AN IMPROVED PERFORMANCE EVALUATION OF CLASSIFICATION ALGORITHM FOR PREDICTION OF LOAN APPROVAL

[1]A.U. Farouk, A. [2]Abdulkadir, A.Y. Abarshi

[1]Department of Statistics, Federal Polytechnic, Bauchi
[2] Department of mathematical Science, Abubakar Tafawa Balewa University, Bauchi
Corresponding Author: Department of Computer science, Federal College of Animal & Production Technology, Vom
Aliyabarshi2@gmail.com
Tel: 0836212010

Abstract

The study aimed at evaluating the performance of five classification models used in predicting loan approval. A new future selection method was introduced to the loan dataset obtained online from kaggle to remove redundant features that can slow the algorithms and also improve the performance of the models. The dataset used consist of 12 variables and 689 data instances. R package v4.1.1 was used in implementing all the analysis conducted in the research work. The evaluation study proved that the new variable selection technique apart from removing the redundant feature also improves the performance of the models. Credit history proved to be best predictor of loan approval. The three performance evaluation metrics used unanimously showed that Naïve Bayes model outperformed Logistic Regression, Decision tree, support vector machine and Random forest algorithms. The overall accuracy and AUC of Naïve Bayes with 6 predictors are 83.2% and 79.2% respectively. Logistic regression with 6 predictors came second with overall accuracy and AUC of 81.6% and 73.7% respectively. Although Random forest with 9 predictors overall accuracy is higher than that of Logistic regression, Logistic regression is chosen to be the second best due to the number of features in the model and also the AUC. Based on the evaluation metric used the study concluded that Naïve Bayes is the best algorithms for predicting loan approval. It is also recommended that our new future selection method should be compare with other classic methods to validate its performance. Also other classification algorithms not used in this work should be compared with Naïve Bayes to authenticate our claim.
Keywords: Features, Machine learning, Loan, Prediction, Preprocessing

Introduction

Nowadays people rely on bank loan to fulfill their needs. The rate of loan applications increases with a very fast speed in recent years. Because of the risk involved in approval of loans, the banking officials are very conscious about the payment of the loan amount by its customers (Kumar et al., 2019).

Because of the risk involve in loan approval, there is need to automate the process to improve the accuracy of the prediction. The use of statistics and machine learning classification algorithms is the best solution. A Prediction Model uses data mining, statistics and probability to forecast an outcome. Every model has some variables known as predictors that are likely to influence future results. A statistical model is made using these predictors. As more data becomes available the model becomes more refined and accuracy of the model increases thereby reduces the risk of wrong prediction and also decreasing the time taken in the process.(Khan et al. 2021). The main target of this research work is to select the best subset of predictors that will give a good model for predicting loan approval using a new method of feature selection. Five machine learning algorithms will be used in training the dataset. Evaluation of the algorithms will be made in other to recommend the best algorithm for prediction of loan approval. This will help the banks in minimizing the risk associated with loan approval by reducing the number of defaulters.

Materials and Methods

Loans are the core business of banks. The main profit comes directly from the loan's interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don't have assurance if the applicant is able to repay the loan with no difficulties. Using Machine learning we predict the loan approval with better accuracy (Meshref 2020). In this study comparison of five machine learning is made. The machine learning algorithms include:

Logistic regression (LR)

Logistic regression (LR) is a widely used multivariable method for modeling dichotomous outcomes. Another concept which is used in logistic regression model is odds ratio. The term "odds" is defined differently according to the situation under discussion, but it is the ratio of the probability of occurrence of an event (P_i) to that of non-occurrence ($1 - P_i$).

The goal of LR is to find the best fitting and most parsimonious model to describe the relationship between the outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. The method is relatively robust, flexible and easily used, and it lends itself to a meaningful interpretation. In LR, unlike in the case of LDA, no assumptions are made regarding the distribution of the explanatory variables.

Logistic regression model is given by

$$logitP = ln\frac{P_i}{(1-P_i)} = \beta_0 + \beta_i X_j \qquad X = 1, 2 \ldots j \tag{1}$$

In order to perform prediction, we needed to estimate the coefficients: $\beta_0, \beta_1, \ldots \beta_n$ using a statistical method such as the maximum likelihood estimation. For the entire sample data, the maximum likelihood function could be given by:

$$L\left(\frac{\beta}{y}\right) = \prod_{i=1}^{n} P_i^{\,y_i} (1 - P_i)^{1-y_i} \tag{2}$$

Decision tree (DT)

Decision tree is a supervised ML technique that can be used for both classification and regression problems, but mostly is preferred for solving classification problems. It is a tree structured classifier, where an internal node represents the features of a dataset, branches represents the decision rules and each leaf node represents the outcome.

To build a decision tree, we use the classification and decision tree algorithm (CART). The decision tree simply asks a question and based on the answer (Yes/No), it further split the tree into sub-trees.

Random Forest (RF)

Random forests or random decision forests are an ensemble learning method for classification and regression. The algorithm operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random forests are frequently used as "black box" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

Naïve Bayes algorithm (NB)

Naïve Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Moreover, the Bayes theorem presumes that the input features also referred as predictors are independent in nature.
Similarly; Naïve Bayes also presumes that the input features are independent in nature. But this is impossible in the realistic procedures.
The Naïve Bayes is used to calculate the set of probabilities by counting the value and frequency of values in a given.

Support Vector Machine (SVM)

A Support Vector Machine is a supervised classification technique that can actually get pretty complicated but is pretty intuitive at the most fundamental level. Let's assume that there are two classes of data. A support vector machine will find a hyper plane or a boundary between the two classes of data that maximizes the margin between the two classes. There are many planes that can separate the two classes, but only one plane can maximize the margin or distance between the classes.

Review of Related Works

A lot of studies have been conducted using various machine learning algorithms to predict loan approval by the baking industries and other money lending institutions.

Aphale and Shinde (2020) carried out a study "Predict Loan Approval in Banking System: Machine Learning Approach for Cooperative Banks". The study compared Neural Network, Discriminant Analysis, Naïve Bayes, K-Nearest Neighbor and Linear Regression. The study concluded that apart from K-Nearest Neighbor and Naïve Bayes, the other three algorithms performs credibly well in terms of their accuracy and other performance metrics.

Chandra and Rekha (2019) conducted a study on "Exploring the Machine Learning Algorithms for Prediction of the Loan Sanctioning Process". The study compared Logistic Regression, Decision Tree, Support Vector Machine and Naïve Bayes. The study came to conclusion with confidence that the Naïve Bayes model is extremely efficient and gives a better result when compared to the other models.

Hamid and Ahmed (2016) carried out a study "Developing Prediction Model for Loan Risks in Banks using Data Mining". The study compared J48, BayesNet and Naïve Bayes. The study concluded that J48 model outperformed the other two based on overall accuracy.

Kadam et al. (2021) carried out a research "Prediction for Loan Approval using Machine Learning Algorithms". The study compared Support Vector Machine and Naïve Bayes algorithms. The study concluded that Naïve Bayes model is more efficient and give more accurate result than Support Vector Machine.
Kumar and Goel (2020) conducted a study on "Prediction of Loan Approval using Machine Learning Techniques". The study used decision tree to predict loan approval. The study concluded that Decision tree is a good algorithm for loan approval prediction modeling.

Madaan et al. (2021) conducted a research on "Loan default Prediction using Decision Tree and Random Forest: Comparative study" The study revealed that Random Forest model out performed the Decision tree with much higher accuracy.

Stiawan and Suharjit (2019) carried out a study “Comparison of Prediction methods for Credit Default on Peer to Peer Lending using Machine Learning Algorithms”. The study proposed a tree base classification method for the loan prediction, binary PSO with support vector machine was used for feature selection, random forest and extremely randomized tree as classifiers. The result showed that extremely randomized tree outperformed random forest in several performance metrics.

Pramod et al. (2021) conducted a study on “An Approach for Prediction of Loan Approval Using Machine Learning Algorithm. The study compared logistic regression and decision using different performance metrics such overall accuracy, sensitivity and specificity. The final results have shown that the model produce different result with decision tree outperforming logistic regression.

Nikam et al. (2021) carried out a research “Prediction for Loan Approval Using Machine Learning Algorithm” the study compared support vector machine, logistic regression and naïve Bayes algorithms. The experimental test found out that Logistic regression model has better performance than the other models in terms of loan forecasting.

PROPOSED METHODOLOGY

The research work is aimed at predicting whether a loan should be approved or not using five machine learning algorithms. To achieve this, the following steps are followed:

Importation of the dataset

The study used bank marketing data sets from kaggle. Bank Marketing data set at Kaggle is one of the frequently analyzed data sets using machine learning techniques. It contains 12 attributes and it has about 689 data instances.

Data cleaning and processing

The imported dataset passed through some pre-processing stages before it was used to train the five machine learning algorithms. The categorical missing values were replaced with the mode and continuous missing value with the mean or median. Normalization of the data was done using Min-max method to reduce the data to a range of 0 to 1.

Feature selection

An efficient and simple method was used to select the best subset of predictors from the set of predictors in the data set. Each algorithm was allowed to select the best subset of predictors to work with. For each algorithm, a total of 12 models were formed using the whole dataset and overall accuracy calculated in each case. The model that contained all the predictors serves as the cut off accuracy. Any model accuracy that is above the cutoff accuracy, the eliminated feature in the model is considered unimportant and so was eliminate. That is if $P_i \leq P_C$ retained the eliminated feature. The selection procedures is describe in table 1

Table 1: Feature selection procedure

Model	Variable eliminated	Variables in the model	Model accuracy	Cut off Accuracy
1	None	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}$	-	P_c
2	X_1	$X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}$	P_1	
3	X_2	$X_1, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}$	P_2	
4	X_3	$X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}$	P_3	
5	X_4	$X_1, X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}$	P_4	

6	X_5	$X_1, X_2, X_3, X_4, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}$	P_5	
7	X_6	$X_1, X_2, X_3, X_4, X_5, X_8, X_9, X_{10}, X_{11}, X_{12}$	P_6	
8	X_7	$X_1, X_2, X_3, X_4, X_5, X_6, X_8, X_9, X_{10}, X_{11}, X_{12}$	P_7	
9	X_8	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_9, X_{10}, X_{11}, X_{12}$	P_8	
10	X_9	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_{10}, X_{11}, X_{12}$	P_9	
11	X_{10}	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{11}, X_{12}$	P_{10}	
12	X_{11}	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{12}$	P_{11}	

P: Predictive accuracy of the model, X_i: are the predictor variables.

Data partitioning

In this study the dataset was split into training set and testing set. 70% of the data was used for training and 30% of the data for testing. After modeling using the training dataset, models were evaluated using the test dataset.

Data mining tool

The experimental data mining tool used was R foundation for statistical computing platform version 4.1.1 (2021). The R is an official part of the free Software environment for statistical computing and graphics.

Performance Evaluation Metrics Used

The study used confusion matrix and area under the receiver operating characteristics curve (AUROC). From the confusion matrix the following machine learning performance evaluation metric were calculated. The metrics are overall accuracy, sensitivity, specificity, misclassification error, F-measure. And AUC was calculated from the ROC. The metrics were obtained from the information on the confusion matrix using the following formulas

The confusion matrix of binary classification is given below.

		Observed	
		True	False
Predicted	True	True positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

Figure 1: Format of a confusion matrix

a. Accuracy:

It measures how often the classifier is correct for both true positives and true negative cases. Mathematically, it is defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

b. Sensitivity or Recall:

Measures how many times the classifier got the true positives correct. Mathematically, it is defined as:

$$Sensitivity = Recall = \frac{TP}{TP + FN} = TPR \tag{4}$$

c. Specificity: It measure how many times did the classifier get the true negatives correct. Mathematically, it is defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

d. Precision:

Precision measures off the total predicted to be positive how many were actually positive. Mathematically, it is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

e. $$F - measure = \frac{2(Precision \times Recall)}{Precision + Recall} \quad (7)$$

RESULTS

a. Feature selection using our proposed method

. The results of the feature selection are as presented below:

Table 2: Logistic regression feature selection

Variable	Associated accuracies (P_i) (%)	Cut off accuracy (P_c) (%)
-		81.73
Sex (X_1)	81.79	
Married (X_2)	81.76	
Number of dependants (X_3)	81.42*	
Education (X4)	81.76	
Working (X_5)	81.76	
Income (X_6)	81.42*	
Spouse Income (X_7)	81.25*	
Loan amount (X_8)	81.59*	
Monthly repayment (X_9)	81.76	
Credit History (X_{10})	69.59*	
Immobile (X_{11})	81.59*	
7 variables retained	X_3, X_6, X_7, X_8, X_{10}, X_{11}	

Source: Author's computation aided by R packages v 4.1.1. *important Features ($P_i \leq P_c$)

Table 3: Naïve Bayes Feature selection

Variable	Associated accuracies (P_i) (%)	Cut off accuracy (P_c) (%)
-		79.90
Sex (X_1)	80.07	
Married (X_2)	79.05*	
Number of dependants (X_3)	79.05*	
Education (X_4)	79.90*	
Working (X_5)	79.90*	
Income (X_6)	80.41	
Spouse Income (X_7)	80.42	
Loan amount (X_8)	79.90*	
Monthly repayment (X_9)	89.08	
Credit History (X_{10})	72.13*	
Immobile (X_{11})	89.08	
6 variables retained	**X_2, X_3, X_4, X_5,X_8, X_{10}**	

Source: Author's computation aided by R packages v 4.1.1. *important Features ($P_i \leq P_c$)

Table 4: Support vector Machine feature selection

Variable	Associated accuracies (P_i) (%)	Cut off accuracy (P_c) (%)
-		82.09
Sex (X_1)	82.19	
Married (X_2)	82.26	
Number of dependants (X_3)	82.09*	
Education (X_4)	82.26*	
Working (X_5)	82.09*	
Income (X_6)	82.09*	
Spouse Income (X_7)	81.09*	
Loan amount (X_8)	82.09*	
Monthly repayment (X_9)	82.26	
Credit History (X_{10})	72.45*	
Immobile (X_{11})	82.19	
7 variables retained	**X_3, X_4, X_5, X_6, X_7,X_8,X_{10}**	

Source: Author's computation aided by R packages v 4.1.1. *important Features ($P_i \leq P_c$)

Table 5: Decision Tree feature selection

Variable	Associated accuracies (P_i) (%)	Cut off accuracy (P_c) (%)
		81.42
Sex (X_1)	81.44	
Married (X_2)	81.44	
Number of dependants (X_3)	81.42*	
Education (X_4)	81.42*	
Working (X_5)	81.42*	
Income (X_6)	81.42*	
Spouse Income (X_7)	81.42*	
Loan amount (X_8)	81.42*	
Monthly repayment (X_9)	81.43	
Credit History (X_{10})	69.43*	
Immobile (X_{11})	81.44	
7 variables retained	**X_3, X_4, X_5, X_6, X_7, X_8, X_{10}**	

Source: Author's computation aided by R packages v 4.1.1. *important Features ($P_i \leq P_c$)

Table 6: Random Forest feature selection

Variable	Associated accuracies (P_i) (%)	Cut off accuracy (P_c) (%)
-		99.66
Sex (X_1)	99.72	
Married (X_2)	99.80	
Number of dependants (X_3)	98.31*	
Education (X_4)	99.49*	
Working (X_5)	99.66*	
Income (X_6)	94.26*	
Spouse Income (X_7)	97.30*	
Loan amount (X_8)	94.76*	
Monthly repayment (X_9)	98.99*	
Credit History (X_{10})	81.83*	
Immobile (X_{11})	98.82*	
9 variables retained	**X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}**	

Source: Author's computation aided by R packages v 4.1.1. *important Features ($P_i \leq P_c$)

Using the efficient feature selection method, twelve (12) models were built for each classifier and the best subsets of the eleven variables were selected in the final model building. Logistic regression: The method revealed that six variables are considered important as present in table 2, Naïve Bayes algorithm considered 6 features as shown in table 3, Support vector machine selected seven features as shown in table 4, Decision tree also considered 7 predictors as important as shown in table 5 and Random forest considered 9 features as important as shown in table 6. Credit history of the applicant has been identified by all the five classifiers as the most important variable in determining loan approval. As stated by Iduseri et al. (2018) that "The goal of variable selection methods is to choose the best subset (or training sample) of relevant variables that typically reduces the complexity of a model and makes it easier to interpret, improves the classification accuracy of the model and reduces the training time. It is our belief that this feature selection method will achieve its on objective.

Building model using the Machine Learning algorithm

After selecting the best subsets of the predictors for each of the machine learning classification algorithms, a comparison of the performances of each of the classifiers before and after the feature selection based on overall accuracy was made. The results are presented in table 8 below.

Table 7: Classification Accuracy before and after feature selection

	Accuracies (%)	
Classifiers	**Before Feature selection**	**After Feature selection**
Logistic Regression	81.8	81.8
Naïve Bayes	79.9	81.6
Support Vector Machine	82.1	82.1
Decision Tree	81.4	81.4
Random Forest	99.7	99.7

Source: Authors' computation aided by R package v4.1.1

Table 7: displays the model building results of the five statistics and machine learning classifiers before and after feature selection. The model building is based on the overall dataset. As it can be seen our method of feature selection has succeeded in improving the overall accuracies of logistic regression (LR) and Naive Bayes. The accuracies of Random forest, Decision tree and support vector machine remain the same with fewer predictors.

Table 8: Performance of the classifiers on training and testing datasets

Model	Model accuracy (%)	
	Train dataset (70%)	Test dataset (30%)
Logistic Regression (LR)	81.6	81.1
Naïve Bayes (NB)	80.4	81.5
Decision Tree (DT)	81.6	82.1
Random Forest (RF)	100.0	100.0
Support Vector Machine (SVM)	82.1	83.6

Source: Authors' computation aided by R package v4.1.1

Table 8: displays the performance of the statistics and machine learning algorithms on training and testing datasets. The model building was done after feature selection using seventy percent (70%) of the data for training and thirty percent (30%) for testing. It can be seen that the models performance on both the training and test data set is encouraging.

Model Performance of the Machine learning algorithms

After training the machine learning models with the training dataset, the test dataset was used to evaluate the performance of the models. The performance evaluation metrics yielded the following result:

Table 9: Performance of different machine learning models for predicting loan approval

	Classifiers				
Metrics	LR	NB	SVM	DT	RF
No of features	6	6	7	7	9
Overall accuracy (%)	81.6	83.2	81.6	81.1	83.2
Sensitivity (%)	97.9	97.7	98.6	97.9	93.5
Specificity (%)	39.3	40.9	37.5	37.5	51.1
F-measure (%)	89.1	89.7	88.5	89.7	89.3
Misclassification (%)	18.4	16.8	18.4	18.9	16.8
AUC (%)	73.7	79.2	69.1	70.8	72.0

Source: Author's computation aided by R package v 4.1.1. LR: Logistic Regression, NB: Naive Bayes, SVM: Support Vector Machine, DT: Decision Tree, RF: Random Forest

By observing the overall accuracy, sensitivity, specificity, F-measure, misclassification error and area under the ROC curve values of the models in table 9, there was a trend indicating that naive bayes and random forest models and logistic regression has a better performance compared with support vector machine and decision tree models.

Conclusion

Based on the findings of this study, we can conclude that the feature selection method introduced in the work proved useful and works well on all the algorithms used in this study. The measure purposes of feature selection which includes improve accuracy, reduced training time, reduce cost of model construction and also reduce over fitting of the model were achieved. All the five models showed significant accuracies in the loan prediction but optimum accuracy was attained by Naïve Bayes in all the three metrics used in the model evaluation. Logistic Regression and Random forest followed sued in that order.

To improve the accuracy of classification models, the study recommends that feature selection method should be used to select the best subset of predictors before training the models. This feature selection method provides effective, stable and minimally biased selection of meaningful predictors to be included in the classification modeling. Although researchers in loan approval prediction area came out with different algorithms for loan prediction, we recommend the banking industries and money lending financial institutions to consider using Naive Bayes after feature selection to predict loan approval. This will reduce the risk of given out loan to wrong customers.

Recommendation

Further research should be conducted by comparing the feature selection method used in this study with existing classic methods used in statistics and machine learning to validate it performance. Also other classification algorithms outside the methods used in this work should be compared with the five methods used to validate our claims and other researchers' claims that Naive Bayes is more effective in loan approval prediction.

References

Aphale & Shide (2020). Predict Loan Approval in Banking System Machine Learning Approach For Cooperative Banks Loan Approval. *International Research Journal of Engineering and technology, 9(08),* 991-995

Chandra & Rekha (2019). Exploring the Machine Algorithm for Prediction the Loan Sanctioning Process. *International Journal of Innovative and Exploring Engineering, 9(1),* 2714- 2719.

Hamid & Ahmed (2016). Developin, G. prediction Model of Loan Risk in Bank using Data Mining. *An International Journal (MLAIJ*), 3*(1),* 1-9

Iduseri & Osemwenkhae (2018). A New Approach for Improving Classification Accuracy in Predictive Discriminant Analysis. *Ann Data Sci. 5(3),* 339-357

Kadam et al. (2021). Prediction for loan approval using Machine Learning algorithm. *International Research Journal of Engineering and technology, 8(04),* 4089-4092.

Khan et al. (2021). Loan Approval Prediction Model: A Comparative Analysis. *Advances and Applications in Mathematical Sciences, 20*(3), 427-435

Kumar et al. (2019). Prediction of Loan using Machine Learning. *International Journal of Advanced Science and Technology*, *28(7),*. 455-460

Kumar & Goel (2020). Prediction of Loan Approval using Machine Learning Technique. I*nternational Journal of Advanced Science and Technology*, *29*(6), 4152-4161

Madaan et al. (2021). Loan default prediction using Decision Tree and Random Forest: A comparative Study. *International conference on Computational Research and Data analytics* (ICCRDA 2020), *1022*

Meshref (2020). Predicting Loan Approval of banks Direct Marketing Data Using Ensemble machine Learning Algorithms. *International Journal of Circuits, System and Signal Processing, 20*

Nikam et al. (2021). Prediction For Loan Approval Using Machine Learning Algorithm. *Resincap Journal of Science and Engineering. 5(5),* 1399-1403.

Pramod et al. (2021). An Approach For Prediction of Loan Approval Using Machine Learning Algorithm. *International Journal of creative Research Thoughts. 9*(6) 568-570.

Setiawan & Suharjito (2019). Comparison of prediction Methods for Credit default on peer to peer lending using Machine Learning. *Procedia Computer science*, *157*(2019), 38-45.