Comparison of Ensemble Learning Methods in Classifying Unbalanced Data on the Bank Marketing Dataset

Yunia Hasnataeni[1*], Kusman Sadik[2], Agus M Soleh[3], and Reka Agustia Astari[4]
[1,2,3,4]Department of Statistics, IPB University, Bogor, Indonesia
*yunia_hasnataeni@apps.ipb.ac.id

Received: 7 Juni 2024
Revised: 24 September 2024
Accepted: 3 January 2025

ABSTRACT – **The banking industry is experiencing rapid growth, particularly in telemarketing strategies to increase product and service sales. Despite widespread use, these strategies need higher success rates due to data imbalance, where fewer customers accept offers than those who reject them. This study evaluates machine learning algorithms, including Random Forest, Gradient Boosting, Extra Trees, and AdaBoost, without and handling imbalanced data using the Random Over-Sampling Examples (ROSE) method. The evaluation covers accuracy, precision, recall, F1-score, and AUC of the ROC curve. Results indicate that Random Forest and AdaBoost consistently perform well, with Random Forest maintaining a high accuracy of 91.00% after handling imbalanced data. Gradient Boosting and Extra Trees improve precision in post-oversampling. All models exhibit high AUC values, close to 0.94, demonstrating excellent differentiation between positive and negative classes. The study concludes that addressing data imbalance enhances model performance, making these models suitable for effective telemarketing strategies in the banking sector.**

Keywords – **Imbalanced Data, Random Forest, Extra Tree, Gradient Boosting, AdaBoost.**

I. INTRODUCTION

The banking industry is experiencing rapid growth, particularly in telemarketing strategies to increase product and service sales. Despite the widespread use of these strategies, their success rates still need to improve. One of the main challenges causing this issue is data imbalance. The number of customers willing to accept offers is often much smaller than those who reject them. This creates a gap in pattern recognition, especially when identifying potential customers. Imbalanced class distribution in a dataset occurs when one class, often the more significant positive or minority class, is underrepresented. Simply put, the number of minority class examples is much smaller than the majority. Rare examples that appear infrequently often lead to predictions of rare events, undetected, ignored, or viewed as noise or outliers. This leads to more classification errors for the minority class than the more common [1].

Researchers have developed numerous methods to address unbalanced datasets, one of which involves using resampling methods on the available data. Resampling entails the repetitive extraction of samples from the original dataset, including oversampling, which involves repeatedly drawing samples from the minority class, and undersampling, which involves randomly selecting samples from the majority class [2]. We will use the Random Over-Sampling Examples (ROSE) technique in this research to effectively address the challenges posed by data imbalance. ROSE stands out for its unique approach to mitigating the impact of class imbalance across both model estimation and evaluation stages. Unlike conventional methods, ROSE integrates ROC curves to gauge classifier performance, transcending the limitations of accuracy metrics alone. Additionally, ROSE offers the flexibility to choose between bootstrap variants or cross-validation as estimation methods, providing a more sophisticated toolkit for effectively and accurately identifying minority classes. This advanced framework holds significant promise in enhancing the efficacy of telemarketing strategies within the banking industry, mitigating inherent biases, and improving the predictive accuracy of models deployed in telephone marketing campaigns [3].

Machine learning models perform better on imbalanced data than experimental tables and other statistical methods. However, the issue of large and imbalanced class sizes makes improving model accuracy difficult, leaving significant potential for further research. In this context, ensemble learning is a practical and continually evolving approach. Ensemble learning involves integrating different models to enhance accuracy [4]. Ensemble learning enhances predictive accuracy by minimizing noise or errors between observed and predicted data. Three categories typically classify ensemble methods: bootstrap aggregation (bagging), boosting, and stacking. These categories aim to align their predictions with observations by mitigating model variance, bias, or both. The primary distinction is that bagging and boosting generally employ homogeneous models, whereas stacking combines heterogeneous models [4]. In this study, we will compare bagging and boosting performance and select two methods from each category for analysis and comparison.

In this research endeavour, we will scrutinise the performance of bagging and boosting methods, selecting two methods from each category for analysis and comparison. Specifically, we will employ bagging-based methods, such as Random Forest and Extra Tree, alongside boosting-based methods, including Gradient Boosting and AdaBoost. Random Forest and Gradient Boosting are renowned for their ability to yield highly accurate predictions. At the same time, Extra Trees and AdaBoost exhibit superior training speed, offering a significant advantage in computational efficiency. Previous research findings provide the rationale for selecting these methods. Ampomah [5] and Nguyen [6] have

conducted studies that consistently demonstrate superior accuracy for Random Forests and Gradient Boosting, and favor Extra Trees and AdaBoost due to their faster training speeds. This strategic selection aims to leverage the strengths of each method to improve the ensemble learning approach's overall predictive performance in addressing the challenges posed by data imbalance in banking telemarketing campaigns.

II. LITERATURE REVIEW

A. Ensemble Learning

Ensemble learning works by combining various machine learning models. Each model has different error levels on the data samples. Through strategic combination, this collection of models can complement each other and correct their respective errors, thereby reducing the total error [4]. Bagging and boosting are the ensemble learning methods used in this research.

1) Bagging, introduced by Breiman in 1996, stands as one of the pioneering ensemble algorithms. The name "Bagging" originates from bootstrap aggregation, a statistical technique that utilizes random sampling with replacement. In statistics, bootstrapping involves evaluating the accuracy of sample estimates and serves as a tool for developing hypothesis tests [7]. Bagging, despite being one of the most straightforward ensemble algorithms, excels in achieving high performance due to its strong generalization capability. Bagging achieves this generalization aspect by generating bootstrap replicas of the training dataset. Essentially, we randomly select various subsets of the training data from the complete training dataset (with replacement), and use each subset to train a distinct model. We then aggregate the predictions from all trained models to derive the final output, thereby leveraging their collective insights [4]. Bagging involves applying the bootstrap method to high-variance machine learning problems. An example is the Random Forest model, which combines bagging and decision trees [6].
2) Boosting encompasses a collection of algorithms designed to enhance the performance of weaker machine-learning models through weighted averaging. Unlike bagging, which combines independently running models at the end, boosting follows a sequential process in which each new model incrementally corrects its predecessor's errors [6]. This approach avoids simplifying assumptions during training, thus increasing model complexity. Boosting is particularly useful when classifiers encounter underfitting, where the model fails to capture the data's underlying patterns, an issue that can occur in bagging models. With its gradual and iterative training, boosting systems learn more intricate data patterns through their sequential error correction process. Additionally, boosting is effective in dealing with imbalanced data problems [4].

B. Random Forest

Random Forest is an ensemble learning technique for classification that uses multiple decision trees, each built independently from different subsets of the data, to make decisions based on the majority vote from all trees [8]. The Random Forest randomly samples the training data with replacement, then averages the results. These sub-trees function independently, without any interdependence. In addition to using different data subsets for each tree, Random Forest differentiates itself in how it constructs the trees. In traditional decision trees, the best choices for all variables are made at each node. The goal is to lower entropy by splitting the dataset linked to the parent node into smaller pieces. In contrast, Random Forest adopts a different approach by randomly selecting the split point for each node from the best split points within a subset of predictors. This random selection aids Random Forest in mitigating overfitting, a prevalent concern with individual decision trees that delve deeply into the dataset [9].

C. Extra Tree

Extra Tree creates a collection of decision trees that do not undergo traditional top-down pruning. When splitting tree nodes, this essentially entails randomizing variables and robustly selecting split points. In extreme cases, this creates entirely random trees with structures that do not depend on the output values of the training samples [5]. The Extra Tree algorithm consists of several decision trees, each having a sequence of decision nodes similar to a tree structure. Based on this sequence, the tree branches out into various branches until it reaches the end (leaf nodes). The leaf nodes derive the prediction result from each decision tree, and several decision trees combine their final results for prediction [10].

D. Gradient Boosting

Gradient boosting is a decision tree-based ensemble learning method suitable for classification tasks. It operates in a sequential manner by progressively incorporating weak predictors into the ensemble, aiming to rectify previous errors. In essence, the ensemble concept involves amalgamating decisions from diverse machine learning techniques, ultimately predicting the class based on the majority consensus. The process of gradient boosting begins by constructing an initial classification tree and iteratively refining subsequent trees through error minimization efforts [11]. The gradient boosting approach employs the descent of gradients to minimize the model's loss function by incorporating weak learners. The model prioritizes misclassified observations by training on residuals. Gradient optimization techniques guide the relative contribution of each weak learner to the final prediction, aiming to reduce the overall errors of more robust and accurate learners [12].

E. AdaBoost

AdaBoost, one of the earliest successful boosting algorithms in classification, played a crucial role in advancing our understanding of ensemble boosting techniques. It operates by incrementally integrating decision trees with single splits, prioritizing samples that the previous model misclassified. The main objective of AdaBoost is to identify the optimal data split, known as the best stump, in each iteration, thus minimizing overall errors. More accurate stumps receive greater weight after the training phase. When presented with a new instance, each stump casts a weighted vote, and a majority vote determines the class label. This methodology aims to alleviate bias rather than variance, although AdaBoost is prone to overfitting due to its sensitivity to noise and outliers [4].

F. Random Over Sampling Examples

In the literature, two widely recognized oversampling techniques are ROSE (Random Over-Sampling Examples) and SMOTE (Synthetic Minority Over-Sampling Technique) [13]. Among these, ROSE is a prominent technique for addressing class imbalance. ROSE diverges from traditional methods by revitalizing the impact of class imbalance not only during model estimation but also in model evaluation. Unlike conventional approaches that rely solely on accuracy metrics, ROSE employs the ROC curve to assess classifier performance, offering a more comprehensive assessment. Additionally, ROSE provides flexibility in selecting estimation methods, allowing for bootstrap variants or cross-validation. This advanced framework enables more effective and accurate identification of minority classes, making ROSE a superior choice for handling class imbalance [3].

G. Model Evaluation

To assess the performance of ensemble learning models, the evaluation criteria used are (1) accuracy, (2) precision, (3) recall, and (4) F1-score [14] and ROC curve [15].

Table 1 Confusion Matrix

		Prediction	
		Intrusion	Normal
Actual	Intrusion	True Positive ($\boldsymbol{tp}$)	False Negative ($\boldsymbol{fn}$)
	Normal	False Positive ($\boldsymbol{fp}$)	True Negative ($\boldsymbol{tn}$)

1) Accuracy is defined as the ratio of accurate predictions to the total dataset. A high accuracy suggests that the model generally makes correct predictions.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{tp} + \boldsymbol{tn}}{\boldsymbol{tp} + \boldsymbol{tn} + \boldsymbol{fp} + \boldsymbol{fn}} \quad (1)$$

2) Precision is defined as the ratio of accurate positive predictions to all positive predictions. A high precision implies that the model seldom generates false positive predictions.

$$\boldsymbol{Precision} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{fp}} \quad (2)$$

3) Recall is defined as the ratio of accurate positive data correctly identified by the model. A high recall indicates that the model rarely overlooks positive data.

$$\boldsymbol{Recall} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{fn}} \quad (3)$$

4) The F1-score is a measure that merges precision and recall to offer a more comprehensive assessment of model performance. A high F1-score suggests an effective equilibrium between precision and recall.

$$\boldsymbol{F1 - Score} = \frac{2 \times \boldsymbol{Presisi} \times \boldsymbol{Recall}}{\boldsymbol{Presisi} + \boldsymbol{Recall}} \quad (4)$$

5) The ROC curve (Receiver Operating Characteristic) is a tool used to assess the performance of classification systems. It displays a graph that compares sensitivity (TPR) on the y-axis and 1-specificity (FPR) on the x-axis to evaluate the overall performance of the model.

III. METHODOLOGY

A. Material and Data

This research uses the Bank Marketing dataset, which is publicly available on the UCI Machine Learning Repository. This dataset contains information about direct marketing campaigns via phone calls from a Portuguese banking institution. The dataset is divided into four parts:

1) The "bank additional full" dataset includes all campaigns with a total of 41,188 examples and 20 input variables.
2) The "bank additional" dataset is a 10% sub-sample of "bank additional full" (4,119 examples), randomly selected for more efficient analysis. This dataset also involves 20 input variables, maintaining consistency with the main data.
3) The "bank full" dataset includes all campaigns with 41,188 examples but with 17 input variables. This dataset is an earlier version with fewer variables but remains sorted by date, offering a valuable perspective.
4) The "bank" dataset is a 10% sub-sample of "bank full," randomly selected and also includes 17 input variables. Although it is an older version with fewer variables, this dataset allows for more computational algorithm testing.

For this research, the "bank additional full" dataset will be utilized, as it includes all features and offers a complete set of data. This decision is based on the comprehensive nature of the information it provides, which makes it easier to predict whether customers will subscribe to a term deposit (y variable).

Table 2 Bank Marketing Data Variables

Data	Variables	Type
Bank Customer	Age	Numerical
	Job	Categorical
	Marital	Categorical
	Education	Categorical
	Default	Categorical
	Housing	Categorical
	Loan	Categorical
Regarding the final contract of the current campaign	Contact	Categorical
	Month	Categorical
	Day_of_week	Categorical
	Duration	Numerical
Social and economic context	Employment variation rate (Emp.var.rate)	Numerical
	Consumer price index (Cons.price.idx)	Numerical
	Consumer confidence index (Cons.conf.idx)	Numerical
	Euro interbank offered rate for 3 months (Euribor3m)	Numerical
	Number of employed (Nr.employed)	Numerical
Other	Campaign	Numerical
	Pdays	Numerical
	Previous	Numerical
	Poutcome	Categorical
Target	Customer's decision to subscribe to a term deposit	Numerical

B. Methods

This research will undertake a series of steps to achieve the study's objectives. Figure 1 provides a detailed illustration of each stage of the process, offering a comprehensive overview of the methodology.

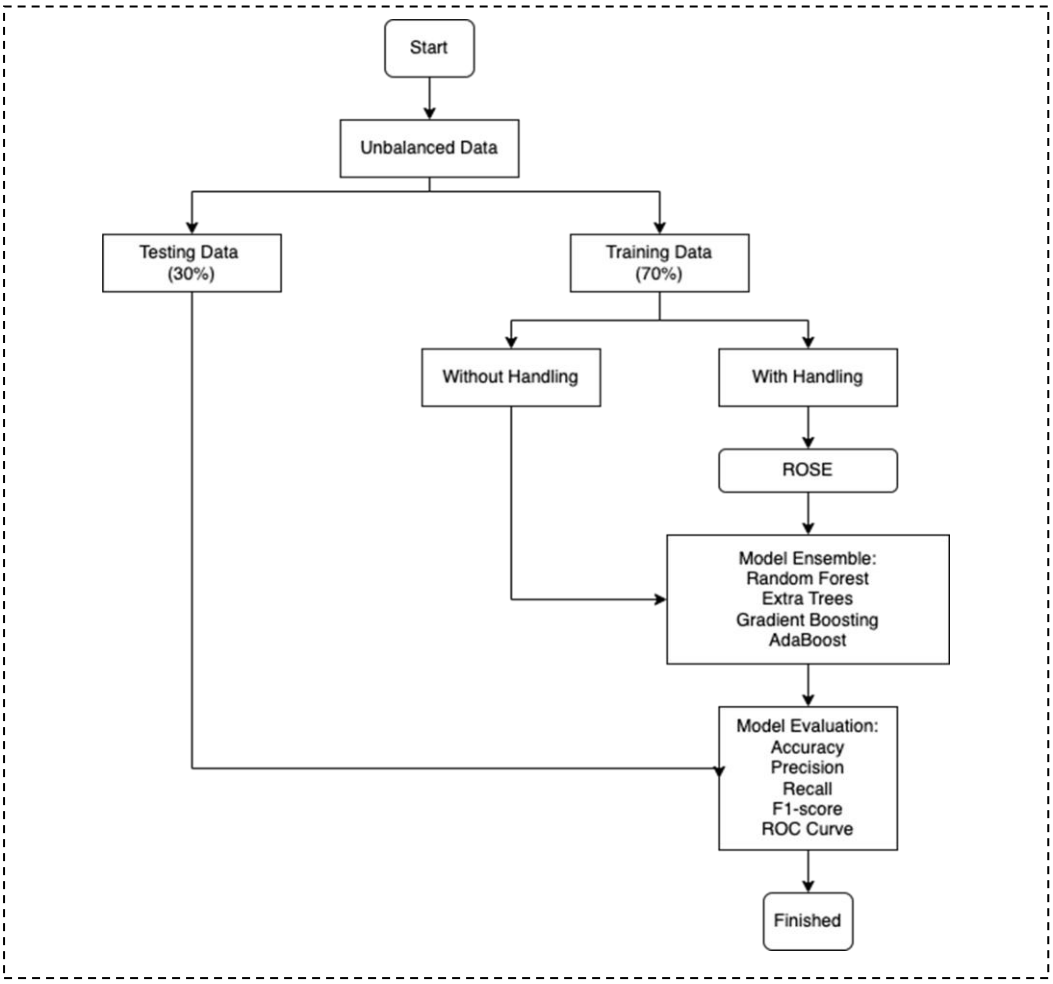

Figure 1 Flow Chart

To ensure the accuracy and reliability of the results, this research uses several stages of data analysis. These stages are described as follows:

1) Identifying unbalanced data.
   The dataset used in this study was thoroughly analyzed and determined to have a significant class imbalance, which required appropriate handling to ensure accurate modeling.
2) Data splitting.
   We systematically divided the dataset into two parts, 70% for training and the remaining 30% for testing, to ensure proper model training and evaluation.
3) Data exploration and Visualization.
   We conducted a comprehensive exploration and visualization of the target distribution using histograms, which clearly depicted the class imbalance within the dataset and highlighted the need for corrective measures.
4) Handling training data.
   The training data was methodically partitioned into two distinct groups; one group remained untreated to serve as a baseline. In contrast, the other group underwent treatment using the ROSE (Random Over-Sampling Examples) technique, which aimed to balance the minority class by augmenting its representation.
5) Ensemble models.
   We meticulously trained a diverse array of ensemble models, including Random Forest, Extra Trees, Gradient Boosting, and AdaBoost, on both the untreated data and the data processed with the ROSE technique to evaluate the impact of class balancing on model performance.
6) Model evaluation.
   We rigorously evaluated the performance of the trained models using a suite of metrics, including accuracy, precision, recall, F1-score, and ROC curves, to provide a comprehensive assessment of their ability to distinguish between majority and minority classes and to measure the overall effectiveness of the class balancing techniques employed.

IV. RESULTS AND DISCUSSIONS

A. Data Exploration

The subsequent Figure 2 depicts the outcomes of exploring the distribution of target classes, offering a comprehensive visualization of the dispersion.

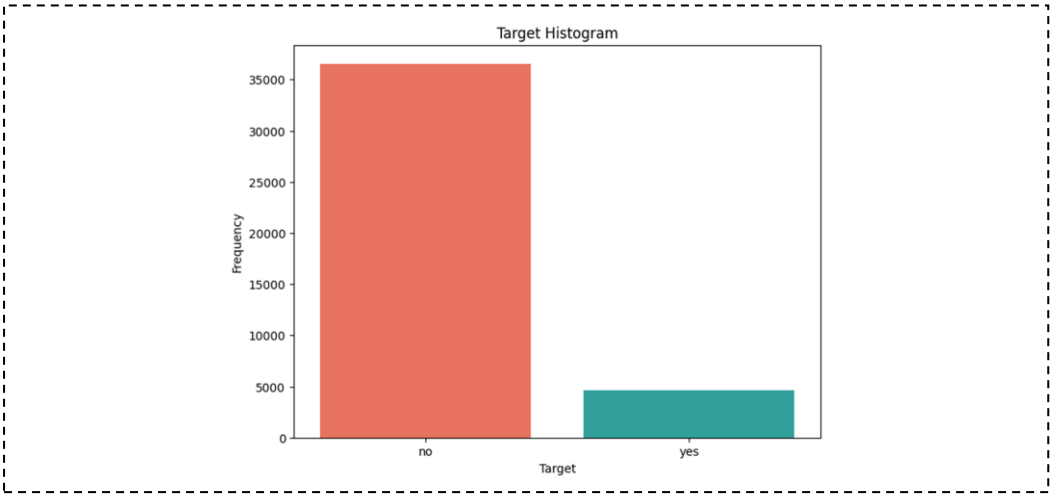

Figure 2 Target Histogram Before Handling

The distribution of target classes shows a significant imbalance, with the "no" class dominating the data at 88.75% and the "yes" class comprising only 11.25%. This imbalance can affect the performance of predictive models, as they tend to be biased towards the majority class. Additionally, correlation analysis among numerical features reveals strong relationships between some features, such as between "emp.var.rate" and "euribor3m" (correlation 0.97) and between "euribor3m" and "nr.employed" (correlation 0.95). We can address these strong correlations of redundancy by removing or combining features to simplify the model. Conversely, features like "age" and "duration" show very weak correlations with other features, indicating that they provide unique information that could be useful for the model. We can improve the model's overall performance by understanding the distribution of target classes and the correlations among features through data preprocessing steps.

B. Data Preprocessing

The data preprocessing process begins with separating the features (x) and the target (y), followed by using One-Hot Encoding to convert categorical variables into binary numerical variables. The target labels are converted into numerical values with 'no' as 0 and 'yes' as 1. The data is then split into training and testing sets in a 70:30 ratio, using stratification to maintain the imbalanced distribution of the target classes. Figure 3 illustrates the outcomes of exploring the distribution of target classes post-smoothing, providing a detailed insight into the dispersion.

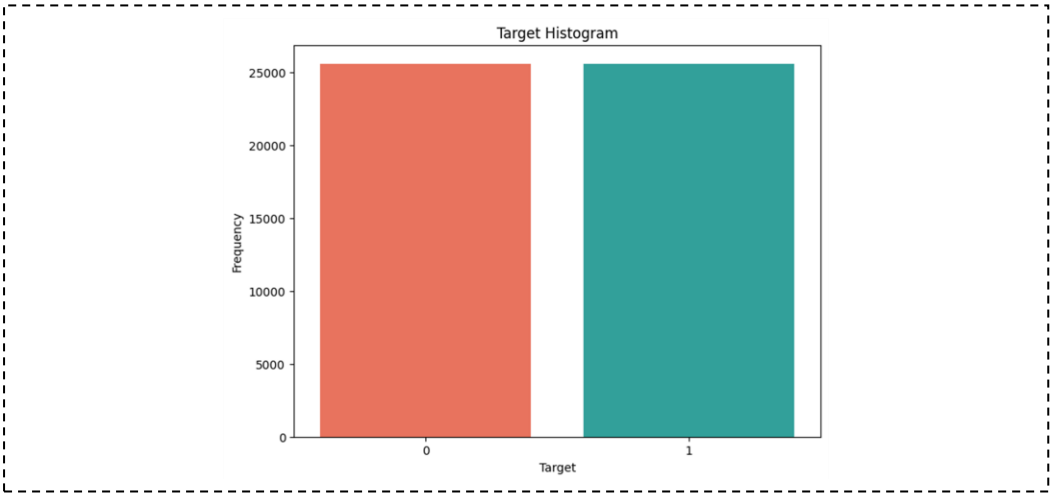

Figure 3 Target Histogram After Handling

The imbalance in the training data is corrected using the Random Over-sampling Examples technique to equalize the number of samples in both classes. After oversampling, the class distribution in the training set is balanced, with each

class comprising 25,583 samples. These steps guarantee that the data preprocessing enhances the model's accuracy and efficiency in tackling class imbalance, thereby reducing the risk of bias towards the majority class.

C. Models and Evaluation

This chapter will discuss the evaluation results of the models developed in this research. We used several machine learning algorithms, both without specific handling of imbalanced data and with handling using Random Over-Sampling Examples methods. The evaluated models include Random Forest, Gradient Boosting, Extra Trees, and AdaBoost. Evaluation results include accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) from the ROC curve. The following Table 3 shows the evaluation results of the models without handling imbalanced data.

Table 3 Evaluation Models without Handling

Models	Accuracy	Precision	Recall	F1-score
Random Forest	0.913167	0.904114	0.913167	0.906446
Extra Trees	0.902565	0.893970	0.902565	0.873706
Gradient Boosting	0.904022	0.890567	0.904022	0.893381
AdaBoost	0.911386	0.900289	0.911386	0.901857

Table 4 presents the evaluation results of the models with imbalanced data handling using the Random Over-Sampler Examples method.

Table 4 Evaluation Models with Handling

Models	Accuracy	Precision	Recall	F1-score
Random Forest	0.910011	0.908514	0.910011	0.909231
Extra Trees	0.826738	0.923789	0.826738	0.854209
Gradient Boosting	0.903536	0.889122	0.903536	0.891518
AdaBoost	0.870762	0.922774	0.870762	0.886762

After addressing the data imbalance, Random Forest continues demonstrating excellent performance with an accuracy of 91.00%. Although the accuracy of Gradient Boosting decreases to 82.67%, the precision of this model increases to 92.38%, indicating that this model is better at identifying positive classes after handling data imbalance. Extra Trees and AdaBoost improve precision and recall metrics after handling data imbalances. Figure 4 depicts the evaluation outcomes through ROC curves, presenting a graphical representation of the trade-off between true positive rate and false positive rate for each model.

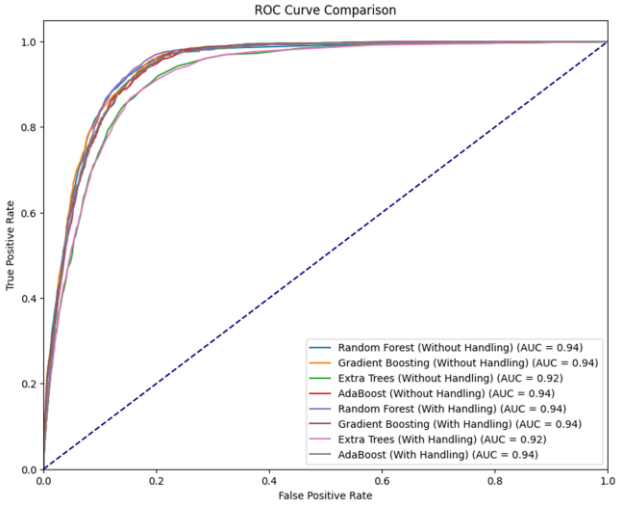

Figure 3 Target Histogram After Handling

Comparing the ROC curves among the evaluated models provides a clearer picture of the model's performance regarding the trade-off between true and false positive rates. From the ROC curves above, we can see that all models have high AUC values, approaching or equal to 0.94, indicating they are excellent at distinguishing between positive and negative classes. Specifically, Random Forest shows excellent ROC curves with and without handling imbalanced data with consistently high AUC values of 0.94. Additionally, AdaBoost also demonstrates excellent performance with an AUC value of 0.94. Without handling imbalanced data, Gradient Boosting and Extra Trees show AUC values of 0.94 and 0.92, respectively. However, after handling imbalanced data, Gradient Boosting and Extra Trees still show the same AUC values, namely 0.94 and 0.92, respectively, reaffirming the consistent ability of these models to distinguish between positive and negative classes.

V. CONCLUSIONS AND SUGGESTIONS

According to the evaluation results and discussions, Random Forest is the most consistent model, delivering the best performance with and without handling data imbalances. The consistently high accuracy values and other evaluation metrics support this claim. Meanwhile, AdaBoost also demonstrates excellent performance, especially in precision and F1- score, indicating its capability to generate accurate predictions. However, Gradient Boosting experiences a decrease in accuracy, and handling data imbalance improves the precision level of this model. Furthermore, handling data imbalances using oversampling methods contributes positively to improving precision and recall metrics, although the effects vary depending on the type of model used. This conclusion emphasizes the importance of appropriate data selection and handling when optimizing the performance of machine learning models on imbalanced datasets.

Based on the presented evaluation results, we can propose several recommendations for further research. Firstly, while Random Forest has demonstrated consistency and optimal performance in imbalanced datasets, it is crucial to explore alternative techniques to fully optimize the model's performance. One approach worth exploring is SMOTE (Synthetic Minority Over-sampling Technique), which has proven effective in handling class imbalances. Additionally, implementing ensemble methods specifically designed for imbalanced data, such as EasyEnsemble or BalancedRandomForest, could provide valuable insights into improving model performance. Therefore, further research can focus on understanding the effects of various techniques for handling imbalanced data on different machine learning models. This will aid in developing more effective and reliable models for applications in multiple fields, from healthcare to finance.

REFERENCES

[1] A. Ali, S. M. Shamsuddin and A. L. Ralescu, "Classification with class imbalance problem: a review," International Journal of Advance Soft Computing Applications, vol. 5, 2013.

[2] T. S. Amelia, M. N. S. Hasibuan and R. Pane, "Comparative analysis of resampling techniques on Machine Learning algorithm," Sinkron: Jurnal dan Penelitian Teknik Informatika journal, vol. 6, 2022.

[3] J. Zhang and L. Chen, "Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis," Computer Assisted Surgery, 2019.

[4] M. Pirizadeh, N. Alemohammad and M. Manthouri, "A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods," Journal of Petroleum Science and Engineering, vol. 198, 2021.

[5] E. K. Ampomah, Z. Qin and G. Nyame, "Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement," Information, vol. 11, 2020.

[6] K. A. Nguyen, W. Chen, B.-S. Lin and U. Seeboonruang, "Comparison of Ensemble Machine Learning Methods for Soil Erosion Pin Measurements," International Journal of Geo-Information, vol. 10, 2021.

[7] Efron, B. and Tibshirani, R. J., An introduction to the bootstrap, Boca Raton: CRC press, 1994.

[8] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan and J. García-Nieto, "Efficient Water Quality Prediction Using Supervised Machine Learning," Water, vol. 11, no. 11, 2019.

[9] W. Wang, G. Chakraborty and B. Chakraborty, "Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm," Applied Sciences, vol. 11, 2021.

[10] Z. Chu, J. Yu and A. Hamdulla, "Throughput Prediction based on ExtraTree for Stream Processing Tasks," Computer Science and Information Systems, 2018.

[11] S. E. Suryana, B. Warsito and Suparti, "Penerapan Gradient Boosting Dengan Hyperopt Untuk Memprediksi Keberhasilan Telemarketing Bank," Jurnal Gaussian, vol. 10, no. 4, pp. 617-623, 2021.

[12] J. Son and S. Yang, "A New Approach to Machine Learning Model Development for Prediction of Concrete Fatigue Life under Uniaxial Compression," Applied Sciences, vol. 12, no. 19, pp. 9766 (1-22), 2022.

[13] S. Demir and E. K. Şahin, "Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes," European Journal of Science and Technology, vol. 34, pp. 142-147, 2022.

[14] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. M. Nasir, N. Shaadan and S. W. Indratno, "Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data," Indonesian Journal of Electrical Engineering and Computer Science, vol. 29, no. 1, pp. 598-608, 2023.

[15] L. Qadrini, A. Seppewali and A. Aina, "Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial," Jurnal Inovasi Penelitian, vol. 2, pp. 1959-1966, 2021.