International Journal of Scientific Research and Management (IJSRM)
||Volume||12||Issue||04||Pages||494-501||2024||
Website: https://ijsrm.net ISSN (e): 2321-3418
DOI: 10.18535/ijsrm/v12i04.m03

Performance of Random Oversampling, Random Undersampling, and SMOTE-NC Methods in Handling Imbalanced Class in Classification Models

Andika Putri Ratnasari[1], Rizky Nur'aini[1]

[1]Universitas Negeri Yogyakarta, Faculty of Mathematics and Natural Sciences,
Colombo Road, Yogyakarta, Indonesia

Abstract:
One common challenge in classification modeling is the existence of imbalanced classes within the data. If the analysis continues with imbalanced classes, it is probable that the result will demonstrate inadequate performance when forecasting new data. Various approaches exist to rectify this class imbalance issue, such as random oversampling, random undersampling, and the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC). Each of these methods encompasses distinct techniques aimed at achieving balanced class distribution within the dataset. Comparison of classification performance on imbalanced classes handled by these three methods has never been carried out in previous research. Therefore, this study undertakes an evaluation of classification models (specifically Gradient Boosting, Random Forest, and Extremely Randomized Trees) in the context of imbalanced class data. The results of this research show that the random undersampling method used to balance the class distribution has the best performance on two classification models (Random Forest and Gradient Boosted Tree).

Keywords: Classification, Imbalanced Class, Random Oversampling, Random Undersampling, SMOTE-NC**.**

1. Introduction

Machine learning methods are divided into two, namely supervised learning and unsupervised learning. Supervised learning involves building a statistical model to predict or estimate output results based on one or more inputs. While unsupervised learning aims to understand the relationships and structure of data [1]. Classification modeling is included in supervised learning, where the algorithm uses one or more inputs to build a model and is then used to predict an output. The classification methods used in this research are Gradient Boosting (GB), Random Forest (RF), and Extremely Randomized Trees (Extra Trees).

Gradient boosting is a supervised learning technique based on decision trees. The GB algorithm creates a classification tree sequentially by minimizing the loss function [2]. Previous research has applied GB to analyze health data. The research results show that GB has better performance and is easier to interpret compared to neural networks and linear models [3].

In contrast to the GB algorithm which creates classification trees sequentially, the formation of classification trees in RF and Extra Trees is done individually or the formation of the next tree is not related to the tree that was formed previously. Both of these methods use majority voting to determine the prediction results. Determining the best splitting criteria in these three methods uses random selection of explanatory variables so that the classification trees formed are not correlated with each other. Apart from using random selection of variables, Extra Trees also uses random selection of cut points to determine the best splitting, this can make computing time on Extra Trees faster [4].

One of the problems often encountered in classification modeling is imbalanced data. Imbalanced data is data that has an unbalanced distribution of response variable classes, the number of one class is less or more than the number of other data classes [5]. Imbalanced class that is not resolved can affect the performance of the model used [6]. The data balancing methods used in this research are Random Oversampling, Random Undersampling, and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC).

Random Oversampling performs random replication on minority samples to balance the class distribution [7]. Meanwhile Random Undersampling used to balance the distribution of each class by randomly removing majority class samples [6]. SMOTE-NC is an oversampling technique that uses K-nearest neighbor characteristics in explanatory variables to produce synthetic data in the minority class [8]. Previous research used the oversampling method for classification. The use of the SMOTE-NC method to balance the class distribution in previous research was carried out on data from heart failure patients. The results of this study show that the heart failure patient data classification model improved the F1 score from 69.39% to 81.90% after class balancing with SMOTE-NC [9].

Based on previous research, classification modeling with balanced classes can improve model performance. Therefore, this research compares methods Random Oversampling, Random Undersampling, and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) on data with imbalanced classes. The data used is Telco customer churn data. Classification model performance (Gradient Boosting (GB), Random Forest (RF), and Extremely Randomized Trees (Extra Trees)) in modeling Telco customer churn data which has balanced class distribution compared to performance using accuracy, sensitivity, and specificity values.

2. Research Method

2.1 Data

The data used in this research is Telco customer churn data downloaded from Kaggle. There are 7043 observations in the Telco customer churn data. This data consists of 19 variables, 18 explanatory variables and one response variable. The response variable in this research is Churn (customers who stop using the service and still use Telco services). A description of the explanatory and response variables in this study is shown in Table 1.

Table1: Description of Variables in Telco Customer Churn Data

Variable Name	Description
Gender	Customer gender (male or female)
Partners	Customer has a partner or not (Yes, No)
Dependents	Customer has dependents or not (Yes, No)
Tenure	The number of months a customer has stayed with the company
Telephone Service	Customer has telephone service or not (Yes, No)
Multiple Lines	Customer has multiple lines or not (Yes, No, No phone service)
Internet Services	Customer internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Customer has online security or not (Yes, No, No internet service)
Online Backup	Customer has online backup or not (Yes, No, No internet service)
Device Protection	Customer has device protection or not (Yes, No, No internet service)
Tech Support	Customer has technical support or not (Yes, No, No internet service)
Stream TV	Customer has streaming TV or not (Yes, No, No internet service)
Streaming Movies	Customer has movie streaming or not (Yes, No, No internet service)
Contracts	Customer contract term (Month to month, One year, Two years)
Paperless Billing	Customer has paperless billing or not (Yes, No)
Payment Method	Customer payment method (Electronic check, Postal check, Bank transfer (automatic), Credit card (automatic))
Monthly Charges	The amount charged to customers each month
Total Charges	The total amount charged to customers
Churn	Whether customers churn or not (Yes or No)

2.2 Random Forest

The ensemble method is a learning method that combines prediction results from several individual models to obtain better performance (accuracy) results [10]. Random Forest is an ensemble method developed by Leo Breiman in 2001 [11]. The individual model used in RF is a classification/regression tree.

This method is a development of Bagging which aims to build trees that are more distinct and not correlated with each other [12]. The process of randomly selecting explanatory variables in RF reduces the correlation between the trees formed, thereby increasing prediction ability and being more efficient. Some of the advantages of RF are that it can overcome overfitting problems, is not sensitive to outliers, and can produce good accuracy [13]. The following are the classification stages using Random Forest [14].

1. Perform bootstrapping on training data.
2. Build a classification tree using bootstrapped data.
3. Choose the best splitting at node t using randomly selected independent variables $m \approx \sqrt{p}$ or $m \approx {}^{p}/_{3}$ with p is all of the independent variables in the data. The splitting selection process is repeated until the stopping criterion has been reached.
4. Determine the prediction results of a classification tree.
5. Steps 1-4 are repeated until b classification trees are obtained.
6. Determine the prediction results from RF by combining the prediction results from each classification tree using majority vote.

2.3 Gradient Boosting Machine

It is included in the supervised learning method which is based on decision trees and can be used for classification modeling [15]. This method was first introduced by Jerome H. Friedman in 2001. The learning procedure in GBMs works sequentially to provide more accurate predictions of response variables [2]. The learning procedures for GBMs are as follows [16].

Input:

1. Data consisting of independent variables (X) and a response variable (Y) with a number of N observations.
2. The number of iterations is M.
3. Loss-function $\Psi(y, f)$
4. Learning base model$h(x, \theta)$

Algorithm:

1. Initialize with a constant value.$\hat{f}_0$
2. For $t = 1$ to M:
3. Calculate negative gradient:$g_t(x)$

$$g_t(x) = E_y\left[\frac{\partial\Psi(y, f(x))}{\partial f(x)}|x\right]_{f(x)=\hat{f}_{t-1}(x)} \tag{2}$$

4. Arrange new learning base functions $h(x, \theta_t)$
5. Determine the best gradient descent step-size(ρ_t):

$$\rho_t = \arg min_\rho \sum_{i=1}^{N} \Psi\left[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)\right] \tag{3}$$

6. Update prediction function:

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x_i, \theta_t) \tag{4}$$

7. Output result: $\hat{f}_t$

2.4 Extremely Randomized Trees (Extra Trees)

Extra Trees was developed by Pierre Geurts, Damien Ernst, and Louis Wehenkel in 2006. As the name suggests, Extremely Randomized Trees, this method carries out extreme randomization. Randomization in Extra Trees is not only carried out when selecting explanatory variables but also when selecting cut points. In addition, Extra Trees does not use bootstrap data to build each classification tree. The data used to build each classification tree in Extra Trees is the entire training data. Extra Trees also does not perform pruning when building a classification tree. The following is the algorithm from Extra Trees [4].

1. The formation of a classification tree in Extra Trees is carried out using all training data.
2. Stages of selecting the best splitting:
   a. Randomly select *m* independent variables.
   b. Randomly select *k* cut points.
   c. Determining the best splitting.
   d. Steps a to c are repeated until the stopping criteria are reached so that prediction results from one classification tree are obtained.
3. Steps 1-2 are repeated until a classification tree is formed.
4. Determine the prediction results from Extra Trees by combining the prediction results from each classification tree using majority vote.

2.5 Measures of Model Performance

The measure of model performance in classification is used to see the accuracy of a model in predicting a class in the data. In classification modeling, the measure of model performance is calculated using a confusion matrix. Confusion matrix is a matrix that shows predicted and actual classification. Table 2 shows the confusion matrix for classification of two classes [17].

Table 2: Confusion Matrix

		Prediction	
		Positive	*Negative*
Actual	*Positive*	*True Positive* (TP)	*False Negatives* (FN)
	Negative	*False Positives* (FP)	*True Negative* (TN)

Confusion matrix can be used to calculate measures of model performance such as accuracy, sensitivity, and specificity. The definition and formula of these measures of model performance are as follows [18].

a. Accuracy

Accuracy is the proportion of the number of observations that are predicted correctly. Accuracy can be calculated using equation 5.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

b. *Sensitivity*

Sensitivity is a measure of the performance of the classification algorithm in classifying data in the positive class. Sensitivity can be calculated using equation 6.

$$Sensitivity = \frac{TP}{TP+FN} \tag{6}$$

c. *Specificity*

Specificity is a measure of the performance of the classification algorithm in classifying data in the negative class. Specificity can be calculated using equation 7.

$$Specificity = \frac{TN}{TN+FP} \tag{7}$$

2.6 Data Analysis Stages

Data analysis in this research was carried out using R software. The stages of data analysis carried out were as follows.

1. Exploring Telco customer churn data.
2. Dividing the data into training data and test data with proportions of 85% and 15%. Training data is used to build the model while test data is used to evaluate the model performance.
3. Handling the problem of class imbalance in Telco customer churn data using Random Oversampling, Random Undersampling, and SMOTE-NC techniques.
4. Carrying out the training process using training data that has not been balanced and training data that has been balanced. The training process is carried out using the GB, RF and Extra Trees methods.
5. The best model from each method that was obtained from stage 4 was then evaluated for its performance using accuracy, sensitivity and specificity.

3. Results and Discussion

3.1 Data Exploration

Telco Customer Churn data has 18 explanatory variables and one response variable consisting of two categories, namely Yes (customer stops using the service) and No (customer does not stop using the service). Figure 1 provides information regarding the percentage of discrete (categorical) and continuous variables contained in the data. Moreover, Figure 1 also shows that there are no missing observations in this data, so missing data was not handled in this study.

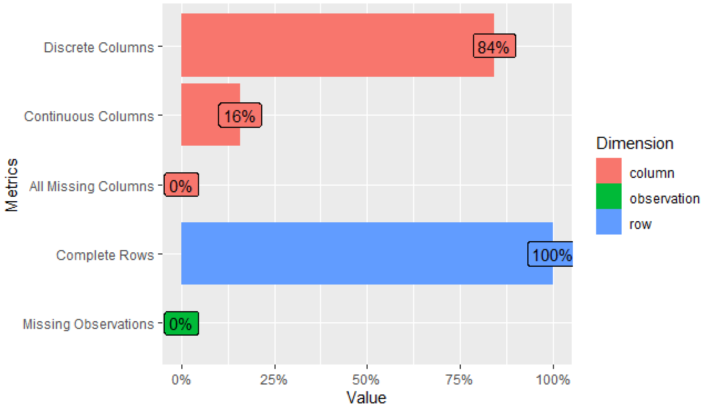

Figure 1: Information related to Variables in Telco Customer Churn Data

Figure 2 shows the proportion of each category in the response variable (Churn). The different percentage values for the two categories indicate that the response variable classes in the Telco Customer Churn data are imbalance. The percentage for the Yes category is 26.6% while the No category has a percentage of 73.4%. This shows that the majority of customers do not stop using the service. Even though the majority of customers are still using the service, if we look at the number of customers who have left the service, there are still 1873 customers.

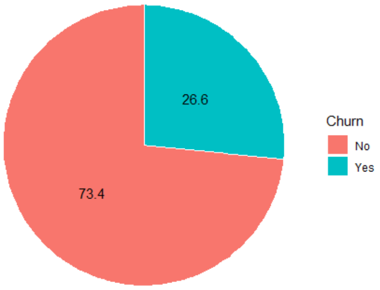

Figure 2: Proportion of Response Variable Categories (Churn)

3.2 Classification Models

The classification model is built using training data that has been balanced and training data before being balanced, so that we can see the effect of imbalanced data on model performance. The following are the performance results of three classification models for each method.

1. **Gradient Boosting (GB)**

Table 3: Performance Measure of the Gradient Boosting Model on Telco Customer Churn Data

Data	Measure of model performance	Value	Average of Accuracy, Sensitivity, and Specificity
Original Data	Accuracy	0.7875	0.7134
	Sensitivity	0.9134	
	Specificity	0.4393	
Balanced class data (SMOTE-NC)	Accuracy	0.7666	0.7289
	Sensitivity	0.8307	
	Specificity	0.5893	
Balanced class data (oversampling)	Accuracy	0.7571	0.7267
	Sensitivity	0.8088	
	Specificity	0.6143	
Balanced class data (undersampling)	Accuracy	0.7486	0.7557
	Sensitivity	0.7364	
	Specificity	0.7821	

2. **Random Forest (RF)**

Table 4: Performance Measure of The Random Forest Model on Telco Customer Churn Data

Data	Measure of model performance	Value	Average of Accuracy, Sensitivity, and Specificity
Original Data	Accuracy	0.7846	0.7111
	Sensitivity	0.9096	
	Specificity	0.4393	
Balanced class data (SMOTE-NC)	Accuracy	0.7619	0.6901
	Sensitivity	0.7619	
	Specificity	0.5464	
Balanced class data (oversampling)	Accuracy	0.7742	0.7280
	Sensitivity	0.8527	
	Specificity	0.5571	
Balanced class data (undersampling)	Accuracy	0.7249	0.7454
	Sensitivity	0.6899	
	Specificity	0.8214	

3. **Extremely Randomized Tree (Extra Trees)**

Table 5: Performance Measure of the Extra Trees Model on Telco Customer Churn Data

Data	Measure of model performance	Value	Average of Accuracy, Sensitivity, and Specificity
Original Data	Accuracy	0.7628	0.6924

	Sensitivity	0.8824	
	Specificity	0.4321	
Balanced class data (SMOTE-NC)	Accuracy	0.7666	0.7289
	Sensitivity	0.8307	
	Specificity	0.5893	
Balanced class data (oversampling)	Accuracy	0.7647	0.6909
	Sensitivity	0.8902	
	Specificity	0.4179	
Balanced class data (undersampling)	Accuracy	0.7201	0.7120
	Sensitivity	0.7339	
	Specificity	0.6821	

3.3 Model Performance Comparison

The results above show that when training data is used without balancing to build a classification model, the specificity value will be low. Specificity is a performance measure of the classification algorithm in classifying data in the negative class. In the Telco Customer Churn classification model, the negative class used is the "Yes" category (customers stop using Telco services). Therefore, the model built using training data without balancing has poor classification ability when used to predict customers in the "Yes" category. The low specificity value in the model without balancing indicates that it is necessary to balance the data before modeling.

The performance of the model built using data with the SMOTE-NC balancing technique apparently still has a low specificity value. A low specificity value indicates that the classification model built is less capable of detecting "Yes" class customers. The random undersampling balancing technique provides the best performance results on two classification models (Random Forest and Gradient Boosting). This is because this model has accuracy, sensitivity, and specificity values that are not much different. So apart from being able to classify customers who have not left the service, this model is also able to classify customers who have stopped using Telco services. Therefore, in the case of Telco Customer Churn, the best balancing technique is random undersampling.

4. Conclusions

Classification modeling on Telco Customer Churn data is carried out using three methods, namely Gradient Boosting (GB), Random Forest (RF), and Extra Trees. The problem of imbalanced data in this analysis was handled using random undersampling, random oversampling, and SMOTE-NC techniques. Models with training data balanced using random undersampling techniques produce better performance compared to random oversampling and SMOTE-NC. This is because the model produced from balanced training data (random oversampling technique and SMOTE-NC) has a low specificity value, which means that the model disable to predict customers who leave Telco services or the "Yes" category accurately.

There are two models with the best performance, the first model is obtained from the Gradient Boosting (GB) model which was built using data that has been balanced using random undersampling techniques. This model has an accuracy of 74.68%, specificity of 78.21%, and sensitivity of 73.64%. The second-best model was obtained from the Random Forest (RF) model which was built using data that had been balanced using random undersampling techniques. This model has an accuracy of 72.49%, specificity of 68.99%, and sensitivity of 82.14%. Therefore, in the case of Telco Customer Churn, the best class balancing technique for data is random undersampling.

References

1. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Application in R*. New York: Springer, 2013. doi: 10.2174/0929867003374372.
2. A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front. Neurorobot.*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.
3. S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nat. Mach. Intell.*, vol. 2, no. January, pp. 56–67, 2020, http://dx.doi.org/10.1038/s42256-

019-0138-9.

4. P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.
5. R. Siringoringo, “Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor,” *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.
6. W. Chaipanha and P. Kaewwichian, “Smote Vs. Random Undersampling for Imbalanced Data-Car Ownership Demand Model,” *Communications*, vol. 24, no. 3, pp. D105–D115, 2022, doi: 10.26552/com.C.2022.3.D105-D115.
7. S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets : A review,” *Science* , vol. 30, no. 1, pp. 25–36, 2006
8. Q. H. Doan, S. H. Mai, Q. T. Do, and D. K. Thai, “A cluster-based data splitting method for small sample and class imbalance problems in impact damage classification,” *Appl. Soft Comput.*, vol. 120, p. 108628, 2022, doi: 10.1016/j.asoc.2022.108628.
9. D. T. Utari, “Integration of Svm and Smote-Nc for Classification of Heart Failure Patients,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 4, pp. 2263–2272, 2023.
10. M. A. Ganai, M. Hu, A. K. Malik, M. Tanvir, and P. N. Suganthan, “Ensemble deep learning: A review,” *Eng. Appl. Artif. Intell.*, vol. 115, 2022, doi: https://doi.org/10.1016/j.engappai.2022.105151.
11. L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
12. M. Aria, C. Cuccurullo, and A. Gnasso, “A comparison among interpretative proposals for Random Forests,” *Machine Learning with Applications*, vol. 6. p. 100094, 2021. doi: 10.1016/j.mlwa.2021.100094.
13. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random forests and decision trees,” *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.
14. S. Han, H. Kim, and Y. S. Lee, “Double random forest,” *Mach. Learn.*, vol. 109, no. 8, pp. 1569–1586, 2020.
15. S. E. Suryana, B. Warsito, and S. Suparti, “Penerapan Gradient Boosting Dengan Hyperopt Untuk Memprediksi Keberhasilan Telemarketing Bank,” *J. Gaussian*, vol. 10, no. 4, pp. 617–623, 2021, doi: 10.14710/j.gauss.v10i4.31335.
16. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
17. R. Kohavi and F. Provost, “Glossary of Terms Glossary of Terms,” *Mach. Learn.*, vol. 30, pp. 271–274, 1998.
18. J. C. Obi, “A comparative study of several classification metrics and their performances on data,” *World Journal of Advanced Engineering Technology and Sciences*, vol. 8, no. 1, pp. 308–314, 2023, doi: https://doi.org/10.30574/wjaets.2023.8.1.0054.