**Scientific Journal of Informatics**
Vol. 11, No. 1, Feb 2024

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

# Comparative Study of Imbalanced Data Oversampling Techniques for Peer-to-Peer Landing Loan Prediction

**Rini Muzayanah[1*], Apri Dwi Lestari[2], Jumanto[3], Budi Prasetiyo[4], Dwika Ananda Agustina Pertiwi[5], Much Aziz Muslim[6]**

[1, 2, 3, 4, 6]Department of Computer Science, Universitas Negeri Semarang, Indonesia
[5, 6]Faculty of Technology Management, Universiti Tun Hussein Onn Malaysia, Johor 86400, Malaysia

**Abstract.**
**Purpose:** Data imbalances that often occur in the classification of loan data on the Peer-to-Peer Lending platform can cause algorithm performance to be less than optimal, causing the resulting accuracy to decrease. To overcome this problem, appropriate resampling techniques are needed so that the classification algorithm can work optimally and provide results with optimal accuracy. This research aims to find the right resampling technique to overcome the problem of data imbalance in data lending on peer-to-peer landing platforms.
**Methods:** This study uses the XGBoost classification algorithm to evaluate and compare the resampling techniques used. The resampling techniques that will be compared in this research include SMOTE, ADACYN, Border Line, and Random Oversampling.
**Results:** The highest training accuracy was achieved by the combination of the XGBoost model with the Boerder Line resampling technique with a training accuracy of 0.99988 and the combination of the XGBoost model with the SMOTE resampling technique. In accuracy testing, the combination with the highest accuracy score was achieved by a combination of the XGBoost model with the SMOTE resampling technique.
**Novelty:** It is hoped that from this research we can find the most suitable resampling technique combined with the XGBoost sorting algorithm to overcome the problem of unbalanced data in uploading data on peer-to-peer lending platforms so that the sorting algorithm can work optimally and produce optimal accuracy.

**Keywords**: P2P lending, Resampling data, Imbalanced data, Machine learning
**Received** February 2024 / **Revised** February 2024 / **Accepted** February 2024



## INTRODUCTION

As data availability continues to grow across various large-scale, intricate, and interconnected systems like surveillance, security, the Internet, and finance, there is an increasing urgency to deepen our fundamental comprehension of knowledge discovery and analysis. This understanding is essential for transforming raw data into actionable insights that can support decision-making processes effectively [1]. While current techniques in knowledge discovery and data engineering have demonstrated considerable success in numerous real-world scenarios, addressing the issue of imbalanced data represents a relatively recent challenge that has garnered increasing interest from both academic and industrial sectors. Imbalanced data is one of the problems that is often encountered in the world of machine learning. Imbalanced data are caused by an error or imbalance in the distribution of the existing dataset. Imbalanced data is a problem that is quite complicated because it can make the algorithm used have difficulty processing the data due to overlapping datasets [2]–[4]. This tends to cause the machine learning system that is built to be biased towards the majority class [5] so that it can cause overfitting and poor generalization [6], [7]. In one study, it was explained that traditional classification strategies assume that there is a slight difference in sample size for each class, while the model tends to learn most classes with similar classification errors [8].

The development of technology today prevides various convenience for various life problems [9]. In the current era of information technology development and financial innovation, the credit scoring process has become a crucial cornerstone in the decision-making of banks and lending institution

*Corresponding author.
Email addresses: rinimuzayanah0415@students.unnes.ac.id (Muzayanah)
DOI: 10.15294/sji.v11i1.50274

[10]. The problem of imbalanced data is a problem that often occurs in the implementation of real data processing [11], one of which occurs in predictions using Peer-to-Peet Lending data. Peert-to-Peer (P2P) Lending is an online platform that handles the loan process. P2P lending is the most effective approach for individuals or small-scale business companies who do not have financial status or history[12]. Many studies have explored how to overcome the problem of imbalanced data, but not many have actually applied it in the world of P2P Lending [13]. New methods have emerged to overcome the problem of imbalanced data [14]. There are two approaches that are usually used to overcome the problem of imbalanced data, namely the algorithmic level approach and the data level approach. The algorithmic level approach is an approach that focuses on improving algorithms, while the data level approach focuses on improving the data used using resampling techniques [15]. Sir & Soepratono in their research in 2022 stated that the data level approach has a better effect than the algorithmic level approach for overcoming the problem of compensated data [15].

Several studies that have been conducted can address the problem of balancing the amount of data [16]. The data resampling method is an approach that works from increasing the sample distribution by copying minority class samples or eliminating some samples in the majority class [14]. The advantage of using resampling techniques, including oversampling, is that the subsequent classification structure will remain the same and can be applied individually before entering the model training process [17]. Oversampling techniques allow the weighting of each class to be more balanced during the machine learning process, so that the model can be more effective in identifying patterns and characteristics of minority classes that may be represented below the level of the original data set. Applying oversampling for high-dimensional gene expression data will cause the data size to be much bigger and take more time to execute [18].

There has been a lot of previous research that discusses handling the problem of imbalanced P2P lending data. One of them is research by Muslim et al. in 2023 [19]. This study proposes a new approach to ensemble learning, known as StackingXGBoost, which combines three basic learning algorithms (KNN, SVM, and Random Forest) into the XGBoost meta-learning algorithm. The effectiveness of the model is evaluated using two different data sets: an online P2P lending data set and a lending club data analysis data set. The evaluation results show that the LGBFS-StackingXGBoost model outperforms the other models and achieves the highest accuracy for both datasets. Specifically, the LGBFS-StackingXGBoost model achieved 99.982% accuracy on the online P2P lending dataset and 91.434% on the lending club loan data analysis dataset. These findings underscore the effectiveness of the LGBFS-StackingXGBoost approach in improving the accuracy of prediction models. Another study conducted by Garcia et al. pada tahun 2022 [20]. This study introduces an intelligent system designed to forecast academic failure using student data from the Industrial University of Santander in Colombia. The predictive model is driven by the XGBoost algorithm, with feature extraction based on TOPSIS methodology and oversampling performed through ADASYN. Additionally, the classifier's hyperparameters were optimized using a cross-validated grid-search algorithm. The findings indicate that the LightGBM algorithm's classification prediction outperforms other methods when applied to multidimensional datasets, achieving an error rate of 19.9% and an accuracy of 80.1%.

Based on the explanation regarding oversampling presented above, the aim of this research is to compare several oversampling techniques to overcome the data imbalance problem in peer-to-peer lending datasets for predicting the risk of default.

## METHODS

This research went through several stages which can be seen in Figure 1.

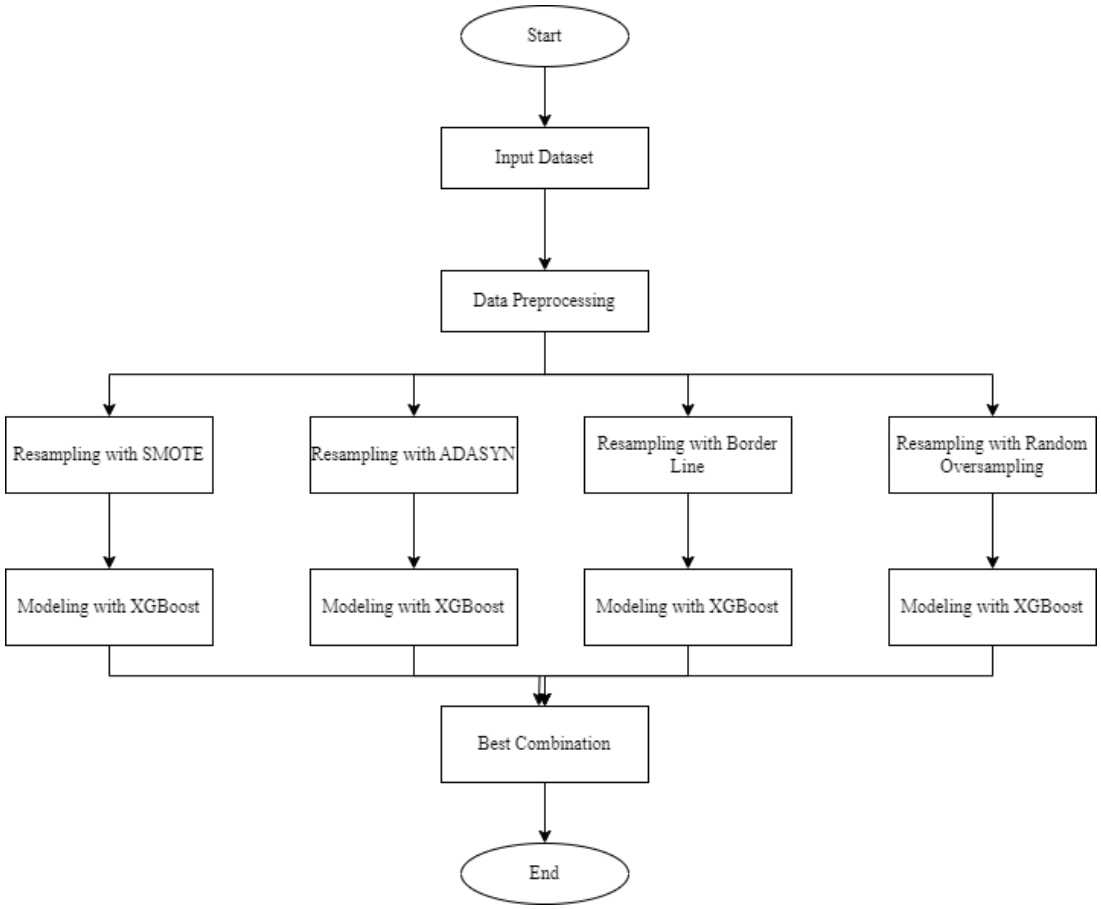


Figure 1. Research flowchart

### Data understanding

This research uses a dataset on the Kaggle dataset platform which contains loan history data from the Lending Club platform for the period 2013 to 2018. This data set includes a total of 1961527 rows of data and 18 variable columns, with one of the variables being the interest variable, namely the lend_data_description variable (variable y). This variable has four different unique values, namely "Completed," "Chargeoff," "Current," and "Defaulted". The value of the lend_data_description variable is the main reference that will be used during prediction and modeling.

### Data preprocessing

Data preprocessing is one of the stages in data mining which includes the preparation and transformation of data into a form that is in accordance that the required procedures [21]. This research implements several data preprocessing techniques, including the following:

a. Checking and handling missing values. Checking for and dealing with missing values is a crucial step in data analysis to ensure the reliability and accuracy of the results. Missing data exist for several reasons, such as survey filters, interviewer mistakes, or anonymization purposes [22]. The first step is to identify where the empty values are located in the dataset. This can be done by looking for signs such as NaN (Not a Number), null, or other markers that indicate empty data. After that, data visualization can help to understand the distribution pattern of missing values, guiding next steps. Analyzing the causes of missing data is an important next step to understand whether the gaps arise due to data collection errors, inappropriate formatting, or indeed reflect relevant gaps such as measurement absences.
b. Checking and handling duplicate data. Checking and handling duplicate data is a crucial process in data analysis to ensure the accuracy and reliability of analysis results. The first step involves

identifying whether there are any duplicate entries within the dataset. Duplicate data refer to entries that have the same values for each attribute or column. For instance, in Python, methods like `.duplicated()` can be utilized to identify duplicate entries. Once identified, duplicate data can be handled by either removing them from the dataset or consolidating them based on specific criteria, depending on the context of the analysis. Removing duplicate data helps to maintain the integrity of the analysis and prevents skewed results that may arise from redundant information. Additionally, it is essential to document the process of checking and handling duplicate data transparently as part of the data analysis workflow. This documentation ensures clarity and reproducibility of the analysis process for future reference.

c. Checking and handling outlier data. From the checks carried out, an outlier was detected in the ammount_borrower variable. Outlier handling is done using z-score. Outliers are data points that significantly deviate from the rest of the dataset. Handling outliers typically involves employing statistical methods such as z-score. Z-score, also known as standard score, quantifies how many standard deviations a data point is away from the mean of the dataset. It is calculated by subtracting the mean from the data point and then dividing by the standard deviation. A threshold value, often set at 3 or -3, is then used to identify outliers. Data points with a z-score beyond this threshold are considered outliers.
d. One Hot Encoding. One hot encoding is a data processing technique used to convert categorical data into a binary vector with a length equal to the number of possible categories (labels) in the data. One hot encoding is particularly useful when dealing with categorical variables with no ordinal relationship between categories, as it prevents the model from interpreting ordinality where there is none. However, it can lead to a significant increase in the dimensionality of the dataset, especially if there are many unique categories in a variable. Therefore, it is important to consider the trade-offs between computational complexity and the benefits of accurately representing categorical data.
e. Encoding Labels. Label encoding is a data processing technique used to change the value of a column (object) in categorical data into a whole number (integer). In label encoding, each unique category or label in a categorical variable is assigned a unique integer. This conversion facilitates the use of categorical data in machine learning algorithms that require numerical input.

**Resampling techniques**

Resampling methods are statistical techniques that utilize sample data for statistical inference without necessitating parametric assumptions, which can be challenging to validate in practical applications [23]. This research will consider and test five different oversampling techniques, namely SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), BorderLine, and Random Oversampling.

SMOTE (Synthetic Minority Oversampling Technique) is a resampling method used in data processing to handle class imbalance. SMOTE created extra training data by performing certain operations onreal data [24]. This technique synthetically generates new samples for the minority class by creating synthetic samples between existing data points in the minority class. This helps increase the representation of minority classes in the dataset, thereby improving the model's performance in predicting minority classes.

ADASYN (Adaptive Synthetic Sampling) is a resampling method used to handle class imbalance in data. In contrast to SMOTE, ADASYN considers the level of difficulty when classifying each minority sample and generates synthetic samples with weights adjusted based on that level of difficulty. The core concept behind ADASYN involves employing a weighted distribution for various examples within the minority class based on their difficulty level in learning. This approach entails generating more synthetic data for minority class examples that pose greater challenges in learning compared to those that are relatively easier to learn [25]. This helps increase the representation of minority classes that are difficult to predict, thereby improving the model's performance to better predict minority classes.

Borderline-SMOTE is a resampling method used to deal with class imbalance in data. This method is a modification of SMOTE which considers minority samples that approach the decision boundary between the majority and minority classes. Borderline SMOTE is an improved oversampling algorithm based on SMOTE, which uses only a few class samples on the border to combine new samples, thus improving the sample category distribution [26]. Borderline-SMOTE selectively generates synthetic samples only for minority samples located around the decision boundary, which is considered more difficult to predict. This

helps increase the representation of minority classes, which is important for improving the model's performance in better predicting minority classes.

Random oversampling is a resampling method used to deal with class imbalance in data. This method works by randomly adding samples from the minority class so that the number is the same as the majority class. In random oversampling, data from the minority class is replicated randomly to a certain extent [27]. By adding samples randomly, random oversampling increases the representation of the minority class in the data set, which helps improve the performance of the model to better predict the minority class.

**Modeling**

This research uses the XGBoost classification algorithm as the main basis for data analysis and making prediction models. XGBoost used widely by data scientists to achieve state-of-the-art results on many machine learning challenges [28]. XGBoost (Extreme Gradient Boosting) is a popular algorithm in the world of machine learning, especially in the context of classification and regression. Extreme gradient boosting (XGBoost), which based on a gradient boosting tree can play a powerful role in gradient enhancement [29]. The advantage of XGBoost lies in its ability to overcome data imbalance problems and improve prediction performance. Using the XGBoost algorithm, this research aims to optimize default predictions on peer-to-peer lending datasets, using oversampling as the main method to overcome data imbalance. An illustration of how the XGBoost algorithm works can be seen in Figure 2.

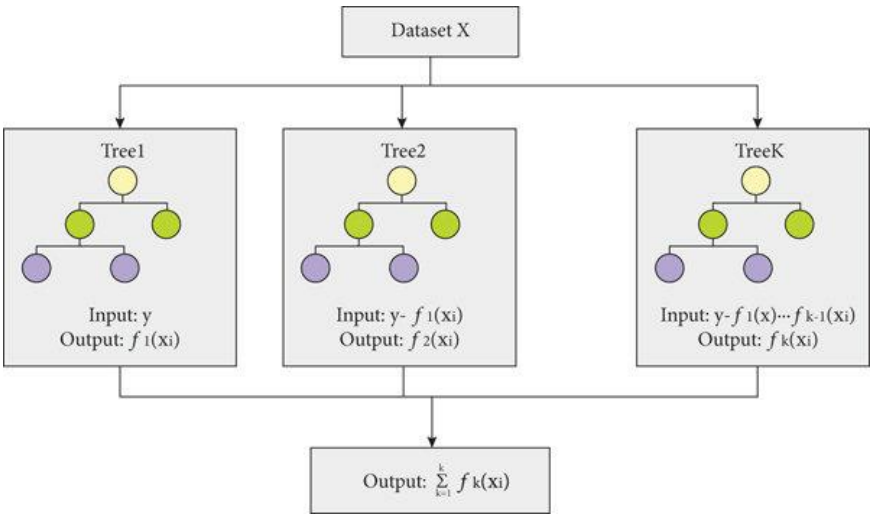


Figure 2. XGBoost algorithm workflow [30]

**Model evaluation**

Model evaluation in this research will be based on the accuracy obtained in the training and testing stages. Training accuracy reflects the extent to which the model can understand patterns in the training data, while testing accuracy measures the extent to which the model can apply learned knowledge to data the model has never seen before. A comparison between training accuracy and testing accuracy will help determine whether the model may be overfitting (strong in training, weak in testing) or underfitting (weak in both stages). Therefore, evaluating a model can help determine the quality of predictions and determine whether improvements or adjustments need to be made to the machine learning process.

## RESULTS AND DISCUSSIONS

**Result of data preprocessing**

Data preprocessing is one of the stages in data mining that includes the preparation and transformation of data into a form that is in accordance with the required procedures. The following are the results of the data preprocessing that has been carried out:

a. After going through the checking process, the dataset has no missing values and has complete data for each row of data.
b. The dataset owned is known to have unique data rows and has no duplicates.
c. From the checks carried out, an outlier was detected in the ammount_borrower variable. Outlier handling is done using z-score.

d. One hot encoding is performed on the "listing_list" column and adds a total of 14 new columns and removes the "listing_list" column.
e. Label encoding is done using the lend_data_description column which initially has a total of 4 value categories, namely "Completed," "Chargeoff," "Current," and "Defaulted" into numeric data values 0 and 1 with "Completed" and "Current" changed to the value 1 and “Chargeoff” and “Deafulted” are changed to value 0. Additionally, label encoding is also carried out on the “grade” column which has a total of 7 unique values, namely values in the A-G categories become columns with values 1-7.

**Result of resampling data**

Data resampling is carried out by implementing several data resampling techniques, including SMOTE, ADASYN, Border Line, and Random Oversampling. Resampling is carried out before the model is built to balance the existing data, namely data on customers who failed to make payments and customers who succeeded or are still making payments.

The following is a comparison of class 0 and class 1 on the data variables in the p2p lending dataset that is owned. A comparison of the numbers of class 0 and class 1 can be seen in Figure 2.

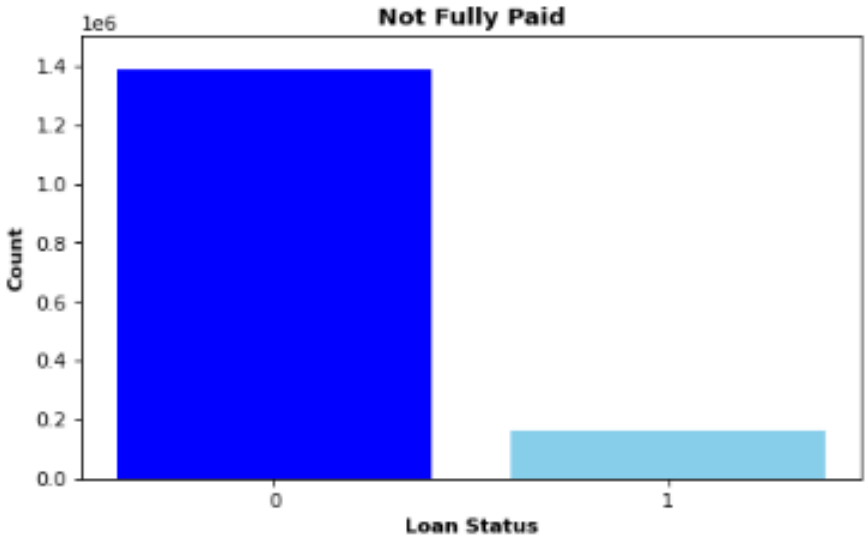


Figure 2. Comparison of the numbers of class 0 and class 1 before resampling

From Figure 2, it is known that the number of class 0 and class 1 in variable y is imbalanced in number. Therefore, resampling must be performed to overcome this problem. To overcome this problem, this research uses SMOTE, ADASYN, Border Libe, and Random Oversampling data resampling techniques. The results of data resampling data can be seen in Figure 3, Figure 4, Figure 5, and Figure 6.

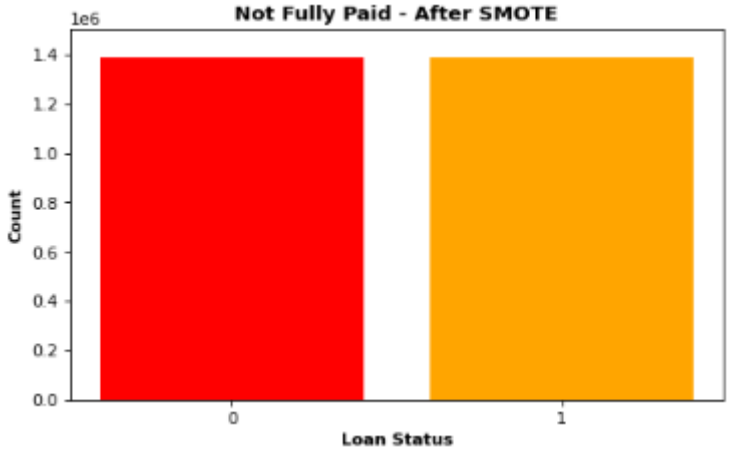

Figure 3. Visualization data after resampling and before resampling with SMOTE

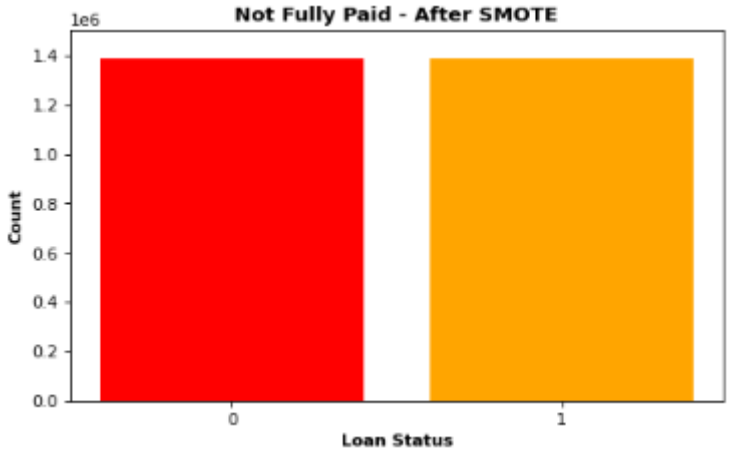


Figure 4. Visualization data after resampling and before resampling with ADASYN

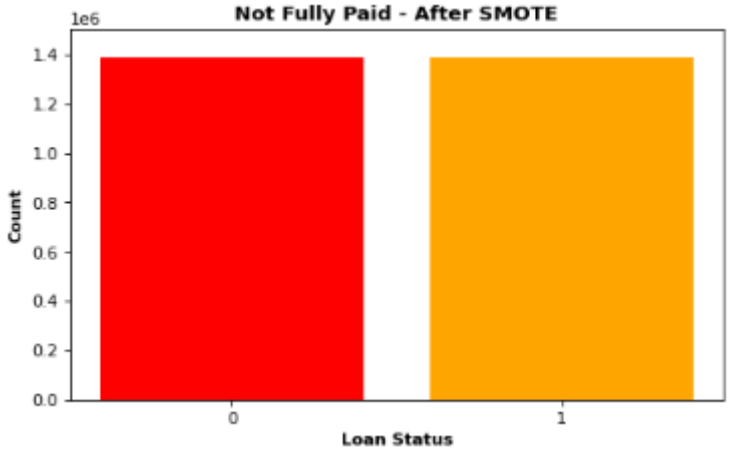


Figure 5. Visualization data after resampling and before resampling with border line

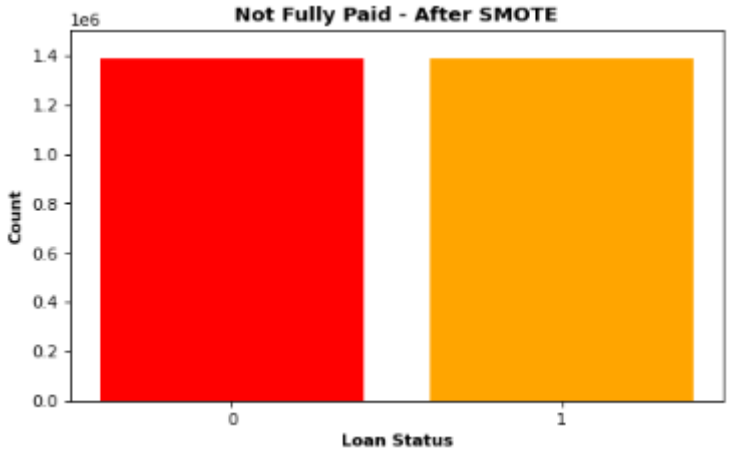


Figure 6. Visualization data after resampling and before resampling with random oversampling

From Figure 2, it can be seen that there is quite a large data imbalance between data labeled 0 and data labeled 1. By using data resampling techniques, fake data will be created to balance the number of class labels 0 and class labels 1. The result of resampling data can be seen in Figure 3, Figure 4, Figure 5, Figure 6.

**Model evaluation**

From the model that has been built, the base model that has been created has produced quite good performance. The base model created succeeded in reaching an accuracy of 0.999372. The performance of the model is measured using the metrics accuracy, recall, precision and f1-score. Evaluation of the base model built can be seen in Table 1.

**Table 1.** Base model evaluation

| | |
|---|---|
| Training Accuracy | 0.99937 |
| Testing Accuracy | 0.99991 |

The model built by Telang has good performance. However, the variable x used in the model is still unbalanced, so the model performance can still be optimized by using resampling techniques to handle unbalanced data. The resampling techniques used in this research are SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), BorderLine, and Random Oversampling. A comparison of model performance using each resampling technique can be seen in Table 2.

**Table 2.** Evaluation of base model + resampling technique

| Resampling Technique | Training Accuracy | Testing Accuracy |
|---|---|---|
| SMOTE | 0.99987 | 0.99933 |
| ADASYN | 0.99986 | 0.99932 |
| Border Line | 0.99988 | 0.99932 |
| Random Oversampling | 0.99964 | 0.99926 |

From Table 1, it is known that the combination of the XGBoost model with the resampling techniques used produces good and consistent model performance. Each combination has a high accuracy score with a slight difference in accuracy. The highest training accuracy was achieved by the combination of the XGBoost model with the Boerder Line resampling technique with a training accuracy of 0.99988 and the combination of the XGBoost model with the SMOTE resampling technique. In accuracy testing, the combination with the highest accuracy score was achieved by a combination of the XGBoost model with the SMOTE resampling technique.

Compared with the testing accuracy produced by the base model, the testing accuracy produced by the combination of XGBoost with the resampling technique is slightly lower. However, the training accuracy produced by the combination of XGBoost with resampling techniques is mostly greater than the training accuracy of the XGBoost base model. This indicates that the combination of the XGBoost model with the resampling technique is able to understand and study the training data better than the base XGBoost model.

## CONCLUSION

The combination of the XGBoost model with the resampling techniques used produces good and consistent model performance. Each combination has a high accuracy score with a slight difference in accuracy. The highest training accuracy was achieved by the combination of the XGBoost model with the Boerder Line resampling technique with a training accuracy of 0.99988 and the combination of the XGBoost model with the SMOTE resampling technique. In accuracy testing, the combination with the highest accuracy score was achieved by a combination of the XGBoost model with the SMOTE resampling technique.

For future research, researchers suggest comparing different resampling techniques using a more complex dataset than the dataset used in this study.

## REFERENCES


[1] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[2] J. Ren, Y. Wang, Y. Cheung, X.-Z. Gao, and X. Guo, "Grouping-based Oversampling in Kernel Space for Imbalanced Data Classification," *Pattern Recognit.*, vol. 133, p. 108992, Jan. 2023, doi: 10.1016/j.patcog.2022.108992.

[3] K. Niu, Z. Zhang, Y. Liu, and R. Li, "Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending," *Inf. Sci. (Ny).*, vol. 536, pp. 120–134, Oct. 2020, doi: 10.1016/j.ins.2020.05.040.

[4] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination

of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, Sep. 2020, doi: 10.52465/joscex.v1i1.5.
[5] S. Chatterjee and Y.-C. Byun, "Highly imbalanced fault classification of wind turbines using data resampling and hybrid ensemble method approach," *Eng. Appl. Artif. Intell.*, vol. 126, p. 107104, Nov. 2023, doi: 10.1016/j.engappai.2023.107104.
[6] E. Artigao, S. Martín-Martínez, A. Honrubia-Escribano, and E. Gómez-Lázaro, "Wind turbine reliability: A comprehensive review towards effective condition monitoring development," *Appl. Energy*, vol. 228, pp. 1569–1583, Oct. 2018, doi: 10.1016/j.apenergy.2018.07.037.
[7] Y. Liu *et al.*, "Imbalanced data classification: Using transfer learning and active sampling," *Eng. Appl. Artif. Intell.*, vol. 117, p. 105621, Jan. 2023, doi: 10.1016/j.engappai.2022.105621.
[8] Q. Gu, J. Tian, X. Li, and S. Jiang, "A novel Random Forest integrated model for imbalanced data classification problem," *Knowledge-Based Syst.*, vol. 250, p. 109050, Aug. 2022, doi: 10.1016/j.knosys.2022.109050.
[9] P. S. Sundari and M. Khafidz Putra, "Optimization house price prediction model using gradient boosted regression trees (GBRT) and xgboost algorithm," *J. Student Res. Explor.*, vol. 2, no. 1, Sep. 2023, doi: 10.52465/josre.v2i1.176.
[10] R. Rofik, R. Aulia, K. Musaadah, S. S. F. Ardyani, and A. A. Hakim, "Optimization of Credit Scoring Model Using Stacking Ensemble Learning and Oversampling Techniques," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, Dec. 2023, doi: 10.52465/joiser.v2i1.203.
[11] H. Kaur, H. S. Pannu, and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–36, Jul. 2020, doi: 10.1145/3343440.
[12] S. C. and J. V. Devasia, "Peer to Peer Lending: Risk Prediction Using Machine Learning on An Imbalanced Dataset," in *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, IEEE, Aug. 2022, pp. 511–519. doi: 10.1109/ICICICT54557.2022.9917708.
[13] L. E. Boiko Ferreira, J. P. Barddal, H. M. Gomes, and F. Enembreck, "Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Imbalanced Learning Techniques," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, Nov. 2017, pp. 175–181. doi: 10.1109/ICTAI.2017.00037.
[14] Y. Yuan, J. Wei, H. Huang, W. Jiao, J. Wang, and H. Chen, "Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring," *Eng. Appl. Artif. Intell.*, vol. 126, p. 106911, Nov. 2023, doi: 10.1016/j.engappai.2023.106911.
[15] Y. A. Sir and A. H. H. Soepranoto, "Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas," *J. Komput. dan Inform.*, vol. 10, no. 1, pp. 31–38, Mar. 2022, doi: 10.35508/jicon.v10i1.6554.
[16] A. Amiruddin, P. N. H. Suryani, S. D. Santoso, and M. Y. B. Setiadji, "Utilizing Reverse Engineering Technique for A Malware Analysis Model," *Sci. J. Informatics*, vol. 8, no. 2, pp. 222–229, Nov. 2021, doi: 10.15294/sji.v8i2.24755.
[17] W. J. XING Yulong SHANGGUAN Wei, PENG Cong, ZHU Linfu, "Track circuit fault diagnosis method for massive imbalanced data," *China Safety Science Journal*, vol. 32, no. 5. pp. 112–118. [Online]. Available: http://www.cssjj.com.cn
[18] P. Kaur and A. Gosain, "Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise," 2018, pp. 23–30. doi: 10.1007/978-981-10-6602-3_3.
[19] M. A. Muslim *et al.*, "New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning," *Intell. Syst. with Appl.*, vol. 18, no. December 2022, p. 200204, 2023, doi: 10.1016/j.iswa.2023.200204.
[20] A. López-García, O. Blasco-Blasco, M. Liern-García, and S. E. Parada-Rico, "Early detection of students' failure using Machine Learning techniques," *Oper. Res. Perspect.*, vol. 11, p. 100292, Dec. 2023, doi: 10.1016/j.orp.2023.100292.
[21] W. Bhaya, "Review of Data Preprocessing Techniques in Data Mining," *J. Eng. Appl. Sci.*, vol. 12, pp. 4102–4107, Sep. 2017, doi: 10.3923/jeasci.2017.4102.4107.
[22] Y. Liu, B. Li, S. Yang, and Z. Li, "Handling missing values and imbalanced classes in machine learning to predict consumer preference: Demonstrations and comparisons to prominent methods," *Expert Syst. Appl.*, vol. 237, p. 121694, Mar. 2024, doi: 10.1016/j.eswa.2023.121694.
[23] M. R. Chernick, "Resampling methods," *WIREs Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 255–262, May 2012, doi: 10.1002/widm.1054.
[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority

Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[25] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, Jun. 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.
[26] Y. Sun *et al.*, "Borderline SMOTE Algorithm and Feature Selection-Based Network Anomalies Detection Strategy," *Energies*, vol. 15, no. 13, p. 4751, Jun. 2022, doi: 10.3390/en15134751.
[27] A. Ghazikhani, H. S. Yazdi, and R. Monsefi, "Class imbalance handling using wrapper-based random oversampling," in *20th Iranian Conference on Electrical Engineering (ICEE2012)*, IEEE, May 2012, pp. 611–616. doi: 10.1109/IranianCEE.2012.6292428.
[28] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[29] Y. Qiu, J. Zhou, M. Khandelwal, H. Yang, P. Yang, and C. Li, "Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration," *Eng. Comput.*, vol. 38, no. S5, pp. 4145–4162, Dec. 2022, doi: 10.1007/s00366-021-01393-9.
[30] J.-J. Liu and J.-C. Liu, "Permeability Predictions for Tight Sandstone Reservoir Using Explainable Machine Learning and Particle Swarm Optimization," *Geofluids*, vol. 2022, pp. 1–15, Jan. 2022, doi: 10.1155/2022/2263329.