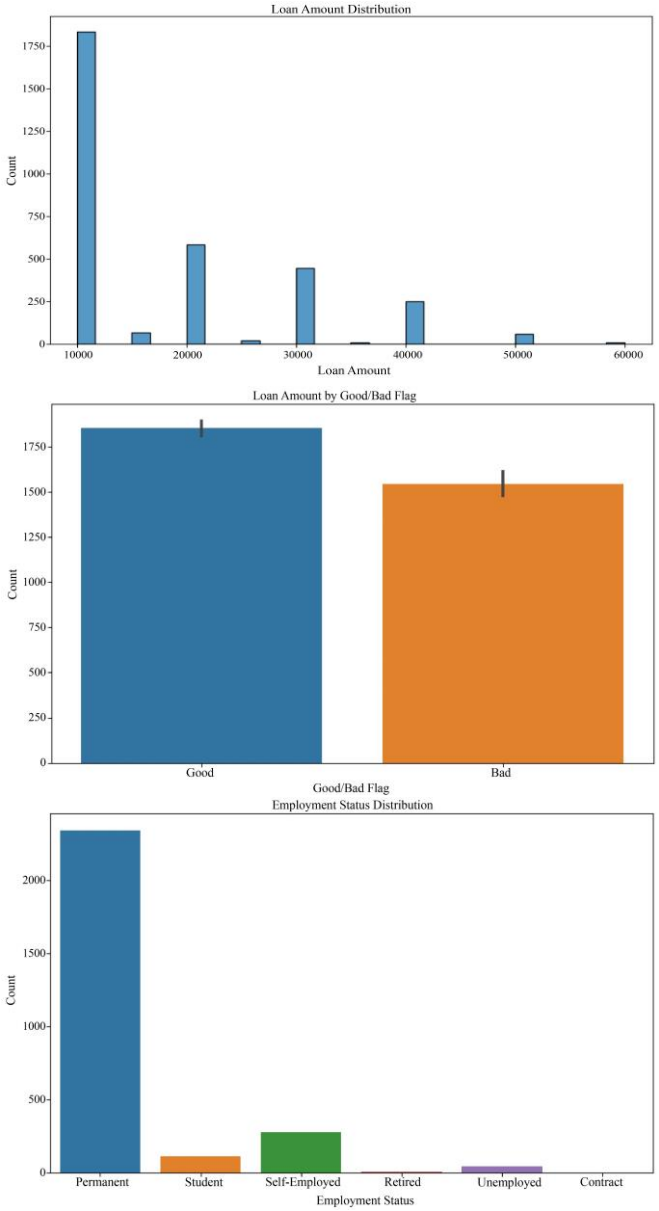

Fig. 3 Plots showing loan amount distribution loan amount by good/bad and Educational level by good/bad

3.3. Model Building

Prior to commencing model training, the dataset was divided into two distinct subsets: a Training set, which encompassed 70% of the data, and a Test set, which encompassed 30% of the data. Our model's predictions needed to be evaluated using a variety of performance metrics, such as accuracy, F1, precision, and confusion matrix. We used four algorithms for our modelling purpose:

3.3.1. Decision Tree

A decision tree classifier variable was set up, and the requisite libraries were imported via Scikit-Learn. Subsequently, the data was adapted to facilitate the training of the decision tree model for both training and test data prediction.

3.3.2. Gradient Boosting Classifier

After importing the necessary libraries, we fit the data to the model to train the Gradient Boosting Classifier. Among the types of Gradient Boosting Classifier, Lightgbm was chosen.

3.3.3. Random Forest

Using the Scikit-Learn library, The necessary libraries were also imported, and a variable was set up for the random forest classifier. For the random forest model, we set the estimator count to 100 and fitted the data accordingly.

3.3.4. Gaussian NB

The necessary libraries were imported to implement sklearn—naive bayes. The data is also employed to train the Gaussian Naive Bayes (GNB) model. The comparative analysis of the four (4) models' performance is presented in the subsequent section.

1. 4. Results and Discussion

This study employed four distinct machine learning algorithms, namely Decision Trees, Gradient Boosting Classifiers, Random Forest, and Gaussian NB, to develop a predictive model for credit risk assessment. The model was designed to evaluate the creditworthiness of borrowers and estimate the probability of loan default. In order to gain a deeper comprehension of the accuracy and other metrics of the models, we hereby provide their corresponding classification reports and confusion matrices.

4.1. Decision Tree

The accuracy of the Decision Tree classifier was calculated to be 65%.

Table 1. Classification report of decision tree

	Precision	Recall	f1-score	Support
0	0.25	0.26	0.26	152
1	0.77	0.77	0.77	502
Accuracy			0.65	654
Macro avg	0.51	0.52	0.51	654
Weighed avg	0.65	0.65	0.65	654

Table 2. Classification Report of Gradient Boosting Classifier

	Precision	Recall	f1-score	Support
0	0.41	0.06	0.10	152
1	0.77	0.97	0.85	502
Accuracy			0.76	654
Macro avg	0.59	0.52	0.48	654
Weighed avg	0.69	0.76	0.69	654

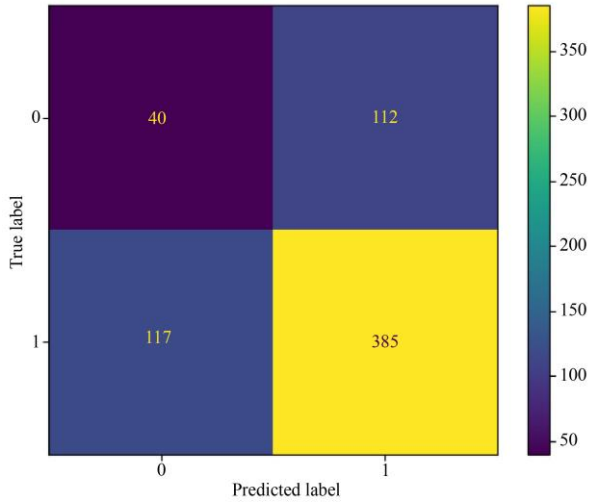

Fig. 4 Confusion matrix of decision tree

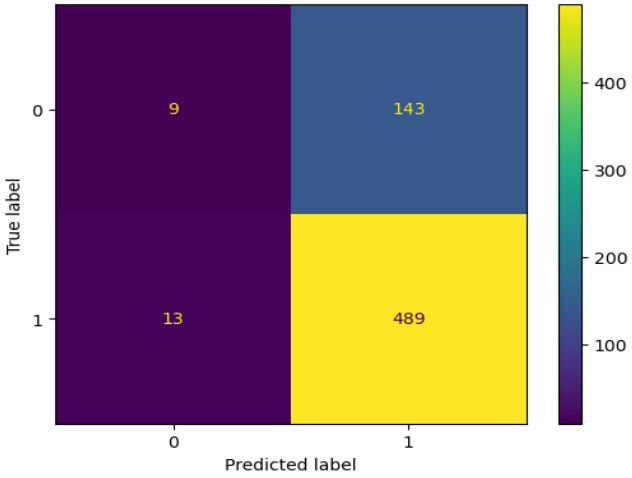

Fig. 6 Confusion Matrix of Random Forest Classifier

4.2. Gradient Boosting Classifier

The Gradient Boosting Classifier gave an accuracy of 76%.

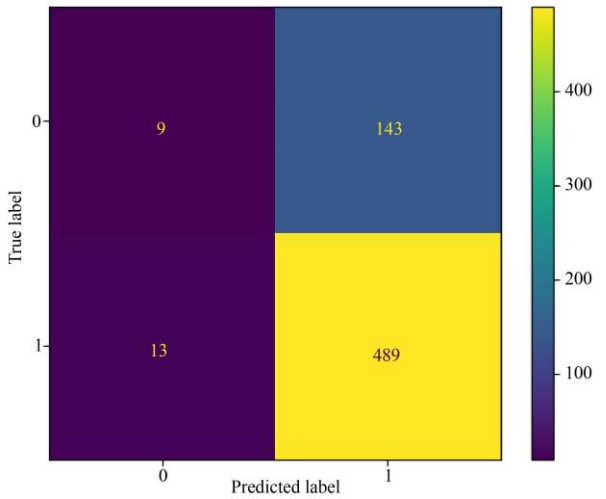

Fig. 5 Confusion Matrix of Gradient Boosting Classifier

4.3. Gaussian Naive Bayes

The Gaussian NB Classifier gave an accuracy of 77%.

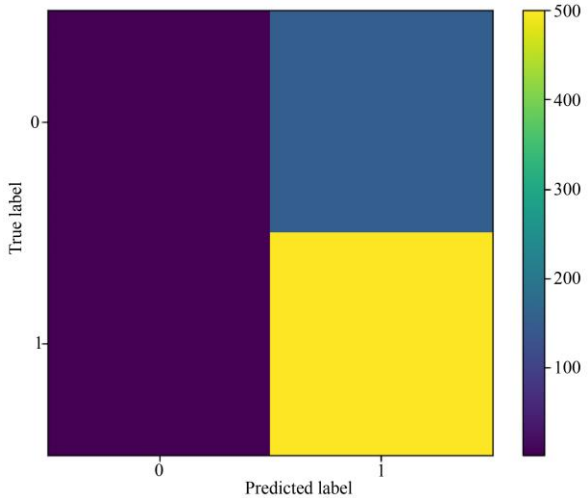

Fig. 7 Confusion Matrix of Gaussian NB Classifier

Table 3. Classification Report of Random Forest Classifier

	Precision	Recall	f1-score	Support
0	0.41	0.06	0.10	152
1	0.77	0.97	0.85	502
Accuracy			0.76	654
Macro avg	0.59	0.52	0.48	654
Weighed avg	0.69	0.76	0.69	654

Table 4. Classification Report of Gaussian NB Classifier

	Precision	Recall	f1-score	Support
0	0.41	0.06	0.10	152
1	0.77	0.97	0.85	502
Accuracy			0.76	654
Macro avg	0.59	0.52	0.48	654
Weighed avg	0.69	0.76	0.69	654

Table 5. Classification Report of the Models

Model	Score
GaussianNB	0.767584
Random Forest Classifier	0.762997
Gradient Boosting Classifier	0.762997
Decision Tree Classifier	0.647615

Upon examining the aforementioned confusion matrices and classification reports for each model. It can be inferred that the Gaussian Naive Bayes Classifier algorithm is the better option compared to the Decision Trees, Gradient Boosting Classifier, and Random Forest algorithms for predicting loans based on the used dataset, as shown in Table 5.

5. Conclusion

This study successfully utilized various classification algorithms to predict credit risk. The aim was to predict the likelihood of a borrower failing to meet their payment obligations and to assess the creditworthiness of customers. Python was used to conduct the analysis, and performance indicators such as accuracy, recall, precision, and f1-score were generated. The accuracy of the Gaussian NB Classifier was 77%, while the accuracy of the Random Forest and Gradient Boosting Classifiers was 76%. At the same time, the Decision Tree method yielded a 65% accuracy. Therefore, The Gaussian NB Classifier appears to be the superior choice for such data. The result of this work will help the financial sector, especially in Nigeria, accurately predict borrowers' creditworthiness before granting loans, thereby reducing risk.

However, some algorithm places some non-defaulters in the default class; in the future, we may wish to investigate this issue further in order to enhance the model's performance and relevance.

Acknowledgements

We want to acknowledge the zindi. Africa for the dataset used for this paper. Also, the contribution of authors: Odeajo Israel contributed to the Machine Learning analysis, proofreading of the methodology, results and final manuscript. Afolabi Moses contributed to the literature, and Jimoh Ibrahim contributed to Machine Learning analysis.

References

[1] Maisa Cardoso Aniceto, Flavio Barboza, and Herbert Kimura, "Machine Learning Predictivity Applied to Consumer Creditworthiness," *Future Business Journal*, vol. 6, no. 1, pp. 1-14, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] [Online]. Available: https://www.statista.com/outlook/dmo/fintech/digital-capital-raising/marketplace-lending-consumer/nigeria

[3] Isa Fatima, and Isa Rehanet, "Treatment of Toxic Asset by Deposit Money Banks in Nigeria: A Review of Literature," *TSU-International Journal of Accounting and Finance*, vol. 1, no. 1, pp. 42-50, 2021. [Google Scholar] [Publisher Link]

[4] [Online]. Available: https://www.cbn.gov.ng/out/2013/ofisd/revised%20guidelines%20for%20primary%20mortgage%20banks%20in%20nigeria.pdf

[5] Fernanda Assef et al., "Classification Algorithms in Financial Application: Credit Risk Analysis on Legal Entities," *IEEE Latin America Transactions*, vol. 17, no. 10, pp. 1733-1740, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Saba Moradi, and Farimah Mokhatab Rafiei, "A Dynamic Credit Risk Assessment Model with Data Mining Techniques: Evidence from Iranian Banks," *Financial Innovation*, vol. 5, no. 1, pp. 1-27, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] G. S. Samanvitha et al., "Machine Learning Based Consumer Credit Risk Prediction," *Sustainable Advanced Computing*, pp. 113-123, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Seyyide Doğan, Yasin Büyükkör, and Murat Atan, "A Comparative Study of Corporate Credit Ratings Prediction with Machine Learning," *Operations Research and Decisions*, vol. 32, no. 1, pp. 25-47, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Swati Tyagi, "Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions," *American International Journal of Business Management*, vol. 5, no. 1, pp. 5-19, 2022. [Google Scholar] [Publisher Link]

[10] Ekaterina V. Orlova, "Methodology and Models for Individuals' Creditworthiness Management Using Digital Footprint Data and Machine Learning Methods," *Mathematics*, vol. 9, no. 15, p. 1820, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Alaba, O.B., Taiwo, E.O., and Abass, O.A., "Data Mining Algorithm for Development of a Predictive Model for Mitigating Loan Risk in Nigerian Banks," *Journal of Applied Sciences and Environmental Management*, vol. 25, no. 9, pp.1613-1616, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Akinwunmi Adeboye A, and Dare Festus Oluwafemi, "Machine Learning Approach to Credit Scoring for Fintech Start-Ups Using Micro Finance Banks in Nigeria," *International Journal of Innovative Research and Development*, vol. 11, no. 8, pp. 97-109, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mehul Madaan, "Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Pagadala Suganda Devi, "Credit Risk Management Practices of Micro Finance Institutions in Ethiopia– A Brief Literature Review," *SSRG International Journal of Economics and Management Studies*, vol. 4, no. 1, pp. 10-16, 2017. [CrossRef] [Publisher Link]

[15] Kush R. Varshney, "Trustworthy Machine Learning and Artificial Intelligence," *XRDS: Crossroads, the ACM Magazine for Students*, vol. 25, no. 3, pp. 26-29, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] Liaqat Ali et al., "A Feature-Driven Decision Support System for Heart Failure Prediction Based on Statistical Model and Gaussian Naive Bayes," *Computational and Mathematical Methods in Medicine*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Breiman, L., Random Forests, Machine Learning, pp. 5-32, 2001.

[18] Iqbal H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, p.160, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Larysa Zomchak, and Viktoria Melnychuk, "Creditworthiness of Individual Borrowers Forecasting with Machine Learning Methods," *Advances in Artificial Systems for Medicine and Education VI*, vol. 159, pp. 553-561, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Alexandru Coser, "Predictive Models for Loan Default Risk Assessment," *Economic Computation & Economic Cybernetics Studies & Research*, vol. 53, no. 2, pp. 149-165, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[21] Lin Zhu, "A Study on Predicting Loan Default Based on the Random Forest Algorithm," *Procedia Computer Science*, vol. 162, pp. 503-513, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Lee Victor, and Mafas Raheem, "Loan Default Prediction Using Genetic Algorithm: A Study Within Peer-to-Peer Lending Communities," *International Journal of Innovative Science and Research Technology*, vol. 6, no. 3, pp. 2456-2165, 2021. [Google Scholar] [Publisher Link]

[23] Pradeep Sudhakaran, and Sujoy Baitalik, "Xgboost Optimized by Adaptive Tree Parzen Estimators for Credit Risk Analysis," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Vinay Padimi et al., "Applying Machine Learning Techniques to Maximize the Performance of Loan Default Prediction," *Journal of Neutrosophic and Fuzzy System,* vol. 2, no. 2, pp. 44-56, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Shrikant Kokate, and Manna Sheela Rani Chetty, "Credit Risk Assessment of Loan Defaulters in Commercial Banks Using Voting Classifier Ensemble Learner Machine Learning Model," *International Journal of Safety and Security Engineering,* vol. 11, no. 5, pp. 565-572, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] [Online]. Available: https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data

[27] Kotsiantis, S.B., Zaharakis, I.D., and Pintelas, P.E., "Machine Learning: A Review of Classification and Combining Techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159-190, 2006. [CrossRef] [Google Scholar] [Publisher Link]