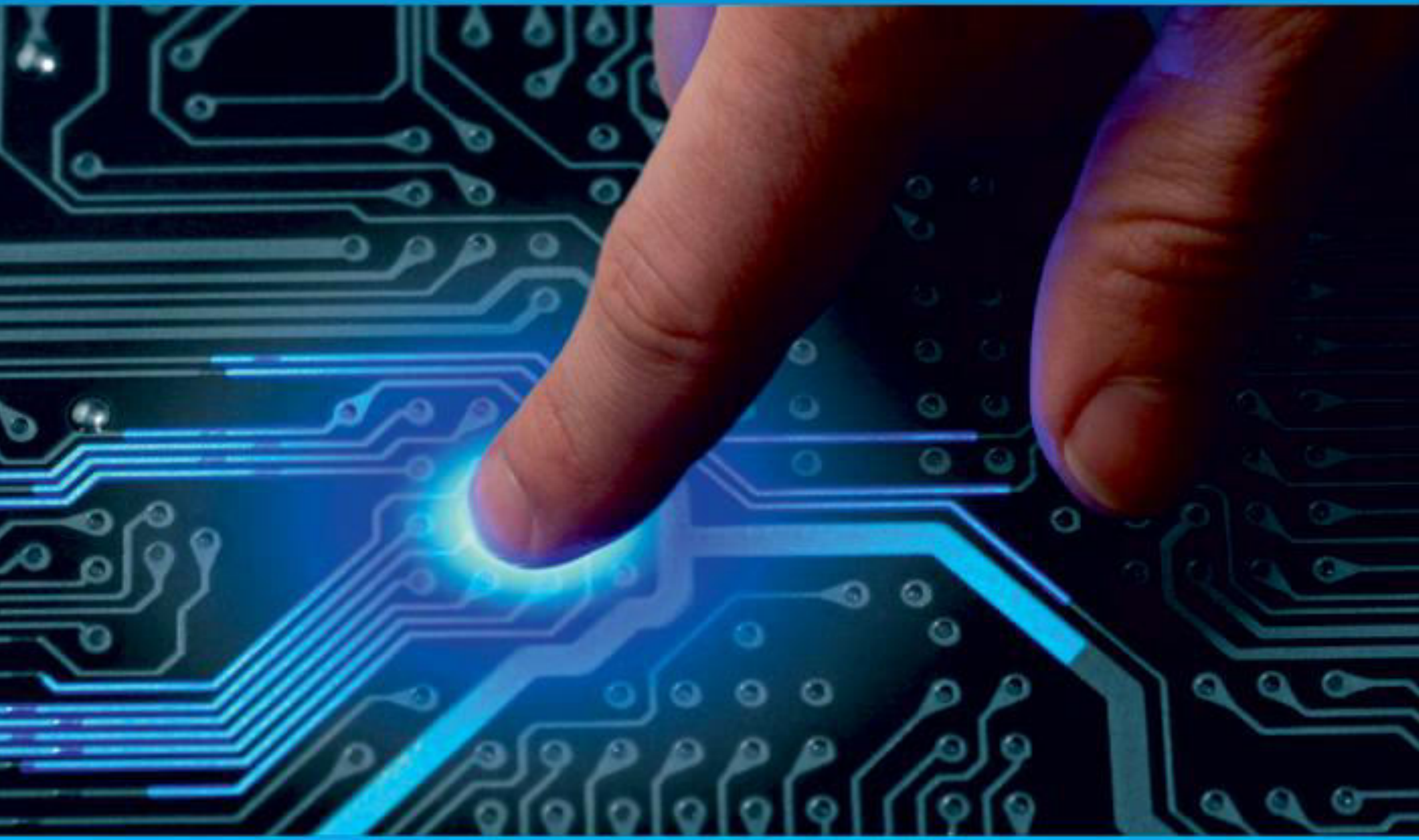
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 7, July 2023

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107001 |

Credit Risk Assessment Using a Hybrid Random Forest Algorithm

Stephen Tatenda Tinofirei, Tawanda Laston Makombe, Obert Chahele, Edington Shumbashava, Lincoln Nhapi

Student, Dept. of S.E., SIST, Harare Institute of Technology, Harare, Zimbabwe

Student, Dept. of S.E., SIST, Harare Institute of Technology, Harare, Zimbabwe

Student, Dept. of S.I., SBMS, Harare Institute of Technology, Harare, Zimbabwe

Lecturer, Dept. of F.E., SBMS, Harare Institute of Technology, Harare, Zimbabwe

Lecturer, Dept. of S. E., SIST, Harare Institute of Technology, Harare, Zimbabwe

ABSTRACT: Financial institutions define their commercial and strategic policies using a variety of methodologies, with credit risk assessment playing a significant role. Many financial institutions are facing a huge challenge of nonperforming loans. This has largely been affected by, among others, changes in macro-economic factors, poor credit culture as well as inadequate credit risk practices. Historically, credit analysts utilized credit score cards to completely assess the consumer's potential to repay the debt, but now it's appropriate to employ machine learning algorithms due to their predictive strength, speed, and effectiveness in learning complex models. Of late, there is now realization of the adoption of machine learning algorithms in many industries. This research aims at exploring the most influential variables in predicting credit risk for various loans, designing the most computationally economic random forest model that can be adopted in credit risk modelling for banks. Random forest has higher predictive power as compared to LightGBM in terms of accuracy. Both overfitting and underfitting have an effect on the accuracy of the model. However, the type of the available data defines the remedial procedures to be performed. Every algorithm must be protected against outliers, since they might distort results and efficiency.

KEYWORDS: Credit Scoring, Machine Learning, random forest algorithm, Peer-to-Peer lending.

I. INTRODUCTION

Because of the rapid development of digital financial services, academics were also compelled to concentrate on credit risk management, which provides effective solutions to decrease credit risk while maintaining a return on investments. The possibility for a creditor to suffer a loss owing to a borrower's inability to repay a loan is known as "credit default risk." .[1]

Despite strong criticism from examiners and auditors, machine learning has lately witnessed a substantial transfer from just academic research to its usage for credit scoring and applications in credit risk. This road has several possible risks and obstacles. Nonperforming loans as a percentage of total loans provided by financial institutions have increased despite the increasing complexity of credit analysis.[2] “In spite of the levels of credit analysis in operation in financial institutions, the percentage of non-performing loans continues to rise.”[3] This is because current credit bureau analytics, like credit scores, are based on slowly fluctuating consumer characteristics and thus are very slow to adapt to ever-changing consumer preferences and behaviours and the market environments over time, such as the policy inconsistencies, for example in Zimbabwe.[4] However, machine learning algorithms have a great deal to offer the area of credit risk assessment owing to their exceptional speed and predictive ability. [1] There is a relatively small amount of effort being done in the banking industry to deploy neural network models and machine learning.[5]

II. RELATED WORK

[6] used machine intelligence to estimate mobile airtime credit risk in 2021. That study recommended predicting airtime credit risk using customer profile education and recharge frequency and subscriber loan amount and frequency from loan information. That experiment used data from 90,000 mobile customers to evaluate the suggested strategy.

This study employed DT, RF, LR, and MLp. In all four models, existing and new features increase airtime credit risk prediction. LR improved 6.97% using new features and client profiles and use. RF saw the greatest improvement (7.8%) for existing loans and new features. Feature ranking algorithms prioritized subscribe loan amount.

[7] conducted research to assess the significance of using expert opinions in credit assessment. The flaw was that numerous credit scoring systems are intended to handle credit applications autonomously. They advocated incorporating expert knowledge with machine software computing approaches. The capabilities of experts to analyze the predictive ability from every attribute in the credit dataset is reinforced by the inclusion of experts in the credit scoring technique, as well as the envisaged wrapper-based feature selection approach, which investigates how the features contributing most to the categorization of borrowers. An unsupervised machine learning approach enables professionals to design one or even more clustering scenarios based on the characteristics by defining the number of clusters to be used in each scenario.

By the year 2020, a machine learning model had been built by [8] that reliably identified which borrowers were qualified for most types of loans. The authors built a logistic regression model to foretell Kaggle users' preferences using data from the platform. Data collection (Kaggle dataset), preprocessing (noisy data and missing values), and feature extraction are all part of the authors' machine learning model creation procedure (techniques used to prepare a proper dataset which is compatible with machine learning model). While this technique is efficient and reduces the likelihood of mistakes, it requires cautious model construction to save banks from losing money due to incorrect predictions.

[9] evaluated algorithm pros and cons in 2020. Support vector machines are used for binary classification tasks. SVMs use supervised learning to classify. Linear sample concerns are decided at the hyper plane edge. SVM optimizes risk framework by adding a regulator term and a loss function to the solution system. SVMs are stable sparse classifiers. SVM is simple for efficient data, adaptable to sparse data, and stable for linear data with binary outputs. SVM's 1964 development makes it less time- and performance-efficient than newer risk estimate methods. Logistic regression analysis may be used for yes/no variables. Logistic regression may predict events like other regression methods. This method reveals differential binary data relationships. Logistic regression is subtle to multivariate collinearity of self-determining features. Decision tree logic dominates machine learning. By analysing training samples, decision trees synthesize ideas and knowledge. However, inconsistent data and irregular sample sizes may skew information gain toward samples with higher values, making decision trees harder to employ. Multilayer Perceptron (MLP) artificial neural networks transform vector inputs to vector outputs while considering the network's design. MLP nodes are like directed, hierarchical, and linked graph vertices. ANN-based, it accepts nonlinear activation functions. MLP needs careful feature selection and data normalization. AdaBoost gradually chooses inferior classifiers from a weighted dataset. AdaBoost's sensitivity to outlier data affects its prediction accuracy, a major drawback. Random Forest chooses characteristics and data to generate numerous decision trees and summarize their findings. Random Forest improves prediction without increasing calculation. It may create several similar decision trees, which may obfuscate the true outcomes. Gradient Boosting Decision Tree integrates multiple poor decision trees into one effective classifier. Emerging all weak classifiers simultaneously improves model performance.

The machine learning model created by [10] in 2019 has substantial potential for usage in a credit risk assessment system. In addition to comparing other machine learning approaches, the authors have determined that the most effective model is a combination of customized Support Vector Machine and Recursive Feature Elimination with Cross- Validation. In addition, the paper includes the authors' proposed five-stage model. Support vector machines often perform better than alternative regression or tree-based models. In addition, their model has shown that recursive feature removal with cross-validation may be superior than other models in the debate on which dimensionality reduction approach to use.

Extreme competition exists amongst financial institutions due to the increased importance of credit risk management [11], which was published in 2018. Due to the difficulties of building a credit risk modelling framework, the research compares risk assessment utilizing the risk-based approach, data dimensionality reduction, and better data categorization to other methods. According to the report, consumers think banks provide money using communal funds. After a thorough client analysis, banks must carefully approve these loans. Accurate risk assessment reduces credit risk and prevents mistakes like rejecting a legitimate consumer. Current credit risk assessment uses so much data to estimate a customer's dependability that soft computing automation is needed. Risk management and analysis are crucial during liberalization in India. The report suggests combining market, credit, and operational risks into one indicator. Cooperative banks must study risk analysis and management according to the Basel Committee Agreements and RBI Guidelines. Accounting, forecasting, and statistics can estimate transactional credit risk. The authors say business intelligence (BI) helps strategic management by identifying, aggregating, and assessing data. Data Mining inside Business Intelligence (BI) systems may help banking and financial institutions find patterns, trends, and qualities that would otherwise be concealed from management due to data volume or velocity. We compare data mining methods for credit card default prediction with the best statistical forecasting methods. Authors employed SVM, Logistic Regression, and Nave Bayes. This article identified key hazards, quantified them, compared risk assessment approaches, and designed a model to better comprehend banks' primary risks and enhance system dependability.

In 2018, a study of credit risk using machine learning and deep learning was provided. [12] Using these models, binary classifiers for loan default were built. The authors emphasize the need of selecting the proper method, parameters, variables, and assessment criteria with attention. This article describes the machine learning and deep learning algorithms, classification standards, approaches for managing unbalanced datasets, outcomes, and model parameters and criteria used by financial institutions to give loans to individuals and businesses. Two machine learning models (a random forest model and a gradient boosting machine) and four deep learning models were ultimately retained. The tree-based models were more trustworthy than the neural network models.

III. PROPOSED ALGORITHM

Description of the Proposed Algorithm

Step 1 – Pick a data set.
Step 2 – Choose arbitrary subgroups of the given data as a second stage.
Step 3 – Third, construct a decision tree for each sample in order to get a prediction from each tree.
Step 4 −To go to Step 3, vote on each projected result.
Step 5 – Choose the most popular result anticipated in the collection.

IV. PSEUDO CODE

Prerequisite: Train_Set T_set = $(X_1, Y_1), \ldots, (X_n, Y_n)$, variables T, and tree numbers in Z.

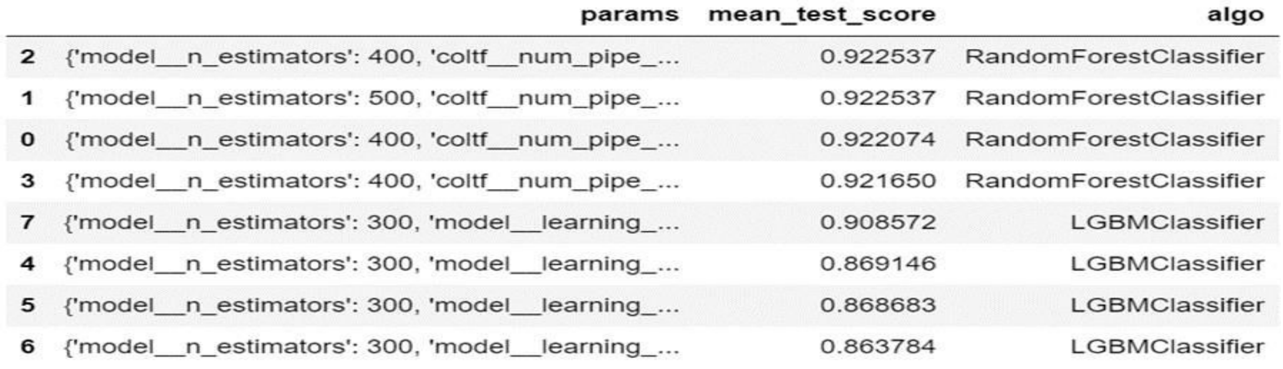
```
function RANDOMFOREST(T_set , T)
$\mathbf{B} \overset{\varnothing}{\leftarrow} \varnothing$
for i ∈ 1, . . . , B do
TS $^{(i)}$ ← A bootstrap sample from T_set
$b_i$ ← RANDOMIZEDTREELEARN (T_set $^{(i)}$, T)
$\mathbf{B} \leftarrow \mathbf{B} \cup \{b_i\}$
end for
return B
end function
function RANDOMIZEDTREELEARN(T_set , T)
        At each node:
m ← very small subset of T
Fragment on best variable in m
return the desired results
end function
```

V. SIMULATION RESULTS

The results of several random forest classifiers and random forests with varied parameters are reported in the table above, where the random forest classifier with 400 estimators emerges as the obvious victor. With a mean test score of 0.925253, this choice is superior than all others. With a mean test score of 0.863784, the classifier with the lowest performance was the LightGBM with 300 estimators. Table 2 displays the values for accuracy, precision, recall, and Fmeasure for the two techniques. The accuracy of the hybrid random forest method is 92%.

The greater the recall, the closer a value is to one on the precision recall curve. In light of the fact that the AUC for the baseline model is just 0.22, the result of 0.86 is acceptable. This shows that the model is operating effectively. The Random Forest Classifier's Learning Curve is shown in fig. 2. Above is a learning curve in which the accuracy of training is near to 1 over a variety of sample sizes. With a maximum score of 0.92, the graph indicates that the validation/testing accuracy is not convincing. The huge disparity between the two lines demonstrates data overfitting. A difference between training accuracy and testing/validation accuracy is one indication of this. Consequently, overfitting generates a very complex model and teaches itself even the undesirable "noise" in the data. Increasing the amount of available training samples is a strategy for resolving the problem of overfitting. This is crucial for the model to increase its learning. However, this is contingent on the facts at hand. Inadequate evidence renders it doubtful. When this happens, one potential option is to use the existing data to simplify the model. To accomplish this objective, the number of features may be reduced, regularization can be raised, or decision trees can be trimmed.

This research included tree trimming. Estimators were set to [100, 200, 300], with [5, 9 13] as the maximum depth and [4, 6 8] as the minimum split. A model with a maximum depth of 13 and a minimum sample split of 4 employing an iterative imputer for linear regression produced the highest mean test score of 0.91160. Since then, the accuracy and AUC have both decreased by 1%. However, recall and accuracy have increased to 0.74 and 0.83, respectively. As seen by the smaller gap between the training and testing curves in the previous learning curve, model quality has improved. It is proved that the quality of the trained model degrades as the number of training data increases.



	params	mean_test_score	algo
2	{'model__n_estimators': 400, 'coltf__num_pipe_...	0.922537	RandomForestClassifier
1	{'model__n_estimators': 500, 'coltf__num_pipe_...	0.922537	RandomForestClassifier
0	{'model__n_estimators': 400, 'coltf__num_pipe_...	0.922074	RandomForestClassifier
3	{'model__n_estimators': 400, 'coltf__num_pipe_...	0.921650	RandomForestClassifier
7	{'model__n_estimators': 300, 'model__learning_...	0.908572	LGBMClassifier
4	{'model__n_estimators': 300, 'model__learning_...	0.869146	LGBMClassifier
5	{'model__n_estimators': 300, 'model__learning_...	0.868683	LGBMClassifier
6	{'model__n_estimators': 300, 'model__learning_...	0.863784	LGBMClassifier

Table.1. The Mean Test Score Table

Classifier	Accuracy	Precision	Recall	F1
Random Forest	0.92	0.94	0.69	0.80
LightGBM	0.90	0.81	0.78	0.80

Table. 2. Performance evaluation table

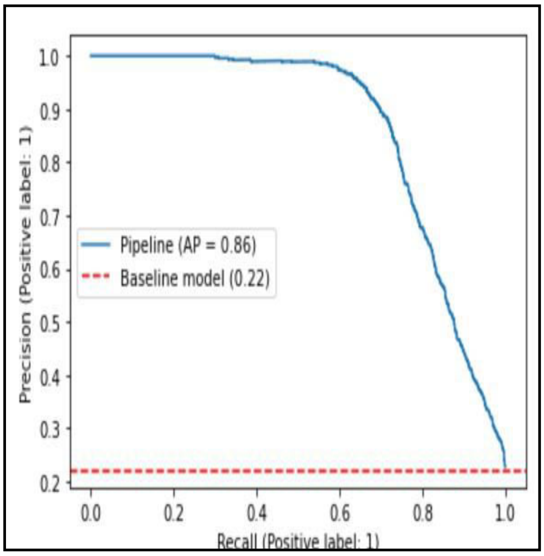


Fig 1. Precision Recall Curve of a Random Forest classifier

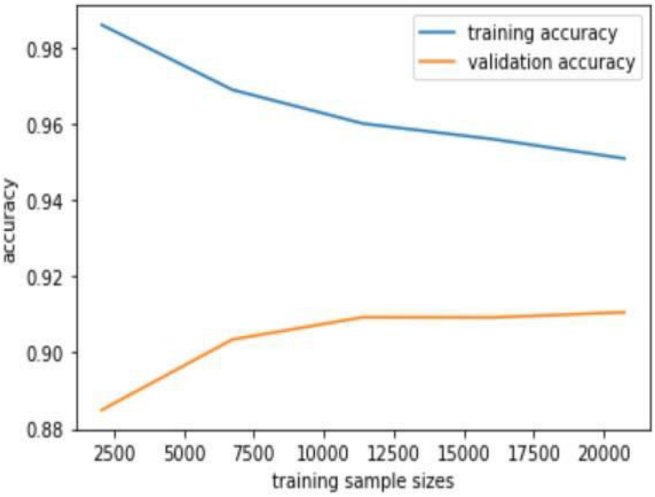


Fig 2. A Learning Curve of the random forest Classifier

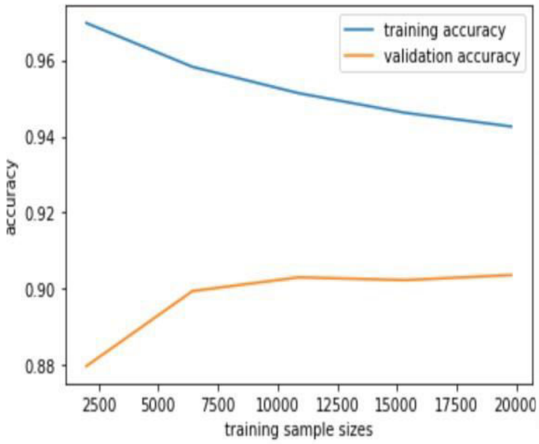


Fig 3. A Learning Curve without outliers

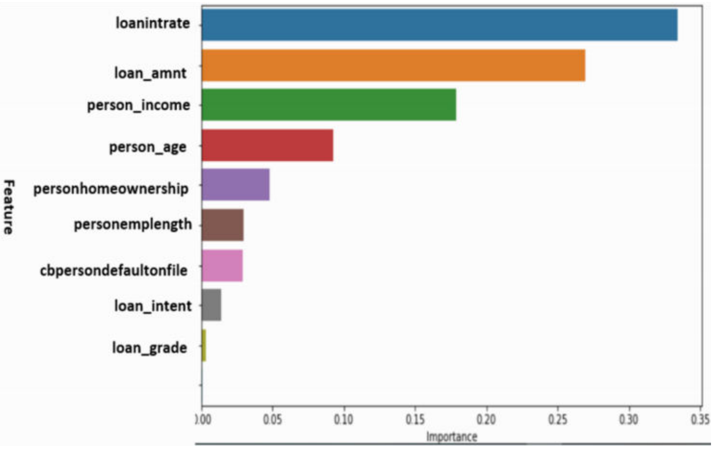


Fig 4. Feature importance

VI. CONCLUSION AND FUTURE WORK

Random forest fared better than the lighter-weight light GBM algorithm. Both overfitting and under fitting have an effect on the accuracy of the model. However, the type of the available data defines the remedial procedures to be performed. Every algorithm must be protected against outliers, since they might distort results and efficiency.

Areas of further research include:

- To provide a more accurate credit rating, a variety of other factors, such as the frequency of 90-day late payments, revolving use of unsecured lines, the frequency of 30-to-50-day late payments that are not worse, and many more, must be included. Nonetheless, this presents a dilemma for many since some of these characteristics vary across institutions.
- Hyper parameter optimization may be used to narrow the gap between training and testing accuracy. You may also test out whole alternative models
- It is crucial to assess the worth of human expertise as a complement to machine learning algorithms for credit default prediction.
- Other machine learning techniques may be applied and their ability to anticipate a client's default assessed.
- Deep learning methods, such as swarm optimization, are emerging and may outperform current machine learning algorithms if implemented.

REFERENCES

[1] S. Beshr, “A Machine Learning Approach To Credit Risk Assessment,” p. 2020, 2020.

[2] J. L. Breeden, “A Survey of Machine Learning in Credit Risk,” researchgate.net, no. May, 2020, doi: 10.13140/RG.2.2.14520.37121.

[3] B. W. Garikai and C. Givemore, “Analysis of Credit Culture in the Zimbabwean Banking Sector,” no. April, 2019.

[4] D. Yuan, “Applications of Machine Learning : Consumer Credit Risk Analysis,” 2015.

[5] A. Petropoulos, V. Siakoulis, E. Stavroulakis, and A. Klamargias, “A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting 1 analysis of large loan level datasets using deep,” no. August, pp. 30–31, 2018.

[6] L. Information and S. B. Berhe, “Addis Ababa Institute of Technology School of Electrical and Computer Engineering Telecommunication Engineering Graduate Program Addis Ababa Institute of Technology School of Electrical and Computer Engineering Telecommunication Engineering Graduate Progr,” no. December, 2021.

[7] G. Palareti et al., “Comparison between different D-Dimer cutoff values to assess the individual risk of recurrent venous thromboembolism: Analysis of results obtained in the DULCIS study,” Int. J. Lab. Hematol., vol. 38, no. 1, pp. 42–49, 2016, doi: 10.1111/ijlh.12426.

[8] M. A. Sheikh, A. K. Goel, And T. Kumar, “An Approach For Prediction Of Loan Approval Using Machine Learning Algorithm,” Proc. Int. Conf. Electron. Sustain. Commun. Syst. Icesc 2020, Vol. 9, No. 6, Pp. 490–494, 2020, Doi: 10.1109/Icesc48915.2020.9155614.

[9] Z. Tian, J. Xiao, H. Feng, And Y. Wei, “Credit Risk Assessment Based On Gradient Boosting Decision Tree,” Procedia Comput. Sci., Vol. 174, Pp. 150–160, 2020, Doi: 10.1016/J.Procs.2020.06.070.

[10] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif, And S. Islam, “Application Of Machine Learning In Credit Risk Assessment: A Prelude To Smart Banking,” Ieee Reg. 10 Annu. Int. Conf. Proceedings/Tencon, Vol. 2019-Octob, Pp. 2023–2028, 2019, Doi: 10.1109/Tencon.2019.8929527.

[11] A. Mittal, A. Shrivastava, A. Saxena, And M. Manoria, “A Study On Credit Risk Assessment In Banking Sector Using Data Mining Techniques,” 2018 Int. Conf. Adv. Comput. Telecommun. Icacat 2018, Pp. 1–5, 2018, Doi: 10.1109/Icacat.2018.8933604.

[12] P. M. Addo, D. Guegan, And B. Hassani, “Credit Risk Analysis Using Machine And Deep Learning Models,” Ssrn Electron. J., No. May 2019, 2018, Doi: 10.2139/Ssrn.3155047.

BIOGRAPHY

Stephen Tatenda Tinofirei is a Student in the Software Engineering Department, Harare Institute of Technology, Harare, Zimbabwe. He received Bachelor of Technology (Hons) in Financial Engineering degree in 2017 from Harare Institute of Technology, Harare, Zimbabwe. He is interested in the following areas research: Machine Leaning, Operations Research, etc.

Tawanda Laston Makombe is a Financial Software Engineer who completed a degree in Master of Technology in Software Engineering in 2022 and a Bachelor of Technology (Hons) degree in Financial Engineering in 2017 at the Harare institute of Technology. His research interests are focused in the Machine Learning, Software Pricing, etc.

Obert Chahele is a Staff Development Fellow at Harare Institute of Technology, Harare, Zimbabwe. He completed a Masters of Technology in Strategy and Innovation in 2022 and a degree in Financial Engineering in 2018 from Harare Institute of Technology. His research interests are focused in: Innovation and Technology, Smart City Development, etc

Edington Shumbashava is a Staff Development Fellow in Financial Engineering department who is Certified Financial Analyst (CFA) 2022 and a holder of Bachelor of Technology (Hons) degree in Financial Engineering in 2017 at the Harare institute of Technology. His research interests are focused in the Financial Modelling, Investment Analyses, etc.

Lincoln Nhapi is a well experienced lecturer in the School of Information Sciences and Technology at the Harare Institute of Technology.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details