https://dx.doi.org/10.4314/dujopas.v11i1b.32
ISSN (Print): 2476-8316
ISSN (Online): 2635-3490

Dutse Journal of Pure and Applied Sciences (DUJOPAS), *Vol. 11 No. 1b March 2025*

An Optimized Light-GBM Based Classification Model for Effective Classification of Loan Defaulter

Adedeji Oluwaseun Bukonla

Department of Computer and Information Science,
Tai Solarin University of Education,
Ijagun,
Ogun state.

Email: adedejiob@tasued.edu.ng

Abstract

Risk analysis techniques are powerful tools that help professionals manage uncertainty and can provide valuable support for decision making. Recent techniques in the area of credit risk modeling have considered and adopted the use of artificial intelligence (AI), machine learning (ML) and deep learning (DL) algorithms. The purpose of this study is to apply an ensemble of machine learning algorithm to the classification of credit worthiness of a loan applicant. Data containing information about features associated with credit worthiness was collected from an online public data repository as a spreadsheet file that was stored in .csv format. The preprocessed dataset was split and fed to Light-GBM ensemble model to develop the classification models for credit worthiness using the holdout method over three simulation runs. The performance of each simulation run for each model was evaluated based on accuracy, recall, precision and f1-score. The study revealed that the ensemble learning model that was adopted in this study achieved very accurate results and proved to be more objective than subjective rule-based models. The results showed that there is a relative degree of importance that the features have with one another relative to the classification of credit worthiness. The study concluded that ensemble models are very effective in the classification of loan defaulter. This study recommended that future study could focus on determining the impact of feature importance on the performance of ensemble learning algorithms adopted for the classification of credit worthiness among loan applicants.

Keyword: Loan, Defaulters, Ensemble, classification, Model

INTRODUCTION

Risk analysis techniques are powerful tools that help professionals manage uncertainty and can provide valuable support for decision making (Bennett, Bohoris, Aspinwall, & Hall, 2020). Risk analysis helps in taking both certain and uncertain elements and include them in a calculation of specific scenarios of the future events. These techniques can be either qualitative or quantitative depending on the information available and the level of detail that is required (Darwish & Abdelghany, 2021 ; (Abdelmoula, 2023).)They are usually estimated with historical data and statistical methods. Many credit scoring models have been developed by researchers for the credit admission decision (Mammadli, 2023).

Traditional financial institutions evaluated creditworthiness of the borrowers based on subjective methods which focus mainly on the 5Cs: character, capacity, collateral, capital, and conditions (Soni & Varghese, 2019). This method largely is unable to assess the borrowers who have no loan history and have limited banking transactions, particularly the customers residing in rural areas (Taiwo *et al.*, 2020). However, these traditional methods failed to provide a comprehensive profile of a potential borrower. In addition to traditional scoring models, complexity and repetitiveness of decisions in the finance area have made financial services sector an area where various kinds of expert systems have found their many applications (Mohammed & Salama, 2023). Recent techniques in the area of credit risk modeling have considered and adopted the use of artificial intelligence (AI), machine learning (ML) and deep learning (DL) algorithms (West, 2020).

Credit risk predictions, monitoring, model reliability and effective loan processing are key to decision making and transparency. As a result of this, banking institutions are adopting more advanced methods for credit assessment by employing ML technologies which are mainly helpful in the prediction of a borrower's repayment behavior (Tijani & Abdullahi, 2021 There are several types of ensemble modeling approaches, each with its unique methodology. Here are some of the main types of ensemble methods (Vinayaka & Gupta, 2020).;Yewale, et al. 2023):

a. **Bagging** (Bootstrap Aggregating): Using various randomly selected portions of the training data, bagging entails training several versions of the same fundamental learning algorithm. By sampling the training data with replacement samples (bootstrap samples), these subsets are produced. By averaging (for regression) or voting (for classification), respectively, the predictions of various models, the final prediction is obtained. Random Forest is a popular example of a bagging ensemble method based on decision trees.
b. **Boosting**: Boosting is an iterative ensemble technique that emphasizes training weak learners consecutively while making an effort to fix mistakes made by earlier models. Giving incorrectly classified occurrences additional weight allows the next models to concentrate more on these challenging cases. AdaBoost (Adaptive Boosting) and Gradient Boosting Machines (GBM) are common boosting algorithms.

Iwasokun, and Olojo, (2022), worked on the development of a rule-based loan eligibility model for agricultural loans. The study identified 15 risk factors that were associated with the assessment of loan eligibility among farmers. The study adopted a neuro-fuzzy model to formulate the predictive model by adopting the identified risk factors to a fuzzy inference engine which was in turn fed to a neural network model. The model was testing using a sample dataset. The results of the study revealed that the predictive model was able to achieve an accuracy of 95%. The study was limited to risk factors that were required for assessing loan eligibility among farmers. Hasan, Elghareeb, Faragat and AboElfotouh (2021), worked on the application of fuzzy logic modeling to credit risk modeling for an organization. The variables considered in this study included information about the personal information of clients and their loan history alongside non-financial managerial factors. The study generated linguistic variables from 84 input variables for which 241 inference rules were extracted based on expert knowledge about the relationship between the variables and credit risk. The model was validated using sample data from selected organization. The results of the study revealed that the validation of the performance of the fuzzy logic model achieved an accuracy of 95.3%.

Hasan, Elghareeb, Farahat, and Elfotouh (2021) worked on the application of fuzzy logic modeling to credit risk modeling in Egypt. The study identified several non-financial factors which are relevant for credit risk scoring. Inference rules were generated from the non-financial factors based on the information provided by the experts at the Bank. The fuzzy model was implemented as a web-based system following which real-life data was used to validate its performance. The results of the study revealed that the validation of the performance of the fuzzy logic model achieved an accuracy of 95.3%. The model is dependent on human knowledge which may likely be biased thus lacking an objective representation of relationship between the factors and credit risk.

Kin, Aizam, Hasan, Ariffin and Mahat (2021), worked on the application of fuzzy logic to the development of a bankruptcy prediction model. The study identified a number of qualitative factors that are associated with the risk of bankruptcy, namely: industrial risk, management risk, financial flexibility, credibility, competitiveness and operational risk. The study adopted the use of 3 trapezoidal fuzzy membership function for representing the linguistic variables: negative, average and positive. Fuzzy inference rules were generated using the identified factors. The performance of the model was validated using real-life dataset. The results of the study revealed that 25 inference rules were generated for the prediction model and the evaluation showed a performance of 99.2%. The study was focused on assessing bankruptcy from the perspective of a corporate body.

There is a need to identify the most relevant features that can improve the identification of loan defaulters using an ensemble of machine learning algorithms, hence this study. This study is aimed at developing an ensemble model with the purpose of improving the performance of the predictive model. This ensemble model will be adopted for the classification of loan defaulters thus reducing errors associated with using traditional machine learning algorithms.

METHODS AND MATERIALS

Identification and Collection of Loan Defaulters' Dataset

In this study, the various features that were required for the classification of loan defaulters were identified from related works (month of payment, Age, Occupation e.t.c.) surrounding the body of knowledge on the assessment of loan defaulters. Table 1 provides a description of the features that were considered for the classification of loan defaulters. According to the features identified in the table, it was observed that the features were generally classified into four main groups, namely: customer profile information, customer credit information, customer payment information and customer account information. The various features were stored using values that were either numeric or categorical in nature. As shown in the table, the study identified seven features including the target class with categorical values and seventeen features with numeric values.

Historical data containing information about the features that are associated with the classification of loan defaulters was collected from an online public data repository that was provided by Kaggle.

Table 1: Identification of features associated with loan defaulters

Class of Variable	Name	Label values
Customer Profile Information	**Month of payment**	Categorical (January to August)
	Age of customer	Numeric – Integer type
	Occupation	Categorical
	Annual income (N)	Numeric – Float type
	Month income (N)	Numeric – Float type
	Number of bank account	Numeric – Integer type
	Number of credit cards	Numeric – Integer type
Customer Credit Information	**Interest rate (%)**	Numeric – Float type
	Type of Loan	Categorical
	Number of Loan	Numeric – Integer Type
	Changed credit limit	Numeric – Float type
	Number of credit inquiries	Numeric – Integer type
	Credit mix	Categorical (Poor, Good, Standard)
	Outstanding debt (N)	Numeric – Float type
	Credit utilization ratio	Numeric – Float type
	Credit history age (in months)	Numeric – Integer type
Customer Payment Information	**Payment of minimum amount**	Categorical (Yes, No, No minimum)
	Delay from due date (in days)	Numeric – Integer type
	Number of delayed payments	Numeric – Integer type
	Payment behavior	Categorical (Spendings & Payments)
Customer Account Information	**Total EMI per month (N)**	Numeric – Float type
	Amount invested monthly (N)	Numeric – Float type
	Monthly balance (N)	Numeric – Float type
Target Class	**Loan Defaulter**	Categorical (Yes, No)

Method of Data-Preprocessing

This section presents the various libraries that were adopted for the purpose of preprocessing the dataset that was collected in this study in order to be acceptable by the deep learning algorithms adopted in this study.

```
from sklearn.feature_selection import mutual_info_classif
# determine the mutual information
mutual_info = mutual_info_classif(x, y)
mutual_info
```

```
mutual_info = pd.Series(mutual_info)
mutual_info.index = x.columns
mutual_info.sort_values(ascending=False)
```

Identification of Feature Importance

The dataset was subjected to a feature selection algorithm called the mutual information metrics for the assessment of feature importance among the identified features in the dataset.

This task was required for providing insights into the relevance of the features towards the classification of loan defaulter.
The dataset was split into two such that the input features were stored as x while the target variable was stored as y. The library *sklearn.feature_selection* was used to implement the *mutual_info_classify* function which was used to implement the features importance from the dataset.

RESULTS AND DISCUSSION

Result of the Identification and Collection of Data

The study initially identified a number of features that are known to be associated with the classification of loan defaulters from related works following which data containing information about the features was collected from credit institutions in Nigeria. The dataset consists of information about 23 features which were collected from 100000 records. The dataset consists of information collected from 46826 poor records and 53174 standard records.

Month	Age	Occupation	Annual Income	Monthly Inhand Salary	Num Bank Accounts	Num Credit Card	Interest Rate	Num of Loan	Type of Loan	Delay from due date	Num of Delayed Payment	Changed Credit Limit	Num Credit Inquiries	Credit Mix	Outstanding Debt	Credit Utilization Ratio	Credit History Age	Payment of Min Amount	Equated Monthly Installments	Amount invested monthly	Payment Behaviour	Monthly Balance	Credit Score
January	23	Scientist	1911412	182484.3	3	4	3	4	ıto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Lo	3	7	11.27	4	Good	80998	26.82262	265	No	49574.94921	21465.38	_Small_valu	31249.41	Poor
February	23	Scientist	1911412	182484.3	3	4	3	4	ıto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Lo	3	4	11.27	4	Good	80998	31.94496	266	No	49574.94921	21465.38	_Large_value	28462.92	Poor
March	23	Scientist	1911412	182484.3	3	4	3	4	ıto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Lo	3	7	11.27	4	Good	80998	28.609352	267	No	49574.94921	21465.38	ledium_val	33120.99	Poor
April	23	Scientist	1911412	182484.3	3	4	3	4	ıto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Lo	5	4	6.27	4	Good	80998	31.377862	268	No	49574.94921	21465.38	_Small_value	22345.13	Poor
May	23	Scientist	1911412	182484.3	3	4	3	4	ıto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Lo	6	4	11.27	4	Good	80998	24.797347	269	No	49574.94921	21465.38	ledium_val	34148.92	Poor
June	23	Scientist	1911412	182484.3	3	4	3	4	ıto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Lo	8	4	9.27	4	Good	80998	27.262259	270	No	49574.94921	21465.38	ledium_val	34047.92	Poor
July	23	Scientist	1911412	182484.3	3	4	3	4	ıto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Lo	3	8	11.27	4	Good	80998	22.537593	271	No	49574.94921	21465.38	_Small_value	24456.53	Poor
February	28	Teacher	3484784	303798.7	2	4	6	1	Credit-Builder Loan	7	1	7.42	2	Good	60503	38.550848	320	No	18816.21457	39684.018	_Large_value	48459.12	Poor
April	28	Teacher	3484784	303798.7	2	4	6	1	Credit-Builder Loan	3	3	5.42	2	Good	60503	39.182656	322	No	18816.21457	39684.018	ledium_val	46567.62	Poor
May	28	Teacher	3484784	303798.7	2	4	6	1	Credit-Builder Loan	3	1	6.42	2	Good	60503	34.977895	323	No	18816.21457	39684.018	_Small_value	44486.7	Poor
June	28	Teacher	3484784	303798.7	2	4	6	1	Credit-Builder Loan	3	0	5.42	2	Good	60503	33.38101	324	No	18816.21457	39684.018	_Large_value	48150.53	Poor
July	28	Teacher	3484784	303798.7	2	4	6	1	Credit-Builder Loan	3	4	5.42	2	Good	60503	31.131702	325	NM	18816.21457	39684.018	ledium_val	46488.07	Poor
August	28	Teacher	3484784	303798.7	2	4	6	1	Credit-Builder Loan	3	4	5.42	2	Good	60503	32.933856	326	No	18816.21457	39684.018	_Small_value	35607.81	Poor
January	34	Engineer	1.4E+07	1218722	1	5	8	3	Auto Loan, Auto Loan, and Not Specified	5	8	7.1	3	Good	130301	28.616735	213	No	246992.3195	168413.7	_Small_valu	104331.6	Poor
February	34	Engineer	1.4E+07	1218722	1	5	8	3	Auto Loan, Auto Loan, and Not Specified	13	6	7.1	3	Good	130301	41.702573	214	No	246992.3195	168413.7	_Small_valu	99886.93	Poor
March	34	Engineer	1.4E+07	1218722	1	5	8	3	Auto Loan, Auto Loan, and Not Specified	8	7	11.1	3	Good	130301	26.519815	215	No	246992.3195	168413.7	_Small_valu	71574.14	Poor
April	34	Engineer	1.4E+07	1218722	1	5	8	3	Auto Loan, Auto Loan, and Not Specified	8	5	9.1	3	Good	130301	39.501648	216	No	246992.3195	168413.7	ledium_val	42651.34	Poor
May	34	Engineer	1.4E+07	1218722	1	5	8	3	Auto Loan, Auto Loan, and Not Specified	10	5	7.1	3	Good	130301	31.37615	217	No	246992.3195	168413.7	_Large_value	81078.22	Poor
June	34	Engineer	1.4E+07	1218722	1	5	8	3	Auto Loan, Auto Loan, and Not Specified	8	6	7.1	3	Good	130301	39.783993	218	No	246992.3195	168413.7	ledium_val	96392.16	Poor
January	31	Lawyer	7392846	598870.5	4	5	8	0	No Data	12	10	10.14	2	Good	54820	39.962685	384	No	0	42635.59	_Large_value	74019.61	Poor
February	31	Lawyer	7392846	598870.5	4	5	8	0	No Data	8	7	10.14	2	Good	54820	42.769864	384	NM	0	42635.59	ledium_val	70593.13	Poor
March	31	Lawyer	7392846	598870.5	4	5	8	0	No Data	8	7	10.14	2	Good	54820	40.712187	385	No	0	42635.59	ledium_val	69881.08	Poor
April	31	Lawyer	7392846	598870.5	4	5	8	0	No Data	8	7	10.14	2	Good	54820	30.201658	386	No	0	42635.59	_Small_value	27066.81	Poor
May	31	Lawyer	7392846	598870.5	4	5	8	0	No Data	11	7	10.14	2	Good	54820	26.33331	387	No	0	42635.59	_Large_value	69091.87	Poor
June	31	Lawyer	7392846	598870.5	4	5	8	0	No Data	7	7	10.14	2	Good	54820	35.275437	388	No	0	42635.59	_Large_value	63308.02	Poor
August	31	Lawyer	7392846	598870.5	4	5	8	0	No Data	8	7	10.14	2	Good	54820	31.58099	390	No	0	42635.59	_Large_value	79623.49	Poor
January	33	Lawyer	1.3E+07	1124278	0	1	8	2	Credit-Builder Loan, and Mortgage Loan	0	3	9.34	2	Good	35216	32.200509	367	NM	137644.6054	86566.388	ledium_val	85846.25	Poor
February	34	Lawyer	1.3E+07	1124278	0	1	8	2	Credit-Builder Loan, and Mortgage Loan	0	2	15.34	4	Good	35216	31.98371	368	No	137644.6054	86566.388	_Small_valu	54776.05	Poor
March	34	Lawyer	1.3E+07	1046921	0	1	8	2	Credit-Builder Loan, and Mortgage Loan	0	3	9.34	4	Good	35216	31.803135	369	NM	911220.1793	86566.388	_Large_value	103856.9	Poor
April	34	Lawyer	1.3E+07	1046921	0	1	8	2	Credit-Builder Loan, and Mortgage Loan	0	2	8.34	4	Good	35216	42.645785	370	No	911220.1793	86566.388	ledium_val	89919.88	Poor
May	34	Lawyer	1.3E+07	1046921	0	1	8	2	Credit-Builder Loan, and Mortgage Loan	0	4	9.34	4	Good	35216	40.902517	371	No	911220.1793	86566.388	_Large_value	96325.48	Poor
July	34	Lawyer	1.3E+07	1046921	0	1	8	2	Credit-Builder Loan, and Mortgage Loan	0	2	9.34	4	Good	35216	26.947565	373	No	911220.1793	86566.388	ledium_val	32624.18	Poor
August	34	Lawyer	1.3E+07	1046921	0	1	8	2	Credit-Builder Loan, and Mortgage Loan	0	2	9.34	4	Good	35216	29.187913	374	No	911220.1793	86566.388	ledium_val	39611.13	Poor
January	23	Doctor	1.1E+07	984386.8	2	5	7	3	Personal Loan, Debt Consolidation Loan, and Auto Loan	13	11	8.24	3	Good	137774	33.664554	256	No	226892.7919	212235.6	_Small_valu	80230.04	Poor
February	23	Doctor	1.1E+07	984386.8	2	5	7	3	Personal Loan, Debt Consolidation Loan, and Auto Loan	14	8	8.24	3	Good	137774	27.626325	257	NM	226892.7919	212235.6	_Large_value	78525.84	Poor
March	23	Doctor	1.1E+07	984386.8	2	5	7	3	Personal Loan, Debt Consolidation Loan, and Auto Loan	11	11	8.24	3	Good	137774	35.141567	258	NM	226892.7919	212235.6	_Small_valu	54710.82	Poor

January	25	Doctor	1.1E+07	845888.4	5	3	13	7	ay Loan, Auto Loan, Credit-Builder Loan, Payday Loan, Payday L	6	10	12.09	5	Standard	41824	36.806276	86	Yes	490574.3386	85529.729	Medium_val	6107.629	Standard
February	25	Doctor	1.1E+07	926696.3	5	3	13	7	ay Loan, Auto Loan, Credit-Builder Loan, Payday Loan, Payday L	10	7	12.09	5	Standard	41824	29.38609	87	Yes	490574.3386	85529.729	Medium_val	6107.629	Standard
March	25	Doctor	1.1E+07	926696.3	5	3	13	7	ay Loan, Auto Loan, Credit-Builder Loan, Payday Loan, Payday L	10	10	12.09	5	Standard	41824	34.009956	88	Yes	490574.3386	85529.729	Small_value	6107.629	Standard
June	26	Doctor	1.1E+07	926696.3	5	3	13	7	ay Loan, Auto Loan, Credit-Builder Loan, Payday Loan, Payday L	10	13	12.09	7	Standard	41824	24.725029	91	Yes	490574.3386	85529.729	Large_value	54334.04	Standard
July	26	Doctor	1.1E+07	845888.4	5	3	13	7	ay Loan, Auto Loan, Credit-Builder Loan, Payday Loan, Payday L	10	10	12.09	7	Standard	41824	35.623143	92	Yes	1298652.476	85529.729	Medium_val	37699.53	Standard
August	26	Doctor	1.1E+07	845888.4	5	3	13	7	ay Loan, Auto Loan, Credit-Builder Loan, Payday Loan, Payday L	10	10	12.09	7	Standard	41824	36.121461	93	Yes	1298652.476	85529.729	Large_value	59059.22	Standard
January	20	Lawyer	5356488	420774	5	3	13	2	Mortgage Loan, and Personal Loan	5	10	18.98	2	Standard	100831	41.03916	190	Yes	80569.83299	46538.148	Medium_val	44341.55	Standard
February	20	Lawyer	5356488	420774	5	3	13	2	Mortgage Loan, and Personal Loan	6	7	19.98	2	Standard	100831	28.206586	190	Yes	80569.83299	46538.148	Small_value	38471.35	Standard
March	20	Lawyer	5356488	420774	5	3	13	2	Mortgage Loan, and Personal Loan	6	10	19.98	2	Standard	100831	30.472553	191	Yes	80569.83299	46538.148	Small_value	26171.6	Standard
April	20	Lawyer	5356488	420774	5	3	13	2	Mortgage Loan, and Personal Loan	6	11	19.98	2	Standard	100831	34.548066	192	Yes	80569.83299	46538.148	Small_value	39127.12	Standard
May	20	Lawyer	5356488	420774	5	3	13	2	Mortgage Loan, and Personal Loan	6	10	23.98	2	Standard	100831	40.343872	193	Yes	80569.83299	46538.148	Medium_val	54366.6	Standard
June	20	Lawyer	5356488	420774	5	3	13	2	Mortgage Loan, and Personal Loan	3	10	19.98	2	Standard	100831	34.321454	194	Yes	80569.83299	46538.148	Small_value	25902.01	Standard
July	20	Lawyer	5356488	420774	5	3	13	2	Mortgage Loan, and Personal Loan	6	10	19.98	2	Standard	100831	31.260939	195	NM	80569.83299	46538.148	Medium_val	44779.84	Standard
August	21	Lawyer	5356488	420774	5	3	13	2	Mortgage Loan, and Personal Loan	6	10	19.98	2	Standard	100831	25.891431	196	Yes	80569.83299	46538.148	Medium_val	43494.19	Standard

Figure 1: Screenshot of loan defaulters dataset

Result of the Simulation and Evaluation of Predictive Models

The result of the simulation of the predictive models using the supervised machine learning algorithm considered in this study based on the dataset that were generated in this study. Each of the dataset was split into a training and testing proportion such that the training set was used to build the predictive models using the deep learning algorithm while the testing dataset was used to evaluate the performance of the model created by the deep learning algorithms. Table 2 shows the proportion of the records that were contained in the training dataset and the testing dataset that was performed over 5 simulations. In simulation 1, 60% of the dataset was used for training and 40% of the dataset was used for testing the predictive model for each dataset using the deep learning algorithms such that 60000 records were used to build the predictive model following which the model was validated using 40000 records in the test set. In simulation 2, 70% of the dataset was used for training and 30% of the dataset was used for testing the predictive model for each dataset using the deep learning algorithms such that 70000 records were used to build the predictive model following which the model was validated using 30000 records in the test set. In simulation 3, 80% of the dataset was used for training and 20% of the dataset was used for testing the predictive model for each dataset using the deep learning algorithms such that 80000 records were used to build the predictive model following which the model was validated using 20000 records in the test set.

Table 1: Results of the number of records stored in the training and testing records

Simulation Runs	Train Data			Test Data		
Simulation# (Train/Test Proportion)	**Standard**	**Poor**	**Total**	**Standard**	**Poor**	**Total**
Simulation1 (60/40)	33104	26896	**60000**	20070	19930	**40000**
Simulation2 (70/30)	38210	31790	**70000**	14964	15036	**30000**
Simulation3 (80/20)	43244	36756	**80000**	9930	10070	**20000**

Results of the Evaluation of the Predictive Model

The results of the evaluation of the predictive models that were generated across the five simulations based on the machine learning and ensemble modeling technique that was adopted in this study. The results are presented for each simulation following which the results of the performance of the algorithms were presented.

Evaluation of predictive models in simulation 1

As stated earlier, simulation 1 was validated using a dataset that composed of 40% of the dataset and consisted of 20070 standard and 19930 poor records. Figure 3 shows the confusion matrices that were used to interpret the results of the evaluation of the ensemble learning models adopted in simulation 1 based on the test dataset. Using LightGBM classifier, it was observed that 20069 out of the 20070 standard records were correctly classified and all 19930 poor records were correctly classified owing to an accuracy of 99.99%. The results revealed that the LightGBM classified showed good performance.

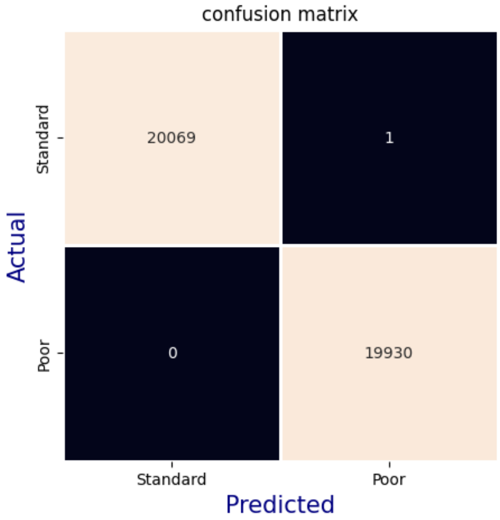

Figure 3: Confusion matrices for the evaluation of LightGBM for simulation 1.

Evaluation of predictive models in simulation 2

As stated earlier, simulation 2 was validated using a dataset that composed of 30% of the dataset and consisted of 14964 standard and 15036 poor records. Figure 4. shows the confusion matrices that were used to interpret the results of the evaluation of both ensemble learning model adopted in simulation 2 based on the test dataset. Using LightGBM classifier, it was observed that all the 14964 standard records were correctly classified and all 15036 poor records were correctly classified owing to an accuracy of 100.00%. The results revealed that the LightGBM performed better than the simulation 1.

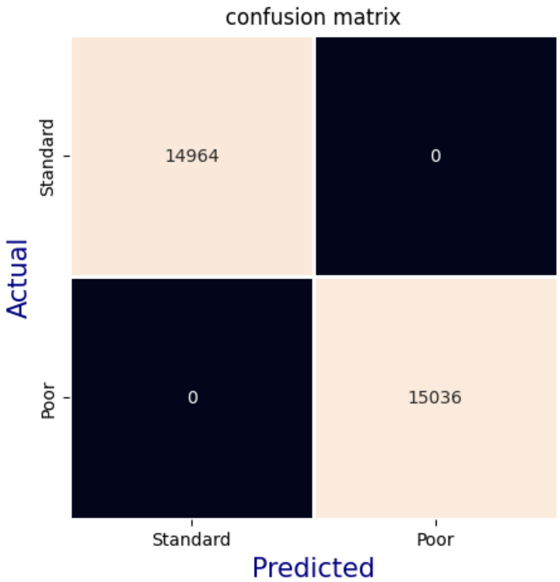

Figure 4: Confusion matrice for the evaluation of LightGBM for simulation 2.

Evaluation of predictive models in simulation 3

As stated earlier, simulation 3 was validated using a dataset that composed of 10% of the dataset and consisted of 9930 standard and 10070 poor records. Figure 5. shows the confusion matrice that was used to interpret the results of the evaluation of ensemble learning model adopted in simulation 3 based on the test dataset. Using LightGBM classifier, it was observed that all the 9930 standard records were correctly classified and all 10070 poor records were correctly classified owing to an accuracy of 100.00%.. The results revealed that the LightGBM classified showed good performance.

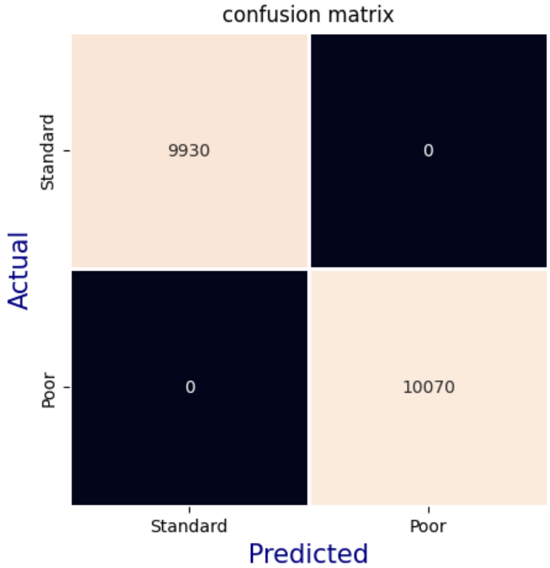

Figure 5: Confusion matrice for the evaluation of LightGBM for simulation 3.

Table 2: Results of the evaluation of the predictive models across five simulations based on performance metrics.

Simulation# (Train/Test Proportion)	Ensemble Model	Number correctly classified	Accuracy (%)	Precision		Recall		F1-score	
				Standard	Poor	Standard	Poor	Standard	Poor
Simulation 1 (60/40)	LightGBM	39999	99.99	1.000	1.000	1.000	1.000	1.000	1.000
Simulation 2 (70/30)	LightGBM	30000	100.00	1.000	1.000	1.000	1.000	1.000	1.000
Simulation 3 (80/20)	LightGBM	20000	100.00	1.000	1.000	1.000	1.000	1.000	1.000

DISCUSSION OF RESULTS

This section presents the discussion of the results of the various analyses that were performed in this study for the development of a predictive model required for the classification of loan defaulters among customers. The study revealed that the ensemble learning model that was adopted in this study achieved very accurate results and proved to be more objective than subjective rule-based models as proposed by Iwasokun, and Olojo, (2022). And proved to show better performance compared to the study by Hasan, Elghareeb, Faragat and AboElfotouh (2021). Unlike the approach that was made in the study by Hasan, Elghareeb, Farahat, and Elfotouh (2021); this study tried to assess the relative level of importance of the features that were identified to be associated with the classification of loan defaulters among customers seeking loans.

CONCLUSION

It was revealed in this study that information collected from a number of limited features can be used as a basis for the classification of the loan defaulters among customers. The study identified that each feature had a relative importance to one another regarding their usefulness in the classification of loan defaulters among customers seeking for loan. The study concluded that ensemble models are very effective in the classification of the academic performance of the student.

REFERENCES

Abdelmoula, A. (2023). Bank Credit Risk Analysis with k-Nearest Neighbour Classifier: Case of Tunisian Banks. *Accounting and Management Information Systems, 14*(1), 79-105.

Bennett, J., Bohoris, G., Aspinwall, E., & Hall, R. (2020). Risk analysis techniques and their application to software development. *European Journal of Operational Research, 95*(3), 467-475.

Darwish, N., & Abdelghany, A. (2021). A Fuzzy Logic Model for Credit Risk Rating of Egyptian Commercial Banks. *International Journal of Computer Science and Infromation Security, 14*(2), 11-18.

Iwasokun A., and Olojo G. (2022). Development of rule based loan eligibility model. *International journal of innovations in engineering and Technology.* 7(20), 532-542.

Kin T., Aizam A., A. H., Hasan A., Ariffin F., A. ., and Mahat N. (2021). Bankruptcy Prediction Model with Risk Factors using Fuzzy Logic Approach. *Journal of Computing Research and Innovation*, *6*(2), 104–113. https://doi.org/10.24191/jcrinn.v6i2.220

Mammadli, S. (2023). Fuzzy Logic based Loan Evaluation System. *Procedia Computer Science, 102*, 495-499.

Mohammed, A., & Salama, A. (2023). A Fuzzy Logic Based Model for Predicting Commercial Banks Financial Failure. *International Journal of Computer Applications, 79*(11), 1-12.

Soni, P., & Varghese, P. (2019). A Fuzzy based Data Mining Approach for the Loan CRedibility Prediction System in Co-operative Banking Sector. *International Journal of Advanced Trends in Computer Science and Engineering, 8*(3), 815-820.

Taiwo, J., Ucheaga, E., Achugamonu, B., Adetiloye, K., Okoye, L., & Agwu, M. (2017). CRedit Risk Management: Implications on Bank Performance and lending growth. *Saudi Journal of Business and Management Studies, 2*(5), 584-590.

Tijani, J., & Abdullahi, M. (2021). Banks and Credit Risk Management in Nigeria. *International Journal of Economics, Commerce and Management, 9*(12), 253-273.

West, D. (2020). Neural Network CRedit Scoring Models. *Computers and Operations Research, 36*(2/1), 1685-1694.

Yewale D., Vijayaragavan S., and Bairagi V. (2023). An effective herat diksease predikction framework based on ensemble techniques in machine learning. *International Journal of Advanced Computer Science and Application,* 14(2), 182-190.