Faktor Exacta

Vol. 17, No. 2, June 2024, pp. 161~166
eISSN: 2502-339X, pISSN: 1979-276X, DOI: 10.30998/faktorexacta.v17i2.21819

Application of Ensemble Tree Algorithm for Installment Payment Arrears Prediction at Makmur Bersama Credit Union

Risanto Darmawan[1], Diajeng Reztrianti[2], Ali Khumaidi[3]
[1,3] Informatics Departement, Faculty of Engineering, Universitas Krisnadwipayana, Indonesia
[2] Management Departement, Faculty of Economic, Universitas Krisnadwipayana, Indonesia

Article Info

Article history:

Received Dec 29, 2023
Revised Jul 15, 2024
Accepted Jul 16, 2024

Keywords:

AdaBoost
Credit Union
Delinquency
Ensemble Tree
Random Forest

ABSTRACT

Some customers in paying installments at the Makmur Bersama Credit Union are not smooth, with a machine learning approach that will predict when customers will be in arrears so that the cooperative can anticipate collection patterns. One of the machine learning techniques used is the ensemble tree method, which is the merger of several classification trees where the final decision is based on the combined predictions of each tree. This approach produces a better accuracy rate than a single classification tree. Two common methods used in the ensemble tree technique are boosting and bagging. This research will predict the status of installment payments at Makmur Bersama Credit Union. The bagging method used is random forest and the boosting method is AdaBoost. To get optimal results, hyperparameter tuning is also carried out. The results showed that the performance of the ensemble method in boosting and bagging was able to handle the classification of cooperative loan installment payment status better than the distance approach, namely kNN (single classification). The performance of the boosting combined tree with the AdaBoost model has an accuracy of 72.89% better than the bagging combined tree with the random forest model whose accuracy is 72.08%.

Corresponding Author:

Ali Khumaidi,
Department of Informatics,
Universitas Krisnadwipayana,
Jl. Kampus Unkris, Jatiwaringin, Pondokgede, Kota Bekasi.
Email: alikhumaidi@unkris.ac.id

1. INTRODUCTION

The Industrial Revolution 4.0 has brought major changes in various fields, including data and intelligent systems. Big data supports an unlimited amount, diverse data forms and high speed of change. This causes classical analysis to be less able to classify big data well [1]. Machine learning techniques are a solution to big data analysis to assist in model construction and inference is done automatically. This technique produces predictive models with excellent accuracy [2]. In addition, machine learning is also able to capture non-linear patterns so that it can provide additional information that classical linear model approaches generally fail to capture [3] and produce more satisfactory predictions.

Machine learning is used in developing models that automatically adapt to identify complex and hidden patterns in data, so as to help decision makers to estimate the impact of several plausible scenarios in real time. One of the techniques is supervised learning [4]. A widely used supervised learning method is the decision tree method, because it is easy to understand where the way of determining decisions in the tree method is similar to the way humans think [5]. The advantages of decision trees over other learning algorithms include noise resistance, low computational cost to generate models, and the ability to handle redundant variables [6]. One of the widely used machine learning techniques is the ensemble tree method, which is the combination of several classification trees where the final decision is based on the combined predictions of each tree [7]. This approach produces better accuracy than a single classification tree [8]. Two common ways to do the ensemble tree technique are boosting and bagging. The difference between the two models is the way

the tree is formed. Tree formation in boosting is done sequentially [9] such as the adaptive boosting method (AdaBoost), while tree formation in bagging is done in parallel such as the random forest method.

Big data analysis with the help of machine learning can improve services and solve problems in various sectors, one of which is the savings and loan cooperative sector. Makmur Bersama Credit Union as a provider of savings and loan business for small and medium scale entrepreneurs or traders has grown quite rapidly. Savings and loan cooperative arrears are a frequent problem and can cause problems for cooperatives, because they disrupt cash flow and affect the ability of cooperatives. It can also lead to the cooperative having to take steps to collect arrears, which can be time-consuming and costly. Cooperatives should have a program or policy to deal with arrears cases quickly and effectively to minimize the impact on the cooperative and other members. Several factors can influence the likelihood of arrears in savings and loan cooperatives, such as: the economic situation; the cooperative's lending policy; and the cooperative's service and management.

Research related to predicting the risk of arrears of savings and loan cooperatives loans given to small community groups including machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest and Gradient Boosting are used to train prediction models and the Gradient Boosting algorithm shows the best performance with an accuracy of 77% [10]. Research to evaluate the performance of Support Vector Machine in predicting the risk of arrears and providing credit ratings in savings and loan cooperatives [11]. The results showed that SVM can be used to predict the risk of arrears and provide credit ratings in savings and loan cooperatives with a fairly high accuracy rate of 86%. Analysis of credit arrears using kNN [12]. Research related to predicting participants in arrears of payments has been conducted on 1,079 customers of the Thailand Provincial Electricity Authority (PEA) in 2020 [13]. In this research, there are two stages, namely customer class grouping using the k-means algorithm and customer prediction; using five machine learning models, namely logistic regression, decision tree, random forests, support vector machine (SVM) and extreme gradient boosted (XGBoost). The results of his research show that random forest is the best model with an F1-Score of 98% in predicting the class of customers who are delinquent in electricity payments. Some recent research using the AdaBoost and random forest methods can be found in [7, 14]. The results of research with these two methods have been carried out in various cases so it is interesting to apply them to other cases, especially in the case of Savings and Loan Credit Union.

Research related to the prediction of participants who are in arrears of payment has never been done using AdaBoost and random forest. Therefore, it is interesting to analyze big data with the right machine learning method using the ensemble tree method, namely AdaBoost and random forest. The general objective of the research is to study patterns and predict cooperative members who will default on dues payments. The results of predicting cooperative members who will default are expected to provide additional insight to cooperative leaders in preventing participants from defaulting. The specific objective is to examine the difference in the level of goodness of the model in the tree method by bagging (random forest) and by boosting (AdaBoost) and single classification (kNN) to get the right and best model in predicting delinquent cooperative members.

2. METHOD

This research uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) method. The CRISP-DM method has 6 stages as shown in Figure 1 in the overall process, namely (1) Business Understanding (Collecting data regarding Business objectives, assessing current conditions, setting the objectives of the data mining process. (2) Data Understanding (Collecting initial data, data description, data exploration, and assessing data quality are stages in this phase). (3) Data Preparation (After the data is obtained, it is necessary to process a selection process, cleansing, made in a certain form, and formatted as needed). (4) Modeling (After the data is cleaned and formed as needed, an appropriate modeling is needed and calibrated regarding the settings to get optimal results). (5) Evaluation (After obtaining the model, an assessment is carried out regarding performance) and (6) Deployment (In this phase, in general, there are 2 activities carried out, namely planning and monitoring the results of the deployment process and completing all activities so as to produce a final report and conduct a review of the project carried out) [15]. Stages 1 to 5 are done in this research.

Business Understanding

At this stage, the goal is to predict the delinquent status of the cooperative customers. The current condition is that almost 31% of customers often experience disobedience in paying installments so the ability to predict the chances of installment payment arrears is needed.

Data Understanding

The data used is a type of primary data from the Makmur Bersama CU Credit Union, Bekasi City, Indonesia consisting of loan and installment data, with a data period of January 2018 to September 2023

totaling 69 loan reports. The response variable or target used is the status of installment payments which consists of 2 classes, namely paid and delinquent.

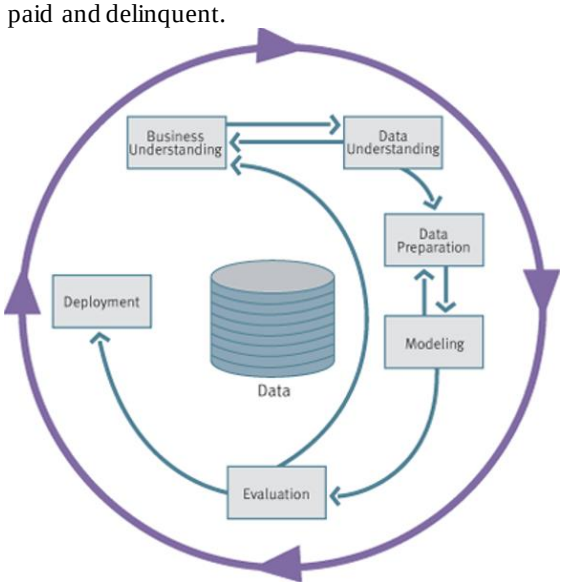

Figure 1. CRISP-DM stages

Data Preparation

Monthly loan and installment reports include 9 attributes as in Table 1. Then the selection of attributes that affect the pay status is carried out, namely name, product, and report date. Adding a status label based on the loan balance, if on the report date the balance decreases from the previous month then the status is paid. If there is no change in the loan balance then the status is in arrears. Table 2 is the attributes that will be used in modeling.

Table 1. Loan and installment report attributes

No	Member ID	Account ID	Name	Product	Borrowing Date	Last Transaction	Loan Balance	Report date

Table 2. Attributes used in modeling

Report date	Name	Product	Status

Modeling

This research applies ensemble tree techniques, namely boosting and bagging. Boosting is done using the adaptive boosting algorithm (AdaBoost) and bagging uses Random Forest. To test both ensemble techniques, the results were compared using the k-Nearest Neighbor algorithm. To obtain optimal performance, hyperparameter tuning is performed on each algorithm. With parameter settings on the three algorithms as in Table 3.

Table 3. Parameter setting

Algoritm	Parameter	Value
Random Forest	Number of tress	1 to 100
	Criterion	Gain ratio, Information gain, Gini index
	Maximal Depth	-1 to 10
	Pruning	True, False
AdaBoost	Iterations	1 to 100

Evaluation

Classification accuracy represents the percentage of correctly classified samples out of all samples, and it is used to estimate model performance [16]. When the number of positive and negative samples is different, the confusion matrix can be used to evaluate different classification models, using information about the correct classification and predicted categories [17]. True positives (TP) are samples that are positive and classified with. True negatives (TN) are samples that are negative and correctly assigned as negative. False positives (FP) are negative samples that are misclassified. False negatives (FN) are negative samples that are incorrectly assigned to other negative categories. In the confusion matrix model for the classification of

installment payment arrears status, the prediction performance used is classification accuracy, precision, and recall.

3. RESULT AND DISCUSSION

Data processing in this study uses the open-source data science tool Rapidminer Studio 10.1 to perform data analysis. The results at the data preparation stage produced a dataset of 3,024 data, with a paid status of 2106 data and a arrears status of 918 data. All data has no missing values. The data type for report date is date, name is polynomial, product is polynomial, and status is polynomial. Figure 2 is a sample view of the data preparation results. In this data processing, the division of training and testing data with a composition of 80% and 20%. To calculate the performance of training data using cross validation with the number of folds = 10.

Status	Tgl	Nasabah	Produk
Bayar	Jan 1, 2019	Afrizal	Pinjaman Umum
Bayar	Feb 1, 2019	Afrizal	Pinjaman Umum
Bayar	Mar 1, 2019	Afrizal	Pinjaman Umum
Bayar	Apr 1, 2019	Afrizal	Pinjaman Umum
Bayar	May 1, 2019	Afrizal	Pinjaman Umum
Bayar	Jun 1, 2019	Afrizal	Pinjaman Umum
Bayar	Jul 1, 2019	Afrizal	Pinjaman Umum
Bayar	Aug 1, 2019	Afrizal	Pinjaman Umum
Bayar	Sep 1, 2019	Afrizal	Pinjaman Umum
Bayar	Oct 1, 2019	Afrizal	Pinjaman Umum

Figure 2. Display of sample data preparation

3.1. Modeling Results without Hyperparameter Tuning

Modeling results without tuning parameters with Random Forest (Number of trees= 100, criterion= gain ratio, maximal depth=10), AdaBoost (Number of trees= 100, criterion= gain ratio, maximal depth=10, Iterasions= 10), and kNN (K=5). The results can be seen in Table 4.

Table 4. Algorithm performance without hyperparameter tuning

Algoritm	Training			Testing			Setting parameter
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	
RF	69.70	64,07	63,95	70.58	65.21	65,24	Number of tree= 100, criterion= gain ratio, maximal depth=10
Adaboost	69.70	64,07	63,95	70.58	65.21	65,24	Number of tree= 100, criterion= gain ratio, maximal depth=10, Iterasions= 10
KNN	63.99	52.14	51.47	63.14	49.7	49.8	K=5

Table 4 provides the results that with the parameters that have been set by default, it gives the same performance to Random Forest and AdaBoost, which has a testing accuracy of 70.58%. The kNN performance with an accuracy of 63.14% has a lower performance than Random Forest and AdaBoost. This proves that the ensemble technique has better performance.

3.2. Modeling Results Using Hyperparameter Tuning

The modeling results provide the best performance in each algorithm according to the parameter settings that have been determined in Table 3. The best parameters in the Random Forest algorithm are Number of trees = 55, criterion = gain ratio, maximum depth = 7. The best parameters in the AdaBoost algorithm are Number of trees = 48, criterion = gain ratio, maximum depth = 6 and Iterations = 12. The kNN has the best performance with a value of k = 90.

Table 5. Algorithm performance using hyperparameter tuning

Algoritm	Training			Testing			Setting parameter
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	
RF	72.6	67.68	66.34	72.07	66.48	65.09	Number of tree= 55, criterion= gain rastio, maximal depth=7

Adaboost	72.7	67.46	65.98	72.89	67.47	65.53	Number of tree= 48, criterion= gain ratio, maximal depth=6 Iterations=12
KNN	70.70	65.24	63.78	69,09	56,79	51,17	K=90

Table 5 shows that the performance of the AdaBoost algorithm with a test accuracy of 72.89% is better than the Random Forest algorithm which is 72.07%. During the hyperparameter tuning process all three algorithms experienced an increase in accuracy. The kNN algorithm with a testing accuracy of 69.09% has increased but its performance is still below the Random Forest and AdaBoost algorithms. From the comparison results using hyperparameter tuning, kNN performance is not better than the performance of the Random Forest and AdaBoost algorithms without hyperparameter tuning. This proves that the Random Forest and AdaBoost algorithms have good performance on the classification of arrears in cooperative installment payments. Overall, the AdaBoost algorithm has the best performance compared to the Random Forest and kNN algorithms.

4. CONCLUSION

This research shows that the boosting and bagging ensemble tree methods are able to handle the classification of cooperative loan installment payment status better than the distance approach, namely kNN (single classification). The results of this study boosting ensemble tree with AdaBoost model has an accuracy of 72.89% better than bagging ensemble tree with random forest model whose accuracy is 72.08%. In the modeling process, AdaBoost builds each tree independently using random data samples, and this randomization makes the model more resistant and reduces overfitting of the training data. The drawback of the AdaBoost model is the large number of trees built, which results in a long processing time.

ACKNOWLEDGEMENT

The author would like to thank the Directorate of Research, Technology and Community Service (DRTPM) of the Ministry of Education and Culture of the Republic of Indonesia for funding research in 2023.

REFERENCES

[1] S. Suthaharan, Machine Learning Models and Algorithms for Big Data Classification, vol. 36. Boston, MA: Springer US, 2016.
[2] J. Boelaert and É. Ollion, "The Great Regression," Rev. française Sociol., vol. Vol. 59, no. 3, pp. 475–506, Sep. 2018, doi: 10.3917/rfs.593.0475.
[3] S. Mullainathan and J. Spiess, "Machine Learning: An Applied Econometric Approach," J. Econ. Perspect., vol. 31, no. 2, pp. 87–106, May 2017, doi: 10.1257/jep.31.2.87.
[4] L. Wu et al., "Robust fall detection in video surveillance based on weakly supervised learning," Neural Networks, Apr. 2023, doi: 10.1016/j.neunet.2023.03.042.
[5] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," Int. J. Comput. Sci. Eng., vol. 6, no. 10, pp. 74–78, Oct. 2018, doi: 10.26438/ijcse/v6i10.7478.
[6] L. Rokach and O. Maimon, "Top-Down Induction of Decision Trees Classifiers—A Survey," IEEE Trans. Syst. Man Cybern. Part C (Applications Rev., vol. 35, no. 4, pp. 476–487, Nov. 2005, doi: 10.1109/TSMCC.2004.843247.
[7] R. Yakkundimath, V. Jadhav, B. Anami, and N. Malvade, "Co-occurrence histogram based ensemble of classifiers for classification of cervical cancer cells," J. Electron. Sci. Technol., vol. 20, no. 3, p. 100170, Sep. 2022, doi: 10.1016/j.jnlest.2022.100170.
[8] W. Liu, H. Fan, and M. Xia, "Tree-based heterogeneous cascade ensemble model for credit scoring," Int. J. Forecast., Aug. 2022, doi: 10.1016/j.ijforecast.2022.07.007.
[9] T. Mildenberger, "Stephen Marsland: Machine learning. An algorithmic perspective," Stat. Pap., vol. 55, no. 2, pp. 575–576, May 2014, doi: 10.1007/s00362-012-0471-0.
[10] K. D. Pradnyana and R. A. Rahadi, "Loan Default Prediction in Microfinance Group Lending with Machine Learning," Int. J. Bus. Technol. Manag., Jan. 2023, doi: 10.55057/ijbtm.2022.4.4.8.
[11] K. Amzile and M. Habachi, "Assessment of Support Vector Machine performance for default prediction and credit rating," Banks Bank Syst., vol. 17, no. 1, pp. 161–175, Apr. 2022, doi: 10.21511/bbs.17(1).2022.14.
[12] A. Setianingrum, A. Hindayanti, D. M. Cahya, and D. S. Purnia, "Perbandingan Metode Algoritma K-NN & Metode Algoritma C4.5 Pada Analisa Kredit Macet (Studi Kasus PT Tungmung Textile Bintan)," Evolusi J. Sains dan Manaj., vol. 9, no. 2, pp. 78–92, 2021.
[13] P. Khansong, J. Karnjana, S. Laitrakun, and S. Usanavasin, "Customer Service Improvement based on Electricity Payment Behaviors Analysis using Data Mining Approaches," in 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical,

Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), 2020, pp. 114–117, doi: 10.1109/ECTIDAMTNCON48261.2020.9090699.

[14] R. Natras, B. Soja, and M. Schmidt, "Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting," Remote Sens., vol. 14, no. 15, p. 3547, Jul. 2022, doi: 10.3390/rs14153547.

[15] A. Khumaidi, "Data Mining For Predicting The Amount Of Coffee Production Using CRISP-DM Method," J. Techno Nusa Mandiri, vol. 17, no. 1, pp. 1–8, Feb. 2020, doi: 10.33480/techno.v17i1.1240.

[16] Q. Qiu et al., "Development and validation of three machine-learning models for predicting multiple organ failure in moderately severe and severe acute pancreatitis," BMC Gastroenterol., vol. 19, no. 1, p. 118, Dec. 2019, doi: 10.1186/s12876-019-1016-y.

[17] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Inf. Sci. (Ny)., vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.