PRINT ISSN 1119-8362
Electronic ISSN 1119-8362

Full-text Available Online at
https://www.ajol.info/index.php/jasem
http://ww.bioline.org.br/ja

J. Appl. Sci. Environ. Manage.
Vol. 25 (9) 1613-1616 September 2021

# Data Mining Algorithm for Development of a Predictive Model for Mitigating Loan Risk in Nigerian Banks

*[1]ALABA, OB; [2]TAIWO, EO; [3]ABASS, OA

*[1]*Department of Computer Science, Tai Solarin University of Education, Ijagun, Nigeria*
[2]*Department of Computer Science, Alvan Ikoku Federal College of Education, Owerri, Nigeria*
[3]*Department of Computer Science, Tai Solarin College of Education, Omu-Ijebu. Nigeria*
**Corresponding Author Emails: alabaob@tasued.edu.ng; Other Authors Email: emmanuel.taiwo@alvanikoku.edu.ng; olaabass@gmail.com*

**ABSTRACT:** The focus of this paper is on the development of data mining algorithm for developing of predictive loan risk model for Nigerian banks. The model classifies and predicts the risk involved in granting loans to customers as either good or bad loan by collecting data based on J48 decision tree, BayesNet and Naïve Bayes algorithms for a period of ten (10) years (2010 2019) from using structured questionnaire. The formulation and simulation of the predictive model were carried out using Waikato Environment for Knowledge Analysis (WEKA) software. The performance of the three algorithms for predicting loan risk was done based on accuracy and error rate metrics. The study revealed that J48 decision tree model is the most efficient of all the three models.

**DOI**:https://dx.doi.org/10.4314/jasem.v25i9.11

**Dates:** Received: 09 May 2021; Revised: 12 August 2021; Accepted: 12 September 2021

**Keywords:** Data Mining, Algorithm, Development, Predictive Model, Loan Risk.

According to Ojo (2010), loan risk is the likelihood that a payment of obtained loan by a debtor will not be fully settled due to insolvency of the debtor. Nigeria's banking industry, as the practice across the globe, provides financial capital to both business community and individual customers. This financial provision by banks is done with the expectation of attaining the focused rates of returns credit facility over a period of time as well as getting back their principal with interest. Any loan grant under the terms of the financial relationship between financial institutions (financier) and individuals or corporate organizations comes with the risk of non-repayment (Olawale *et al.*, 2015). As a result of this, banks show strong interest with due diligence before granting loans on a regular basis to maximally reduce the loan risk to achieve an improved organizational financial value. Since loans remain the core banking assets, the assessment of quality of bank loan as it affects the bank's financial stability is crucial. As observed by Sulaimon (2001), there remains a difficult task by the management of Nigerian banks to identify problems associated with loans obtain by customers. This, according to Ojo (2010), is due to ignorance and the intense desire to pronounce profits at the end of the financial year. This makes banks' balance sheets usually not reflecting the banks' authentic financial condition. Risk occurs due to obligor's refusal to perform its obligations and respect the loan contract or the power to carry out the obligations is hindered leading to an economic loss to the bank (CBN, 2000). This is not only due to the occurrence of borrower defaulting in the loan in terms of either re-payment of principal and interest, but the occurrence of decline in the repayment capability. There are different categories of loan a customer needs to consider when approaching banks for loan in terms of attached criteria. Loan categorization deals with the process of critically evaluating loan collections, assigning loans to the beneficiaries putting into consideration the perceived risk and other related loans attributes (Ogawa *et al.*, 2013). The act of continual examining and classifying loans allows proper assessment of the quality of loan. Hence, the adoption of highly sophisticated internal classification techniques instead of current standardized schemes which bank managers need to report reasons for granting loans with a view to make easy of observing criteria for forestall non-payment of loans and real interbank evaluation. Data mining (DM), a decision support process, is used to observe the patterns and relationships hidden in the datasets and then generate information with potential value. Amanze and Onukwugha (2017) developed an intelligent system for detecting loan fraud in Nigeria's banking industries using data mining approach. Ogwueleka (2011) applied DM-based neural network (called self-

**Corresponding Author Emails: alabaob@tasued.edu.ng* organizing map neural network, SOMNN) architecture to develop a credit card fraud detection system using unsupervised learning method. The system was applied on the transactions data to generate four clusters (i.e. low, high, risky and high-risk). The SOMNN technique was used to solve the problem of making optimal classification for each transaction into its associated group because a prior output is unknown. The receiver-operating curve (ROC) for credit card fraud (CCF) detection watch detected over 95% of fraud cases without causing false alarms unlike other statistical models and the two-stage clusters. The result showed that the performance of CCF detection watch is in agreement with other detection software, but performs better. The objective of this paper is therefore to develop data mining algorithm for predictive loan risk model for Nigerian banks by collecting data based on J48 decision tree, BayesNet and Naïve Bayes algorithms for a period of ten (10) years (2010 -2019).

## METHODS AND METHODS

*Sample Collection:* Eight hundred and Seventy-Five (875) data of loan applicants were collected from Access Bank Plc for a period of ten (10) years from 2010 to 2019. The data collected contained information like: Name, Date of birth, Gender, Nationality, Gender, Marital status, Address, City, State, Country, Occupation/Job, Business category, Business sector, Borrower type, Date of loan (facility) disbursement/Loan effective date, Maturity date, Credit_amount, Credit_history, Instalment amount, Currency, Loan (Facility) type, Loan (Facility) tenor, Housing, Repayment frequency and Loan classification. These were converted into electronic format and stored as Microsoft Excel files from their paper-based storage. The data were processed and attribute selection processes were also performed on the data to identify eight most important attributes (input variables): Age, Gender, Purpose, Credit_history, Credit_amount, Housing and Job. Class was used as the output. Class is the “loan” with two options of “good” or “bad”.

*Description of Method used for Data Analysis:* Three different supervised learning algorithms: J48 Decision Trees classification, BayesNet and NaiveBayes were used. The purpose of using supervised learning algorithms is that the study centered on classification in which the output of prediction is already known to either be good or bad. The Waikato Environment for Knowledge Analysis (WEKA) software was used in developing the predictive models. The data which consists of 875 loan applicants was used to develop the predictive models while a test data was used in testing the models developed. Each model was compared and their limitations identified and the most efficient model was chosen based on evaluation criteria and results.

**Table 1:** Dataset

| S/N | Variable | Description | Measurement |
|---|---|---|---|
| 1 | Credit_history | Previous history of customer credit | Nominal |
| 2 | Purpose | The loan purpose | Nominal |
| 3 | Gender | Male or female | Nominal |
| 4 | Credit_amount | The amount of credit | Numeric |
| 5 | Age | Customer Age | Numeric |
| 6 | Housing | Rent, own or for free | Nominal |
| 7 | Job | Is the customer has job | Nominal |
| 8 | Class | The class of loan good/bad | Nominal |

## RESULTS AND DISCUSSION

*J48 Decision Tree Prediction Model***:** From the result of the analysis made on the dataset using J48 Decision Tree Classifier to train the data and validate the model developed using 10-fold cross validation, it was discovered that the J48 Decision Tree prediction model made 868 correct classifications and 7 incorrect classifications of the output of the loan risk.. From the results of the analysis made on the dataset using Decision

Tree classification in developing the predictive model for loan risk in Bank, the following are also discovered: that out of the 722 classified as GOOD, 719 are actually correctly classified as GOOD while 3 data was misclassified BAD, and out of the 153 data classified as BAD, 149 was correctly classified as BAD while 4 was misclassified as GOOD. Table 2 below shows that J48 Decision Tree prediction model has an accuracy of 99.2%, 868 data were correctly classified and 7 incorrectly classified. The true positive (TP), false positive rate (FP), precision, area under ROC, mean absolute error (MAE), root mean square error (RMSE) and relative absolute error (RAE) values are 0.992, 0.022, 0.992, 0.975, 0.0158, 0.0896 and 5.4653% respectively.

*Bayesnet Prediction Model:* From the results of the analysis made on the dataset using BayesNet classifier to train the data and validate the model developed using 10-fold cross validation, it was discovered that the BayesNet prediction model made 864 correct classifications and 11 incorrect classifications of the output of the loan risk. From the results of the analysis made on the dataset using BayesNet classification in developing the predictive model for loan risk in Bank, the following are also discovered: that out of 722 data classified as GOOD, 716 are actually correctly as GOOD while 6 data which ought to be classified as GOOD was misclassified as BAD and out of 153 data classified as BAD, 148 was correctly classified as BAD while 5 was misclassified as GOOD.

**Table 2:** Summary of Results for J48 Decision Tree Prediction Model

| | |
|---|---|
| Number of data supplied | 875 |
| Correct Classification | 868 |
| Incorrect Classification | 7 |
| Accuracy | 99.2% |
| TP rate | 0.992 |
| FP rate | 0.022 |
| Precision | 0.992 |
| Area under ROC | 0.975 |
| Mean Absolute Error | 0.0158 |
| Root Mean Squared Error | 0.0896 |
| Relative Absolute Error | 5.4653% |

Table 3 below shows that BayesNet prediction model has an accuracy of 98.7429%, 864 data were correctly classified and 11 incorrectly classified. The true positive rate (TP), false positive rate (FP), precision, area under ROC, mean absolute error (MAE), root mean square error (RMSE) and relative absolute error (RAE) values are 0.987, 0.028, 0.987, 0.993, 0.0293, 0.1152 and 10.1416% respectively.

**Table 3:** Summary of Results for BayesNet Prediction Model

| | |
|---|---|
| Number of data supplied | 875 |
| Correct Classification | 864 |
| Incorrect Classification | 11 |
| Accuracy | 98.7429% |
| TP rate | 0.987 |
| FP rate | 0.028 |
| Precision | 0.987 |
| Area under ROC | 0.993 |
| Mean Absolute Error | 0.0293 |
| Root Mean Squared Error | 0.1152 |
| Relative Absolute Error | 10.1416% |

*Naïve Prediction Model:* From the results of the analysis made on the dataset using Naïve Bayes classifier to train the data and validate the model developed using 10-fold cross validation, it was discovered that the Naïve Bayes prediction model made 858 correct classifications and 17 incorrect classifications of the output of the loan risk. From the results of the analysis made on the dataset using Naïve Bayes classification in developing the predictive model for loan risk in Bank, the following are also discovered: that out of 722 data classified as GOOD, 715 are actually correctly as GOOD while 7 data which ought to be classified as GOOD was misclassified as BAD and out of 153 data classified as BAD, 153 was correctly classified as BAD while 10 was misclassified as GOOD.

Table 4 below shows that Naïve Bayes prediction model has an accuracy of 98.0571%, 858 data were correctly classified and 17 incorrectly classified. The true positive rate (TP), false positive rate (FP), precision, area under ROC, mean absolute error (MAE), root mean square error (RMSE) and relative absolute error (RAE) values are 0.981, 0.056, 0.98, 0.992, 0.0332, 0.1254 and 11.4755% respectively.

The development of predictive model for loan risk in banks using the different methods: J48 Decision Tree, BayesNet classifier and Naïve Bayes classifier proved to be very effective in determining the loan risk in Banks having provided information on the credit_history, purpose, gender, credit_amount, age, housing, job and class of the loan applicant. The three prediction models were able to predict the loan risk in bank for the data provided via training and the developed models with the results identified were evaluated for their respective performance as shown in table 5. It is shown from table 5 that J48 algorithm proved to be the best algorithm for the classification of loan risk because the algorithm has high accuracy and low mean absolute error.

Equally, J48 algorithm has the potential of correctly classifying the instances than the other techniques. In iteration of the experiment, different set sizes of training and test were used (70% training 30% test set, 60% training 40% test and 80% training 20% test) and the same results were obtained in each round showing that J48 algorithm presented the optimal results in classifying loans as either good or bad loan. This finding is in agreement with Aboobyda & Tarig (2016) who established that the best algorithm for loan classification is J48 algorithm.

**Table 4** Summary of Results for Naïve Bayes Prediction Model

| | |
|---|---|
| Number of data supplied | 875 |
| Correct Classification | 858 |
| Incorrect Classification | 17 |
| Accuracy | 98.0571% |
| TP rate | 0.981 |
| FP rate | 0.056 |
| Precision | 0.98 |
| Area under ROC | 0.992 |
| Mean Absolute Error | 0.0332 |
| Root Mean Squared Error | 0.1254 |
| Relative Absolute Error | 11.4755% |

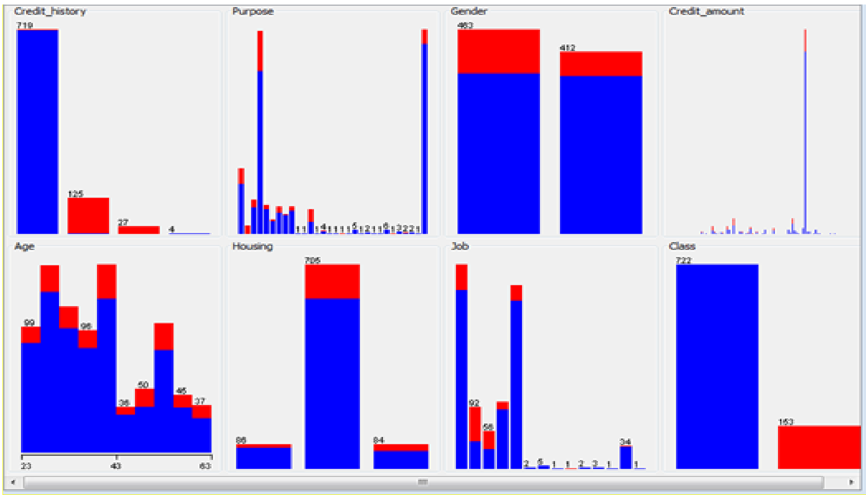

**Fig 1:** Data Visualization

**Table 5:** The Result from the three Algorithms

| | J48 Decision Tree Model | BayesNet Model | Naïve Bayes Model |
|---|---|---|---|
| Accuracy | 99.2% | 98.7429% | 98.0571% |
| Correct Classification | 868 | 864 | 858 |
| Incorrect Classification | 7 | 11 | 17 |
| TP rate | 0.992 | 0.987 | 0.981 |
| FP rate | 0.022 | 0.028 | 0.056 |
| Precision | 0.992 | 0.987 | 0.98 |
| Area under ROC | 0.975 | 0.993 | 0.992 |
| Mean Absolute Error | 0.0158 | 0.0293 | 0.0332 |
| Root Mean Squared Error | 0.0896 | 0.1152 | 0.1254 |
| Relative Absolute Error | 5.4653% | 10.1416% | 11.4755% |

Figure 1 showed data visualization of variables used in this paper. These include Credit-history, Purpose, Gender, Credit-amount, Age, Housing, Job and Class.

*Conclusion:* In this paper, we presented a data mining-based three algorithms (J48 Decision Tree, BayesNet and Naïve Bayes) to build a model that can be used to classify and predict the applications of loans from the bank customers as either good or bad. The system presents an automated way of investigating the behaviour of the customers using previous credit history. The predictive model was implemented on WEKA software. We find out that the best algorithm for loan risk classification is J48 Decision Tree as it has high accuracy and low mean absolute error.

## REFERENCES

Aboobyda, JH; Tarig, MA (2016). Developing Prediction Model of Loan Risk in Banks using Data Mining. Machine Learning and Applications: *An Intl. J. (MLAIJ)* Vol.3, No.1.

Amanze, BC; Onukwugha, CG (2017). Loan Fraud Detection System for Banking Industries in Nigeria Using Data Mining and Intelligent Agents: The Way Forward. *Intl. J. Innov. Rsrc. in Technol. Basic. Appl. Sci.*

Central Bank of Nigeria (2000). Risk Management Guidelines for Commercial Banks and Deposit Financial Institutions. Retrieved from www.cenbank.org 2020.

Ogawa, S; Park, J; Nita, T (2013). Financial Interconnectedness and Financial Sector. Reforms in the Caribbean. No. 13-175. International Monetary Fund.

Ogwueleka, FN (2011). Data Mining Application in Credit Card Fraud Detection System. *J. of Eng. Sci. Technol.* 6 (3): 311-322.

Ojo, AT (2010). The Nigerian Maladapted Financial System: Reforming Tasks and Development Dilemma, CIBN Press Limited, Lagos.

Olawale, FK; Tomola, MO; James, AO; Felix, AA (2015). Credit Risk and BankPerformance in Nigeria. *IOSR J. of Econ. Fin.*. 6 (2): 21-28.

Sulaimon, OM (2001). Asset Quality: A major Determinant of the performance of Bank Management, M.Sc. Degree Dissertatrion submitted to the Department of Finance, University of Calabar, Nigeria.