SLUJST
Sule Lamido University Journal of Science & Technology Vol. 5 No. 1&2 [Dec, 2022], pp. 1-12
https://doi.org/10.56471/slujst.v5i.222
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Default Prediction for Loan Lenders Using Machine Learning Algorithms

Awuza Abdulrashid Egwa[1*], Bello Habeeb[2], Ahmad Ajiya Ahmad[3], and Suleiman Musa Bizi[4]
[1,3,4]Department of Computer Science, Federal University Gashua, Yobe State, Nigeria
[2]Department of Electronics and Telecommunications Engineering, Ahmadu Bello University Zaria, Nigeria
talk2awuza@yahoo.com[1*], bellohabeebu@yahoo.com[2], ahmad.ahmad@aun.edu.ng[3], bizi00065@gmail.com[4]

*Corresponding Author

Abstract

Credit loans are considered most essential aspect of most financial institutions. All loan mortgagees or lenders are demanding to identify out effective commercial and business approaches to encourage customers to apply their credit loans. There are numerous business patrons who act negatively after their requests got approval. To avert this condition, lenders have to discover some techniques to forecast customer's behaviors. This resulted to the usage of machine learning algorithms by the financial lending institutions for accessing loan applicants. Despite advancements in automating decision-based loan systems, most existing models do not consider the "early loan repayment" attribute as a factor in resolving this prediction error. In reality, the amendment for preliminary loan reimbursement in model building is obligatory, since a larger numbers of timely loan reimbursement observed during the loan period, reduces default rate. For effective model's comparison based on accuracy and minimum errors of prediction, six supervised machine learning algorithms i.e. Random Forest, Artificial Neural Network, Classification and Regression Tree, Support Vector Machine, Logistic Regression, and Naïve Bayes were adopted to develop a default prediction models which include the early loan repayment attribute. The models were trained and tested on a loan dataset consisting of attributes with, and without early loan repayment attribute and were evaluated using five performance metrics. The results of the performance evaluation show that models that account for early loan repayment have higher accuracy, recall, precision, Root Mean Square Error and Receiver Operative Characteristics curve values than models trained without the early loan repayment attribute. The Random forest model proofed to be the best predictive model having 93% accuracy, 11% RMSE, 90% precision, 89% recall and 81% ROC value over others models.

Keywords: *Machine learning; Prediction; loan default; early repayment;.*

1. Introduction

Business firms and households sometimes seek for extra-funding to fulfill certain needs. The need which ascends from the demand of additional funds is satisfied by the credit market. Banks and financial lending institutions are the key players in this market (Gaigaliene & Cesnys, 2018).

Loan is the key and most imperative attribute of most financial institutions. Most financial lenders try to find operational business schemes for persuading customers to be interested for loans. Though, there are some borrowers that default in loan payments (Begum & Deniz, 2019). Throughout the facility period, default might arise when the borrower fails to fulfill his/her payment obligation. Consequently, evaluation of a borrower's risk

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

of default with time is essential for appropriate risk handling. Credit officers determine whether borrowers can fulfill their requirements using manual analysis of borrower's credit history. In the last decade, this trend has changed over time with technological advancement (Rehman, 2017).

In recent years, according to Bao et al., (2019) financial lending institutions are using automated loan default models as credit risk scoring tools when granting loans to potential borrowers. Machine learning (ML) algorithms have remained effective technique used to assess the credit risk of borrowers in financial lending institutions (Djeundje & Crook, 2018). Reliable schemes for credit risks play a significant role in forfeit control and profits growth (Luo et al., 2016). Earlier research treated loan default prediction as a binary discrete classification, where an applicant is considered creditworthy or non-creditworthy (Rosenberg & Gleit, 1994). Linear Discriminant Analysis (LDA) and LR and are two most used tools for building credit scoring models (Liu et al., 2015). Subsequently, other classification algorithms such as, SVM (Alaka et al., 2018), ANN (Gulsoy & Kulluk, 2019), Decision Trees (DT) (Liu et al., 2015), and Bayesian classifier (BC) (Carta et al., 2020), have been widely adopted to predict borrowers' probability of default.

In recent times, time-to-default approach also fascinated and increasing research attention (Dirick et al., 2017). Time-to-default statistics spill into the group of wide-ranging lifetime data, which is generally considered by survival analysis (SA) (Malekipirbazari & Aksakalli, 2015).

In loan prediction, two types of errors inevitably lead to inefficiency in prediction accuracy i.e. the rejection of a true credit worthy applicant (also known as a "false positive"), and the non-rejection of a false credit worthy applicant (also known as "false negative") (Awuza et al., 2022). Most loan default prediction models revolve around the minimization of one or both of these errors (Nok, 2017).

Despite advancements in automating credit decision-based systems, there continues to be issues of default over estimation (Jiang et al., 2019). Several researchers have applied ensemble techniques and integration of two or more ML algorithms in resolving the default over estimation problem. However, they could not identify the factors that significantly minimize the errors of classifying good borrowers as bad ones or bad borrowers as good ones i.e. false positive and false negative rate (Awuza et al., 2022, Malekipirbazari & Aksakalli, 2015 and Begum & Deniz, 2019).

Jiang et al., (2019) identified the "early loan repayment" attribute as a factor that could drastically minimize classification errors. However, they could not use it as an attribute in developing their prediction model because of its exclusion in the dataset they collected. However, this attribute was applied in a research by (Awuza et al., 2022) but the result could not be comprehensive as only fewer algorithms were used for classification.

2. Related Works

From literature, numerous research have been conducted on credit risk analysis, classification and survival analysis method were used to determine default pattern of borrower (Lessmann et al., 2015). Chen et al. (2016) conducted the default prediction using Information Gain technique to shade out the eventual indicator variables from the novel sample and then construct a credit risk evaluation model of Chinese Online P2P Lending using LR. Chow (2018) in his study explored the efficacy of several ML algorithms in expressing business well-being. He used two very unlike datasets of companies in the industrial sector for analyses and matched his result to differenr researchers' outcomes. He applied all eminence control of several ML methods on the dataset when joint with other dimensionality reduction mechanism to obtain Accuracy (ACC), Recall (Rec), Precision (Prec), and F1 score for all the classifier, though his result does not specify the supremacy of one classifier over the other in

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

bankruptcy prediction. But Malekipirbazari and Aksakalli (2015) in their research were able to show supremacy of one classifier over the other. They propose a RF based method for borrower default status. Using FICO credit score and LC grade as two performance metrics, their experimental comparison discloses that RFs meaningfully surpass in both FICO scores and LC grades in identifying default probability of individual borrower.
Furthermore, a combination of two or more classifier could give improved performance. This was shown in a study Fu (2017) combined RF and ANN, discriminant analysis were conducted on data including approximately 1320 borrowers records. The performance metrics shows improvements in the "ANN + RF" Combination, with a higher ACC than a single Neural Network. Also, pre-processing makes great difference when using the combination, i.e. the ACC with and without the pre-processing of removing the average of features and dividing the variance differed from each other. Though "RF + ANN" grouping resolved at lower ACC than single RF, it was static faraway better than particular NN.

Similarly Huo et al. (2017) applied back-propagation neural network (BPNN) combination with LR on customers' business transactions data. Their results presented that the classification ACC of the BPNN-LR model was better compared to LR method. Indicating that the new inclusive valuable produced by BPNN was possible to extremely read the specific valuables that delivered technical references to differentiate 'safe' borrowers from 'unsafe' one's as basis for decisions making.

In another research by Begum and Deniz (2019) adopted six classification techniques: Bayesian Networks, Multilayer Perceptron, NB, LR, RF, and J48 for default prediction. These algorithms were matched in view of the RMSE, ROC, ACC, Prec, F-measure, and Rec as performance measures. Their study shows that LR is the top model according to the experiments, and was used to determine the effect of covariant on the variables with higher default factor using Chi-square.

Also, (Ying, 2018) used RF, LR, SVM and other appropriate techniques via python to review and examine the bank loan dataset. Base on five model effect evaluation matrixes: ACC, Rec, Prec, F1-score and ROC, RF algorithm proof to be more appropriate for bank credit default prediction model, because of its high classification effect, specifically with larger or highly dimensional data set. Padimi et al. (2022) considered P2P lending data to forecast the credit loan default. They adopted five naive-bias algorithm, LR, DT, RF and KNN algorithms. The authors were committed on the usage of probability of default in a default point to minimize the credit risk and obtained an accuracy of 94.6%. The article determine that random forest is best-suited model for the adopted model implementation. Wu (2022) proposed a model for Credit risk management in banking organizations to improved prediction and normalize pre-lending assessment. This author uses the RF technique to implement a default forecast based on previous loan information. The experimental data obtained in this article indicated that RF technique outperforms other algorithms, such as DT and LR based on ACC and REC rate.

Early repayment was identified and used as a factor that could greatly minimize classification error in a research by (Awuza et al. 2022). However, their work uses only three algorithms in models building with Logistic regression proven to be the best classification model. Therefore, this work will further explore the use of more algorithms with the more dataset for predicting probability of default from list of loan applicant.

3.0 Design Methodology of The Proposed Model

The design flow of the models is presented in fig. 1 below.

Sule Lamido University Journal of Science & Technology Vol. 5 No. 1&2 [Dec, 2022], pp. 1-12
https://doi.org/10.56471/slujst.v5i.222
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

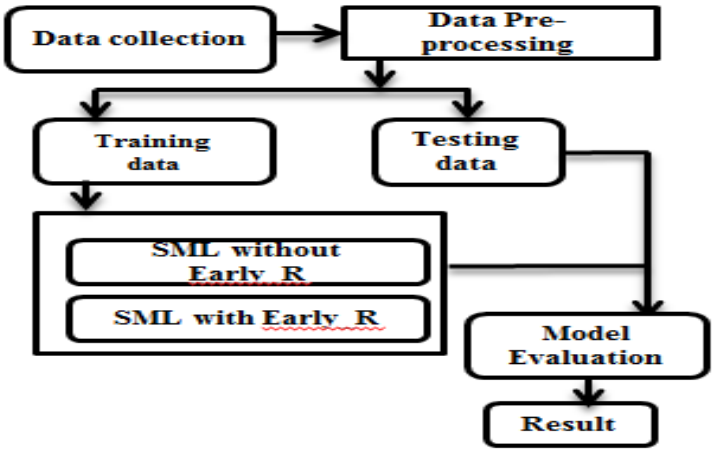

Figure 1: Design Methodology (Awuza et al. 2022)

i. Data Collection

The dataset adopted for this research is a company housing finance (CHF) dataset involving of 960 borrower's record. It comprises both demographic and socioeconomic characteristics of separate borrower and was obtained from Kaggle data repository.

Attribute Description

The attributes used in our model are described below.

Attribute name	Description
gender	male = 1, female = 0
married	married = 1, others = 0
education	graduate = 1, non-graduate = 0
dependent	no dependents = 0, only one dependent = 1, dependents = 2 and three & above = 3+
self-employed	yes = 1, no = 0
applicant income	high = 1, low = 0
co-applicant income	high = 1, low = 0
loan amount	amount applicant borrow
loan term	loan duration
credit history	Record of current and previous loan
property area	rural = 0, semi-urban = 1 urban = 2,
early repayment	early repayment = 1, non-early repayment = 0
loan status	non-default = N, default = Y

ii. Data Pre-Processing

The loan dataset is made up of a target variable (Loan Status) and 13 independent covariant all represented in an object format, int64 and float64 data types. Majority of ML models are prone to errors when handling non-numeric columns, and missing values in dataset. Therefore, it is important to undergoes some pre-processing steps. This is carried out on the raw dataset in order to have quality data with few but effective attributes for classification (Gulsoy & Kulluk, 2019).

Sule Lamido University Journal of Science & Technology Vol. 5 No. 1&2 [Dec, 2022], pp. 1-12
https://doi.org/10.56471/slujst.v5i.222
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

i. Missing Value

The dataset contains missing value which can either be removed or filled. The missing value approach is done using the mean, median or mode. Thus, this research adopted imputation of mean on numerical variables while mode for categorical variables using a Scikit-learn pipeline. The pipeline enables series of transformative steps, fit and predict on an estimator with a single line of code.

ii. Dummy Variables for Non-numeric Data

The categorical variables are replaced with dummy variables. Dummy operation turns categorical data into a discrete number of 0 and 1, enabling them much more flexible to measure and associate variables. The dataset were all encoded into numerical format to form a new pre-processed dataset.

iii. Correlation Analysis

The correlation analysis is a paramount metric which quantify the relationships between dataset variables. Therefore it defines how directly or inversely related are the dataset variables. The correlation of the dataset can easily be visualized using the heat map as shown in fig. 2 below.

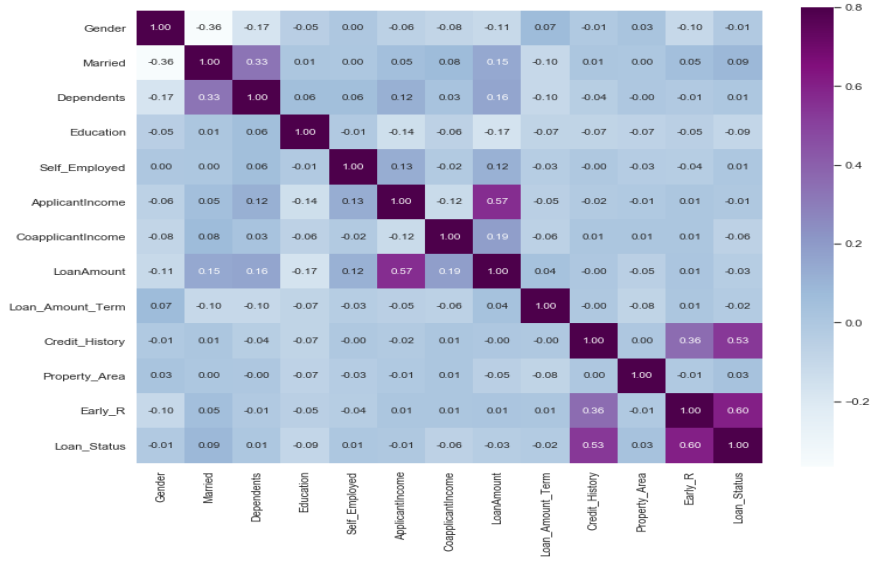

Figure 2: Loan Dataset Numerical Correlation heat map (Awuza et al. 2022)

iv. Data Splitting

The pre-processed loan dataset consisting of 960 rows and 14 columns was randomly split into 70% training dataset and 30% testing dataset.

v. Algorithm used for Model Building

Six algorithms were used and trained on dataset with and without attribute of early repayment.

a. Support Vector Machines (SVM)

For a data set with a discrete target variable, SVM classified the data into two parts in the p-dimensional feature space through hyper-plane with high boundary width among object of the two classes (Cortes & Vapnik, 1995). Exploiting the boundary width cuts the difficulty of the model and the general risk of error. Before the data got separated by a hyper-plane, which is typically the object in reality, a soft boundary is used. In this rule, a positive slack is supplemented to the objects on the erroneous side of the boundary. The slack growths comparatively to

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 5 No. 1&2 [Dec, 2022], pp. 1-12
https://doi.org/10.56471/slujst.v5i.222
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

how far the corresponding object is from the boundary. The aim is to reduce the sum of these slacks while increasing the width of the boundary. A regularization parameter C rules the relative cost of each goal in the process of optimization, which is taken as 1.0 in our application. SVM is chosen a classifier as it have been used effectively in related applications before (Alaka et al., 2018).

b. Logistic Regression (LR)

LR is a parametric technique and has been deliberated as the industrial routine in the field of default prediction (Lessmann et al., 2015). It is adopted to resolve binary and regression classification difficulties (default and non-default).

The objective of LR model is to predict the conditional probability of a certain sample of a particular class using an estimator. The estimates are calculated using the maximum likelihood method (Thomas et al., 2017).

c. Naive Bayes (NB)

NB algorithm is a supervised learning algorithm, which is centered on **Bayes theorem** and adopted for resolving classification difficulties. It is one of the modest and most current Classification algorithms which support in constructing profligate machine learning models that can create rapid predictions.

d. Random Forest (RF)

RF is an adaptable ML algorithm capable of performing both regression and classification tasks (Chow, 2018). In this algorithm, numerous trees are grown. To categorize a new object based on attributes, each tree gives a classification in form of voting for the class. The forest chooses the classification having the most votes and in case of regression, it takes the average of outputs by different trees.

e. Artificial Neural Network (ANN)

ANN is non-parametric approach that has a wider range of regression and classification problems (Gulsoy & Kulluk 2019). It is encouraged by its biological neuron structure and multifaceted pattern of relating inputs and outputs like human brain. There are many types of ANNs and one of the most common designs is the multilayer perceptron (MLP), which consist of an input layer, a one or more embedded layers and an output layer. For default prediction in loan context, ANN model is initiated ascending the variables of each instances via the input layer, then processes these features through the hidden layers, and lastly getting the output layer where the final status will be known based on its weights. These weights are assigned to each variable based on its relative prominence. Then sigmoid, tangent-sigmoid (activation function) collects the entire weighted variable to produce outputs. The procedure of regulating weights is repeated in many loops, aiming to increase errors among true class and predicted class.

f. Classification and Regression Tree (CART)

CART is a non-parametric technique widely adopted for credit scoring (Lee et al., 2006). The prediction output of CART is represented as an acyclic or directed graph in a tree pattern, where its output can visibly be accepted. CART model classifies inputs sample by first grouping it over the tree and then assigning it to the most suitable leaf node. In a CART graph, each node signifies some feature of the sample and each subdivision signifies a likely value of that feature. In this study, we adopted C4.5 DT (Quinlan, 1993) an extension of Quinlan's earlier Iterative Dichotomiser 3 (ID3) algorithm. C4.5 uses information entropy to choose features as decision nodes, improvements in terms of supervision of missing data points and pruning trees to avoid over-fitting problems.

3.2 Parameter Optimization

In this work, all the models were developed using Scikit-learn and grid search was used to enhance the grouping of hyper-parameters. Because grid search has exhaustive search of pre-defined hyper-parameter space, the search space were set as thus: For SVM model, we adopted RBF kernel, examining $\log_2 c$ and $\log_2 g$ ranges from -10 to 10; CART model's tree depth ranges from 5 to 15 and minimum samples spit between 5 to 20; for RF model, tree

Sule Lamido University Journal of Science & Technology Vol. 5 No. 1&2 [Dec, 2022], pp. 1-12
https://doi.org/10.56471/slujst.v5i.222
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

number ranges from 100 to 1000, minimum samples leaf between 3 to 10 and maximum depth between the range of 5 to 25; ANN model, the hidden nodes were searched between the range of 10 to 500 and the log(alpha) in range of (8,8,8); Logistic Regression and Naïve Bayes were both set to default.

3.3 Model Evaluation

Confusion matrix analysis is the most up-front method to evaluate the classification model performance. It is a concept from ML that involved data about real classifications and predicted classification (Deng et al., 2016).
In this research work, sequence of performance measures were engaged including Accuracy (acc) rate, Precision (prec), Recall (rec), Root Mean Square Error (RMSE) and Receiver Operation Characteristic curve (ROC).

4.1 Model Implementation Result

The various algorithms (LR, RF, SVM, ANN, NB and CART) were both subjected to the same Loandata.csv dataset and evaluated. Model performance results after series of experiments using the training and testing data are presented in tabular forms and chart.
The results obtained on training the various models using the training dataset are presented in Table 1 and Table 2. Table1 shows results of various models with attribute of Early_R while Table 2 shows results of the models without attribute of Early_R.

Table 1: Performance Results with Early_R

Algorithms	Accuracy	RMSE	Precision	Recall	AUC for ROC
LR	0.9211	0.1189	0.8922	0.8801	0.7913
RF	**0.9299**	**0.1051**	**0.9043**	**0.8929**	**0.8145**
SVM	0.8365	0.2811	0.5168	0.7189	0.5000
ANN	0.7297	0.6284	0.5973	0.6055	0.5031
Naïve B	0.9185	0.1244	0.8783	0.8757	0.8023
CART	0.8614	0.2051	0.7894	0.7944	0.7281

Table 2: Performance Results without Early_R

Algorithms	Accuracy	RMSE	Precision	Recall	AUC for ROC
LR	0.8572	0.2250	0.7578	0.7724	0.6410
RF	**0.8781**	0.2000	**0.8295**	**0.8001**	**0.7944**
SVM	0.8365	0.2811	0.5168	0.7189	0.5000
ANN	0.6711	0.6617	0.6061	0.5611	0.5124
Naïve B	0.8716	0.2062	0.7911	0.7941	0.6635
CART	0.7618	0.3511	0.6306	0.6471	0.5425

4.2 Discussions

The results presented above explained the overall models performances with and without Early_R attribute, since this research is focus on finding the significant impact of Early_R on models performances as well as determining best predictive model as shown in Table 1 and Table 2. Our research outputs are similar to work of (Awuza et al., 2022: Malekipirbazari & Aksakali, 2015). Table 1 shows a general models performance improvement for all base learner algorithms with respect to the five performance metrics compared to Table 2 with exception of SVM which have same performance rate for both instance.

In terms of accuracy of prediction, models with Early_R shows an increase of 7% for LR, 6% for RF, 6% for ANN, 5% for Naïve B. and 10% for CART as shown in Fig. 3.

Sule Lamido University Journal of Science & Technology Vol. 5 No. 1&2 [Dec, 2022], pp. 1-12
https://doi.org/10.56471/slujst.v5i.222
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

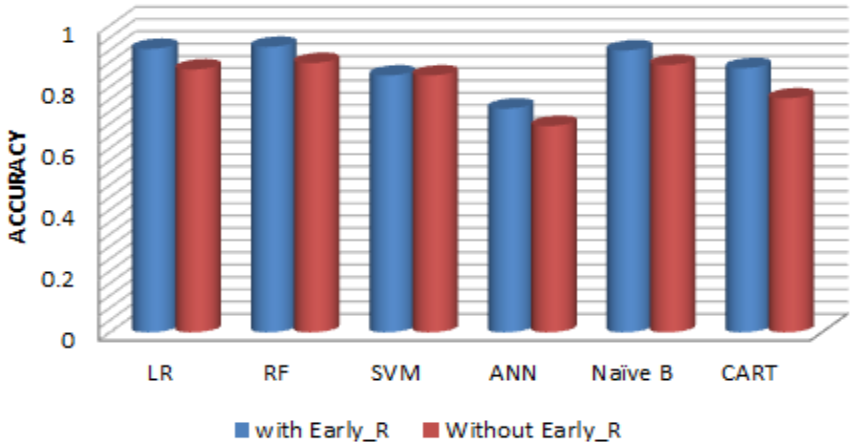

Figure 3: Models Accuracy Comparison

For minimal error of prediction, the RMSE of models with Early_R shows a decrease of 11% for Logistic Regression, 9% for RF, 3% for ANN, 9% for Naïve Bayes and 14% for CART as shown in Fig. 4.

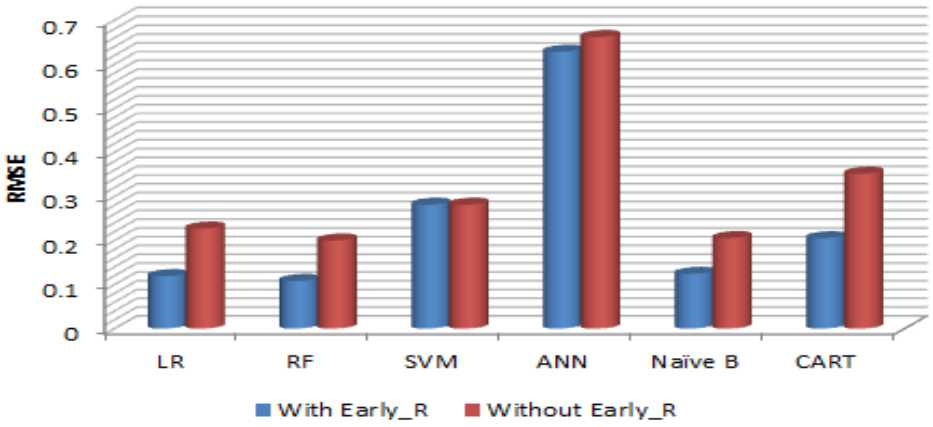

Figure 4: Model RMSE Comparison

To check how well the models can predict borrower as default or Non-default, models with Early_R shows increase in Precision value of 13% for LR, 7% for RF, 9% for Naïve Bayes and 16% for CART as shown in Fig. 5.

Sule Lamido University Journal of Science & Technology Vol. 5 No. 1&2 [Dec, 2022], pp. 1-12
https://doi.org/10.56471/slujst.v5i.222
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

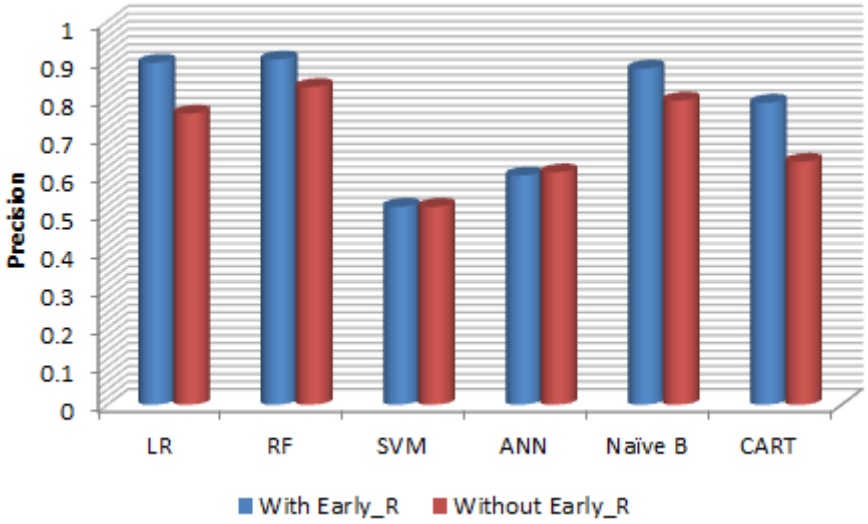

Figure 5: Model Precision Comparision

For Recall value, models with Early_R shows increase of 11% for LR, 9% for RF, 5% for ANN, 9% for Naïve Bayes and 14% for CART as shown in Fig. 6.

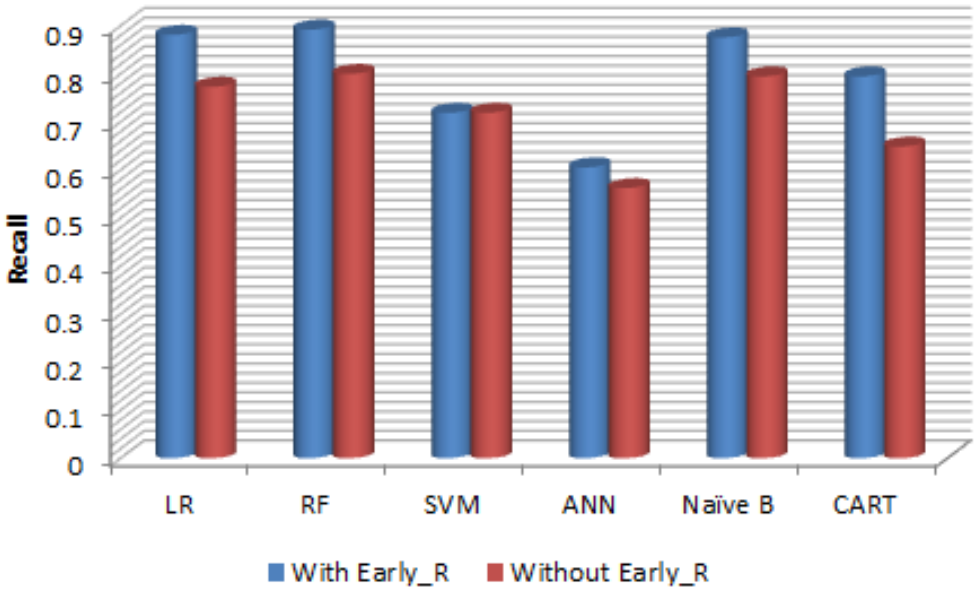

Figure 6: Models Recall Comparison

The ROC curve of the models with Early_R shows improvement compared to models without Early_R as presented in Figure 7. The AUC values increases with 15% for LR, 2% for RF, 14% for Naïve Bayes and 19% for CART as shown in Fig. 7.

Sule Lamido University Journal of Science & Technology Vol. 5 No. 1&2 [Dec, 2022], pp. 1-12
https://doi.org/10.56471/slujst.v5i.222
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

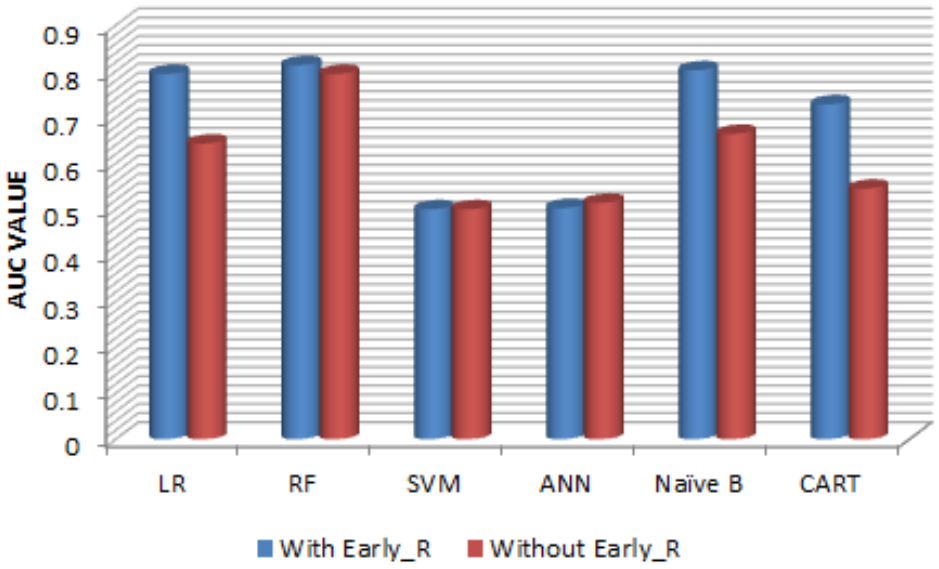

Figure 7: AUC Values comparison

Aside from general models performance comparison, individual model shows some efficacy in prediction over the other. From Table 1 and 2, Random forest classifier outperformed other classifiers in terms of the five performance metrics used.

4.3 Research Findings

After conducting various experiments, the following findings were made from this research work:

- Inclusion of Early loan repayment attribute in a loan dataset improved the performances of machine learning algorithms and reduces classification error.
- Random forest shows supreme performance rate for all parameter matrix used. Hence it is more suitable for loan default prediction models. This is also consistent with the work of (Awuza et al., 2022).

4.4 Conclusion, Limitation and Recommendations

Machine learning is an influential approach for financial forecasters to make forecasts and estimation to discover patterns in the data with consistency. Various models and validation techniques be existent and adopted to assist data mining and decision making in financial institutions. It is hard to regulate if one machine learning technique is better to others. Actually, as a financial data scientist or professional it is possibly more valuable to yield the powers of diverse approaches to make enhanced business judgments. Conceivably, this research work has done justice to that effect. Six supervised learning algorithms were trained and tested on dataset with and without attribute of early repayment. Models with early loan repayment show performance improvement compared to models without early loan repayment attribute. However, because of privacy issues with accessing domestic data sets for bank credit, this research used only foreign loan data. More so, in future, it is necessary to collect domestic dataset to compare borrower's behavior to default rate in Nigeria and Foreign countries. Another gap could be the use of data balancing techniques for improves performance because of the number of datasets used.

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

References

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., and Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert Systems with Applications, 94, 164-184.

Awuza, A. E., Habeebah, K. A., Ahmad, A. A., Abubakar, M. B. and Muhammad, A. M. (2022). Prediction Model for Loan Default Using Machine Learning. The International Journal Of Science & Technoledge. DOI No.: 10.24940/theijst/2022/v10/i2/ST2202-009

Bao, W., Yue, K., Yongtao, Z., Dapeng, L. and Lianju, N.(2019) Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment, Expert Systems With Applications https://doi.org/10.1016/j.eswa.2019.02.033`

Begum, C. and Deniz, U. (2019) Comparison of Data Mining Classification Algorithms: Determining the Default Risk. Research Article, Hindawi Scientific Programming Volume 2019, Article ID 8706505, 8 pages https://doi.org/10.1155/2019/8706505

Carta, V., Ferreira, A., Recupero, D. R., Saia, M. and Saia, R. (2020) A combined entropy-based approach for a proactive credit scoring. *Engineering Applications of Artificial Intelligence*. *87*. 103292.

Chen, N., Ri, B., kijbeiro and Chen, A. (2016) Financial credit risk Assessment: a recent review. Artificial Intelligence Review, 45th ed. 1, pp. 1–23.

Chow, J. C. (2018) Analysis of Financial Credit Risk Using Machine Learning. 2018 *arXiv preprint arXiv:1802.05326*.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. Machine Learning,. 20, pp. 273–297.

Deng, X. Liu Q. and Deng,Y. (2016) An improved method to construct basic probability assignment based on the confusion matrix for classification problem . Information Sciences, 340, pp. 250-261.

Dirick, L. Claeskens, G. and Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study, Journal of the Operational Research Society. 68, pp. 652–665

Djeundje, V. B. and Crook, J. (2018) Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. European Journal of Operational Research, 2nd ed. 271, pp. 697-709.

Fu, Y. J. (2017). Combination of Random Forests and Neural Networks in Social Lending. Journal of Financial Risk Management, 6, 418-426. https://doi.org/10.4236/jfrm.2017.64030

Gaigaliene, A. and cesnys, D. (2018). Determinants of default in Lithuanian peer-to-peer platforms. Organizacijų vadyba: sisteminiai tyrimai= Management of organizations: systematic research. Kaunas: Vytauto Didžiojo universitetas.

Gulsoy, N. and Kulluk, S. (2019). A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 9 (3), e1299.

Huo, Y.J., Chen, H.Z. and Chen, J.C. (2017) Research on Personal Credit Assessment Based on Neural Network-Logistic Regression Combination Model. Open Journal of Business and Management, 5, 244-252. https://doi.org/10.4236/ojbm.2017.52022

Jiang C., Wang Z. and Zhoo, H. (2019). A Prediction-driven Mixture cure model and its Application in Credit Scoring. European Journal of operational Research. 277, pp20-31

Lessmann, S., Baesens, B., Seow,H. V. and Thomas, L. C. (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. European Journal of Operational Research. 1(247) pp. 1-32.

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, *50*(4), 1113-1130.

Liu, F., Hua, Z. and Lim, A. (2015). Identifying future defaulters: A hierarchical Bayesian method. European Journal of Operational Research. 241, pp. 202–211.

Luo, S., Kong, X. and Nie, T. (2016). Spline based survival model for credit risk modeling. European Journal of Operational Research. 3(253) pp. 869–879

Malekipirbazari, M. and Aksakalli, V. (2015) Risk assessment in social lending via random forests. *Expert Systems with* Applications. 42(10), pp. 4621-4631

Nok, W. M. (2017). Bankruptcy Prediction of Industrial Industry in the UK. Sriwijaya international journal of dynamic Economics and Business. 1(1), pp. 1-26.

Padimi V., Venkata S. T. and Devarani D. N. (2022). Applying Machine Learning Techniques To Maximize The Performance of Loan Default Prediction. *Journal of Neutrosophic and Fuzzy Systems (JNFS)*. Vol. 2, No. 2, PP. 44-56.

Quinlan, J.R. (1993) C4.5: Programs for machine learning. Morgan Kaufmann Publishers, Massachusetts.

Rehman, N. (2017). Data mining techniques, method, algorithms and tools. International Journal of computer science and mobile computing . 6, pp. 227-231.

Rosenberg, E. and Gleit, A. (1994) Quantitative methods in credit management: a survey. Operations Research. 42(4) pp. 589-613.

Thomas, L., Crook, J. and Edelman, D. (2017). Credit scoring and its applications. Society for industrial and Applied Mathematics.

Wu Q. (2022). "Real-time Predictive Analysis of Loan Risk with Intelligent Monitoring and Machine Learning Technique,"*2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2022, pp. 852-856, doi: 10.1109/ICPICS55264.2022.9873618.

Ying, L. (2018). Research on bank credit default prediction based on data mining algorithm. The International Journal of Social Sciences and Humanities Invention . 5(6) pp. 4820-4823.

