JOIN (Jurnal Online Informatika)
Volume 7 No. 2 | December 2022: 177-183
DOI: 10.15575/join.v7i2.842

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Random Forest Method to Customer Classification Based on Non-Performing Loan in Micro Business

Muhammad Muhajir[1], and Julia Widiastuti[2]

[1]Department of Statistics Universitas Islam Indonesia, Indonesia
[2]Service Operation Electronic Payment Artajasa, Indonesia

Article Info

Article history:

Received December 09, 2021
Revised May 22, 2022
Accepted October 26, 2022
Published December 26, 2022

Keywords:

Classification
Imbalanced Data
Improved random forest
Oversampling Technique

ABSTRACT

This study aims to classify potential customers' characteristics based on non-performing loans through the random forest method. This research uses data obtained from Syariah Mandiri Bank branch in Jambi, which includes data on micro-financing customers in years 2016–2020. The random forest method is used for analysis. The novelty of this work is that, unlike existing researches that used other soft-computing methods, we employ Random Forest method, specifically using an imbalanced class sampling technique. The obtained results show that credit risk can be estimated by taking into account factors such as age, monthly installments, margin, price of insurance, loan principal, occupation, and long installments. The research results indicate that the sensitivity, precision, and G-mean value increase compared to using the original data. Random forest with oversampling technique has the high Area Under the ROC Curve score that is equal to 66.69%.

Corresponding Author:

Muhammad Muhajir
Department of Statistics
Universitas Islam Indonesia, Indonesia
Email: mmuhajir@uii.ac.id

1. INTRODUCTION

The Micro, Small, and Medium Enterprises (UMKM) is a national economic landmark, because it has conquered most of the enterprises in Indonesia. The UMKM created employment opportunities in Indonesia, which reduced unemployment and advanced Indonesian economic activities by providing the 99.5 % employment opportunities [1]. The UMKM still had problems, and solving them is essential for sector growth. One issue is the lack of, or limited, access to financing for capital and business owners [2]. A loan of money from the bank is a solution for the UMKM sector. The Syariah Bank program provides financing to fund their customers, who needed and deserved it [3]. One of the Syariah Bank branches that offer financing for UMKM development is *Syariah Mandiri Bank* (BSM). Their products are called *WARUNG MIKRO*. The provision of funding based on the syariah principle, contained in UU no. 10 1998 article 8 of Republic of INdonesia, conducted based on analysis using the principle of caution against customers who are unable to settle debts or restore funding in accordance with the agreement so the risk of failure or congestion in payment can be avoided [4].

According to Samti in [5], financing risky customers can affect the Syariah Bank's performance. The risk is caused by the counterparty of customers in satisfying an obligation. Kumar and Sheshadri explained that, in order to minimize the risk, financing could be conducted by examining debtors' demographics based on credit status [6]. Using information technology which may automate business processes, so that the Syariah Bank can store daily data transactions in large quantities. In the era of big data, Syariah Bank data deals with large quantities. Humans have limited data analytical abilities, especially in terms of using vast amounts of data to produce useful information that will help in decision-making processes [7].

A non-performing loan (NPL) is a situation where customers are not able to pay part or all of their obligations to the bank. The random forest method is used to build decision trees, consisting of a root node, node internal, and leaf nodes by taking attributes based on random data [8]. This method helps Syariah Bank

p-ISSN: 2528-1682
e-ISSN: 2527-9165

determine variables that influence consumer revenue in financing micro-businesses, making lending targeted and effective while avoiding NPLs. Credit assessment is one of the main components a financial institution uses to assess credit risk, increase cash flow, manage risk, and make decisions [9]. Credit assessment is also a process of recognizing bank customers and extending credit based on certain criteria [10]. The purpose of credit assessment is to classify the credit applicant into one of two categories, good and bad applicants. Credit assessment refers to the applicant's ability to pay their obligations using characteristics such as age, type of customers, type of microfinance, loan principal, margin, long installments, monthly installments, occupation, and the price of insurance [11].

Motivated from above literatures, this study is aimed to analyze and evaluate customer credit analysis using the Random Forest method with an imbalanced class sampling technique to find the classification of characteristics of potential customers based on NPLs.

2. METHOD

In this study, the model that proposed to solve the problem class imbalance i.e. with apply a combination of approach techniques data level by resampling method and using an algorithmic approach with ensemble method based on random forest. The proposed model framework is shown in figure 1.

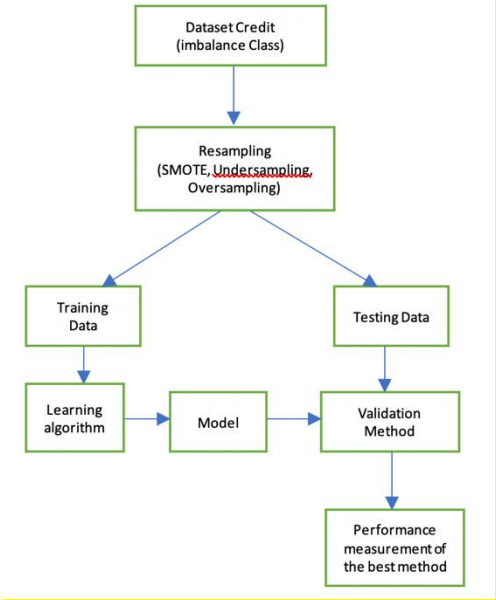

Figure 1. Architecture system of Improved Random Forest

2.1. Data Research

This research is a quantitative work, where the data used is secondary data obtained from Syariah Mandiri Bank branch of Jambi. The data include information on customers who received micro financing stalls between 2016 and 2020. In total, there are 232 customers, 55 performed good credit and 177 performed bad credit. The research variable consisted of dependent variable and independent variables. The dependent variable used into balance sufficient or insufficient information. There are nine independent variables in this research, including age, customer type, type of microfinance, loan principal, margin, long instalments, monthly instalments, occupation, and the price of insurance.

2.2. Random forest

Random Forest is one of an ensemble method for improving data classification accuracy of an unstable single separator through a combination of multiple segments, similar to the voting process, to obtain a final classification prediction [12], [13].

This is the random forest algorithm, according to Breiman & Cutler, and Liaw & Wiener [12], [14]:

1. Determining the data train (training) and data test (test) where the training data used to train the performance of methods in machine learning.
2. Carries out a random sample of size n with recovery in the data cluster where this stage is a bootstrap stage.

3. Determine the Mtry or explanatory variables and Ntree or the best number of trees. There are three ways to get the value of m to observe the OOB error is.

$$m = \frac{1}{2}\left|\sqrt{p}\right| \quad (1)$$

$$m = \left|\sqrt{p}\right| \quad (2)$$

$$m = 2 \times \left|\sqrt{p}\right| \quad (3)$$

where: p = variables total.

The proper use of m will result in a random forest with fairly small tree-to-tree correlations but each tree's strength is large, as indicated by the acquisition of a small value OOB error [10].

4. Predict the classification of sample data based on the classification tree formed.
5. Repeat steps 1 through step 4 to form a forest consisting of k trees.

2.3. Accuracy size of classification

The measures of classification performance used a confusion matrix, which provides the decisions obtained in training and testing. Table 1 below shows the confusion matrix [15]**.**

Table 1. Heading and text fonts

Observation	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	TP (True Positive)	FN (False Negative)
Actual Negative Class	FP (False Positive)	TN (True Negative)

As intended:

TP (True Positive) : The number of positive observations with right predicted.
TN (True Negative) : The number of precise negative observations predicted.
FP (False Positive) : The number of wrong positive observations predicted as negative.
FN (False Negative) : The number of wrong negative observations predicted as positive.

A False Positive is known as a type 1 error, occurring when the case should be classified as a negative classified positive while false negative is known as a type 2 error occurs if the case that should be classified as positive is classified negatively [16].

Minority class accuracy could be the true positive rate or recall (sensitivity) matrix. G-mean and AUC are more comprehensive predictor evaluations in the context of imbalance. By using matrices such as true negative rate (specificity), true positive rate (sensitivity), APER, total accuracy rate (1APER), G-mean, precision, and F-measure to be able to evaluate machine learning performance at the time of imbalanced data, such as that shown in Table 1. The formula is as follows [17]–[19]:

True Negative Rate (Acc^-) or specificity $= \frac{TN}{TN + FP}$ (4)

True Positive Rate (Acc+) or sensitivity / Recall $= \frac{TP}{FN + TP}$ (5)

Precision or PPV $= \frac{TP}{TP + FP}$ (6)

APER $= \frac{FP + FN}{N}$ (7)

1-APER $= \frac{TP + TN}{N}$ (8)

G-mean $= \sqrt{\text{specificity} \times \text{sensitivity}}$ (9)

F-measure $= \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$ (10)

False Positive Rate $= \frac{FP}{FP + TN}$ (11)

Area Under the ROC Curve $= \frac{1 + \text{Sensitivity} - \text{FPrate}}{2}$ (12)

Accuracy $= \frac{TP + TN}{TP + TN + FP + FN}$ (13)

p-ISSN: 2528-1682
e-ISSN: 2527-9165

AUC is a good way to get the performance value of the classifier in general and to compare it with other classifications.

2.4. Sampling technique class imbalance

One of the most popular methods for the problem of class imbalance is the sampling technique. The imbalanced dataset can cause problems, such as an accuracy paradox where predictive models with certain accuracy levels may have greater predictability than models with higher levels of accuracy, where better classification accuracy measures are used, such as precision and recall [20]. One of the sampling techniques for overcoming the imbalance class in machine learning is oversampling, which involves the resampling of minority classes randomly to the number of samples of other classes or the majority class [21].

2.5. Measure of importance variable

One of the measures of importance variable independent on random forest is Mean Decrease Gini (MDG). Suppose, contained p variable independent with h = 1, 2…p. Therefore, MDG measure of independent variables Xh in ways [22]:

$$MDG_h = \frac{1}{k}\sum_t [d(h,t)I(h,t)] \quad (14)$$

Where:

d(h, t) : Decrease of Gini Index of Independent variables Xh

I(h, t) : node of t

k : number of trees on random forest

3. RESULTS AND DISCUSSION

3.1. Classification of random forest

The first step in the classification methods is to determine training data and testing data. Training data is needed on machine learning algorithms, whereas testing data used to train performance on the machine learning. The data of 232 customers were divided into 55 with good credit and 177 with bad credit. Researchers used training data of 70% and testing data of 30%. The distribution of training data and testing data are shown in Table 2.

Table 2. Training Data and Testing Data

Data Class	Training Data	Testing Data	Total
Good Credit	44	11	55
Bad Credit	125	52	177
Total	169	63	232

The next step used experiments without resampling methods. To get an optimal parameter, we tried seven input parameters of Ntree (25, 50, 100, 250, 300, 500, dan 1000). After running the process with seven parameters, we obtained an optimal parameter in total of 250 trees, and OOB estimate of error is 25.44% when Mtry value of 3. The results of testing data shown in Table 3.

Table 3. Measure of accuracy random forest

Evaluation Measure	Random Forest
Specificity	100%
Sensitivity	0%
Precision	0%
APER	17.46%
1-APER	82.54%
G-mean	0%
AUC	50%

According to Table 3, the random forest obtained specificity or proportion measurement of true negative (bad credit of customers) was 100% accurate, whereas sensitivity or proportion measurement of true positive (good credit of customers) identified 0% correctly. 1-APER calculation on the random forest method is worth 82.54%, which means that 82.54% of the sampled data is classified appropriately. However, since the data used is an imbalanced dataset, then other classification accuracies should be noted as well. G-mean is

useful for determining the balance of prediction accuracy to measure an imbalanced problem, where the G-mean obtained when using the random forest method is zero, which denotes a predictive disproportion [23]. G-mean is necessary because the classification method tends to be good at predicting classes with more but bad sample data in predicting classes with few sample data. AUC is a good way to get general classifier performance values, where AUC is a popular performance measure in class imbalances [24]. The following confusion matrix a method of random forest can be seen in Table 4.

Table 4. Confusion matrix of random forest

Data		Prediction	
		Good Credit	Bad Credit
Actual	Good Credit	0	11
	Bad Credit	0	52

The accuracy result showed that the random forest method is good for classifying the bad credit class, but not the good credit class. This indicates a biased result, called the accuracy paradox. In this case, then it is better to use precision classification sizes other than accuracies, such as precision and sensitivity. To overcome imbalance prediction data, researchers use a random forest with the oversampling technique

3.2. Improved random forest

Experiments conducted by applying the random oversampling technique to the random forest. Where data training into the balance between class "good credit" and class "bad credit" to the number of instances in each class being 125 instances. The optimal parameter is obtained in 250 trees, and the OOB estimate of error is 14.8%. The test results are shown in Table 5.

Table 5. Measure of accuracy between random forest and Oversampling Technique Method

Evaluation Measure	Random Forest	Oversampling
Specificity	100%	78.85%
Sensitivity	0%	54.55%
Precision	0%	35.29%
APER	17.46%	25.39%
1-APER	82.54%	74.60%
G-mean	0%	65.58%
AUC	50%	66.69%

According to Table 5, the 1-APER value (total accuracy rate) in oversampling techniques are less accurate than the random forest method without the imbalanced data sampling technique. Although the accuracy value of the random forest method is highest, for the classification of the true positive class, there is no precision. The oversampling sensitivity value is higher than the random forest method.

Of the oversampling techniques, the imbalanced data class provides better precision and sensitivity values than using the random forest method without the imbalanced data sampling technique. The precision value with the oversampling technique is higher than the random forest method, which is equal to 35.29%. Furthermore, the AUC value of the oversampling technique is higher than the random forest method, which is equal to 66.69%. The following confusion matrix a method of improved random forest, can be seen in Table 6.

Table 6. Confusion matrix of Improved random forest

Data		Prediction	
		Good Credit	Bad Credit
Actual	Good Credit	6	5
	Bad Credit	11	41

Based on the AUC, researchers will use a method of improved random forest with an oversampling technique to see the importance of the independent variable. The next step determined the measure of importance from each independent variable. Following the result of MDG:

p-ISSN: 2528-1682
e-ISSN: 2527-9165

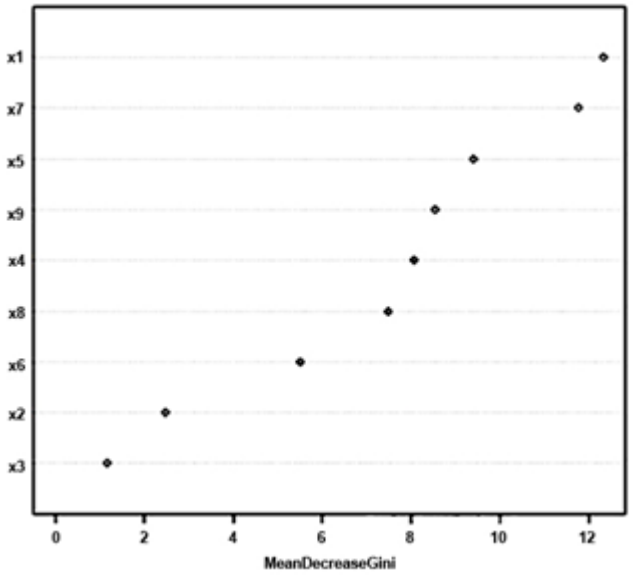

Figure 2. Measure of Importance

It can be seen that, in Fig. 2, the independent variable has a measure of importance MDG In a sequence that is X1 (age), X7 (monthly installments), X5 (margin), X9 (price of insurance), X4 (principal of loan), X8 (occupation), X6 (long installments), X2 (Type of customers), and X3 (Type of microfinance). It can conclude that variables X2 and X3 are not important for determining credit risk because the variables' importance values are below the Mtry value with the smallest OOB error.

On the financing Warung Mikro of Syariah Mandiri Bank branches Jambi, it can be seen that customers over 41 years old had a lower risk of bad credit because they had become established at work with steady incomes. On monthly instalments, the variable is known that customers with less than Rp 2.500.000 tend to pay in steady in instalments every month. It is closely related to the principal of the loan, where less than Rp 50.000.000 credit risk was less than a higher loan value. For variables, the margin lowest of 16% and the highest equal to 36%. For the majority of customers who were classified as having good credit, that is less than 22.05% with a margin. Most customers have long instalments of 36 months, it is very dependent on monthly instalment where the longer instalments, then the smaller the payment, which minimize the risk of NPLs. The highest amount of NPLs were among customers with a job employee, and price of insurance less than Rp 100.000.000, indicating that a potentially large customer employees having stuck credit, and customers who have high-value collateral are more likely to repay their loans on time.

3.2. Comparison Test with other methods

The researcher tested what method was suitable for the data by compare the methods used, namely random forest, random forest using oversampling, random forest using undersampling, and random forest using SMOTE. The test results are shown in Table 7.

Table 7. Measure of accuracy between random forest, Oversampling, Undersampling, and SMOTE technique Method

Evaluation Measure	Random Forest	Oversampling	Undersampling	SMOTE
Specificity	100%	78.85%	63.46%	69.23%
Sensitivity	0%	54.55%	45.45%	45.46%
Precision	0%	35.29%	20.83%	23.81%
APER	17.46%	25.39%	39.68%	34.92%
1-APER	82.54%	74.60%	60.32%	65.08%
G-mean	0%	65.58%	54.08%	56.09%
AUC	50%	66.69%	54.46%	57.34%

According to Table 7, oversampling technique gives the largest value of sensitivity, precision, and G-mean, namely the value of respectively 54.55%, 35.39%, and 65.58%. In addition, the highest AUC value is obtained at random oversampling technique forest of 66.69%, this shows that the performance to compare popular classifications in class imbalance.

4. CONCLUSION

This research uses the random forest method to classify potential customers using the oversampling method. Based on research, the improved random forest with oversampling is better than random forest, random forest using undersampling, and random forest using SMOTE. It can be seen from the confusion matrix and AUC score that is equal to 66.69%, describing the method's performance. Moreover, credit risk can be done by taking certain factors into accounts, such as age, monthly installments, margin, price of insurance, loan principal, occupation, and long installments.

ACKNOWLEDGEMENTS

This paper is our original work and has not been published or submitted simultaneously elsewhere. All authors have agreed to the submission and declared that they have no conflict of interest. This paper was supported in part by the department of development academic Islamic University of Indonesia.

5. REFERENCES

[1] Bank Indonesia, "Profil Bisnis Usaha Mikro, Kecil, dan Menengah," 2015. www.bi.go.id.
[2] Geev, "Mengenal Apa Itu UMKM dan Perkembangannya di Indonesia," 2017. .
[3] Z. Arifin, *Dasar-dasar Manajemen Bank Syari'ah*. Jakarta: Alfabeta, 2002.
[4] Bank Indonesia, *Undang-Undang Nomor 10 Tahun 1998 tentang Perubahan Undang-Undang No. 7 Tahun 1992 tentang Perbankan*. Jakarta: Gramedia, 1998.
[5] Y. H. Fahmi, I and Lavianti, *Pengantar Manajemen Perkreditan*. Bandung: Bandung, 2010.
[6] A. KumarM.N and H. S. Sheshadri, "On the Classification of Imbalanced Datasets," *Int. J. Comput. Appl.*, vol. 44, no. 8, 2012, doi: 10.5120/6280-8449.
[7] P. Trkman, K. McCormack, M. P. V. De Oliveira, and M. B. Ladeira, "The impact of business analytics on supply chain performance," *Decis. Support Syst.*, vol. 49, no. 3, 2010, doi: 10.1016/j.dss.2010.03.007.
[8] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[9] L. Lin, F. Wang, X. Xie, and S. Zhong, "Random forests-based extreme learning machine ensemble for multi-regime time series prediction," *Expert Syst. Appl.*, vol. 83, pp. 164–176, Oct. 2017, doi: 10.1016/j.eswa.2017.04.013.
[10] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *J. Retail. Consum. Serv.*, vol. 27, 2015, doi: 10.1016/j.jretconser.2015.07.003.
[11] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: Adaption of different imbalance ratios," *Expert Syst. Appl.*, vol. 98, 2018, doi: 10.1016/j.eswa.2018.01.012.
[12] L. Breiman, "Manual on setting up, using, and understanding random forests v3. 1," *Tech. Report, http//oz.berkeley.edu/users/breiman, Stat. Dep. Univ. Calif. Berkeley, …*, 2002.
[13] P. Singh, S. and Gupta, "Comparative study ID3, cart and C4 . 5 Decision tree algorithm: a survey," *Int. J. Adv. Inf. Sci. Technol.*, vol. 27, no. 27, pp. 97–103, 2014.
[14] A. Liaw and M. Wiener, "Classification and Regression with Random Forest," *R News*, vol. 2, 2002.
[15] D. Ramyachitra and P. Manikandan, "Imbalanced Dataset Classification and Solutions: a Review," *Int. J. Comput. Bus. Res. ISSN (Online*, vol. 5, no. 4, 2014.
[16] K. Santra and C. J. Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering," *Int. J. Comput. Sci.*, vol. 9, no. 1, 2012.
[17] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, 2013.
[18] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, 2015, doi: 10.5121/ijdkp.2015.5201.
[19] J. M. Johnson and T. M. Khoshgoftaar, "Deep learning and data sampling with imbalanced big data," 2019, doi: 10.1109/IRI.2019.00038.
[20] M. Bramer, *Principles of data mining fourth edition*, vol. 30, no. 7. 2020.
[21] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 3, 2015.
[22] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in Forests of randomized trees," 2013.
[23] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Trans. Reliab.*, vol. 62, no. 2, 2013, doi: 10.1109/TR.2013.2259203.
[24] X. Y. Liu and Z. H. Zhou, "Ensemble methods for class imbalance learning," in *Imbalanced Learning: Foundations, Algorithms, and Applications*, 2013.