Jurnal Informasi dan Teknologi

https://jidt.org/jidt

2023 Vol. 5 No. 3 Hal: 26-30 e-ISSN: 2714-9730

Implementation of Gradient Boosted Tree, Support Vector Machinery and Random Forest Algorithm to Detecting Financial Fraud in Credit Card Transactions

Ferdinand Salomo Leuwol[1]✉, Asri Ady Bakri[2], Muhsin N. Bailusy[3], Hari Setia Putra[4], Ni Ketut Sukanti[5]

[1]Universitas Pattimura
[2]Universitas Muslim Indonesia
[3]Universitas Khairun
[4]Universitas Negeri Padang
[5]Universitas Ngurah Rai

eddieleuwol0@gmail.com

Abstract

According to Google Trends data, machine learning-based credit card identification has grown over the last five years, at the very least, across all nations. In order to detect credit card fraud in this study, the authors will use machine learning methods such random forests, support vector machines, and gradient-boosted trees. The authors used the Synthetic Minority Oversampling Technique (SMOTE) and Random Under Sampling (RUS) sampling methods in each algorithm to compare because there was a class imbalance in this investigation. The research findings demonstrate that the author's algorithm and sample technique were successfully used, as shown by the AUC values obtained for each being > 0.7. The top score in RUS was 0.7835 using the Random Forest algorithm, whereas the greatest score in SMOTE was 0.73 with the Gradient Boosted Trees approach. The Random Forest algorithm and the Random Under Sampling (RUS) technique are developed as a result of this research, and they are useful for identifying fraudulent credit card transactions.

Keywords: Credit Card, Machine Learning, Random Forest, Vector Machine, Gradient-Boosted Tree.

1. Introduction

Based on data from the Indonesian Credit Card Association (AKKI), the use of credit cards has increased. In 2015, the number of users was 16,863,842, and in 2020, it increased to 16,940,040. The growth of credit card users in Indonesia is in line with the level of crime that occurs in banks, especially transaction fraud. Based on the statistics reported, countries with the highest risk of fraud show that Indonesia is in second place with the highest fraud rate of 18.3%. This is a type of crime in the field of information technology that is very disturbing to credit card users around the world, including in Indonesia. The unlawful use of a credit card to get something of value with the purpose to defraud is known as credit card fraud. The types of fraud themselves can be divided into several types, starting with the most well-known, namely fraud in credit card transactions, money laundering, fraud in calling or using telecommunications services, and insurance system fraud. Banking fraud is often found in transactions and accounting practices [1].

The Financial Services Authority issued regulation number 39/POJK.03/2019 regarding the implementation of an anti-fraud strategy for commercial banks. This strategy has at least four pillars, including prevention, detection, investigation (with reporting and sanctions), and evaluation (monitoring and follow-up). According to evidence from earlier research findings, fraud in transactions explains why some banks in Indonesia have not yet adopted a machine learning-based anti-fraud approach, despite the fact that banks in Indonesia have nearly completely done so [2]. The results of an interview conducted by the author with one of the Conventional Bank employees indicated that they had not yet implemented machine learning in fraud detection on credit card transactions. Another interview with an employee of an Islamic bank in Indonesia said that machine learning had not been applied to a credit card transaction [3]. According to Google Trends data, from 2017 to 2022, there has been an increase in the use of machine learning for credit card identification. So based on some of the things above, it is necessary to apply machine learning to detect fraudulent credit card transactions [4].

Machine learning is the adoption of computer programs and mathematical formulas that make predictions for the future while learning from historical data. The learning process in question has two steps, involving testing and training, in an effort to develop intelligence [5]. Three categories of machine learning exist: reinforcement learning, unsupervised learning, and supervised learning. Credit card transaction fraud detection is included in the supervised learning classification, in which data sets are labeled to classify unknown classes. Classification has

Diterima: 26-06-2023 | Revisi: 01-07-2023 | Diterbitkan: 21-08-2023 | doi: 10.60083/jidt.v5i3.386

many uses in fraud detection, marketing targets, performance prediction, manufacturing, and medical diagnosis [6]. However, in the classification of supervised learning, class imbalance needs to be considered. Class imbalance, or class imbalance, is a condition of unequal distribution between classes in a dataset where one class has a very large amount of data compared to other classes (majority class). One technique used to address the issue of unbalanced classes is sampling [7]. To balance the amount of data for each class, the sampling approach adjusts the training dataset's data distribution between the majority and minority classes. To address class imbalances at the data level, several sampling strategies are used. Synthetic Minority Oversampling Technique (SMOTE) and Random Under Sampling (RUS) are two methods that are frequently employed [8].

For datasets with class imbalances, the Synthetic Minority Oversampling Technique (SMOTE) technique is capable of improving accuracy by 0.73% and reducing loss by 0.0179 during training as well as improving accuracy by 2.2% and reducing loss by 0.039 during validation. Able to handle class imbalances in datasets using the NN+RUS method with a higher AUC of 0.88 compared to the NN method, which does not use RUS 0.83. As a result, the writer will employ the Synthetic Minority Oversampling Technique (SMOTE) and Random Undersampling (RUS) sampling approach for cases of class imbalance [9]. Support vector machines (SVM) are one of the machine learning algorithms that are most frequently utilized in fraud detection techniques. With an accuracy rate of more than 80%, the support vector machine algorithm performs well in detecting fraud. It also improves the TP Rate (fraud catching rate), has a low FP Rate (false alarm rate) value, and has a low error rate. The gradient-boosted trees approach, which is the best algorithm with an accuracy value of 99.85% and an AUC value of 1, was used in research on fraud detection to produce the results [10]. Regarding other research demonstrating that random forest delivers accuracy and nearly flawless precision of 99%, they clarify that it performs better with more training data. The authors will employ gradient-boost trees, support vector machines, and random forests as three machine learning methods based on the aforementioned research. The research mentioned above just haven't been incorporated into the web interface yet [11] [16]. Because, basically, building web-based interfaces requires extra effort when using frameworks such as Flask, Django, and others. Currently, there is a framework that makes it easy to build web interfaces in the field of data science and machine learning, namely Streamlit [12] [13]. A Python-based open-source framework called Streamlit was developed to make it simpler to create web apps for data science and machine learning [14] [15]. Implementation and comparison of the three algorithms and sampling strategies are done during the process to assess performance and determine which algorithm and sampling strategy is best (most appropriate) for use in detecting fraudulent credit card transactions.

2. Research Methods

Finding and gathering information about study-related topics, such as theoretical underpinnings, writing methodologies, process methodologies, and relevant research references, is the goal of data collection. The author's chosen strategy for gathering data for this study was a literature review. In order to complete the steps of data gathering through literature review, references pertinent to the topic are sought out. Online reference searches are performed in places like e-books and official websites. The many pieces of information required to support this research were chosen after the references had been gathered. The author's technique of implementation adheres to the Data Science Methodology's stages. It's just that the stages used by the author start from business understanding to evaluation. The last two stages, namely deployment and feedback, were not used by the authors because the scope of the research was limited to model experiments with several techniques used to obtain or compare model performance from the results of evaluations made later so as to find out which algorithms and sampling techniques are suitable for use. From the data science methodology above, we get the experimental flow that the author made using the Atom and Terminal Anaconda Navigator tool.

3. Results and Discussion

The author takes an analytical approach, namely by breaking down the problem into the elements needed to answer it. Based on existing literature studies in previous studies, the author's solution is to answer the problems that have been defined previously. How can we determine which algorithm and sampling method are suitable for identifying transaction fraud on credit cards by comparing the performance of random forest algorithms, support vector machines, and gradient-boosted trees with the Synthetic Minority Oversampling Technique (SMOTE) and Random Under Sampling (RUS) methods? Answer: The author uses a sampling method with the synthetic minority oversampling technique (SMOTE) and random under sampling (RUS) as comparisons to identify fraudulent credit card transactions. To allow for a fair comparison, the three classification machine learning algorithms that were used random forest, support vector machines, and gradient boosted tree were each coupled to one of the two methods (SMOTE and RUS). To find out the performance of the algorithm and sampling technique used, the authors tested the performance of the model. The test will be carried out by comparing the measuring parameters in the form of AUC Score and ROC Curve, Fraud Catching Rate (Recall), False Alarm Rate (FP), Matthew Correlation Coefficient (MCC), and F-Measure. Meanwhile, to determine which algorithms and sampling

techniques are suitable for use, the authors will use the values of AUC Score, Fraud Catching Rate (Recall), False Alarm Rate (FP), and MCC.

The authors use a dataset of credit card transactions to address the topic of what information is required to address current issues. The 2018 FINHACKS competition's open dataset of credit card transactions were used by the authors of this work. A dataset of credit card transactions was gathered by the author from the 2018 FINHACKS competition hosted by one of Indonesia's traditional banks. To get the dataset, it can be accessed via the following link: https://github.com/rezafaisal/FinHack2018. The data used is synthetic and not real. And this dataset is already in CSV form. The number of available data points (rows) is 13125, with 28 data features (columns), one of which is a class or label. Each measuring parameter (AUC Score, Recall (Fraud Catching Rate), FP (False Alarm Rate), F-Measure, and Matthew Correlation Coefficient (MCC), which are compared based on the SMOTE and RUS techniques in each algorithm) is used to assess the performance of the classification algorithm. The AUC score value is represented by the ROC curve, therefore when compared, they will have the same value.

AUC values of the three modeling algorithms used with the SMOTE or RUS techniques. The Random Forest method yields the lowest AUC value in the SMOTE approach, with a value of 0.7092, and the Gradient Boosted Tree strategy generates the greatest AUC value at 0.7269. Regarding the RUS approach, the Random Forest algorithm had the highest AUC value with a value of 0.7742, while the Support Vector Machine algorithm achieved the lowest with a value of 0.7096. The AUC value is calculated using a scale with values ranging from 0 to 1. The better a model is at predicting a classification, the higher its AUC value. Accordingly, the random forest algorithm with the RUS approach receives the greatest score, or 0.7742, based on the overall AUC value based on the SMOTE and RUS procedures. In addition, the three modeling algorithms using the SMOTE or RUS techniques used by the authors yielded a value > 0.7, which indicates that the classification performance carried out in this study was quite good. The fraud catching rate, recall rate, or TF rate is the number of fraudulent transactions predicted as fraud. The higher the value obtained, the better the fraud-catching rate, recall rate, or TF rate. For the SMOTE technique, the Gradient Boosted Trees algorithm is obtained with the highest value of 0.59. While the Random Forest algorithm was obtained using the RUS approach with the greatest value of 0.75. In light of this, it is evident that the random forest algorithm with the RUS approach, which has a value of 0.75, performs well in terms of fraud detection rate and recall when compared to the greatest score between SMOTE and RUS. The SMOTE random forest algorithm has the lowest fraud-catching rate, namely 0.48.

The number of regular (non-fraud) transactions that are projected to be fraudulent is the false alert rate itself. Therefore, if the outcome value is low, the false alarm rate's (FP) performance is regarded to be good. The best value, according to the SMOTE method, was produced by the random forest algorithm at 0.06. The optimal value for the RUS method, however, is 0.07 as determined by the support vector machine algorithm. The SMOTE random forest method, with a value of 0.06, performs well in terms of false alarm rate when compared to the RUS technique. The F-measure value, often known as the F1-score, is used to address issues with class imbalance. The performance of F-Measure improves as the value generated rises. The highest f-measure value generated based on the SMOTE technique is 0.41 with the random forest algorithm. The highest f-measure value generated based on the RUS technique is 0.4 with the support vector machine algorithm. The support vector machine algorithm with the SMOTE technique and the random forest and gradient-boosted trees with the RUS technique both produced values of 0.33 for the f-measure, with the random forest algorithm with the SMOTE approach producing the highest result. The highest score for the SMOTE method is +0.37 for the random forest algorithm. The support vector machine algorithm receives the highest RUS score of +0.35 in the meantime. A value close to +1 means the performance of the classification algorithm is getting better, and a value closer to -1 means the performance of the classification algorithm is getting worse. As a result, the support vector machine algorithm with the SMOTE technique yielded the lowest value of +0.29 and the random forest algorithm with the SMOTE technique produced the best MCC value of +0.37. The SVM algorithm used in the RUS approach received the second-highest score, with a value of +0.35.

The results of comparing the values of the measuring parameters used to assess how well the algorithm and sampling technique performed on the generated test data. The performance matrix for each method is then obtained using the SMOTE and RUS approaches. Support vector machines and gradient-boosted trees are produced based on the SMOTE and RUS approaches after knowing the performance of each random forest methodology. Next, we will describe the calculation of each measuring parameter used. The predictions generated to detect the number of normal cases (0) and fraud (1) from the application of the RF, SVM, and GBT algorithms and the sampling techniques, namely SMOTE and RUS, are used on test data (testing data). The model that has been built with test data (testing) shows that the majority can correctly predict both the normal (0) and fraud (1) labels. The highest AUC value and fraud-catching rate were obtained by the RUS technique random forest algorithm, with respective scores of 0.7742 and 0.75. Using the SMOTE approach and the random forest algorithm, the best value for the false alarm rate was 0.06. The random forest algorithm used in the SMOTE technique produced the highest MCC score, which was +0.37. Then it is obtained: The RUS technical random forest algorithm excels twice in AUC value and fraud detection rate. The SMOTE technique random forest algorithm excels twice in the values of false

alarm rate and Matthew Correlation Coefficient (MCC). Because there is no other superior algorithm based on these four parameters other than the random forest algorithm, it was chosen as a suitable algorithm for use in credit card transaction fraud detection because it is equally superior in RUS and SMOTE techniques. So, to choose a suitable sampling method, the authors will compare the random forest algorithms of the SMOTE and RUS techniques based on the ratings obtained from the four measuring parameters.
The SMOTE technique has two measuring parameters that rank sixth, namely the AUC with a value of 0.71 and the fraud-catching rate with a value of 0.48. From the RUS technique, it has one measuring parameter in the fifth rank, namely the false alarm rate, with a value of 0.2. Also has one measuring parameter: MCC in the fourth rank is +0.31. In terms of ranking, it was found that the RUS technique was superior to SMOTE because two of the RUS measuring parameters received better ratings than the two SMOTE measuring parameters, both of which were ranked sixth. So, it was found that a suitable technique for detecting credit card transactions in this study using the FINHAKS 2018 dataset was the Random Under Sampling (RUS) technique with the Random Forest Algorithm.

4. Conclusion

It was discovered that the Random Forest, Support Vector Machines, and Gradient Boosted Trees algorithms with the Synthetic Minority Oversampling Technique (SMOTE), and Random Under Sampling (RUS) were successfully applied in detecting credit card transaction fraud based on the results of the research discussion that the author has explained. As evidenced by the AUC score of each algorithm and sampling technique obtained, which is > 0.7 and is included in the classification performance group, this is quite good. In the evaluation based on the SMOTE and RUS techniques of each algorithm, the highest AUC values were obtained, namely the SMOTE Gradient Boosted Trees algorithm with a value of 0.73 and the Random Forest algorithm with the RUS technique with a value of 0.78. In this study, based on the four measuring parameters and the Fin hacks 2018 credit card dataset used, the Random Forest algorithm with the Random Under Sampling (RUS) technique proved to be suitable for fraud detection of credit card transactions with each value obtained, namely 0.77 for the AUC Score, 75% Fraud Catching Rate, 20% False Alarm Rate, and +0.31 Matthew Correlation Coefficient (MCC). Using the Random Forest algorithm without sampling and research conducted by the author, it was concluded that the Random Forest algorithm with and without sampling works well in detecting fraud in credit card transactions. The writer realizes that this research still has its weaknesses. Consequently, recommendations for further study include: Using hyperparameter tweaking for machine learning algorithms to enhance the effectiveness of machine learning models contrasting with other sampling techniques including T-Link, near miss, random oversampling, and adaptive synthetic (ADASYN). Conduct several tests to determine how test and train data are distributed. In order to simplify developing and testing in the machine learning process, it is envisaged that it can be integrated into other cloud computing platforms like Azure Machine Learning Studio, Amazon Sage Maker, IBM Watson Studio, or Google Cloud AutoML/AI Platform.

References

[1] Ardiyansyah, & Rahayuningsih, P. A, "Application of Sampling Techniques to Overcome Class Imbalances in Online Shoppers Intention Classification," JTIK (Jurnal Teknik), vol. 4, no. 1, pp. 7–15, 2020.
[2] Putra, H. S., Huljannah, M., Anis, A., & Azhar, Z. (2021). Debit and Credit Cards: Money Velocity Risks. *Jurnal Ekonomi & Studi Pembangunan*, *22*(2), 228-243.
[3] Dheepa, V., & Dhanapal, R, "Behavior Based Credit Card Fraud Detection Using Support Vector Machines," ICTACT Journal on Soft Computing, vol. 02, no. 04, pp. 391–397, 2012.
[4] Putra, H. S., Huljannah, M., & Putri, M. (2021). Analysis Of The Demand For Money And The Velocity Of Money In The Digital Economy Era: A Case Study In Indonesia. *Jurnal REP (Riset Ekonomi Pembangunan)*, *6*(1), 110-125.
[5] Huang, G. Bin, Zhu, Q. Y., & Siew, C. K, "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, no. 1, pp. 489–501, 2016.
[6] Niveditha, G., Abarna, K., & Akshaya, G. V, "Credit Card Fraud Detection Using Random Forest Algorithm," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 301–306, 2019, https://doi.org/10.32628/cseit195261
[7] Prasetyo, S. N, "Formulation of Credit Card Fraud Arrangements in Indonesian Criminal Law Viewed from the Principle of Legality," Jurnal Ilmiah Hukum LEGALITY, vol. 24, no. 1, pp. 101, 2017, https://doi.org/10.22219/jihl.v24i1.4260
[8] Sahin, Y., & Duman, E, "Detecting credit card fraud by ANN and logistic regression," INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications, pp. 315–319, 2011, https://doi.org/10.1109/INISTA.2011.5946108
[9] Somantri, O., & Khambali, M, "Feature Selection Short Story Category Classification Using Naïve Bayes

and Genetic Algorithm," Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI), vol. 6, no. 3, pp. 301–306, 2017, https://doi.org/10.22146/jnteti.v6i3.332

[10] Merlin, N. M., & Vanchapo, A. R. (2021). Readiness Management in Handling COVID-19 Pandemic and Early Detection in The Referral Hospital in East Nusa Tenggara Province. *KEMAS: Jurnal Kesehatan Masyarakat*, *17*(2), 279-286.

[11] Somvanshi, M., & Chavan, P, "A review of machine learning techniques using decision tree and support vector machine. Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA," 2016. https://doi.org/10.1109/ICCUBEA.2016.7860040

[12] Sutoyo, E., & Fadlurrahman, M. A, "Application of SMOTE to Overcome Class Imbalance in Television Advertisement Performance Rating Classification Using Artificial Neural Networks," JEPIN (Jurnal Edukasi Dan Penelitian Informatika), vol. 6, no. 3, pp. 379–385, 2020.

[13] Vanchapo, A. R. (2020). Pengaruh Upah terhadap Motivasi Kerja Karyawan Sukarela di Puskesmas Se Kabupaten Sikka. *CHMK Nursing Scientific Journal*, *4*(1), 157-161.

[14] Syukron, A., & Subekti, A, "Application of Random Over-Under Sampling and Random Forest Methods for Credit Rating Classification," Jurnal Informatika, vol. 5, no. 2, pp. 175–185, 2018, https://doi.org/10.31311/ji.v5i2.4158

[15] Zhang, Y., & Haghani, A, "A gradient boosting method to improve travel time prediction," Transportation Research Part C: Emerging Technologies, vol. 58, pp. 308–324, 2015, https://doi.org/10.1016/j.trc.2015.02.019