Ingénierie des Systèmes d'Information
Vol. 27, No. 5, October, 2022, pp. 815-821
Journal homepage: http://iieta.org/journals/isi

Dynamic Effectiveness of Random Forest Algorithm in Financial Credit Risk Management for Improving Output Accuracy and Loan Classification Prediction

Afolashade Oluwakemi Kuyoro, Olufunmilola Adunni Ogunyolu*, Thomas Gbadebo Ayanwola, Folasade Yetunde Ayankoya

Department of Computer Science, Babcock University, Ogun State 121003, Nigeria

Corresponding Author Email: Ogunyolu0363@pg.babcock.edu.ng

https://doi.org/10.18280/isi.270515

Received: 24 June 2022
Accepted: 6 October 2022

Keywords:
credit scoring, decision tree, default, feature selection, forecasting, random forest, loan

ABSTRACT

With technology impacting several sectors, it can be imagined that the financial sector has a lot to benefit from the increasing level of technological innovations. These institutions take from the surplus of the economy and lend to the deficit sectors of the economy. Individuals and organizations obtain credit facilities from financial institutions to meet basic needs and boost their businesses. However, the stability of the economy is better guaranteed when borrowers pay back the loans availed to them rather than default. This study aims to identify the effectiveness of Random Forest in credit scoring using 32,581 observations. The study proved that Random Forest provides better output accuracy of 91% based on Gini Index for variable selection according to the level of importance when compared to Decision Tree with an output of 83%. It offers better credit scoring accuracy and credit rating as a result of its classification power. The objective of the study is to point out the random forest predictive strength using an unprocessed German credit dataset from Kaggle and to provide an explainable framework sufficient for Financial Institutions and banks to make decisions when granting loans to existing and new applicants.

1. INTRODUCTION

Credit scoring techniques are used to monitor and evaluate systematic risk when customers apply for loans in financial institutions and banks [1]. Using data the bank has per time-based on Irish banks data, logistic regression recursion parting (RP), conditional influence trees (CIT), SVM, least absolute shrinkage select operation (LASSO) approaches and it was stated that support vector machine (SVM) has higher performance than other algorithms based on area under the curve (AUC). Credit scoring objectives are to assess the economics, and credit risk through the use of immediate warning techniques to forecast possible defaults and remove items that could be a risk to the process.

Several works have discussed machine learning loan default prediction noting that it uses logistic regression for classifying its purpose because of its large sample size and relationship among variables and also deep identification of relationships in the set of data [2]. Other researchers have focused on the role of artificial intelligence and big data on loan decisions in Saudi Arabian banks which rely more on traditional means of making loan decisions by adhering to stipulated guidelines of the bank but this however led to identifying the statistical and major relationship between AI and decision making on loan availed noting that there was a relationship between experience and use of big data [3].

The focus of this work is to provide a better, understandable, and interpretable credit scoring model for financial institutions and banks using the Random Forest technique.

2. RELATED WORK

Several machine learning algorithms have been used in credit scoring while some machine learning algorithms provide better accuracy than traditional credit scoring methods as a result of their limitations which can introduce other errors and can affect the result of the model. Random Forest is an ensemble decision tree that helps in data analysis where variables of the dataset have multicollinearity and variable relationships and it comprises a selection of trees to make a forest [4]. According to Madaan et al. [5], Random Forest was compared with decision tree and it was observed that Random forest gave 80% better accuracy than Decision trees, however, the result can be better given a default probability. Other works like Zhang et al. [6] provided optimization of random forest algorithm using grid and feature reliable scoring model.

Most banks require the loan status of the clients to identify their ability to pay back using machine learning techniques like Random Forest and according to the study by Vanara et al. [7], Random Forest help minimise the risk inherent in classifying the loan applicants having a result of 85.75%. Another researcher mentioned having used Random Over Sampling with random forest improved the accuracy to 90.1% and without the Random over-under sampling gave 76% [8]. Other methods used in credit scoring entail weighted random forest based on Gini Index and the outcome gave a forecasting accuracy of over 70% on an imbalanced credit dataset better than just a random forest [9]. There are several machine learning methods like Naïve-Bayesian Forest, Decision Tree, and KNN classifier and they all have their advantages and disadvantages, not as good as Random Forest as mentioned by Wang et al. [10]. Several ways in which the creditworthiness

of customers can be improved to identify a customer's possibility of a default or non-default.

An ensemble decision tree algorithm can help minimise overfitting by improving the output result [11]. Another way to improve the accuracy of models based on machine learning in credit scoring is to eliminate unwanted features by implementing parallel random forest and also enhancing its performance having 76.2% on a German dataset and 89.4% on an Australian credit dataset according to Van Sang et al. [12]. An ensemble method like AdaBoost, Bagging, and Random Forest are three algorithms that produced a higher performance by classifying good from bad loan customers [13].

Another comparison of several machine learning classifiers using information gain as a feature selection method along with Random Forest and the output provided better efficiency and minimal positive rate that provided a higher-performing credit scoring output of 80.70% better than Chi-Square and gain ratio [14] while it is believed that risk can be minimised in credit scoring by aggregating feature selection models like Principal Component Analysis and Genetic Algorithm and Random Forest as well as Support Vector Machine [15].

Credit Scoring is a technique adopted by various banks and financial institutions to classify the customers who are likely to default, or pay based on existing or historical credit features like Interest rate, repayment amount, Credit and Debit Turn over, Tenor, the purpose of the loan, age of applicant among others.

2.1 Definition of random forest

They are ensembles of decision trees as displayed in Figure 1; it handles datasets from minimal to medium volume, both classification and regression problems, and usually result in a very good outcome. Random Forest can proffer solutions to business or organisations pressing real-life problems like in credit scoring, object identification like in Traffic among others. In Machine learning, Random Forest being a supervised algorithm is fast during computation, and forecasting, and based on a statistical model, it is useful in visualization and it can identify the relationship and relevance of variables. It can provide solutions to bagging problems and proffers solutions to decision tree problems [16].

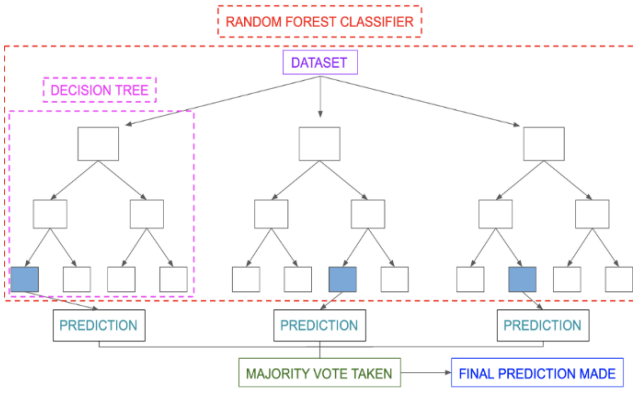

Figure 1. Explains the combination of decision trees as a Random Forest and based on the mode of the outcome, the best output is accepted [17]

2.2 Bagging and boosting ensembles in random forest

The ensemble is building a couple of forecasting models and merging their results into a stronger individual forecast. Ensembles use two methods which include Boosting and Bagging. As the name signifies Bagging is a combination of bootstrap combinations and its example is Random Forest while boosting combines lesser training models with each other to develop a better model and then selects the end model with the highest accuracy. An example is Ada Boost and XG Boost.

2.3 Nodes: Architecture of decision tree

Decision Trees consist of three significant nodes namely Root Node, Leaf Node, and Decision Node as shown in Figure 2. These nodes of each subset of the dataset operate based on the splitting possibility that occurs on each node, the Root node begins by splitting into homogenous mini sets [18] The decision node and leaf nodes are the end output of the tree and it can comprise of two or more branches [19]

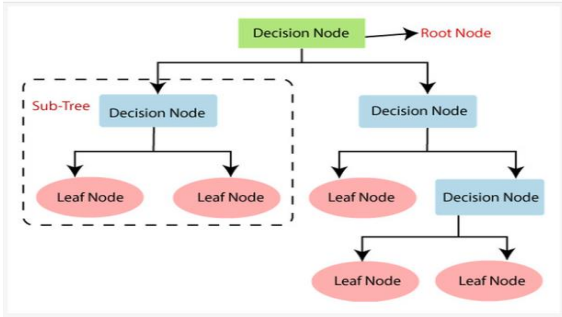

Figure 2. Depicts the various nodes that make up a decision tree [20]

3. METHODOLOGY

This section describes the methodology behind Random Forest for its positive impact on credit scoring classification, and its performance using 32,581 observations of a credit dataset comprising 12 variables.

3.1 Random Forest algorithm

This work emphasis how Random Forest is better used for classification problems since the Decision tree has several limitations, maximises the entropy gain, it is highly sensitive to training data and so susceptible to high variance which might not generalise. Random Forest on the other hand is not sensitive to training data sets since new dataset sets can be developed from an existing dataset with the variables unaltered and this process is called Bootstrapping while the Decision tree is trained on each new dataset. The output of the trained new dataset is derived based on a portion of the dataset used which makes up the Random Forest.

To make predictions, a new data point is used to predict by sending it through each of the trees producing an output of each tree. The output predictions are combined and the output with the highest occurrence between 0 and 1 is selected as the best, this is called Aggregation. The combination of Bootstrapping and Aggregation is called Bagging [21]. While Random Forest is training, several actions take place like sampling data and sampling the variables during a split. It also tackles during training out of bag problems and it helps tune the parameters in the dataset. Out-of-bag occurs when subsets of observations are not picked as a result of bootstrap

aggregation by an individual tree in Random Forest during training [22]. The study gave oob score of 0.905 which is close to the performance based on test data. It shows that it is used to validate and confirm the Random Forest Model as also mentioned by Enes [23].

Some researchers believe the number of trees needed in a Random Forest relies on the number of rows available for better results [24] while others believe enhancing the power of prediction by increasing the number of trees improves the outcome but minimises computation [25]. In this work, the number of trees used is a minimum of 50 estimated trees with a random state set to zero (0) for consistency in the output in every code runs for train-test split and to have a balanced processing time. The study showed that deciding on using more trees did not have any significant impact on the output but rather increased time complexity [26]. Furthermore, feature selection was used to enhance the parameters evaluated during the training process which also contributed to the reduction in computation as mentioned also by Hassine et al. [27].

3.2 About dataset

Data was taken from Kaggle, consisting of 32,581 observations with 12 Variables. The credit dataset is an unprocessed German dataset and the Variables are person_age, person_income, person_home_ownership, person_emp_length, loan_intent, loan_grade,loan_amnt, loan_int_rate, loan_status, loan_percent_income, cb_person_default_on_file, cb_person_cred_hist_length as shown Figure 4 and the Dataset was split into 70% Training and 30% Test data.

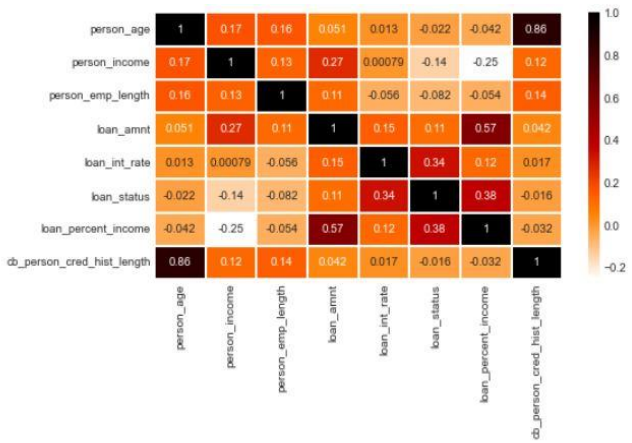

Figure 3. Pearson correlation used to determine the statistical relationship between the variables ranging from -1 to +1, where +1 denotes a positive correlation while -1 denotes a negative correlation

This figure shows minimal relationship between the variables

The Credit dataset is a financial data that is used for predicting the loan status of applicants where the Target Variable Y is the loan status comprising of 0 and 1 which makes it a binary classification problem (Where 0 denotes Paid and 1 denotes default) as explained in Figure 3.

Step 1: Importation of Dataset and Data exploratory analysis: Dataset is imported and data Exploratory is done to have an idea of the dataset visually also from the data it was observed that the variables consist of integer, character, and double category.

```
Columns: 12
$ person_age                <int> 22, 21, 25, 23, 24, 21, 26, 24, 24, 21, 22, 21, 23, 26, 23, 23~
$ person_income             <int> 59000, 9600, 9600, 65500, 54400, 9900, 77100, 78956, 83000, 10~
$ person_home_ownership     <chr> "RENT", "OWN", "MORTGAGE", "RENT", "RENT", "OWN", "RENT", "REN~
$ person_emp_length         <int> 123, 5, 1, 4, 8, 2, 8, 5, 8, 6, 6, 2, 2, 4, 2, 7, 0, 7, 8, 8, ~
$ loan_intent               <chr> "PERSONAL", "EDUCATION", "MEDICAL", "MEDICAL", "MEDICAL", "VEN~
$ loan_grade                <chr> "D", "B", "C", "C", "C", "A", "B", "B", "A", "D", "B", "A", "A~
$ loan_amnt                 <int> 35000, 1000, 5500, 35000, 35000, 2500, 35000, 35000, 35000, 16~
$ loan_int_rate             <dbl> 16.02, 11.14, 12.87, 15.23, 14.27, 7.14, 12.42, 11.11, 8.90, 1~
$ loan_status               <int> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0,~
$ loan_percent_income       <dbl> 0.59, 0.10, 0.57, 0.53, 0.55, 0.25, 0.45, 0.44, 0.42, 0.16, 0.~
$ cb_person_default_on_file <chr> "Y", "N", "N", "N", "Y", "N", "N", "N", "N", "N", "N", "N", "N~
$ cb_person_cred_hist_length <int> 3, 2, 3, 2, 4, 2, 3, 4, 2, 3, 4, 2, 2, 4, 4, 3, 4, 4, 4, 2, 4,~
```

Figure 4. The different variables and their classes in the unprocessed dataset before converting categorical columns to binary

One of the efficiencies of Exploratory Data Analysis is that it helps identify errors in the dataset like nulls which are handled using the interpolation technique of known values to estimate unknown values. EDA gives insights into the different data types available in datasets as shown in Figure 4 which are converted from categorical to numerical using dummy encoding.

Step 2: Data Cleaning and Data preprocessing

As part of Data analysis, data cleaning helps identify missing values and handle errors like outliers. In other to improve the performance, factors were considered to enhance data processing and feature selection. The number of input variables was reduced using filter feature selection. In supervised learning, the filter method used Feature Importance technology to remove irrelevant variables based on their relationship with the target variable to enhance the efficiency of the model [28]. The nulls and missing values were handled using the interpolation method.

To avoid overfitting the number of trees estimated is 50 while Bagging/Boosting trained the model on remaining training data based on bootstrapped samples.

Step 3: Feature Importance: This is used to determine the hierarchy of importance of the variables needed for the model. Random Forest uses Gini Index to select the best present split. It identifies the mean gain of purity of an identified variable based on the split. The maths behind this work is to identify the features based on relevance and remove redundant features in Gini impurity.

$$\text{Gini(d)} = 1 - \sum_{i=0} \text{r(Pi)2}$$

Formula 1: Gini Index ranges from 0 to 1, it calculates for all columns based on available conditions and picks the column with the best minimum Gini index as a criterion for the split. Where P is the probability of splits that happens in the tree, by calculating the impurity of the node using classification [29] that occurred during training.

Feature Importance is a resident property in tree Based classifiers used to display the score of individual variable in a dataset such that the higher the score the better its relevance to the target variable in determining the loan status [30]:

i. Identify target variables and a random dataset.

ii. Data Partitioning into training and test data.

iii. Establishing the Random Forest Classifier by training the data.

iv. Analysis and impurity computation is achieved

v. Weight average is derived from node impurities having values ranging from 0 to 1 [31].

This step helped to minimise the dimensionality of the model and enhance performance as well as manage processing time by removing features that will not contribute to the performance of the model as displayed in Figure 5.

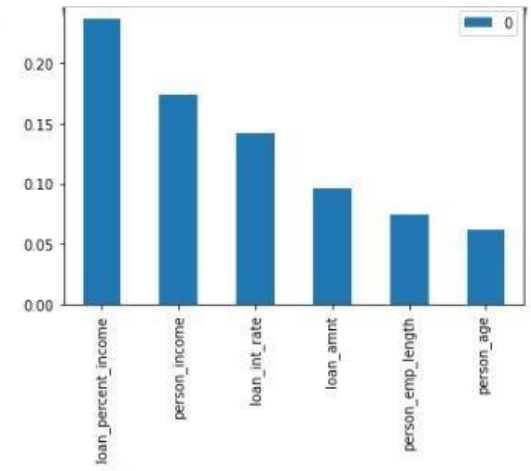

Figure 5. The variable based on the level of importance as a result of Gini Index computation

4. RESULTS AND DISCUSSIONS

The methodology compared Random Forest and Decision tree modeling on Intel Core (TM) i3-10110U CPU@2.10GHZ with 12.0GBof RAM under windows 10-64 bits, processorx64.

4.1 Random forest and decision tree comparison

In financial institutions and banking, non-performing loans caused by internal factors started as a result of the inability of financial institutions to run a proper credit check of clients before loans or credits are availed. Rather than adopting the traditional means of evaluating credit datasets, Random Forest has shown better fairness and performance in classifying loan applicants. From this work, the outcome of the study showed that the random forest model produced approximately 91% better performance while Decision Tree gave an outcome of 83%. Table 1 showed that the feature importance technique gave insights into the calculated score of the features loan_percent_income, person_income, loan_int_rate, loan_amnt, person_emp_length and person_age have higher impact scores than others, and based on the confusion matrix as shown in Table 2 it was able to truly predict 7390 true observations correctly and 771 correctly negative. Table 3 displayed the classification report performances of the two models. The result clearly shows that Random Forest outperformed the Decision tree.

4.1.1 Random forest and decision tree evaluation using confusion matrix

According to Figure 6 and Figure 7, it visually explains the combination of the predicted and actual classes. The confusion Matrix shows how well the model has performed for each class type and clearly shows the number of classifiers predictions that were correctly classified and also shows when the classifier was in doubt. The diagram (Figure 6) shows here the classifier struggled at classifying the defaulters having a recall value of 0.67 approximately. Figure 7 shows here the classifier for the Decision tree struggled at classifying the defaulters having a recall value of 0.38 on the Confusion Matrix.

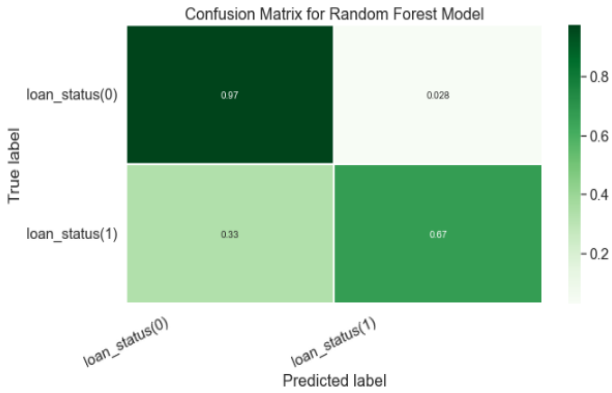

Figure 6. Confusion matrix for random forest

Table 1. Indicates loan percent income as the most important variable having 0.23

Variables	Score
loan_percent_income	0.237518
person_income	0.174074
loan_int_rate	0.142518
loan_amnt	0.096405
person_emp_length	0.07469
Person_age	0.051

Table 2. Shows the default and non-default percentages of the Random and Decision Tree algorithms

	Random Forest		Decision Tree	
	0	1	0	1
0	7524	220	7390	771
1	677	1354	1260	354
Accuracy=	sum(diag(matrix)/sum(matrix))			
		0.91		**0.83**

Table 3. Summary of accuracy measure for random forest

	Random Forest classification report				Decision Tree Classification report			
	Precision	Recall	F1 Score	Support	Precision	Recall	F1 Score	Support
0	0.92	**0.97**	0.94	7744	0.85	**0.95**	0.90	7744
1	0.86	**0.66**	0.75	2031	0.69	**0.38**	0.49	2031
Accuracy			**0.91**	9775			**0.83**	9775
Macro avg	0.89	0.82	0.85	9775	0.77	0.67	0.70	9775
Weighted avg	0.90	0.91	0.90	9775	0.82	0.83	0.82	9775

where Precision = True Positive/(True Positive+False Positive)=7524/(7524+677) = 0.92

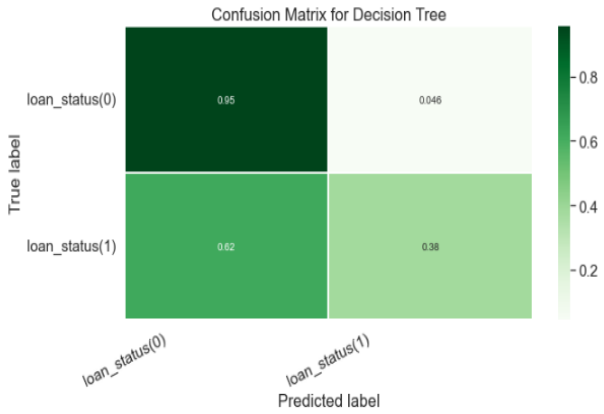

Figure 7. The confusion matrix for decision tress

The classifier correctly picks loans paid back (0) as 7524-Random Forest and 7390 for Decision trees, it also shows defaulters (1) correctly classified for 1354 for Random Forest and 354 for Decision Tree. This represents 30% of the test set from the dataset.

Row of Matrix=Instances of class predicted
Column of Matrix=Instances of actual class
TPR: True Positive rate(recall) details on correctly classified predicted as True, an outcome that the applicants will pay back loans availed to them [32].

$$\text{(Accuracy)TPR} = \frac{\text{Positive Outcome identified}=7524+1354}{\text{Number of Positive Outcomes } 7524+1354+220+677} = 0.91 \text{ (Random Forest)}$$

Random forest and decision tree classification report: Shows the relationship between the Precision, Recall, and F1 implication of the model.
with Recall = True Positive/(True Positive + False Positive) = 7524/(7524+220)= 0.97, it shows the number correctly predicted while F1 score is the combination of both Precision and Recall.

$$\text{F1 Score} = 2*\frac{(\text{Precision} * \text{recall})}{(\text{Precision} + \text{recall})} = 0.94$$

which means F1 at 1 is perfect and with 0.94 its fair.

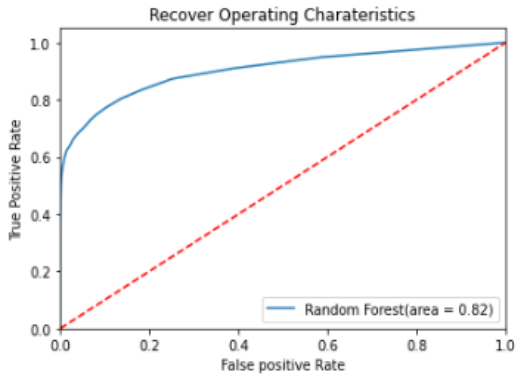

Figure 8. ROC shows the performance of the model based on several thresholds, the nearer the curve approaches the upper left side corner, the better the Random Forest model is and it showed an area under the curve of 0.82

Recover operating characteristics/area under curve (ROC/AUC): Assess the performance of a model and it shows the different threshold which identifies the trade-off for true positive rate and False Positive Rate (Figure 8).

4.1.2 Dynamic effectiveness and usefulness of Random forest

Random forest is a popular machine-learning technique that can be used to solve both regression and classification tasks. It works well with large datasets and can handle lots of input variables in a dataset. It's capacity to estimate missing data and enhance accuracy. As a result of its predictive power, it can be used to detect diseases like Parkinson accurately [33] and also help with fraud detection. In this study, Random Forest gained insights into financial data, it showed that it can easily be trained [34], and gave a better prediction accuracy than decision Trees. It perfectly handles bias and it is not prone to overfitting because of its ability to average trees across iterations.

5. RECOMMENDATIONS

One of the challenges in credit scoring is the inability to develop a simple yet performing model that can be used to accurately predict defaulters and non-defaulters. Random Forest has proved time and time again the efficiency of its classification power in credit scoring as a result of its ensembling decision trees power and ability to overcome errors inherent with decision trees.

Random Forest can provide computationally minimised cost solutions for financial Institutions' expected accuracy. The outcome of the study shows that Random Forest can help prevent overfitting and overcome limitations as a result of choosing variables based on level of importance.

The application of the Gini Index for computing feature importance also assisted in the removal of inefficient variables from the dataset. This result can also be enhanced in the future by improving Random Forest using Hybrid Modelling techniques for higher prediction accuracy and easier interpretation for financial institutions and banks.

6. CONCLUSION

A major driving sector of any economy is the financial and banking sector which provides services to individuals, and corporate organizations to meet basic needs and improve business drive. Machine Learning has been adopted and implemented over the years to enhance loan prediction in credit scoring. Loans availed to customers can affect the stability and growth of this sector if not repaid as at when due. To meet the obligation of stakeholders of any financial sector, it is necessary to have a standard, interpretable, and easy-to-understand model which can help enhance the accuracy of loan classification, and handle large datasets provided by financial sectors. This will help classify defaulters and non-defaulters better thereby minimizing financial loss. In this study, the efficiency and predictive power of Random Forest is compared with that of the Decision Tree. As a result of this comparison, the output of the study showed that Random Forest gave better performance than Decision Tree based on the German dataset used. Future study on the comparison of different feature selection techniques will be done to reduce the irrelevant features that will not contribute to the efficiency

of the model and comparison of different techniques of converting categorical to numerical variables will be explored so as to improve the model output.

REFERENCES

[1] Ilter, D., Kocadagli, O., Nalini, R. (2019). Feature selection approaches for machine learning classifiers on yearly credit scoring data. In book: y-BIS 2019 Conference Book: Recent Advances in Data Science and Business Analytics. Publisher: Mimar Sinan Fine Arts University.

[2] Kumar, A., Manager, T. (2018). Machine learning application in loan default prediction. JournalNX - A Multidiscip. Peer Rev. J., 4(5): 1-5.

[3] Abuhusain, M. (2020). The role of artificial intelligence and big data on loan decisions. Accounting, 6(7): 1291-1296. http://dx.doi.org/10.5267/j.ac.2020.8.022

[4] Dewani, P., Sippy, M., Punjabi, G., Hatekar, A. (2020). Credit scoring: A comparison between random forest classifier and K-nearest neighbours for credit defaulters prediction. International Research Journal of Engineering and Technology (IRJET), 1887-1892.

[5] Madaan, M., Kumar, A., Keshri, C., Jain, R., Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering, 1022(1): 012042. http://dx.doi.org/10.1088/1757-899X/1022/1/012042

[6] Zhang, X., Yang, Y., Zhou, Z. (2018). A novel credit scoring model based on optimized random forest. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp. 60-65. http://dx.doi.org/10.1109/CCWC.2018.8301707

[7] Vanara, R., Wani, P., Pawar, S., More, P., Patil, P. (2021). Predication approval for bank loan using random forest algorithm. International Journal of Progressive Research in Science and Engineering, 2(7): 137-142.

[8] Syukron, A., Subekti, A. (2018). Penerapan metode random over-under sampling dan random forest untuk klasifikasi penilaian kredit. Jurnal Informatika, 5(2): 175-185. http://dx.doi.org/10.31294/ji.v5i2.4158

[9] Wong, T.T., Yeh, S.J. (2019). Weighted random forests for evaluating financial credit risk. Proc Eng Technol Innov, 13: 1-9.

[10] Wang, Y., Zhang, Y., Lu, Y., Yu, X. (2020). A Comparative assessment of credit risk model based on machine learning——A case study of bank loan data. Procedia Computer Science, 174: 141-149. https://doi.org/10.1016/j.procs.2020.06.069

[11] Ampountolas, A., Nyarko Nde, T., Date, P., Constantinescu, C. (2021). A machine learning approach for micro-credit scoring. Risks, 9(3): 50. https://doi.org/10.3390/risks9030050

[12] Van Sang, H., Nam, N.H., Nhan, N.D. (2016). A novel credit scoring prediction model based on Feature Selection approach and parallel random forest. Indian Journal of Science and Technology, 9(20): 1-6. https://doi.org/10.17485/ijst/2016/v9i20/92299

[13] Devi, C.D., Chezian, R.M. (2016). A relative evaluation of the performance of ensemble learning in credit scoring. In 2016 IEEE International Conference on Advances in Computer Applications (ICACA), pp. 161-165. https://doi.org/10.1109/ICACA.2016.7887943

[14] Pal, R., Kapali, S., Trivedi, S. (2020). A study on credit scoring models with different feature selection and machine learning approaches. In e-journal-First Pan IIT International Management Conference–2018. https://doi.org/10.2139/ssrn.3743552

[15] Lenka, S.R., Bisoy, S.K., Priyadarshini, R., Hota, J., Barik, R.K. (2021). An effective credit scoring model implementation by optimal feature selection scheme. In 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 106-109. https://doi.org/10.1109/ESCI50559.2021.9396911

[16] Seo, J.Y. (2020). Machine learning in consumer credit risk analysis: A review. International Journal of Advanced Trends in Computer Science and Engineering, 9(4): 6340-6345. https://doi.org/10.30534/ijatcse/2020/328942020

[17] Mbaabu, O. (2020). Introduction to random forest in machine learning. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning, accessed on May 25, 2022.

[18] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2): 612-619. https://doi.org/10.14569/ijacsa.2020.0110277

[19] Loh, W.Y. (2011). Classification and regression trees. Data Mining and Knowledge Discovery, 1(1): 14-23. https://doi.org/10.1002/widm.8

[20] Pramod, P. (2020). Decision tree algorithm-Pianalytix-Machine learning. https://pianalytix.com/decision-tree-algorithm/, accessed on May 25, 2022.

[21] Normalized, N. (2021). Random forest algorithm clearly explained! - YouTube. https://www.youtube.com/watch?v=v6VJ2RO66Ag, accessed on May 28, 2022.

[22] Divya, C. (2021). Bootstrapping and OOB samples in random forests by divya Choudhary analytics Vidhya medium. https://medium.com/analytics-vidhya/bootstrapping-and-oob-samples-in-random-forests-6e083b6bc341, accessed on Sep. 06, 2022.

[23] Enes, Z. (2021). Out-of-bag error in random forests baeldung on computer science. https://www.baeldung.com/cs/random-forests-out-of-bag-error, accessed on Sep. 07, 2022.

[24] Piotr, P. (2020). How many trees in the Random Forest? https://mljar.com/blog/how-many-trees-in-random-forest/, accessed on May 28, 2022.

[25] Niklas, D. (2021). Random forest algorithm: A complete guide. https://builtin.com/data-science/random-forest-algorithm, accessed on May 28, 2022.

[26] Sharoon, S. (2022). Random forest hyperparameter tuning in python. Machine learning. https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/, accessed on Sep. 07, 2022.

[27] Hassine, K., Erbad, A., Hamila, R. (2019). Important complexity reduction of random forest in multi-classification problem. In 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 226-231. https://doi.org/10.1109/IWCMC.2019.8766544

[28] Jason, B. (2019). How to choose a feature selection method for machine learning.

https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/, accessed on Sep. 08, 2022.
[29] Aman. (2020). (633) Gini Index and Entropy|Gini Index and Information gain in Decision Tree|Decision tree splitting rule - YouTube. https://www.youtube.com/watch?v=-W0DnxQK1Eo, accessed on Jun. 11, 2022.
[30] Rahil, S. (2018). Feature selection techniques in machine learning with python by Rahil shaikh towards data science. https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e, accessed on Sep. 08, 2022.
[31] Soumendu, C. (2022). Understanding feature importance using random forest classifier algorithm by Soumendu Chatterjee Medium. https://medium.com/@soumendu1995/understanding-feature-importance-using-random-forest-classifier-algorithm-1fb96f2ff8a4, accessed on Sep. 07, 2022.
[32] Martin, D. (2019). What is a ROC Curve? A visualization with credit scores. https://kiwidamien.github.io/what-is-a-roc-curve-a-visualization-with-credit-scores.html, accessed on Jun. 11, 2022.
[33] Great Learning Team, "Random forest Algorithm in Machine learning | Great Learning," Feb. 19, 2020. https://www.mygreatlearning.com/blog/random-forest-algorithm/, accessed on Sep. 07, 2022.
[34] Chitambira, B. (2022). Credit Scoring Using Machine Learning Approaches. http://www.diva-portal.org/smash/get/diva2:1664698/FULLTEXT01.pdf.