Handling Problems of Credit Data for Imbalanced Classes using SMOTEXGBoost

Heru Mardiansyah1, Rahmat Widia Sembiring2, Syahril Efendi3
1Graduate Program of Computer Science,
2Department of Information Technology,
3Department of Computer Science,
Faculty of Computer Science and Information Technology,
Universitas Sumatera Utara, Medan, Indonesia

Email: heru.mardiansyah@students.usu.ac.id

Abstract. Some researchers find data with imbalanced class conditions, where there are data with a number of minorities and a majority. SMOTE is a data approach for an imbalanced classes and XGBoost is one algorithm for an imbalanced data problems. This research uses SMOTE and XGBoost or abbreviated as SMOTEXGBoost for handling data with an imbalanced classes. The results showed almost the same accuracy value between SMOTE and SMOTEXGBoost at 99%. While the value of AUC SMOTEXBoost has a more stable value than SMOTE that is equal to 99.89% for training and 98.51% for testing.

Keyword—SMOTE, SMOTEXGBoost, AUC, Accuracy.

1. Background

Some researchers have found data with conditions referred to as imbalanced classes, where a small amount of data is referred to as a minority class and many data are referred to as the majority class. This is usually found in data such as credit, health and other data [12]. Liu et al. [10] state that learning algorithms that do not consider imbalances in the majority class tend to be overwhelmed by minority classes.

Some traditional approaches or classification methods such as logistic regression, discriminant analysts, and decision trees, show that these techniques are generally less suitable for handling data with the imbalanced class conditions [4,11]. This is because these techniques are more likely to classify objects into the majority class and not into minority classes [17].

Liu et al. [8] and He et al. [6] introduced several techniques for classification imbalances known as oversampling and under-sampling concepts. Oversampling technique is by multiplying minority classes so that the number approaches the majority class. And conversely, the under-sampling technique is to reduce the majority class so that the number approaches the minority class [16,19].

The Synthetic Minority oversampling Technique (SMOTE) introduced by [1] is an over-sampling technique that generates pseudo samples based on the characteristics of the object and the nearest neighbor. Assembling pseudo samples has different procedures for each numerical and categorical variable.

Chen, T. & Carlos, G. [2] introduces one of the tree boosting techniques that is very scalable for large data and is known as Extreme Gradient Boosting (XGBoost). In many studies show that the XGBoost technique has better results than some other boosting techniques.

In this study, the authors used the SMOTE algorithm in the pre-modeling step then with the XGBoost boosting technique used when modeling for the imbalanced class problems. Next will be compared the results of predictions between XGBoost and SMOTEXGBoost. SMOTE uses the closest neighbor to create synthetic samples from minority classes and SMOTEXGBoost then injects the SMOTE method to each boosting iteration.

2. Literature Review

a. The Imbalanced Classes

The imbalanced class is a condition if there is one class having more instances than the other classes. If there is an imbalanced class occurring in the classification of the two classes (binary class) then the
class that has the number of instances is called the majority class. Whereas classes that have a fewer number of instances are called minority classes [14]. The imbalanced classes can be seen in Figure 1[8].

![Figure 1 Non-balanced class](image)

There are 2 (two) methods used to solve an imbalanced class problems, namely the first data level approaches such as undersampling and oversampling techniques and both algorithmic approaches such as bagging, boosting and stack.

b. Synthetic Minority Oversampling Technique

SMOTE is a very popular method and is widely used by researchers to deal with an imbalanced class problems. This technique synthesizes new samples from minority classes to balance datasets by creating new instances of minority classes with the convex formation of combinations of adjacent instances. As shown in Figure 2 below[8], it effectively describes the lines between minority points in the feature space and samples along the lines. With this method it can make the dataset balanced without overfitting, this is by making synthetic samples rather than duplicating the sample.

![Figure 2 SMOTE](image)

Syntetic samples can be generated in smart way to balance minority class with majority class. Good example of such type of methods is syntetic minority oversampling technique (SMOTE) presented in [3]. This approach uses KNN to create artificial examples.

c. XGBoost (Extreme Gradient Boosting)

Friedman [5] conducted a study that used the relationship between boosting and optimization to create Gradient Boosting Machine (GBM). GBM is a technique in machine learning to help solve the problems of regression and classification which produce predictive models in the form of an ensemble of weak prediction models. The model is built using the boosting method, which is by creating a new model to predict the error/residual from the previous model. A new model is added so that no more errors can be made. This algorithm is called gradient boosting because it uses gradient descent to minimize errors when creating new models.

d. Performance Measure

In this study, the confusion matrix and Receiver Operating Characteristic (ROC) curves were used to measure the performance classifier [6,18]. The confusion matrix is a classification record table as in the table below.

Performance measurement usually uses accuracy. However, for the imbalanced class distribution problems, specificity and sensitivity are used as additional measurements. In this case, sensitivity is a measure that measures how well a model predicts bad credit customers. Whereas specificity is a measure that measures how well a model predicts good credit customers. While the
ROC curve is a graph of sensitivity (true positive rate) on the Y axis with 1-specificity on the X-axis (false positive rate). From this curve, it also shows the Under Curve Area (AUC), which is the area under the ROC curve.

e. Cross Validation

Cross Validation is a model validation technique to assess how the model performs against other independent datasets [13]. There are 2 (two) types of cross validation, namely K-Fold Cross Validation and Monte-Carlo Cross Validation. In this paper use the K-Fold Cross Validation method with a value of $k = 10$.

3. Method
   a. Model Frame Work

   ![](image)

   Figure 3 Research Flow Chart

   b. SMOTEXGBoost

   SMOTEXGBoost is a hybrid solution that combines the SMOTE approach for unbalanced data settlement which then uses the XGBoost algorithm for its prediction model calculation, which includes: Training/training data that is processed using the SMOTE method so that training data/training has a better data balance. Then using the data is done building model Furthermore, by using test/testing data, a prediction calculation process from the model used is carried out.

4. Result
   a. Data

   The research uses several imbalanced datasets obtained from the UCI Machine learning [9], namely German Data Credit Cards, Winconsi, Glass and E-coli.

   b. Result

   Using the Logistic Regression (LR) algorithm, SMOTE Logistic Regression (SLR), Random Forest (RF), Random Forest SMOTE (SRF), Support Vector Machine (SVM), SMOTE Support Vector Machine (SSVM), XGBoost (XGB) and XGBoost SMOTE (SXGB) is produced by the value of AUC as seen on the screen.
The test dataset used in the research for the SMOTE Logistic Regression model has a better predictable outcome than the Logistic Regression model, with the AUC difference in the range 0.16 to 1.06. And for test datasets used in research for the Random Forest SMOTE model has a better predictable outcome than the Random Forest model, with a value of AUC in the range 0.39 to 1.10. Whereas for the test dataset used in the research for the SMOTE model Support Vector Machine has a prediction that is better than the Support Vector Machine model, with a value of AUC in the range 0.91 to 2.87. And for the test dataset that is used for research for the XGBoost SMOTE model, the results are predicted to be better compared to the XGBoost model, with a value of AUC in the range 0.08 to 1.53.

The model or algorithm is classified as Logistic Regression, Random Forest and Support Vector Machine for the test dataset above the AUC value, which is always lower than the model that uses the SMOT method of the Data OTOTGBGBost and the approach to the algorithm that is XGBoost [3].

Based on the above, algorithms/model classifications such as Logistic Regression, Random Forest, Support Vector Engine have predicted AUC values that are lower if for data with the unbalanced if they are mixed with XGBoost and SMOTEXGBoost. The proposed solution is the SMOTEXGBoost which has the highest ASUC value combined with the data level approach (the SMOTE method) in balancing the data and the approach of the XGBoost algorithm used in predicting the model that produces an "excellent classification" value.

c. Model Stability

To ensure that the resulting model has performance stability, in this test data with an imbalanced class will be divided into 5 (five) groups of data with variations in proportions, namely (0.6: 0.4), (0.7: 0.4): (0.8: 0.2): (0.95: 0.05).

The comparison of AUC SMOTEXGBoost performance as shown in table 7 shows that for Data1 has a training value of 99.92% and a testing value of 98.06%, for Data2 has a training value of 99.95% and a testing value of 97.38%, for Data3 training values is 99.92% and testing value is 98.94%, for data4 training value is 99.80% and testing value is 99.05% and for data 5 training value is 99.85% and testing value is 99.14%. So it can be concluded that the value is very stable for training at the level of 99.89% and testing 98.51%.

### Table 1: Comparison of Models Based on AUC

| Dataset   | LR    | SLR   | RF    | SRF   | SVM   | SSVM  | XGB   | SXGB  |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| German CC | 97.60 | 97.76 | 97.79 | 98.40 | 96.41 | 98.32 | 97.35 | 98.88 |
| Wisconsin | 98.04 | 99.03 | 98.03 | 98.45 | 96.35 | 98.32 | 99.63 | 99.93 |
| Glass     | 98.43 | 99.03 | 98.01 | 98.40 | 96.91 | 97.82 | 99.63 | 99.80 |
| E-coli    | 94.65 | 95.69 | 95.15 | 96.25 | 94.95 | 97.82 | 99.32 | 99.40 |

### Table 2: Comparison of AUC SMOTEXGBoost Performance

| Data   | Proportion | AUC       |
|--------|------------|-----------|
|        |            | Training  | Testing  |
| Data1  | 0.6 : 0.4  | 99.92%    | 98.06%   |
| Data2  | 0.7 : 0.3  | 99.95%    | 97.38%   |
| Data3  | 0.8 : 0.2  | 99.92%    | 98.94%   |
| Data4  | 0.9 : 0.1  | 99.80%    | 99.05%   |
| Data5  | 0.95 : 0.05| 99.85%    | 99.14%   |
5. Conclusion

Some researchers often find the problem of class data imbalance, which makes classification techniques difficult to produce bad values and instability in performance with the model proposed in this study, namely SMOTEXGBoost has the highest level of accuracy or AUC value among the other models with predictions of "elasticity classification". The model for classification algorithms (LR, RF, SVM) has a weakness for predicting unbalanced data, then for German credit card datasets with F1-Measure> 0.5 and AUC values> 90, then SMOTE LR, SMOTE RF, SVM SMOTE models, XGBoost and SMOTEXGBoost have more accurate predictive values. Using the SMOTE method in the LR, RF, SVM models, XGBoost can increase predictive values so that they are more accurate. Wisconsin dataset, Glass.0.1.2.3_vs_4.5.6, E-coli.0_vs_1 the prediction of the resulting model is not the same as the German Credit Card dataset so that different datasets differ in the level of accuracy prediction of the model. In the future, it should be considered to conduct research for an imbalanced class data using several other data approaches algorithms and algorithmic approaches such as bagging and stacking.

References
[1] Arwan, et. al. 2018. Synthetic Minority Over Sampling Technique (SMOTE) Algorithm for Handling Imbalanced Data.
[2] Chen, T. &Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22\textsuperscript{nd} ACM SIGKDD international conference on knowledge discovery and data mining. ACM. Pp 785-794.
[3] Chawla, Nitesh V., et al. 2002. SMOTE: Synthetic Minority Over Sampling Technique. Journal of Artificial Intelligence Research.
[4] Dwivedi, P. & Kumar, P.U. 2017. A Review on Classification Algorithm for Imbalance Class Dataset. International Journal of Scientific Research Engineering & Technology (JNSRET). ISSN 2278-0882 vol 6 issue 5
[5] Friedman, J.H. 2008. A Gradient Boosting Machine. The Annals of Statistics 2001, Vol. 29 No. 5, 1189-1232
[6] Flach, P. A. 2007. ROC analysis for ranking and probability estimation. Notes from the UAI tutorial on ROC analysis.
[7] He, H. and Garcia, E.A. 2009. Learning from Imbalanced Data. IEEE Transaction on Knowledge and Data Engineering 21:1263-1284.
[8] Jordan, J. 2018. Learning from Imbalanced Data. https://www.jeremyjordan.me/imbalanced-data/. 15 February 2018. (acces 1 Desember 2018)
[9] Linchan, M. 2013. UCI Machine Learning Repository. German Credit. Data.
[10] Liu, X.Y., Wu, J.,and Zhou, Z.H. 2009. Exploratory Undersampling for Class-Imbalance Learning, IEEE Transaction on System: 965-969
[11] Louzada, F. Ara, A. &Fernandes, G.B. 2016. Classification methods applied to credit scoring: Systematic review and Overall Comparison. Survey in Operations Research and Management Science.
[12] Nikitin, N.O, Kalyuzhnaya, A.V., et al. 2018. Evolutionary ensemble approach for behavioral credit scoring. ICCS Camera Ready Version 2018.
[13] Nurzhaputra, A. & Muslim, M.A. 2017. PeningkatanAkurasiPadaAlgoritma C4.5 menggunakanADABoostUntukMeminimalkanResikoKredit. Prosiding SNATIF ke-4
[14] Magegie, M. 2018. A Comparison of machine learning algorithms using insufficient number of labeled observation. Thesis. VrijeUniversiteit Amsterdam.
[15] Patil, P.S., Aghav, J.V. &Sareen, V. 2016. An Overview of Classification Algorithms and Ensemble Methods in Personal Credit Scoring. IJCST vol7 issue2
[16] Salunkhe, U.R. & Mali, S.N. 2016. Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach. Procedia Computer Science 85 725-732
[17] Wang, B.X & Japkowicz. 2010. *Boosting support vector machine for imbalanced dataset.* Knowledge and Information System, vol. 25, no. 1, pp. 1-20

[18] Wang, Q. 2014. *A Hybrid Sampling SVM Approach to Imbalanced Data Classification.* Abstract and Applied Analysis Volume 2014. article ID 972786

[19] Xiao, H. Xiao, Z. Wang, Y. 2014. *Ensemble classification based on supervised clustering for credit scoring.* Accepted Manuscript. Applied Soft Computing.