Contents lists available at www.journal.unipdu.ac.id
Register
Journal Page is available to www.journal.unipdu.ac.id/index.php/register

Research article

An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset

Prasetyo Wibowo [a], *Chastine Fatichah* [b]

[a,b] *Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*

email: [a] pras.wibowo96@gmail.com, [b] chastine@if.its.ac.id

ARTICLE INFO

Article history:
Received 23 December 2020
Revised 13 February 2021
Accepted 18 February 2021
Available online 28 February 2021

Keywords:
classification
imbalanced dataset
oversampling
performance analysis

Please cite this article in IEEE style as:
P. Wibowo and C. Fatichah, " An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 7, no. 1, pp. 63-71, 2021.

ABSTRACT

Class imbalance occurs when the distribution of classes between the majority and the minority classes is not the same. The data on imbalanced classes may vary from mild to severe. The effect of high-class imbalance may affect the overall classification accuracy since the model is most likely to predict most of the data that fall within the majority class. Such a model will give biased results, and the performance predictions for the minority class often have no impact on the model. The use of the oversampling technique is one way to deal with high-class imbalance, but only a few are used to solve data imbalance. This study aims for an in-depth performance analysis of the oversampling techniques to address the high-class imbalance problem. The addition of the oversampling technique will balance each class's data to provide unbiased evaluation results in modeling. We compared the performance of Random Oversampling (ROS), ADASYN, SMOTE, and Borderline-SMOTE techniques. All oversampling techniques will be combined with machine learning methods such as Random Forest, Logistic Regression, and k-Nearest Neighbor (KNN). The test results show that Random Forest with Borderline-SMOTE gives the best value with an accuracy value of 0.9997, 0.9474 precision, 0.8571 recall, 0.9000 F1-score, 0.9388 ROC-AUC, and 0.8581 PRAUC of the overall oversampling technique.

Register with CC BY NC SA license. Copyright © 2021, the author(s)

1. Introduction

Class imbalance occurs when the distribution of classes between the majority and the minority classes is not equal. In real-world applications, the class imbalance may vary from minor to severe [1]. The dataset can be said to be unbalanced if each class is not represented in a balanced manner. The majority class consists of a large portion of the dataset, while the minority class consists of a small portion of the dataset, which is often the object to be modeled. The consequence of high-class imbalance will impact the classification's overall performance when the model is most likely to predict most of the data falling within the majority class. This model will yield biased results, and the performance predictions for the minority class frequently have little impact on the model. The research conducted by He and Garcia [2] explains that unbalanced data can be seen based on the majority and minority class ratios ranging from 100:1 to 10,000:1. The creation of a model using a high level of imbalance is considered challenging by experts due to the complexity of the processing of balancing data [3].

There are two methods commonly used in modeling data imbalances, namely supervised and semi-supervised [4]. The model will estimate based on the dataset's existing sample to be classified in the supervised approach. Several algorithms can be used, such as Random Forest (RF) [4, 5], K-Nearest Neighbors (KNN) [6, 7], and Support Vector Machine (SVM) [8, 9]. In the semi-supervised approach, data classification will be identified using label and unlabeled data in the dataset. One of the algorithms included in the semi-supervised approach is K-Means [10, 11]. Modeling in the high-class data imbalance is important because not all models can provide the actual classification. The unbalanced characteristics of the dataset will result in bias in the model so that a process at the data level is required, such as resampling minority data.

Resampling is a process that tries to balance the distribution of data between minority data and majority data. There are two groups in resampling, namely under-sampling and oversampling. Under-sampling is a way to remove or reduce existing samples in the data majority, and oversampling tries to flatten data distribution by duplicating minority data. Oversampling is the focus of this research because all data in the dataset is crucial to be modeled by machine learning.

Currently, the state-of-the-art oversampling technique developed by Chawla [12] is Synthetic Minority Oversampling Technique (SMOTE). This method produces a sample based on the closest rivalry by duplicating it along the lines between minority classes. There are several developments in the oversampling technique method based on the SMOTE method. Each method development has its characteristics, such as Borderline-SMOTE [13], which focuses on duplication on the margins of the minority class, and Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) [14], which calculates based on the density of minority data. Even though many oversampling techniques have been developed, only a few have been applied to high-class data imbalances.

Many imbalanced datasets can be accessed publicly, such as wine quality [15], cancer diagnosis [16], and credit card fraud detection [17]. The wine quality dataset contains data for classifying the quality of red wine based on wine quality classes. This dataset has 1,599 data with the percentage ratio of minority/majority data of 3.31%/96.69%, which indicates that only about 53 minority data are available in this dataset. Furthermore, cancer diagnosis data contains data to classify whether the patient has cancer or not. This dataset has a total of 1,056 data with a minority/majority data percentage ratio of 18.09%/81.9%, which indicates that there are about 191 minority data that can be modeled. The last dataset is the credit card fraud dataset, which contains data to detect credit card fraud. This dataset's total data is 284,807, with a percentage ratio of the minority/majority data of 0.17%/99.83%. The ratio data shown in the credit card fraud dataset indicates that this dataset experiences a high-class imbalanced dataset with a minority data of 492 from the existing data population. The credit card fraud dataset will be selected in this study because it has the characteristics of a high-class imbalanced dataset.

Several studies use this high-class data imbalance dataset, such as Makki [18], which performed fraud detection using several machine learning models, namely Support Vector Machine and Artificial Neural Network with random oversampling technique. Meng [19] used a combination of XGBoost with oversampling and under-sampling techniques to detect credit card fraud. The obtained result is that the SMOTE technique gives a better score than the under-sampling technique. Almhaithawi [20] used several machine learning methods such as Logistic Regression, Random Forest, and XGBoost by calculating the cost-sensitive Minimum Bayes Risk, which is balanced using SMOTE, which shows good results. Awoyemi [21] used a hybrid technique of under-sampling and oversampling to solve the data imbalance problem. Several machine learning methods used in this test, namely Naïve Bayes, K-NN, and Logistic Regression, showed quite good results. However, all of the above studies use SMOTE as the main oversampling, and a few use other oversampling techniques to solve the problem of high-class data imbalance.

This study aims for an in-depth performance analysis of the oversampling techniques to address the high-class imbalance problem. We compare the performance of Random Oversampling (ROS), ADASYN, SMOTE, and Borderline-SMOTE techniques. All oversampling techniques are combined with machine learning methods such as Random Forest, Logistic Regression, and K-NN. Utilizing the oversampling technique will balance each class's data to provide unbiased evaluation results in modeling. Each characteristic of the oversampling technique has its respective advantages, which will be explained in this study.

2. Material and Methods

2.1. Research diagram system

Fig. 1 shows the diagram system used in this study. The initial workflow in the system model is to split up the data using the hold-out method so that it becomes two parts, namely training set, and test set. In the training set, oversampling techniques are used to increase the number of minority class so that when a model fitting is carried out, a balanced dataset can be obtained.

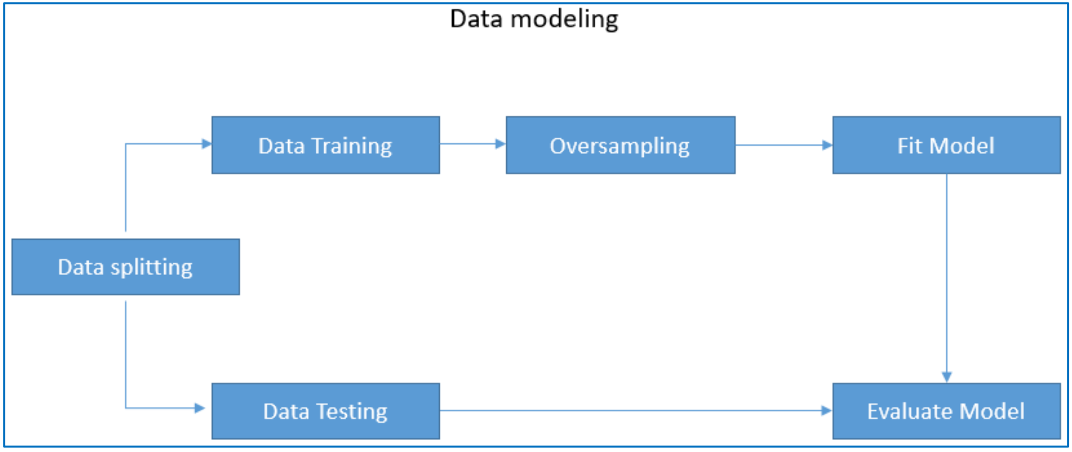


Fig. 1. System diagram

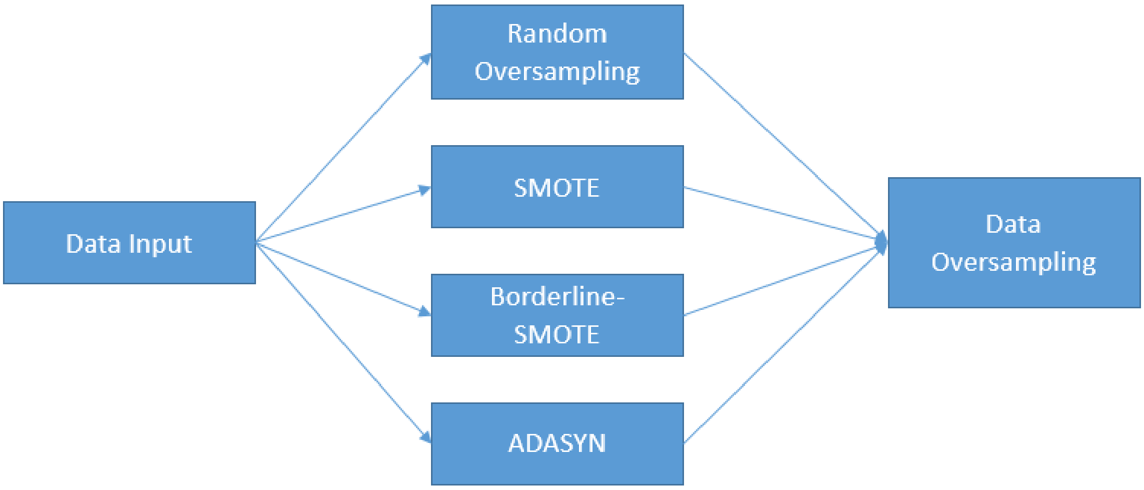


Fig. 2. Chart of the oversampling technique used

Fig. 2 shows the oversampling technique used in this study. The methods used are Random Oversampling, SMOTE, Borderline-SMOTE, and ADASYN. These four methods have different characteristics, as follows:

- Random oversampling method (ROS) performs duplication by selecting a random set of minority classes for random data replication [22]. Because the sampling process is carried out randomly, random oversampling has a disadvantage, namely that it requires a long training time, and there is the possibility of overfitting.
- SMOTE duplicates data by measuring the similarity between neighboring minority class samples [12]. Each data will be reproduced based on the nearest neighbor line.
- Borderline-SMOTE duplicates data by making the closest neighbor line to the sample data on the margins of the minority class [13]. Duplication on this border will strengthen the difference between the minority and majority class.
- ADASYN calculates the density distribution in minority classes before duplicating data based on the criteria obtained when calculating class density [14]. This approach helps to focus on the midpoint of the minority class area depending on the dataset.

2.2. *Description of the dataset*

The dataset used in this study is obtained from European cardholder transactions for two days in September 2013 [17]. The dataset is collected by two research teams from Worldline and Libre Brussels University. The dataset contains 30 attributes containing transaction time, number of transactions, and 28 other attributes labeled V1 to V28 through the transformation results using Principal Component Analysis (PCA). This transformation is performed to hide sensitive attribute data from the user. Fig. 3 shows the distribution of the minority and majority classes. There were 284,807 transactions with a positive fraud class percentage of 0.17% and a negative fraud class of 99.83% of the total transactions. Credit card transactions contain numeric and categorical data so that this combination will be processed to detect credit card fraud.

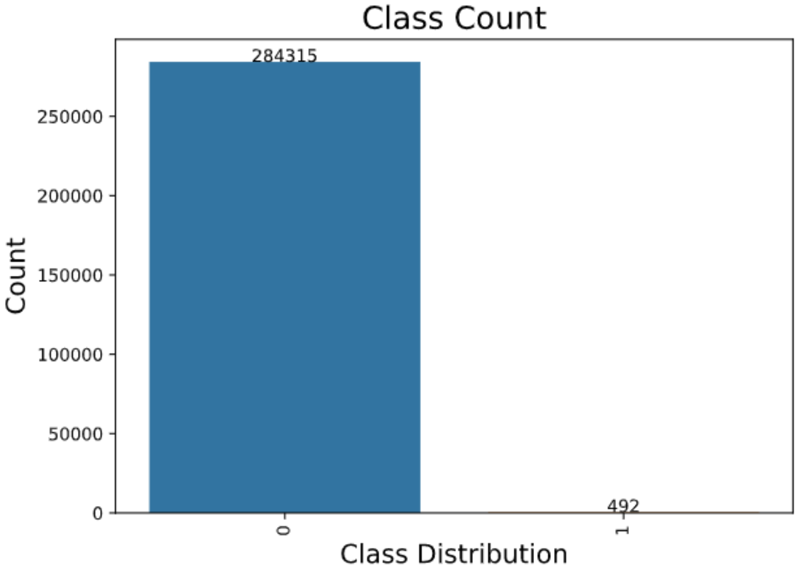


Fig. 3. The number of distribution of minority and majority classes

2.3. *Evaluation metrics*

There are four basic metrics used for the evaluation of this study, namely the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP and TN is the number of samples from the test set that are classified correctly and incorrectly, respectively. In contrast, FN and TP represent the number of samples from the test set, classified incorrectly as negative and positive. A commonly used evaluation metric uses accuracy to get the ratio of how many samples of data were correctly classified. The accuracy is shown in Eq. 1.

The information provided by accuracy is sometimes useless when measuring how reliable a model using imbalanced data is. So, in this case, we can measure it using the evaluation metric F1-score. F1-score evaluation requires two metric evaluations: precision and recall, which can be seen in Eq. 2 and Eq. 3.

The F1-score will balance the two metrics so that they are harmonious with each other through Eq. 4. An evaluation called Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is needed to find out how well the model can differentiate between positive and negative classes (ROC-AUC), which is shown in Eq. 5. In the context of an imbalance, the ROC AUC evaluation data sometimes provide misleading information so that an evaluation metric called Precision-Recall Curve (PRAUC) is used.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}, \tag{1}$$

$$Precision = \frac{TP}{TP+FP}, \tag{2}$$

$$Recall = \frac{TP}{TP+FN}, \tag{3}$$

$$F1 - Score = 2\,\frac{precision.recall}{precision+recall}, \tag{4}$$

$$ROC - AUC = \frac{1+TP_{rate}-FP_{rate}}{2}, \tag{5}$$

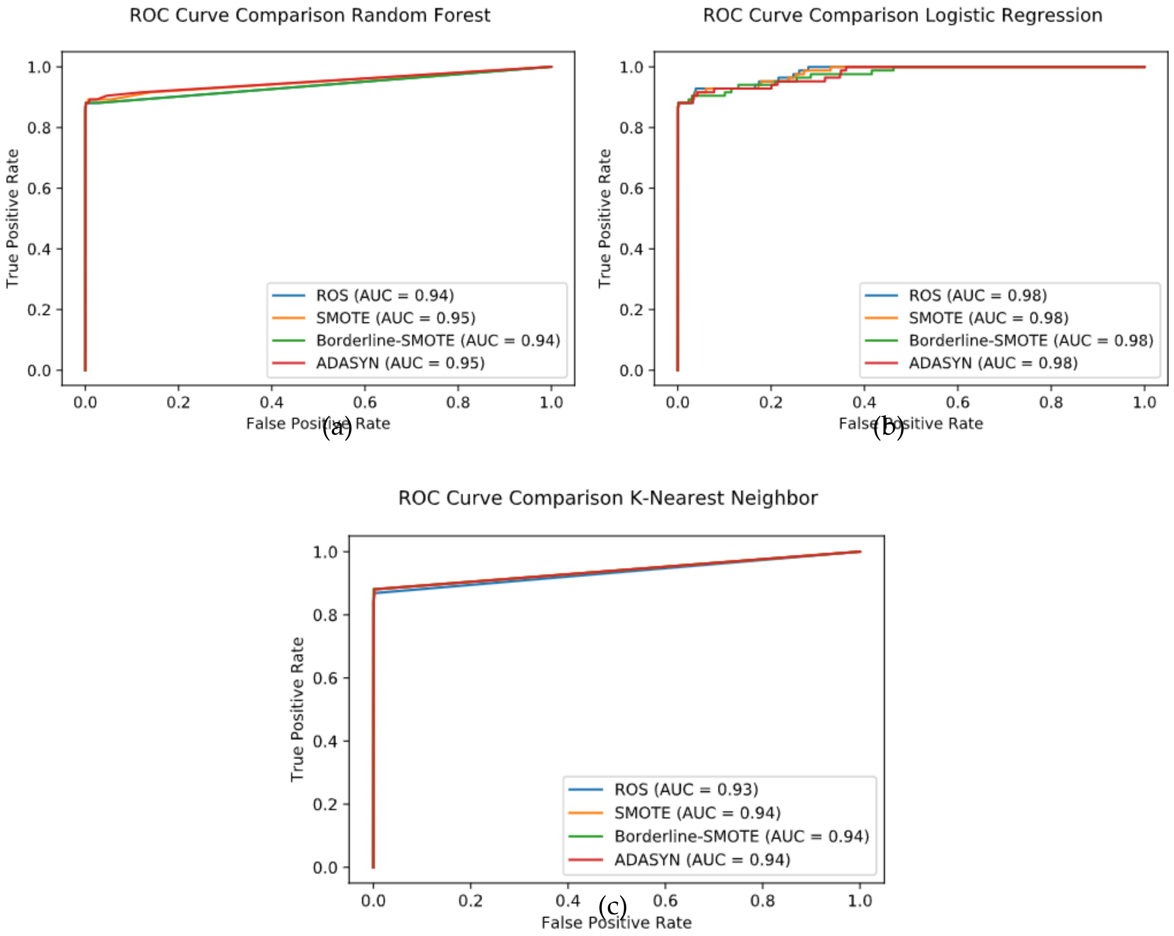
3. Results and Discussion

Four oversampling techniques are carried out depending on the formal diagram scheme being used: ROS, SMOTE, Borderline-SMOTE, and ADASYN. Three machine learning methods are employed to determine the best candidate to solve credit card fraud, namely Random Forest, Logistic Regression, and K-NN.

Table 1 shows the classification accuracy, precision, recall, F1-score, ROC-AUC, and PRAUC of the four oversampling techniques on the credit card fraud dataset. In the use of raw data, the results are shown to be very good. However, this is the main problem because the data's composition is highly unbalanced, resulting in a bias in the classification model. The oversampling technique is needed to increase the number of minority classes in data imbalances [23].

Table 1. Comparison of experimental results

Model	Oversampling	Accuracy	Precision	Recall	F1-score	ROC-AUC	PRAUC
Random Forest	Raw Data	0.9997	0.9583	0.8214	0.8846	0.9391	0.8663
	ROS	0.9997	0.9342	0.8452	0.8875	0.9389	0.8603
	ADASYN	0.9996	0.8780	0.8571	0.8675	0.9518	0.8383
	SMOTE	0.9996	0.8889	0.8571	0.8727	0.9500	0.8487
	Borderline-SMOTE	0.9997	0.9474	0.8571	0.9000	0.9388	0.8581
Logistic Regression	Raw Data	0.9994	0.8769	0.6786	0.7651	0.9787	0.7823
	ROS	0.9759	0.0515	0.8810	0.0974	0.9822	0.7553
	ADASYN	0.8955	0.0129	0.9286	0.0255	0.9764	0.7432
	SMOTE	0.9762	0.0521	0.8810	0.0983	0.9808	0.7519
	Borderline-SMOTE	0.9845	0.0780	0.8810	0.1433	0.9758	0.7608
K-NN	Raw Data	0.9996	0.9444	0.8095	0.8718	0.9344	0.8445
	ROS	0.9991	0.6460	0.8690	0.7411	0.9344	0.7728
	ADASYN	0.9976	0.3700	0.8810	0.5211	0.9399	0.5753
	SMOTE	0.9976	0.3719	0.8810	0.5230	0.9400	0.5916
	Borderline-SMOTE	0.9994	0.7475	0.8810	0.8087	0.9403	0.8206



(a) (b) (c)

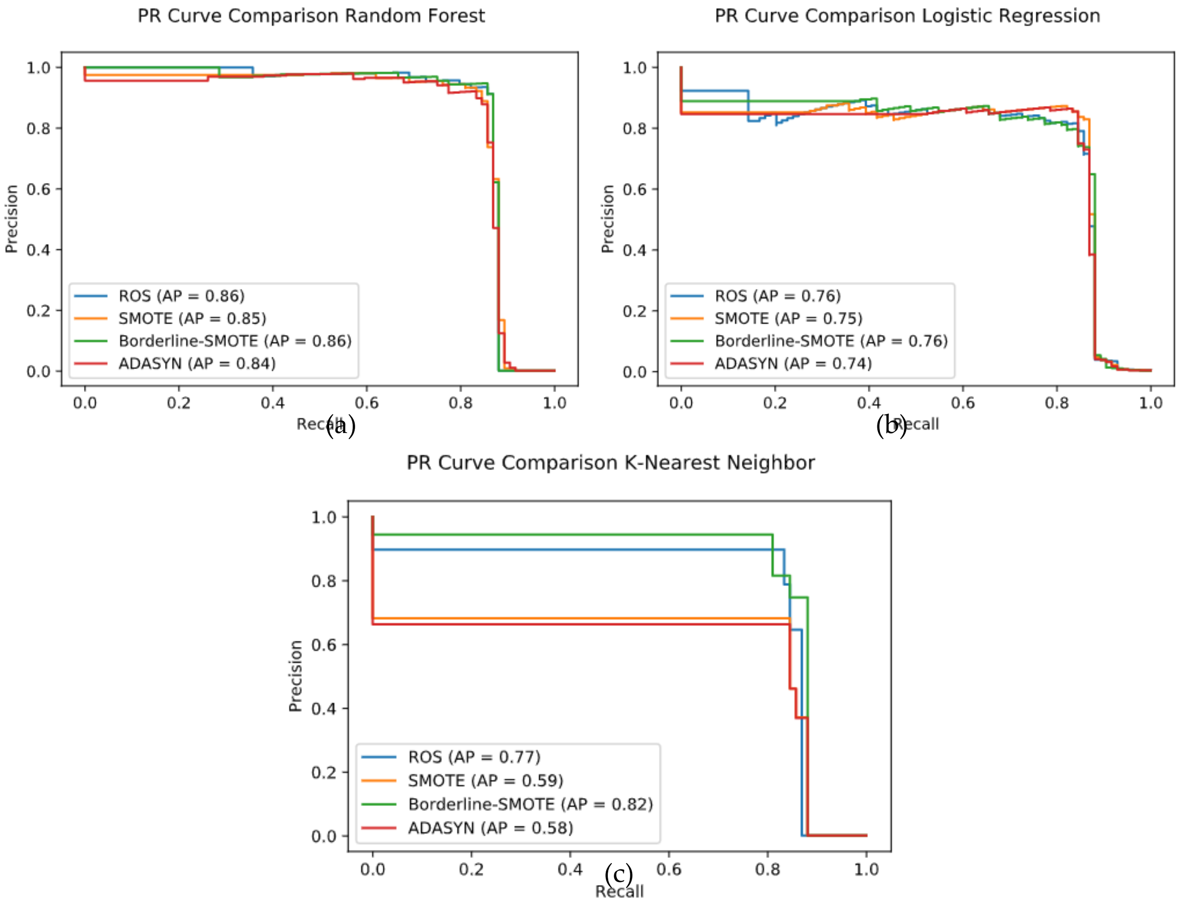
Fig. 4. ROC-AUC comparison chart of all experiments: (a) Random Forest; (b) Logistic Regression; (c) K-NN

Accuracy has an important role in finding out how well the predicted value and the actual value in the dataset. ROS and Borderline-SMOTE obtain the best accuracy results with a value of 0.9997. Even though it gets high scores, accuracy still does not represent whether the used technique is the best, so it is necessary to check the precision and recall values.

The precision value determines how many relevant predictive values were correctly predicted. The best score is Borderline-SMOTE, with a result of 09474. Interestingly, when observing the model, Random Forest's results give the highest average precision compared to other models because Random Forest uses bagging approach in the decision tree process to provide lower variance [24].

The next metric is recall, which determines how many the actual predictive value of the data. ADASYN held the best score with a result of 0.9286. ADASYN has the advantage of calculating the density of data distribution so that the results of the duplication data provided can be focused on one particular area and strengthen the results of recall evaluation.

Precision and recall have mutually dependent characteristics, which are called trade-offs. Through this, the F1-score plays an important role in knowing the alignment between the precision and recall values by calculating the average value for each evaluation value from the precision and recall metrics. The best F1 score is Borderline-SMOTE, with a score of 0.9000. This high result indicates that Borderline-SMOTE provides the most balanced results compared to other oversampling values because the resulting data duplication is on the margins of the minority data to strengthen the differentiating value between the minority class and the majority class.



(a)

(b)

(c)

Fig. 5. PRAUC comparison chart of all experiments: (a) Random Forest; (b) Logistic Regression; (c) K-NN

Furthermore, the ROC-AUC value will determine how well the machine learning model distinguishes between negative and positive classes. Among all the oversampling techniques in this study, the highest result is obtained by ROS with a value of 0.9822. It can be seen in Fig. 4 that the combination of machine learning and oversampling techniques can distinguish negative and positive classes with satisfactory results with a value range of 0.93-0.98. Even though the ROC-AUC result is relatively high, there is still a possibility that the results presented could be misleading when using a highly unbalanced dataset. Therefore, PRAUC is used to find out the reliability of the resulting classification model.

PR-AUC identifies how well the average precision is generated at each threshold result of the recall. The best result is ROS, with a score of 0.8603, beating all other competitors in the oversampling techniques. According to Fig. 5, it can be seen that the K-NN gives results between 0.58-0.82, which shows that this model is not reliable for credit card fraud cases. Furthermore, the Logistic Regression results show that the value between 0.74-0.76, giving an unreliable average result. In Table 1, the precision results generated by Logistic Regression are terrible. Finally, Random Forest yields a result value between 0.84-0.86, which shows that this model is reliable for credit card fraud cases.

4. Conclusion

The effect of high-class imbalance can affect the overall classification performance. The machine learning model will predict most of the data included in the majority class so that the evaluation of the evaluation results gives biased results. Results without using oversampling give high evaluation results indicating bias in the model. Bias in the model occurs because the prediction model for the minority class has no impact on the model so that an oversampling technique is needed to avoid bias in the results of the model evaluation. In this study, the oversampling technique can avoid bias in the overall evaluation results. The test results of the oversampling technique on the imbalance of credit card fraud data gave quite good results with an accuracy of above 0.89. However, the accuracy results sometimes give misleading results, so that the evaluation of other metrics are needed, such as precision, recall, F1-score, ROC-AUC, and PRAUC. Borderline-SMOTE gave the best results with an average PRAUC of 0.8131. Borderline-SMOTE achieves the highest results because it duplicates the margins of the minority data so that it strengthens the difference between the minority and majority class data. ROS obtained the second position with an average PRAUC result of 0.7961. This result is obtained by duplicating based on minority data and randomly. Even though it has a high result, ROS's biggest drawback is that there is little variation in the duplicated data due to the duplication of data based on the original data. SMOTE gives an average PRAUC result of 0.7307. SMOTE duplicates data based on the closest neighbor to minority data so that the characteristics of SMOTE are sensitive to outliers, which can result in minority data being duplicated into the majority data area. Finally, ADASYN with an average PRAUC result of 0.7189. ADASYN works by duplicating data based on the density of the data distribution for the minority class. The more scattered the minority data, the higher possibility of resulting one sample data of minority class.

The best machine learning model is generated by Random Forest with an average accuracy value of 0.9996, precision of 0.9121, recall of 0.8541, F1-score of 0.8819, ROC-AUC of 0.9448, and PRAUC of 0.8513. The Random Forest has a characteristic of being more resistant to overfitting, usually when minority data are accidentally duplicated into the majority data area. Interestingly, Logistic Regression gives poor precision results with an average value of 0.0486. Poor precision results from Logistic Regression affects the resulting F1 score because there is a trade-off between precision and recall. The best combination of oversampling and machine learning method is Borderline-SMOTE with Random Forest with 0.9997 accuracy, 0.9474 precision, 0.8571 recall, 0.9000 F1 score, 0.9388 ROC-AUC, and 0.8581 PRAUC. Through the results of this test, it can be seen that the oversampling technique can provide actual results from data imbalances and can avoid bias in the entire machine learning model. Selecting the right oversampling technique is critical to do in order to maximize the evaluation results of the predictive model.

Declaration of Competing Interest

We declare that we have no conflict of interests.

References

[1] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder and N. Seliya, "A survey on addressing high-class imbalance in big data," *J Big Data,* vol. 5, no. 42, 2018.

[2] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 21, no. 9, pp. 1263-1284, 2009.

[3] I. Triguero, S. d. Río, V. López, J. Bacardit, J. M. Benítez and F. Herrera, "ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem," *Knowledge-Based Systems,* vol. 87, pp. 69-79, 2015.

[4] H. Kaur, H. S. Pannu and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions," *ACM Comput. Surv.,* vol. 52, no. 4, 2019.

[5] D. J. Dittman, T. M. Khoshgoftaar and A. Napolitano, "The Effect of Data Sampling When Using Random Forest on Imbalanced Bioinformatics Data," in *2015 IEEE International Conference on Information Reuse and Integration*, San Francisco, CA, USA, 2015.

[6] I. Indrajani, Y. Heryadi, L. A. Wulandhari and B. S. Abbas, "Recognizing debit card fraud transaction using CHAID and K-nearest neighbor: Indonesian Bank case," in *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, Yogyakarta, 2016.

[7] A. G. Pertiwi, N. Bachtiar, R. Kusumaningrum, I. Waspada and A. Wibowo, "Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm without smote in diagnosis of diabetes disease in balanced data," *Journal of Physics: Conference Series,* 2020.

[8] S. Cui, D. Wang, Y. Wang, P.-W. Yu and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," *Computer Methods and Programs in Biomedicine,* vol. 166, pp. 123-135, 2018.

[9] R. Pruengkarn, K. W. Wong and C. C. Fung, "Imbalanced data classification using complementary fuzzy support vector machine techniques and SMOTE," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Banff, AB, 2017.

[10] F. Last, G. Douzas and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," *Information Sciences,* vol. 465, 2018.

[11] J. Zhang, L. Chen and F. Abid, "Prediction of Breast Cancer from Imbalance Respect Using Cluster-Based Undersampling Method," *Journal of Healthcare Engineering,* vol. 2019, 2019.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, p. 321–357, 2002.

[13] H. Han, W.-Y. Wang and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang DS., Zhang XP., Huang GB," in *Advances in Intelligent Computing. ICIC 2005*, Berlin, Heidelberg, 2005.

[14] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, China, 2008.

[15] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems,* vol. 47, no. 4, pp. 547-553, 2009.

[16] S. Ali, A. Majid, S. G. Javed and M. Sattar, "Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data," *Computers in Biology and Medicine,* vol. 73, pp. 38-46, 2016.

[17] A. D. Pozzolo, O. Caelen, Y.-A. L. Borgne, S. Waterschoot and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," *Expert Systems with Applications,* vol. 41, no. 10, pp. 4915-4928, 2014.

[18] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. Hacid and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," *IEEE Access,* vol. 7, pp. 93010-93022, 2019.

[19] C. Meng, L. Zhou and B. Liu, "A Case Study in Credit Fraud Detection With SMOTE and XGBoost," *Journal of Physics: Conference Series,* vol. 1601, 2020.

[20] D. Almhaithawi, A. Jafar and M. Aljnidi, "Example-dependent cost-sensitive credit cards fraud detection using SMOTE and Bayes minimum risk," *SN Appl. Sci.,* vol. 2, no. 1574, 2020.

[21] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 International Conference on Computing Networking and Informatics (ICCNI)*, Lagos, Nigeria, 2017.

[22] W. Han, Z. Huang, S. Li and Y. Jia, "Distribution-Sensitive Unbalanced Data Oversampling Method for Medical Diagnosis," *J Med Syst,* vol. 43, no. 39, 2019.

[23] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog Artif Intell,* vol. 5, no. 221–232, 2016.

[24] S. Wager and S. Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association,* vol. 113, no. 523, pp. 1228-1242, 2018.