CREDIT RISK MODELING USING MACHINE LEARNING

By

DEVANSH SHARMA

A research capstone submitted
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

Faculty Advisor: Marc Velasco
Department of Computer Science
California State University, Fullerton
FULLERTON, CALIFORNIA, USA

May 11 , 2023

Abstract

Credit Risk Modeling Using Machine Learning. Sharma, Devansh, 2023. Research Capstone Paper, California State University Fullerton.

Credit risk modeling using machine learning is a technique lenders use, to find out the level of credit risk, related to extending credit to a borrower. Organizations use credit risk modeling using ML, including insurance corporations, banks, investment enterprises, and government treasuries. Sometimes, individual people use credit risk modeling to loan away their own money strategically. Credit risk modeling is crucial anywhere people are borrowing money. Machine Learning models are being used to protect against increasingly sophisticated fraud attempts. This research aims to analyze and ensure that the models created for credit risk using machine learning methods such as K-Nearest Neighbors, Decision Trees, Random Forests, Neural Networks, and SVM produce data that are both accurate and scientific. And then dive into the ongoing research of the most accurate model from our analyzed results.

Table of Contents

Introduction

To learn as humans, understand and solve problems through real-world interactions and observations, various machine learning algorithms help computer systems to systematically generate results close to a human's cognition. One such crucial problem to be addressed is credit risk modeling which deals with recognizing patterns in credit card fraud and loan defaults. As the number of loan defaulters is on the rise, the necessity of an accurate machine learning model becomes very important since banking institutions face huge losses with assets getting frozen and transactions being halted (Shoumo et al., 2019).

This project aims at understanding the concept of credit risk modeling, its necessity in the real-world scenario, and the algorithms used to build credit risk models. Moreover, the main idea behind this research is to create a hypothesis, 'Comparative study of credit risk modeling approaches to find the two best working algorithms', and analyze critical points of this hypothesis, along with collecting enough data points that could validate the proposed hypothesis. For the purpose of conducting research, different experiments conducted in the past are referenced and analyzed. Classifiers like the random forest, SVM, Logistic regression, and XGBoost would be taken into account to conduct the research, and they would be trained and tested on datasets from personal history with the previous credit history of loan-seeking individuals. This research could help financial institutions in accurately assessing the risk of lending loans to individuals and assessing the ability of a loan seeker to pay it back to them.

Related Work

The topic of credit risk assessment is one that receives a lot of attention in the world of banking and finance. More importantly, this topic has become much more important and well-known as a result of the recent growth of data science and a number of significant developments in the field of machine learning. Numerous important discoveries have been produced in this area, serving as stepping stones for ongoing research and upcoming investigations. As shown in (Attigeri et al., 2017; Bekhet & Eletter, 2014; Stergiou & Siganos, 2010), two suitable methods for assessing credit risk are Artificial Neural Networks (ANN) and Logistic Regression (LR) models. In both situations, the LR model fared better at predicting loan defaulters than the ANN model. It should be emphasized that for best results, the number of nodes in each layer of models based on neural networks (NN) should be tweaked and optimized. In order to address this issue, numerous neural networks with various numbers of nodes in each layer were given in (Khashman, 2010), and the optimal model was ultimately chosen after experimentation. Models based on Support Vector Machine (SVM) are also used in credit risk modeling. An SVM model built on fuzzy logic was used by the authors (Wang et al., 2005) and performed better than NN and regression models. Additionally, multi-agent models and ensemble (bagging, boosting, stacking) models offer an intriguing approach to this issue (Yu et al., 2008; Yu et al., 2010). The studies (Addo, 2018; Khandani et al., 2010; S & T, 2015; Twala, 2010) demonstrate the use of tree-based models in the identification of defaulters, where they significantly outperformed NN and other linear models. For the optimum outcome, it is always advised to tune the hyperparameters of the learning algorithms. If applied with proper tuning, the SVM kernel method can outperform many classifiers on massive datasets. Furthermore, a significant problem in credit risk assessment is the situation of imbalanced datasets. Using weighting, undersampling, or oversampling can all be used to address this. This problem has been covered in the paper (Khandani et al., 2010). Therefore, it is clear from earlier related works that great progress has been made and still has to be made in the area of credit risk modeling. According to the studies stated above, adequate dimensionality reduction techniques, parameter tuning, noise treatment, and imbalanced dataset handling are essential for creating a model that is intended to aid in credit risk assessment. The cost and complexity of the calculation should also be given considerable consideration.

The first Research observation and experiments propose to create such a model which addresses all these issues. For our first research observation, grid search with cross-validation will be used for parameter tuning and noise handling, and the issue of an unbalanced dataset will also be successfully resolved. Different dimensionality reduction techniques will be used to compare various supervised classification algorithms, and the best combination will be chosen for credit risk assessment. Overall, our goal is to present a model that takes into account all the aforementioned issues and develop an accurate model for estimating credit risk in the banking industry (Shoumo et al., 2019).

The second Research observation and experiments examine the use of the most popular and best-supervised machine learning methods, including Logistic Regression and the XGBoost algorithm, an improved algorithm of the Gradient Boosted Decision Trees algorithm that has fast calculation speed and better model performance, for credit risk modeling. It also conducts a thorough analysis of forecasting techniques and their impacts (Li, 2019).

The third Research observation and experiment's purpose is to help companies to predict customers who have the potential to become debtors. The study is conducted by comparing Logistic regression model and XGBoost. The accuracy of the testing data on the logistic regression model and the XGBoost model is compared. The comparison results are based on four evaluation indicators, namely accuracy, sensitivity, specificity, and precision (Dwidarma et al., 2021).

Project Plan

For this project, two experiments on credit risk modeling were taken into account. The first study makes a comparison of algorithms like SVM, Random Forest, XGBoost, and Logistic regression. The model is trained on a medium-sized dataset and tested for all these algorithms, and the results are compared by using different evaluation metrics such as F1 score, precision, recall, AUC score, and prediction accuracy. This study concludes that SVM outperformed all the other approaches.

For the next experiment, a comparison between Logistic regression and XGBoost has been conducted on a large dataset, with the intention of identifying bad customers who do not pay the money back from the good customers. Metrics such as AUC score, KS, and PSI values have been used to evaluate each model. The results show that the XGBoost algorithm works much better.

The third experiment aims to assist businesses in identifying clients who may eventually become debtors. The study is conducted by comparing Logistic regression model and XGBoost. The comparison results are based on four evaluation indicators, namely accuracy, sensitivity, specificity, and precision. This study concludes that XGBoost outperformed Logistic Regression.

Therefore a hypothesis is proposed for the above problem, and further research is done to collect enough data points for the purpose of validating the hypothesis. The results are finally summarised in a separate section.

For further graduate research, a comparison of the two best approaches obtained from this project would be conducted based on self-made observations through building a model. The main goal of future research would be to analyze the performance efficiency of XGBoost as compared to SVM when these two models are run on the same large-sized dataset.

Methodology

In the past years, a significant number of experiments have been conducted on improving the efficiency of credit risk modeling methodologies. The main aim of this project is to derive the two best working algorithms for credit risk modeling.

For the purpose of conducting the research, a hypothesis is proposed. At first, the hypothesis 'Comparative study of credit risk modeling approaches to find the two best working algorithms' is analyzed using studies done by other researchers. The results obtained in these experiments are compared to verify if they validate the above-stated hypothesis, followed by deriving two best working approaches. These two approaches are further used for building a model as a part of future research.

Analysis methods:

The analysis methods used for this research are primarily studying various experiments done by researchers and comparing their observations. For every research experiment taken into consideration, firstly, the dataset used is mentioned along with the evaluation metrics. The research methodology for every experiment is then scrutinized to understand the approaches used for solving problems with existing methods. This is followed by judging the results based on the various evaluation metrics. For this research, more than one evaluation metric has been used in order to draw rock-solid conclusions that are capable of validating the hypothesis. With various experiments being analyzed, two best working algorithms are derived, which could be used for future project plans.

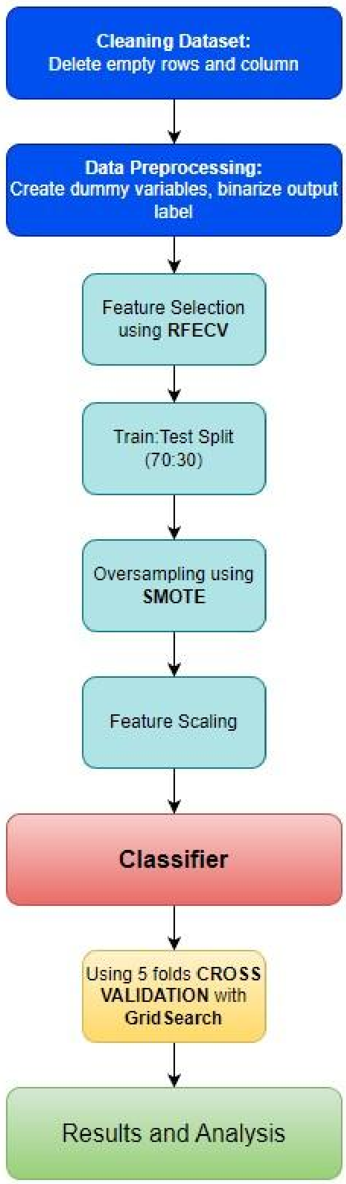
Research Observations and Experiments #1

Research Methodology:

The dataset used in this study was collected from Lending Club, a well-known and reputable peer-to-peer lending institution based in the United States. First, dummy variables will be created to accommodate categorical values, and the output label will be binarized to start the dataset preprocessing. Now, the desired output will either be a "1" for someone who successfully repaid the loan or a "0" for someone who failed to do so. Models will be created using dimensionality reduction techniques such as Recursive Feature Elimination with Cross-Validation (RFECV) which is the optimal dimensionality reduction procedure as compared to PCA, as depicted in (Shoumo et al., 2019, p. 2025). When building models, dimensionality reduction approaches like Recursive Feature Elimination with Cross-Validation (RFECV), which is superior to PCA in terms of dimensionality reduction, will be used, as depicted in (Shoumo et al., 2019, p. 2025). Additionally, the dataset will be divided into a training set and a test set with a 70:30 split between them. The model will also be tested using a 5-fold cross-validation process, with the training set serving as the domain. GridSearchCV will only be utilized with the 5-fold cross-validation when tweaking the hyperparameters. This is done to demonstrate how tuned classifiers outperform classifiers in their basic form. In addition, the minority class will be oversampled using the Synthetic Minority Oversampling Technique (SMOTE), given how strongly unbalanced the dataset is. Standardization will be used to scale the features. Finally, various assessment metrics will be used to assess the anticipated outcomes, including F1 score, precision, recall, AUC score, and prediction accuracy. The findings will be displayed in a tabular and graphical format (Shoumo et al., 2019).

The process that was used to develop the suggested model is depicted in the flow chart Figure 1 (Shoumo et al., 2019, p. 2026).

Figure 1: Flowchart illustrating the workflow of the model



Note. Adapted from 'Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking' by Syed Zamil Hasan Shoumo, Mir Ishrak Maheer Dhruba, Sazzad Hossain, Nawab Haider Ghani, Hossain Arif, Samiul Islam, 2019, *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, p. 2024, 10.1109/TENCON.2019.8929527. Copyright, 2019 by IEEE.

Every machine-learning model needs an accurate classifier as its foundation. The analysis will use four supervised algorithms: Support vector machine (SVM), Logistic Regression (LR), Extreme Gradient Boosting (XGB), and Random Forest (RF). To find the ideal set of values for each model, the hyperparameters of these algorithms will be tuned using GridSearchCV.

Results:

The accuracy has been measured using 5-fold cross-validation, along with fine-tuned hyperparameters using GridSearchCV. 5-fold cross-validation is used to generate more stable results that are trained and tested on the entire dataset. Moreover, hyperparameter tuning will help in better model performance.

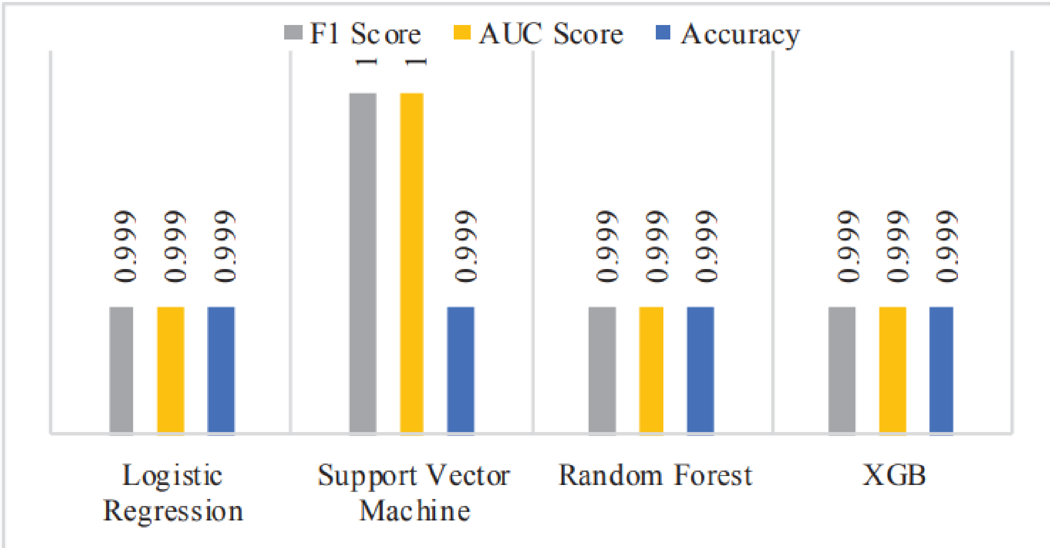
Table 1: RFECV with Cross-Validation

Details	RFECV			
	Logistic Regression	Support Vector Machine	Random Forest	XGB
Precision	0.998	1.0	0.999	0.999
Recall	1.0	1.0	1.0	1.0
F1 Score	0.999	1.0	0.999	0.999
AUC Score	0.999	1.0	0.999	0.999
Accuracy	0.999	0.999	0.999	0.999

Note. From 'Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking' by Syed Zamil Hasan Shoumo, Mir Ishrak Maheer Dhruba, Sazzad Hossain, Nawab Haider Ghani, Hossain Arif, Samiul Islam, 2019, *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, p. 2026, 10.1109/TENCON.2019.8929527. Copyright, 2019 by IEEE.

Table 1 (Shoumo et al., 2019, p. 2026) showcases the evaluation of the models after applying GridSearchCV and a 5-fold cross-validation. Among all the models, the SVM model with RFECV has attained the highest scores.

Figure 2: Model Comparison using RFECV and Cross-Validation



Note. From 'Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking' by Syed Zamil Hasan Shoumo, Mir Ishrak Maheer Dhruba, Sazzad Hossain, Nawab Haider Ghani, Hossain Arif, Samiul Islam, 2019, *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, p. 2026, 10.1109/TENCON.2019.8929527. Copyright, 2019 by IEEE.

Figure 2 (Shoumo et al., 2019, p. 2026) illustrates how the SVM model based on RFECV overcame all other models after hyperparameter tuning using GridSearchCV.

It was observed that SVM generated a maximum of AUC and F1 scores. In fact, it was found to be the highest score among all other models in all the experiments. This clearly indicates that the SVM model trained using RFECV generates the most optimal results and is the best combination to produce efficient risk analysis.

Table 2: Variance Comparison using RFECV

RFECV	
Algorithm	**Variance**
Logistic Regression	0.00078
Support Vector Machine	0.00012
Random Forest	0.00104
XGB	0.00074

Note. From 'Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking' by Syed Zamil Hasan Shoumo, Mir Ishrak Maheer Dhruba, Sazzad Hossain, Nawab Haider Ghani, Hossain Arif, Samiul Islam, 2019, *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, p. 2026, 10.1109/TENCON.2019.8929527. Copyright, 2019 by IEEE.

Table 2 (Shoumo et al., 2019, p. 2026) shows the variance of each of the 5 folds that were created during the evaluation of every model.

In a nutshell, based on all the metrics, namely, prediction accuracy, precision, recall, F1 score, AUC score, and variance, the SVM model combined with RFECV has consistently performed the best among all other classifiers. Therefore, after going through the performance of all the classifiers based on the different evaluation metrics, this study concludes that SVM with RFECV has outperformed other tree-based approaches and can be safely used to analyze the risk involved while lending loans to individuals.

Conclusion:

The SVM model based on recursive feature elimination with cross-validation has achieved an accuracy of 99.9 percent with an AUC score of 1, a precision score of 1, a recall score of 1, and an F1 score of 1 when handling a sum total of 38,661 instances. It can be concluded that the proposed model in this experiment has achieved notable results and is capable of assessing the risk in loan lending and targeting the defaulters.

Research Observations and Experiments #2

Research methodology:

In this experiment, supervised machine learning methods are used for credit risk modeling, and a detailed analysis of the performance of these approaches based on various accuracy metrics has been taken into account. A comparison between Logistic Regression and XGBoost is made using a dataset that has an observation period of 2 months prior to a loan application. XGBoost, which stands for Extreme Gradient Boosting, is an implementation of the Gradient Boosted Decision Tree algorithm. It is faster in calculation and performs better than other classifiers. Therefore, this classifier is compared with Logistic Regression to analyze the performance efficiency of Logistic Regression with XGBoost.

For the purpose of testing the models, two evaluation metrics, namely, Discrimination Capability and Stability, have been used. Discrimination Capability, which is indicated by metrics like AUC or KS value, refers to the model's ability to differentiate bad samples from good samples, whereas Stability, which is judged using PSI, indicates if the distribution of the population has changed for different samples at different times after classification in accordance with the probability or score (Li, 2019).

The working methodology for building models is divided into 6 major steps:

1. Cleaning the data and preprocessing the missing, duplicate, single, or abnormal values
2. Processing relevant variables and removing low-level features using feature selection techniques.
3. Splitting dataset into training, testing, and validation data
4. Training both the classifiers

5. Evaluating the performance of these classifiers based on metrics like AUC, KS, and PSI values with repeatedly iterating over steps 4 and 5 to enhance the performance.
6. Testing the model performance on the validation set (Li, 2019).

Results:

1. Logistic regression

   For logistic regression, it was observed that the KS value of the training and testing data is the same, and it achieves a value of 52% on the validation data. Whereas the PSI value reached 0.0458 on the testing and validation dataset, indicating that the stability is quite high (Li, 2019).
2. XGBoost

   In the case of XGBoost, a difference of 0.01 was observed between the KS values of the training set and test set, whereas on the validation set, it reaches a value of 0.61. However, the discrimination effect of XGBoost is higher than logistic regression by a value of 0.08. The XGBoost model was observed to be quite stable on the training set and test set, with a PSI value of 0.0565 on the testing and validation set (Li, 2019).

Conclusion:

After making a comparison between the model discrimination and stability metrics of both the models, it was observed that XGBoost performed significantly higher than logistic regression and is found to be capable of enhancing the identification ability of possible risk in lending loans (Li, 2019).

Research Observations and Experiments #3

Research methodology:

In this experiment, the classification method is used to predict potential debtor customers. The dataset used in this research was 30304 customers, consisting of 14883 customers who became debtors and 15421 non-debtor customers recorded in the data warehouse of one of the largest private banks in Indonesia. After obtaining influential variables, the analysis is continued by making predictions and knowing the variables which influence a customer to potentially take a credit facility using two algorithms, namely Logistic Regression and XGBoost method. For the purpose of testing the models, confusion matrix method is used as the evaluation metrics, and four evaluation indicators, namely accuracy, sensitivity, specificity, and precision, have been used. AUC metric, which refers to the model's ability to differentiate bad samples from good samples, is also used (Dwidarma et al., 2021).

The working methodology for building models is divided into several stages:

1. Data preprocessing: The R software's nearZeroVar function is used to eliminate predictor variables with very low variances (variances that are close to zero). Multivariate Imputation by Chained Equations (MICE) is used to impute the data. The sampling method is used to impute the data, meaning that blank entries in the data will be filled with random values taken from the observation data.
2. Data preparation: Using the createDataPartition function supplied by the caret package in R software, the data in this study are split into a composition of 80% training data and 20% testing data. After the data are separated, this function employs random sampling in the dependent variable category to keep the data balance.

3. Logistic regression modeling: At this point, modeling is done via the logistic regression method. The categorical dependent variable may be predicted using logistic regression, and the significant independent variables can be identified.
4. XGBoost modeling: Extreme Gradient Boosting (XGBoost) is the modeling technique being used at this point. The dependent categorical variable is predicted using XGBoost.
5. Comparison: To choose the best model, the results of the model evaluation utilizing accuracy, sensitivity, specificity, and precision values between the logistic regression model and the XGBoost model are compared (Dwidarma et al., 2021).

Results:

According to Table 3 (Dwidarma et al., 2021), the XGBoost model has higher values for accuracy, sensitivity, specificity, and precision for training data than the logistic regression model. The XGBoost model outperforms the Logistic Regression model in terms of accuracy, sensitivity, specificity, and precision in the testing data. The XGBoost model, which is superior to the Logistic regression model and yields the best forecasts for probable debtors of one of the biggest private banks in Indonesia, can therefore be drawn.

Table 3: Comparison of Logistic regression model and XGBoost model

Criterion	**Logistic Regression**		**XGBoost**	
	Train	***Test***	***Train***	***Test***
Accuracy	0.8788	0.8804	**0.9994**	**0.9267**
Sensitivity	0.8243	0.8233	**0.9988**	**0.9158**
Specificity	0.9313	0.9355	**0.9999**	**0.9376**
Precision	0.9205	0.9249	**0.9999**	**0.9362**

Note. From "Comparison of logistic regression and xgboost for predicting potential debtors" by Revina Dwidarma, Syarifah Diana Permai, Jeklin Harefa, 2021, 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), (pp. 1-6), 10.1109/AiDAS53897.2021.9574350. Copyright, 2021 by IEEE

Conclusion:

As a consequence, 87.88%, 82.43%, 93.13%, and 92.05%, respectively, are the values for the accuracy, sensitivity, specificity, and precision of the logistic regression model. The accuracy, sensitivity, specificity, and precision values offered by XGBoost modeling are also higher than those of Logistic Regression. The corresponding percentages are 99.95%, 998.8%, 999.99%, and 999.99%. The XGBoost model's tuning parameters revealed that the best values for eta, max depth, subsample, column sample by tree, and gamma are 0.06, 0.8, and 1. The highest AUC test value is generated by these factors. These findings indicate that the logistic regression in this study is inferior to the XGBoost (Dwidarma et al., 2021).

Conclusion and Future Work

In the first experiment, a comparison between four algorithms Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost was made. It was found that the SVM model based on recursive feature elimination with cross-validation has achieved an accuracy of 99.9 percent with an AUC score of 1, a precision score of 1, a recall score of 1, and an F1 score of 1 when handling a sum total of 38,661 instances. In the second experiment, a comparative study between Logistic Regression and XGBoost was conducted and it was observed that XGBoost performed significantly higher than logistic regression and is found to be capable of enhancing the identification ability of possible risk in lending loans. Finally, in the third experiment, the performance of Logistic Regression and XGBoost were compared on a different dataset and different sets of evaluation metrics. It was the XGBoost model that had greater accuracy, sensitivity, specificity, and precision values than the logistic regression model for training data. In the testing data, the XGBoost model also produces greater accuracy, sensitivity, specificity, and precision values than the Logistic regression model. Then, it can be concluded that the XGBoost model is better than the Logistic regression model and produces the best predictions for the potential debtors of one of the largest private banks in Indonesia. Therefore, the proposed hypothesis ‘Comparative study of credit risk modeling approaches to find the two best working algorithms’ gets validated from the above experiments, and their observed results with the two best working algorithms, namely SVM and XGBoost, have been derived for future experiments as a part of the proposed project plan.

Reference

[1] Shoumo, S. Z. H., Dhruba, M. I. M., Hossain, S., Ghani, N. H., Arif, H., & Islam, S. (2019, October). Application of machine learning in credit risk assessment: a prelude to smart banking. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (pp. 2023-2028). IEEE.

[2] Li, Y. (2019, August). Credit risk prediction based on machine learning methods. In *2019 14th International Conference on Computer Science & Education (ICCSE)* (pp. 1011-1013). IEEE.

[3] Wang, Y., & Ni, X. S. (2019). A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433*.

[4] C. Stergiou and D. Siganos, “NEURAL NETWORKS,” rep., 2010.

[5] H. A. Bekhet and S. F. K. Eletter, “Credit risk assessment model for Jordanian commercial banks: Neural scoring approach,” Review of Development Finance, vol. 4, no. 1, pp. 20–28, 2014.

[6] G. V. Attigeri, M. M. M. Pai, and R. M. Pai, "Credit Risk Assessment Using Machine Learning Algorithms," Advanced Science Letters, vol. 23, no. 4, pp. 3649–3653, Jan. 2017.

[7] A. Khashman, "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes," Expert Systems with Applications, vol. 37, no. 9, pp. 6233–6239, 2010.

[8] Y. Wang, S. Wang, and K. Lai, "A new fuzzy support vector machine to evaluate credit risk," IEEE Transactions on Fuzzy Systems, vol. 13, no. 6, pp. 820–831, 2005.

[9] L. Yu, S. Wang, and K. Lai, "Credit risk assessment with a multistage neural network ensemble learning approach," Expert Systems with Applications, vol. 34, no. 2, pp. 1434–1444, 2008.

[10] L. Yu, W. Yue, S. Wang, and K. Lai, "Support vector machine based multiagent ensemble learning for credit risk evaluation," Expert Systems with Applications, vol. 37, no. 2, pp. 1351–1360, 2010.

[11] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer Credit Risk Models Via Machine-Learning Algorithms," SSRN Electronic Journal, 2010.

[12] S. M. S and R. S. T, "Loan Credibility Prediction System Based on Decision Tree Algorithm," International Journal of Engineering Research and, vol. V4, no. 09, 2015.

[13] P. M. Addo, D. Guegan, and B. Hassani, "Credit Risk Analysis Using Machine and Deep Learning Models," SSRN Electronic Journal, 2018.

[14] B. Twala, "Multiple classifier application to credit risk assessment," Expert Systems with Applications, vol. 37, no. 4, pp. 3326–3336, 2010.

[15] Dwidarma, R., Permai, S. D., & Harefa, J. (2021, September). Comparison of logistic regression and xgboost for predicting potential debtors. In *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 1-6). IEEE.