Journal of Soft Computing Exploration

Homepage: https://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Using Genetic Algorithm Feature Selection to Optimize XGBoost Performance in Australian Credit

Dwika Ananda Agustina Pertiwi[1*], Kamilah Ahmad[2], Shahrul Nizam Salahudin[3], Ahmed Mohamed Annegrat[4], Much Aziz Muslim[5]
[1,2,3,5]Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
[4]Faculty of Economics, University of Bani Waleed, Libya
[5]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received March 30, 2024
Revised April 3, 2023
Accepted April 3, 2024

Keywords:

Credit risk
Australian credit
Genetic algorithm
XGBoost

ABSTRACT

To reduce credit risk in credit institutions, credit risk management practices need to be implemented so that lending institutions can survive in the long term. Data mining is one of the techniques used for credit risk management. Where data mining can find information patterns from big data using classification techniques with the resulting level of accuracy. This research aims to increase the accuracy of classification algorithms in predicting credit risk by applying genetic algorithms as the best feature selection method. Thus, the most important feature will be used to search for credit risk information. This research applies a classification method using the XGBoost classifier on the Australian credit dataset, then carries out an evaluation by measuring the level of accuracy and AUC. The results show an increase in accuracy of 2.24%, with an accuracy value of 89.93% after optimization using a genetic algorithm. So, through research on genetic algorithm feature selection, we can improve the accuracy performance of the XGBoost algorithm on the Australian credit dataset.

Corresponding Author:

Dwika Ananda Agustina Pertiwi,
Faculty Technology Management and Business,
Universiti Tun Hussein Onn Malaysia,
Persiaran Tun Dr. Ismail, 86400 Parit Raja, Johor, Malaysia
Email: dwikapertiwi13@gmail.com

https://doi.org/10.52465/joscex.v5i1.302

1. INTRODUCTION

Credit problems in banks in the last few decades have required a deepening of existing risk management [1]. Therefore, to help strengthen the reliability of credit institutions and improve significantly. So, in this case credit management becomes an important issue [2-4]. Being one of the competitive financial institutions for profit, Bank provides a variety of financial services such as credit to individuals and businesses as well as managing various types of risk. From this fact, that taking a risk is synonymous with being profitable, banks derive a significant portion of their profits from their lending activities [5]. As a result, they are keenly interested in creating credit risk assessment models that are constantly more accurate in order to optimize the performance of loans that they have granted.

The risk of credit has been predicted using a variety of ways. Probability of default shows the possibility of default at a certain time, and is the main parameter in the credit risk evaluation system Since credit scores determine the probability of default, credit risk evaluation has emerged as an aid credit risk management tool by identifying “good” or “bad” applicants. By adopting a data mining strategy to examine applicant data, the degree of credit risk may be decreased [6], [7]. Many studies of risk management for credit risk evaluation have been carried out, this important issue is interesting to study by utilizing a classification algorithm with an optimal level of accuracy [8-10].

Supervised machine learning models have been widely applied in credit risk assessment, in particular they are used in credit scoring models to find default probabilities and then predict default classification usually in binary format. Various types of classification algorithms have been applied to credit risk prediction studies such as LightGBM [11], Logistic Regression (LR) [12], Gradient Boosting [13], XGBoost [10], Neural Network (NN) [14]. This paper applied eXtreeme Gradient Boosting Tree (XGBoost) proposed recently by Chen and Guestrin [15]. Due to its speed and accuracy, XGBoost has attracted interest in some significant global big-data contests including Kaggle and DataCastle. We believe a robust yet effective solution is a potential advance in this area instead of going for a complicated way of creating a model for financial institutions to use in practice.

This research applies the selection of the best features to improve classification accuracy by removing redundant and unnecessary features from the dataset [16], [17]. The most popular feature selection technique is Genetic Algorithm (GA), and has been proven effective in several scopes of computer science [18], including data mining [19], and industrial applications [20]. Genetic algorithms have been used to obtain optimum values and demonstrate their superiority in increasing the accuracy of classification models.

In this study proposes a Genetic Algorithm (GA) to improve XGBoost classifier accuracy of Australian credit dataset. Then, to handle imbalanced class data using a synthetic minority oversampling technique (SMOTE) model [21], [22]. At the evaluation stage we look at the value of accuracy, and AUC.

2. METHOD

In this study, a method was developed consisting of data collection, preprocessing, data split, classification method, and evaluation using a confusion matrix. The benchmarks and suggested models are validated using a real-world credit dataset called Australian credit. The method design in this study is shown in Figure 1.

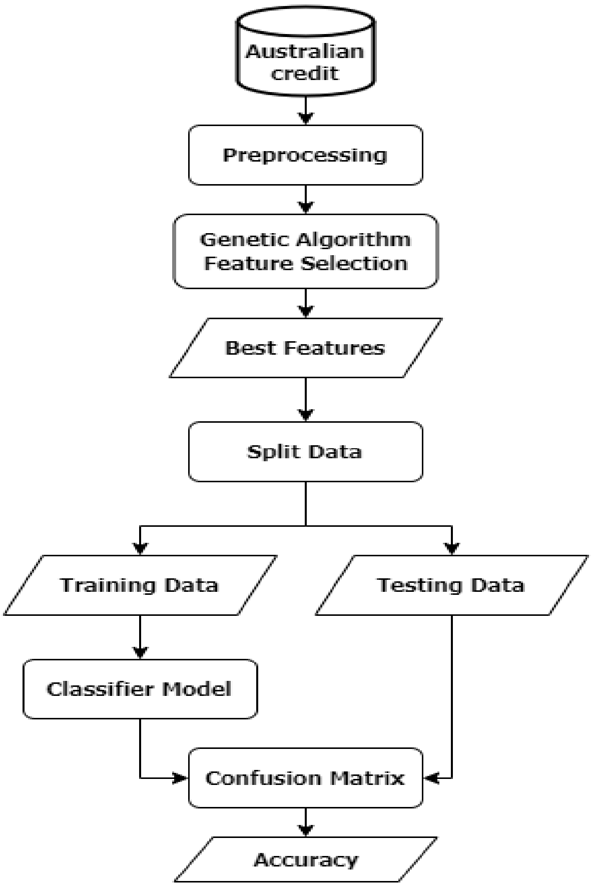

Figure 1. Stage of the Research

The Australian credit dataset is a collection of real-world credit data used in this study to test the effectiveness of the classification model. This dataset was obtained from the UCI Machine Learning Repository, which consists of 690 instances and 14 features, with 8 numeric features and 6 feature categories. Then the data preprocessing stage is an important stage in modeling which can make the data ready for the classification process in data mining [23]. This research performs normalization of data, and balancing of data classes using the SMOTE oversampling. Then, the features of categorical type are normalized using the dummy variable method, which transfers n attributes from the original features to n-1 features with only two attributes "0" and "1" [24], [25]. The normalization equation is as the following Equation (1).

$$\mathrm{x}' = \frac{\mathrm{x} - \min(\mathrm{x})}{\max(\mathrm{x}) - \min(\mathrm{x})} \tag{1}$$

Where x stands for the original feature value, x' represents the feature value after normalization, $\max(\mathrm{x})$ and $\min(\mathrm{x})$ stand for the maximum and minimum values of the original feature. After preprocessing, the dataset is divided into a training set and a testing set with a total proportion of 70:30, widely used in many studies [8], [26], [27]. After going through the data normalization process, feature selection is carried out by applying a genetic algorithm. This method represents the process that propels biological evolution by using limited and unconstrained optimization problems to solve natural selection-based optimization issues [25]. It can be implemented in selecting the best feature from a dataset, where there are five phases in the genetic algorithm, which is initial populations, fitness function, selection, crossover, and mutation [28]. From selecting features using a genetic algorithm to produce a subset of data that contains the best features to proceed to the classification modeling process, which in this study applies the XGBoost algorithm. Xgboost is one classifier that can be enabled to predict decision tree based [29], [30]. Algorithm it is possible to optimization 10 times faster compared to other GBM [15]. The last stage is the evaluation of the model to find out the accuracy value of XGBoost after being optimized using GA, the research applies a confusion matrix which can be shown in Table 1.

Table 1. Confusion matrix

		Predicted	
		Positive	**Negative**
Real	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

In accordance with **TABLE 1**, True Positive (TP) indicates that the prediction is correct if the actual values are 0. A False Negative (FN), however, denotes the possibility that the forecast value of 1 may come from the real value of 0. Additionally, the True Negative (TN) exhibits the same outcomes as the False Positive (FP), which predicts a value of 0 even when the actual value is 1. Then, the accuracy formula, can be seen in Equation (2).

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FN+FP+TN}} \tag{2}$$

3. RESULTS AND DISCUSSIONS

This study uses Australian datasets to test the XGBoost classifier model on credit risk prediction problems. The Australian credit dataset is unbalanced data with 307 good credit data and 383 default data. Thus, the SMOTE technique to deal with this data imbalance problem has been applied and produce a balanced proportion, where the amount can be seen in the graph provided in Figure 2.

.

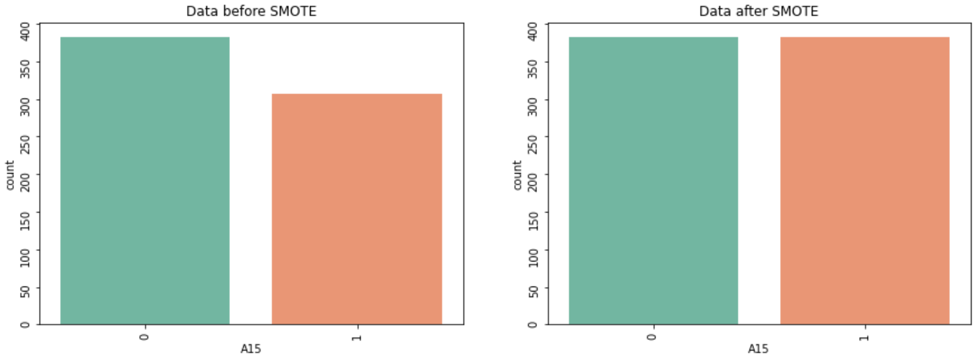

Figure 2. Results of SMOTE-oversampling from Australian credit dataset

Figure 2 shows that the number of classes in the dataset becomes balanced after SMOTE-oversampling with the number of classes 1 and 0 each totaling 383 data. Then, the data is ready to be processed in the feature selection process using the GA algorithm, in the feature selection process the result is getting the 8 best features in the Australian credit dataset from the previous number of 14 features.

Next, after producing a data subset with the 8 best features and a balanced data class, the data sample is entered into the stage distribution of training and testing data which is then processed at the modeling stage using the XGBoost classifier. In this modeling, the performance of the classification model evaluated using a confusion matrix that obtains a value, which can be seen in Table 2.

Table 2. Confusion matrix of XGBoost in Australian credit dataset

		Predicted	
		Positive	**Negative**
Real	**Positive**	225	37
	Negative	29	245

So, the accuracy calculation is as follows,

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

$$\text{Accuracy} = \frac{225+245}{225+37+29+225} = 87.69\%,$$

The results of the performance accuracy of the XGBoost classifier for predicting credit risk in the Australian dataset as shown in Table 2 obtained a result of 87.69%. Then, we will see the performance of the XGBoost classifier which is optimized using the Genetic Algorithm as a feature selection method, where the results of the confusion matrix are presented in Table 3.

Table 3. Confusion matrix of XGBoost+GAFS in Australian credit dataset

		Predicted	
		Positive	**Negative**
Real	**Positive**	228	34
	Negative	20	254

So, the accuracy calculation is as follows,

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

$$\text{Accuracy} = \frac{228+254}{228+34+20+254} = 89.93\%,$$

In Table 4, it shows the increased accuracy of the XGBoost classifier after being optimized using the Genetic Algorithm. The accuracy result is 89.93%, an increase from the accuracy value before optimization, which was only 87.69%. The resulting increase was 2.24% from the XGBoost and Genetic Algorithm feature selection trials. This shows that the selection of the best features can affect the accuracy of credit risk predictions based on the Australian credit dataset, and the diagram of comparison model performance, presented in Figure 3.

TABLE 4. Comparison of model performance *in Australian credit dataset*

Model	Accuracy
XGBoost	87.69%
XGboost+GAFS	**89.93%**

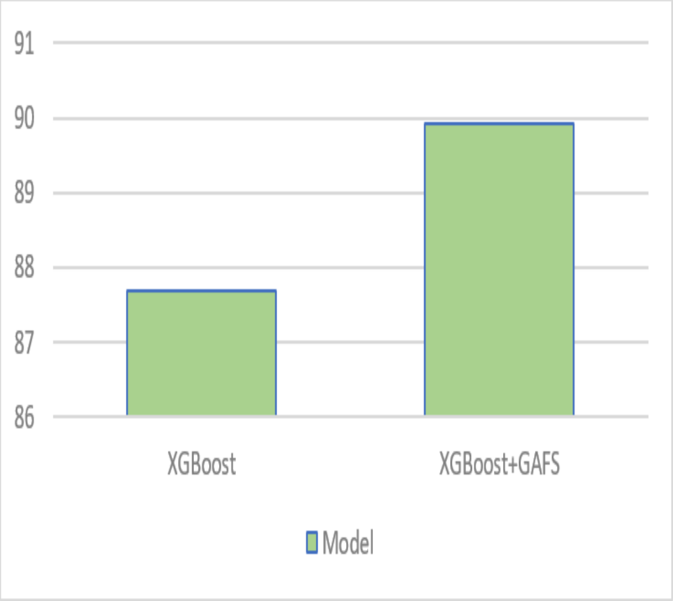

Figure 3. Results of comparison model in Australian credit dataset

4. CONCLUSION

Based on this research which aims to provide recommendations for this credit risk prediction model, where the model is classified as one of the data mining techniques for credit risk management, the probability of default has been carried out. The application of the genetic algorithm as a feature selection method was successful increase the accuracy value of the XGBoost algorithm as a classification model with an increase of 2.24% in predictions credit risk based on Australian credit dataset. As a future study, research on credit risk prediction can be continued by optimizing the Genetic Algorithm as a tuning hyperparameter, and testing other credit datasets.

REFERENCES

[1] J. Witzany and J. Witzany, *Credit risk management*. Springer, 2017.

[2] S. Claessens, J. Frost, G. Turner, and F. Zhu, “Fintech credit markets around the world: size, drivers and policy issues,” *BIS Quarterly Review September*, 2018.

[3] I. Irwansyah and Y. Ahsan, “Police Efforts In Tackling The Crime Of Theft By Children In The Jurisdiction Of The Tampan Police Sector,” *Jurnal Kajian Ilmu Hukum*, vol. 1, no. 1, pp. 40–61, 2022.

[4] S. Zulkarnain, “Penggunaan Upaya Paksa Oleh Penegak Hukum dalam Perspektif Hukum Acara Pidana Indonesia,” *Jurnal Mahkamah*, no. Oktober, 2014.

.

[5] G. N. Al-Eitan and T. O. Bani-Khalid, "Credit risk and financial performance of the Jordanian commercial banks: A panel data analysis," *Academy of Accounting and Financial Studies Journal*, vol. 23, no. 5, pp. 1–13, 2019.

[6] J. Luo, X. Yan, and Y. Tian, "Unsupervised quadratic surface support vector machine with application to credit risk assessment," *Eur J Oper Res*, vol. 280, no. 3, pp. 1008–1017, 2020.

[7] H. Hassani, X. Huang, and E. Silva, "Digitalisation and big data mining in banking," *Big Data and Cognitive Computing*, vol. 2, no. 3, p. 18, 2018.

[8] M. Soui, I. Gasmi, S. Smiti, and K. Ghédira, "Rule-based credit risk assessment model using multi-objective evolutionary algorithms," *Expert Syst Appl*, vol. 126, pp. 144–157, 2019.

[9] M. Ala'raj and M. F. Abbod, "Classifiers consensus system approach for credit scoring," *Knowl Based Syst*, vol. 104, pp. 89–105, 2016.

[10] Y. Chang, K. Chang, and G. Wu, "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions," *Applied Soft Computing Journal*, vol. 73, pp. 914–920, 2018, doi: 10.1016/j.asoc.2018.09.029.

[11] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electron Commer Res Appl*, vol. 31, pp. 24–39, Sep. 2018, doi: 10.1016/j.elerap.2018.08.002.

[12] A. Levy and R. Baha, "Credit risk assessment: a comparison of the performances of the linear discriminant analysis and the logistic regression," *International Journal of Entrepreneurship and Small Business*, vol. 42, no. 1–2, pp. 169–186, 2021.

[13] 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.

[14] X. Huang, X. Liu, and Y. Ren, "Enterprise credit risk evaluation based on neural network algorithm," *Cogn Syst Res*, vol. 52, pp. 317–324, 2018.

[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[16] A. L. Bluma and P. Langley, "Selection of relevant features and examples in machine," *Artif Intell*, vol. 97, no. 97, pp. 245–271, 1997.

[17] S. Jadhav, H. He, and K. Jenkins, "Information gain directed genetic algorithm wrapper feature selection for credit rating," *Applied Soft Computing Journal*, vol. 69, pp. 541–553, 2018, doi: 10.1016/j.asoc.2018.04.033.

[18] L. Zhuo, J. Zheng, X. Li, F. Wang, B. Ai, and J. Qian, "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine," in *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images*, SPIE, 2008, pp. 503–511.

[19] H. Chen, W. Jiang, C. Li, and R. Li, "A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm," *Math Probl Eng*, vol. 2013, pp. 1–6, 2013.

[20] C. Liu, D. Jiang, and W. Yang, "Expert Systems with Applications Global geometric similarity scheme for feature selection in fault diagnosis," *Expert Syst Appl*, vol. 41, no. 8, pp. 3585–3595, 2014, doi: 10.1016/j.eswa.2013.11.037.

[21] M. Mukherjee and M. Khushi, "SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features," *Applied System Innovation*, vol. 4, no. 1, p. 18, 2021.

[22] R. Muzayanah, A. D. Lestari, B. Prasetiyo, and D. A. A. Pertiwi, "Comparative Study of Imbalanced Data Oversampling Techniques for Peer-to-Peer Landing Loan Prediction," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 245–254, 2024, doi: 10.15294/sji.v11i1.50274.

[23] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowl Based Syst*, vol. 98, pp. 1–29, 2016.

[24] M. A. Muslim and Y. Dasril, "Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, p. 5549, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5549-5557.

[25] M. A. Muslim, Y. Dasril, H. Javed, W. F. Abror, D. A. A. Pertiwi, and T. Mustaqim, "An Ensemble Stacking Algorithm to Improve Model Accuracy in Bankruptcy Prediction," *Journal of Data Science and Intelligent Systems*, vol. 1, no. 1, 2023.

[26] N. Arora and P. D. Kaur, "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment," *Applied Soft Computing Journal*, vol. 86, p. 105936, 2020, doi: 10.1016/j.asoc.2019.105936.

[27] T. M. Alam *et al.*, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201173–201198, 2020.

[28] N. Maleki, Y. Zeinali, S. Taghi, and A. Niaki, “A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection,” *Expert Syst Appl*, vol. 164, no. September 2020, p. 113981, 2021, doi: 10.1016/j.eswa.2020.113981.

[29] Y. Jiang, G. Tong, H. Yin, and N. Xiong, “A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters,” *IEEE Access*, vol. 7, pp. 118310–118321, 2019.

[30] D. A. A. Pertiwi, T. Mustaqim, and M. A. Muslim, “Prediksi Rating Aplikasi Playstore Menggunakan Xgboost,” in *Proceedings of SNIK*, Semarang, 2020, pp. 108–112.

.