WHICEB 2022 Proceedings

Wuhan International Conference on e-Business

Summer 7-26-2022

Prediction of Customers' Subscription to Time Deposits Based on SMOTEENN-XGBoost Model

Ying Li
College of Economics and Management, China Jiliang University, China

Zengyuan Wu
College of Economics and Management, China Jiliang University, China, wuzengyuan@cjlu.edu.cn

Full Research Paper

Prediction of Customers' Subscription to Time Deposits Based on SMOTEENN-XGBoost Model

Ying Li[1], *Zengyuan Wu*[1*]

[1]College of Economics and Management, China Jiliang University, China

Abstract: In the fierce competitive banking industry, accurate prediction of customers' subscription to time deposits is vital for banks. This can reduce unnecessary time and energy spent on targeted customer service to improve bank efficiency. Traditional prediction methods do not handle the imbalanced data problem very well. In this paper, in order to minimize the impacts from imbalanced data, we combine Synthetic Minority Oversampling Technique (SMOTE)and Edited Nearest Neighbor Technique (ENN) to make the data as balanced as possible. Then, Extreme Gradient Boosting algorithm (XGBoost) is adopted as classification algorithm to improve the accuracy of data classification results. For clarity, this model is called SMOTEENN-XGBoost model. A bank customers dataset published on the Kaggle platform is used to demonstrate its effect by numerical experiments. We compare the performance of the SMOTEENN-XGBoost in this paper with Decision Tree (DT), Adaptive Boosting (AdaBoost), XGBoost, SMOTE-XGBoost in terms of Accuracy (ACC), Area Under ROC Curve (AUC), and Geometric-mean (G-mean). The results show that the mean ACC, AUC and G-mean of SMOTEENN-XGBoost model are 0.92, 0.97, and 0.92, which are better than the other models. It indicates that this model has good classification performance and can effectively dig out potential customers.

Keywords: imbalanced data, time deposits, SMOTEENN, XGBoost

1. INTRODUCTION

The survival of banking industry is closely linked to its deposits, which are the foundation of banks. The amounts of deposits or loan-to-deposit ratio directly affect the profits of banks. The banks will lose its liquidity if they don't have deposits. Furthermore, their financial activities will be greatly affected. To sum up the above, deposits are very important to the banks. In the past few decades, when banks sell time deposits, it is difficult to find out who their target customers are. Many factors could affect customers' subscription to bank deposits. At the same time, more and more customers start to complain about irrelevant calls they received from the banks. In order to solve these business problems, banks are beginning to leverage their vast customer data to gain insight into customer behaviors and purchasing preferences. Actually, their marketing efficiency has really improved. In such a market environment, banks are necessary to enhance their competitiveness and obtain customers who are interested in subscribing to time deposits. These can help banks maintain a certain level of competitiveness and avoid making them eliminated by the general environment. From the perspective of the banks' development path, the banking industry must build superior banking transaction products and service capabilities to maintain high-value customer relationships. Besides, they should select targeted customers to improve efficiency. Therefore, it is self-evident for banks to predict whether customers will subscribe time deposits.

It is a major trend for the banking industry to carry out accurate marketing based on customers' classification. Autoregressive models are proposed to the prediction of savings deposit by Petropoulos [1]. Mitat [2] used the ARIMAX model to predict whether customers will subscribe to time deposits. These traditional forecasting methods are complex, time-consuming and difficult to adapt to complex data. Therefore, some scholars propose to use machine learning to predict whether customers will subscribe to bank deposits, such as Logistic Regression

[*] Corresponding author. Email: wuzengyuan@cjlu.edu.cn(Zengyuan Wu)

[3], Support Vector Machine (SVM) [4], Neural Networks [5]. However, experiments show that these single classifiers are not very effective when dealing with large amounts of data and complex problems. In order to further improve the classification performance, scholars propose to use ensemble methods for classification prediction. Kumar [6] used Hybrid Weighted Random Forests Method to predict online consumer purchasing behaviors, which is better than using a single algorithm for prediction. Actually, prediction of customers' subscription to time deposits is an imbalanced data binary classification. Imbalanced data problem is one of the challenges for machine learning to perform classification.

The problem of imbalanced data classification is widely existing in real life, such as medical diagnosis [7] ,credit fraud [8], and spam filtering [9]. Existing research solving imbalanced data classification problem mainly focus on two main aspects. One is external data-level processing, the other is internal algorithm-level processing [10]. The main data-level processing method is the resampling technique, which includes random under-sampling and random oversampling. The random under-sampling technique is used to even out the sample distribution by randomly removing majority class samples, but it will lose important information. The main random under-sampling techniques are K-means Adacost bagging (KACBag), ENN and Tomek links. The random oversampling technique balances the data by randomly replicating minority class samples, but it will make the information redundant, increasing the complexity of model training and causing over-fitting problems. A typical oversampling method is SMOTE [11]. These resampling methods are easy to operate and have good adaptability. However, deleting or expanding data don't follow the original data distribution. In this case, they may lead to the loss of valuable information or model overfitting problems.

At the algorithmic-level, scholars try to improve the algorithm directly. It mainly includes Clustering, Cost-sensitive Learning, and Ensemble Learning. Currently, Ensemble Learning is favored by most scholars. It cascades multiple classifiers to improve the classification performance. Bagging and Boosting are its typical representatives. Boosting is to cascade multiple weak classifiers to form a strong classifier. AdaBoost [12] belongs to the Boosting algorithm, and it is widely used. However, the AdaBoost itself is overfitting and not robust, besides, it is time-consuming when training the algorithm.

Therefore, the research focus on the following question: how to devise an algorithm based on machine learning to improve predict accuracy? In this paper, at the data level, we propose to use SMOTEENN of hybrid sampling technique to retain valuable information and overcome the overfitting problem. First, we use SMOTE to oversample the imbalanced data. Second, we remove the noise presented in the new generated samples by ENN. At the algorithmic level, we choose XGBoost which is a kind of Boosting algorithm to further improve the classification performance. Finally, we combine SMOTEENN and XGBoost to improve the classification accuracy.

Based on the above analysis, we conclude that single sampling has many shortcomings and a single prediction model is more difficult to meet the requirements of the imbalanced dataset. In order to solve these problems, we propose a combined prediction model to predict whether customers will subscribe to time deposits. It's named SMOTEENN-XGBoost.At the data level, we choose the hybrid sampling technique SMOTEENN to reduce the data imbalance rate. At the algorithm level, we choose the XGBoost to improve the overall prediction accuracy. The model is evaluated by a ten-fold cross-validation method.

2. SMOTEENN-XGBOOST MODEL

2.1 The Synthetic Minority Oversampling Technique (SMOTE).

SMOTE is a kind of resampling technique, and it is more commonly used. The SMOTE does not simply copy the minority class samples. It generates a new sample by adding K adjacent minority class samples along the linearity for each minority sample. Then taking the center of the line segment, and performing median interpolation process. The general flow of the algorithm is as followsAs shown in Figure 1, the oversampling process is as follows:

● Step 1: For the minority sample S, calculate each sample's nearest neighbors by Euclidean distance. It can obtain k-nearest neighbors in the minority sample set.

● Step 2: The sampling ratio is specified according to the proportion of sample imbalance in the dataset. Then determine the sampling multiplicity N. For each minority sample x, randomly selected some samples among its k-nearest neighbors, and denoted as $x_{i\ (near)}, near \in \{1,2,\dots,k\}$.

● Step 3: Generate a new sample by constructing $x_{i(near)}$ separately according to equation (1). In this equation, x is a minority sample, $\tilde{x}$ is the nearest neighbor sample of x, $\|\tilde{x}-x\|$ is the distance formula. Calculate the difference between the nearest neighbor sample and the feature vector x .

$$x_{new} = x + rand(0,1) \times \|\tilde{x}-x\| \tag{1}$$

● Step 4: Repeat step 3 N times. New samples of N are synthesized, denoted them as $x_{i(new)}$,$new \in \{1,2,\dots N\}$.

● Step 5: Perform the above operation for all samples in the minority sample S. This will result in a new sample $S \times N$.

The principle of ENN is that for each sample in the training set, we find its three nearest neighbors. If the sample is a majority class sample and more than two of its three nearest neighbors are minority class samples, we remove it. Conversely, if the sample is a minority class sample and more than two of its three nearest neighbors are majority samples, we remove the majority class samples from its nearest neighbors. It is more aggressive than Tomek Links in reducing the majority class samples and provides a deeper cleanup.

The SMOTEENN is similar to SMOTETomek. As a hybrid sampling method, it generates a new minority class samples by using the SMOTE to obtain an expanded data set T. Then predict each sample in T by KNN. Here K is generally taken as 3. If the predicted result does not match the actual class, the sample is rejected. It tends to remove more noisy samples compared to SMOTETomek.

2.2 XGBoost algorithm.

XGBoost algorithm is a kind of Boosting ensemble algorithm mainly applied in supervised learning. Based on GBDT (Gradient boosted tree). The tree model used in the XGBoost is the CART (Classification and Regression Tree). CART is smaller in size but more efficient compared to other decision trees. It assumes that the decision tree is a binary tree, and the features of the internal nodes are taken as "yes" or "no" only. XGBoost uses a gradient boosting approach to ensemble weak evaluators. The following are the steps of the XGBoost.

● Step1: The prediction model of XGBoost for the sample results can be expressed as follows:

$$\hat{y}_i = \sum_{r=1}^{R} g_r\,(x_i), g_r \in G \tag{2}$$

In this equation, R is the number of decision trees in the model, x_i is the i-th input sample, g_r is the r-th decision tree, $g_r(x_i)$ denotes the leaf weight, which is the prediction score of the r-th tree for sample i. $\hat{y}_i$ denotes the predicted value, and G is the set of all possible CARTs.

● Step2: XGBoost performs a second-order Taylor expansion of the loss function in the optimization of the algorithm, and its optimization objective and loss function are as follows:

$$M(k) = \sum_{i=1}^{h} m(y, \hat{y}(k-1) + g_k\,(x_i)) + \Omega(g_k) \tag{3}$$

In this equation, $M(k)$ is the objective function of the k-th iteration, the total number of samples is h, M is the loss function, $\hat{y}(t-1)$ is the predicted value of the previous iteration, $g_k(x_i)$ is the newly added function, and $\Omega(g_k)$ is the regularization term. The loss value is used to measure the magnitude of the difference between the true value y_i and the predicted result $\hat{y}_i$, and the regularization is used to measure the complexity of the model and to avoid the phenomenon of model overfitting.

A Taylor second-order expansion of the above equation is performed and simplified to obtain the following objective function:

$$\bar{M}(k) = \sum_{i=1}^{h}\left[f_i g_i(x_i) + \frac{1}{2} j_i g^2(x_i)\right] + \Omega(g_k) \tag{4}$$

In this equation, f_i is the first-order derivative and j_i is the second-order derivative.

- ● Step3: The last step is to perform the solution of the optimal value.

$$Z = -\frac{F_i}{J_i+\lambda} \tag{5}$$

$$\bar{M}(k) = -\frac{1}{2}\sum_{i=1}^{K}\frac{F_i^2}{J_i+\lambda} + \lambda K \tag{6}$$

In this equation, Z is the optimal value and $\bar{M}(K)$ is the value of the objective function.

2.3 Overview of SMOTEENN-XGBoost model

Although the XGBoost has a good improvement in classification performance over a single classifier, it has a deficiency in handling imbalanced data. Therefore, we proposed SMOTEENN-XGBoost model to improve the XGBoost from the data level. Typical resampling methods from the data level either reduce the majority class samples or increase the minority samples. These methods have disadvantages, they will lose important data information, make the information redundant and have overfitting problems. The SMOTEENN as a hybrid resampling technique retains the features of majority class samples, and it increases the features of minority class samples to avoid overfitting problems to improve the classification accuracy. The main flowchart of the model is shown in Figure 1.

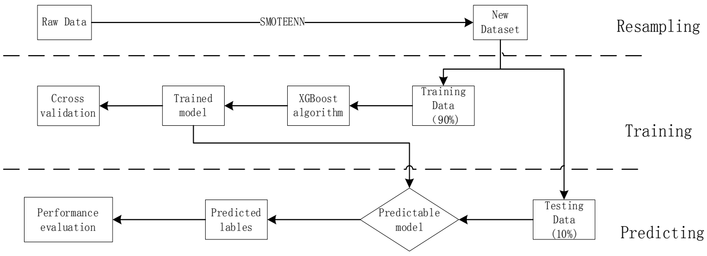

Figure 1. The flowchart of SMOTEENN-XGBoost

The model algorithm flow is roughly as follows:

- ● Step1: Input category imbalanced data set T
- ● Step2: Random selection of minority samples x.Using the Euclidean distance as a criterion, calculate the distance between this sample and the k nearest samples. Randomly draw a number of samples from the k samples, assuming that the selected samples are x_n. Take the random number 0-1 and generate a new sample for each randomly selected nearest neighbor according to Equation (1). Repeat this step until data balance, then put the new samples into the original dataset to obtain T_1.
- ● Step3: For each sample in T_1, find its three nearest neighbors, and if the sample is different from two of its three neighboring samples, delete the sample. The algorithm ends and the new dataset T_2 is obtained.
- ● Step4: Using the k-fold cross-validation method. The training set is proportionally divided into a test set and a training set, and the XGBoost is used to the training set for classification:

(a)Initialize the predicted values for each sample.

(b)Define the loss function:

$$J(f_t) = \sum_{i=1}^{n} L(y_i\,, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C \tag{7}$$

(c)Compute the derivative function of the loss function with respect to the predicted values of each sample.

$$J(f_t) \approx [L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2}\mathrm{h_i} f_i^2(x_i)] + \Omega(f_t) + C \tag{8}$$

$$g_i = \frac{\partial L(y_i,\hat{y}_i^{(t-1)}}{\partial \hat{y}_i^{(t-1)}}, h_i = \frac{\partial^2 L(y_i,\hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \tag{9}$$

(d)Build a new decision tree based on the derivative information

$$J(f_t) = -\frac{1}{2}\sum_{j=1}^{T_t}\frac{G_j^2}{H_j+\lambda} + \gamma^{T_t} \tag{10}$$

(e)The sample values are predicted based on the new decision tree and accumulated to the original values for n iterations until the condition is satisfied and stopped.

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{11}$$

- Step5: Classification prediction of the trained model on the test set.
- Step6: Return prediction results.

3. EMPIRICAL ANALYSIS

3.1 Data pre-processing.

This section includes the description and the missing values procession present in this dataset. It can improve the quality of the data and perform better operations.

3.1.1 Data description

The data set was selected from Kaggle. The dataset is provided by a Portuguese banking institution to classify customers by predicting whether they would subscribe to time deposit or not. There are 42,639 data in total, of which 3,961 are positive samples (subscribing to time deposit) and 38,678 are negative samples (not subscribing to time deposit). The dataset contains 17 attributes and its imbalance ratio is 1:10.

Table 1. Sample Attributes

Attributes name	means	Attributes type
AGE	age of customer	continuous
JOB	type of job	text
MARITAL	marital status	discrete
EDUCATION	educational situation	text
DEFAULT	has credit in default?	discrete
BALANCE	account balance	continuous
HOUSING	has housing loan?	discrete
LOAN	has personal loan?	discrete
CONTACT	contact communication type	text
DAY	last contact day of the week	continuous
MONTH	last contact month of year	continuous
DURATION	last contact duration	continuous
CAMPAIGN	number of contacts performed during this campaign and for this client	continuous
PDAYS	number of days that passed by after the client was last contacted from a previous campaign	continuous
PREVIOUS	number of contacts performed before this campaign and for this client	continuous
POUTCOME	outcome of the previous marketing campaign	discrete
TERM-DEPOSIT	has the client subscribed a term deposit?	discrete

3.1.2 Missing value handling

In Table 2, we can see the dataset used in this study has some missing values. POUTCOME has more missing values, the values of the feature are differentiated under different categories, so the feature cannot be deleted. Therefore, we delete the data that is missing this feature. After it, there are still three attributes, JOB, EDUCATION and CONTACT, with missing values, so we use plural to fill them. After the missing values are processed, there are 6554 data left. There are 972 positive samples and 5582 negative samples, and the imbalance ratio is 1:6.

Table 2. Sample Missing Values

Attributes Name	Number of missing values
JOB	264
EDUCATION	1690
POUTCOME	36085
CONTACT	12776
OTHER ATTRIBUTES	0

3.2 Model interpretability.

XGBoost is an ensemble algorithm consisting of regression trees. The inherent interpretability of the decision tree itself reduces the complexity of the model and enhances the interpretability of the overall model. An overall interpretation of the model decisions can be obtained by ranking the importance of the feature weights. The higher the value of the feature weight, the greater the contribution of the feature to improving the prediction model. In Figure 2, we can see the 16 attributes in descending order of importance, are HOUSING、DURATION、LOAN、POUTCOME、PDAYS、MONTH、DAY、CAMPAIGH、MARITAL、JOB、BALANCE、CONTACT、EDUCATION、AGE、PREVIOUS、DEFAULT.

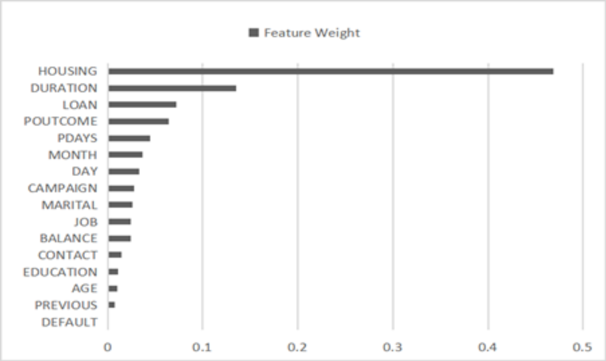

Figure 2. Feature weight importance ranking chart

3.3 Comparison of classification results.

The data partitioning method used in this study is the ten-fold cross-validation method. The method divides all data into ten mutually exclusive subsets of equal number and similar size. The nine of the obtained data are used as the training set and one as the test set. Iteratively, ten training and testing sessions are performed and evaluated both model accuracy and stability against four classification algorithms, Dt, Adaboost, XGBoost and SMOTE-XGBoost.

3.3.1 Evaluation indicators

The prediction of the experiment is a binary classification problem, so the three evaluation indicators, ACC, AUC and G-mean, are used to evaluate the effectiveness of the model. For the binary classification problem, the true kinds in the dataset and the categories obtained from the classifier predictions can be combined and divided into four categories, which can be represented by the confusion matrix as Table 3.

Table 3. Confusion matrix for binary classification problems

Actual Category	Prediction Category	
	Prediction is a positive case	Prediction is a negative case
Actual is a positive case	TP	FN
Actual is a negative case	FP	TN

(1) ACC

ACC represents the number of correctly classified samples as a percentage of the total number of samples. It is a more commonly used performance metric in classification tasks with imbalanced data. From the Table3, the ACC can be expressed as:

$$\text{ACC} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \tag{12}$$

(2) AUC

AUC is the area under the ROC curve and it is not affected by classification imbalance. In general, the higher the value of AUC, the better the performance of the classifier is.

(3) G-mean

The G-mean is a common evaluation indicator in the performance evaluation of imbalance classification. When the sample distribution may change over time or differ between the training and test set sample distributions, the G-mean has good robustness. The higher its value, the better the model's combined classification effect.

$$\text{G-mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{FN+FP}} \tag{13}$$

3.3.2 Model accuracy analysis

In this study, a ten-fold cross-validation method is used to compare DT, AdaBoost, XGBoost, SMOTE-XGBoost, and SMOTEENN-XGBoost. In addition to this, we use ACC, AUC, and G-mean to evaluate model performance. The classification performance results of the five models can be visually seen by Table 4.

Table 4. Comparison of mean values of five algorithms experimental results

Model	ACC	AUC	G-mean
DT	0.7207	0.7811	0.7207
AdaBoost	0.7670	0.8303	0.7614
XGBoost	0.8106	0.8678	0.8033
SMOTE-XGBoost	0.8125	0.8783	0.8049
SMOTEENN-XGBoost	0.9208	0.9711	0.9215

About the performance comparison of the five algorithms, we compare the mean values of the three evaluation indicators in Table 4. It is obvious that the ensemble algorithm has better performance than the single algorithm. Besides, it is verified that the classification performance of the XGBoost is better than AdaBoost, so we choose the XGBoost as the classification algorithm. The mean ACC, AUC, and G-mean of SMOTEENN-XGBoost are significantly better than SMOTE-XGBoost. It confirmed that the hybrid sampling method can make up for the lack of single sampling. In Table 4, it also confirmed that combined classifier model is better than single classifier. At the seam time, it can further improve the classification performance of the classifier.

By observing the mean ACC comparison of five algorithms in Table 4, the mean ACC of SMOTEENN-XGBoost reaches 0.92, but the other four algorithms are around 0.80. It indicates that the model has stronger differentiation ability and can better classify customers.

The value of AUC between 0 and 1. If its value is close to 1, it means the classification accuracy of the model is high. From Table 4, we can see that the mean AUC of SMOTEENN-XGBoost reaches 0.97. The mean AUC of DT, AdaBoost, XGBoost and SMOTE-XGBoost are only about 0.85. Therefore, the model with predict whether customers will subscribe to time deposits or not accurate.

The G-mean is a composite of positive case accuracy and negative case accuracy. The higher the value of G-mean, the more accurate the model discriminates positive or negative cases. The G-mean of SMOTEENN-XGBoost is 0.92, but the G-mean of other four algorithms are less than 0.81. It indicates that the model can accurately classify customers who will subscribe or not time deposits.

In summary, SMOTEENN-XGBoost has higher accuracy in predicting whether customers will subscribe to time deposits.

3.3.3 Model stability analysis

To some extent, stability analysis of classifier models can determine the goodness of the models. We use the standard deviation calculation to further capture the dispersion of the model classification results in ten experiments. Based on this, the values of each indicator in ten-fold cross-validation are plotted in a line figure. We can see the classifier models' stability more intuitive by the line figures.

Table 5. The stabilities of five models

	ACC			AUC			G-mean		
Model	Min	Max	SD	Min	Max	SD	Min	Max	SD
DT	0.7078	0.7818	0.0224	0.7381	0.8096	0.0234	0.6837	0.7520	0.0216
AdaBoost	0.7283	0.8024	0.0271	0.8056	0.8609	0.0164	0.7282	0.7949	0.0248
XGBoost	0.7818	0.8436	0.0193	0.8346	0.9102	0.0228	0.7781	0.8391	0.0179
SMOTE-XGBoost	0.7860	0.8559	0.0206	0.8575	0.9145	0.0169	0.7795	0.8502	0.0217
SMOTEENN-XGBoost	0.9029	0.9477	0.0150	0.9526	0.9859	0.0097	0.9036	0.9486	0.0151

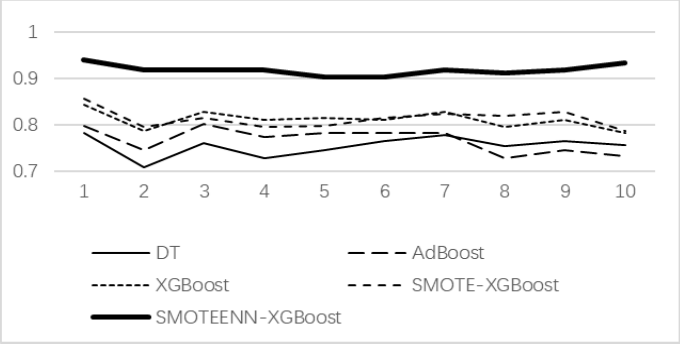

Figure 3. Line chart of overall ACC using ten-fold cross-validation

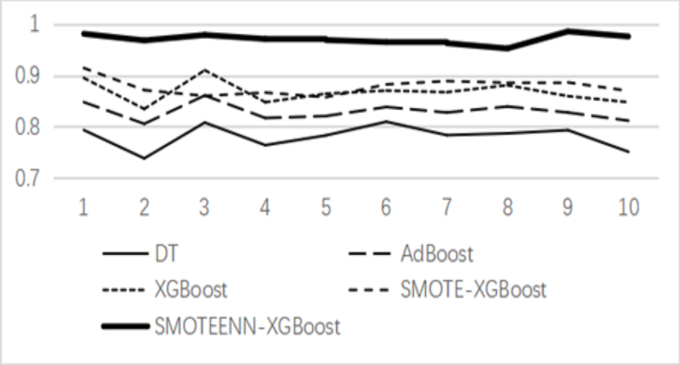

Figure 4. Line chart of AUC using ten-fold cross-validation

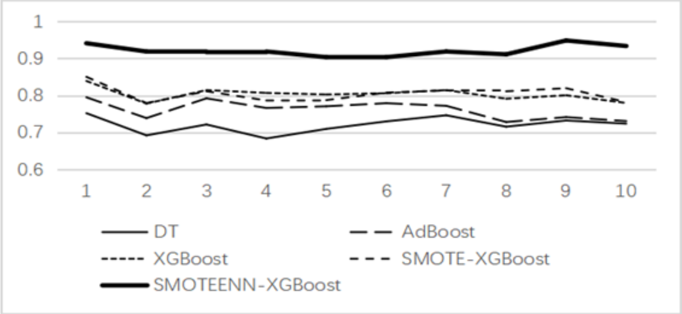

Figure 5. Line chart of G-mean using ten-fold cross-validation

Figures 3, 4, and 5 show the fluctuations of the evaluation indicators values about DT, AdaBoost, XGBoost, SMOTE-XGBoost, and SMOTEENN-XGBoost in ten cross-validation experiments. The horizontal axis showing the number of experiments and the vertical axis showing the values of the evaluation indicators. It is obvious that the SMOTEENN-XGBoost has the least fluctuation of ACC, AUC and G-mean. It indicates that the model has better stability.

4. DISCUSSION

In this paper, there are two significant theoretical contributions. First, we use a hybrid sampling algorithm to solve the imbalanced data processing problem. Although some ensemble algorithms can improve the classification accuracy of the model, there is still a certain shortcoming in the processing of imbalanced data. The existing studies mainly adopts single sampling method, such as SMOTE[13], ENN to solve the imbalanced data problem in terms of sampling. These single samplings can solve the impact of imbalanced data on classification accuracy to a certain extent, but they will lose important information and make the model produce the problem of overfitting. Therefore, we propose a hybrid sampling technique that combines SMOTE and ENN for imbalanced data. It named SMOTEENN. The hybrid sampling technique can not only retain valuable information but also avoid the problem of model overfitting. As confirmed by the data in Table 4, SMOTEENN-XGBoost compared with SMOTE-XGBoost, ACC, AUC and G-mean values are greatly improved. It indicates that the hybrid sampling is superior to single sampling. Second, we use XGBoost to overcome the deficiency of single classifier, and combines XGBoost with the SMOTEENN to further improve classification performance. The existing studies mainly use a single classifier [14] to predict whether bank customers will subscribe to time deposit. However, a single classifier only can be trained on a small amount of data. Besides, the single classifier has a small hypothesis space and can only solve locally optimal problems. Therefore, some scholars propose to use ensemble algorithm to improve the classification performance, mainly using the AdaBoost [15]. In Table 4, Compared with the single classifier DT, Adaboost improves the classification performance, but its training process is time-consuming. Therefore, this paper proposes to use XGBoost as the classification algorithm, which is significantly better than AdaBoost through the comparison in Table 4. On this basis, we combined SMOTEENN and XGBoost to overcome the shortcomings of a single model. The above indicates that SMOTEENN-XGBoost can make the prediction of customers' subscription to time deposits more accurate.

5. CONCLUSIONS, LIMITATIONS AND FUTURE RESEARCH

In today's information age, the banking industry is very competitive. Traditional marketing methods are eliminated by this era of big data. Machine learning can solve the problems of classification and prediction well. In the existing literature about time deposits prediction, a single sampling method is usually used at the data level. As for the algorithm level, the single classifier or ensemble algorithms based on AdaBoost is mostly used. However, combined predictive models are seldom used. Based on this background, we propose the SMOTEENN-XGBoost model to make predictions about whether customers will subscribe to time deposits. The model is suitable for solving the imbalance classification problem in the context of big data. First, the data is preprocessed by using the hybrid sampling of SMOTEENN. The SMOTEENN overcomes the shortage of single sampling effectively. Besides, it can reduce the impact of data imbalance to improve the prediction accuracy. Second, we combined XGBoost with the sampling technique. The combined model can further improve the model prediction accuracy. It is experimentally demonstrated that the ACC, AUC, and G-mean of the combined model are significantly higher than single model. Furthermore, we use ten-fold cross-validation method confirms that the model has good stability and classification performance compared with other models.

There are still three shortcomings in this paper. First, we only choose the features of one data source for the experiment. Actually, the features are most closely related to the classification effect, so future studies can choose new features for classification prediction, we can use the Customers' Subscription data to Time Deposits of China commercial bank to verify the effectiveness of the proposed method. Second, we only use SMOTEENN to process the imbalanced data. Therefore, in the future, scholars can consider using CNNTomek, SMOTETomek or other hybrid sampling methods to classification studies. Third, we do not improve the XGBoost algorithm directly, so future studies can improve the XGBoost algorithm by combing with cost-sensitive learning.

ACKNOWLEDGEMENT

This research was supported by the Natural Science Foundation of Zhejiang Province under Grant number LY20G010008 and the National Natural Science Foundation of China under Grant number 71472169, 71572187.

REFERENCES

[1] Petropoulos A, Vlachogiannakis N E, Mylonas D. (2018). Forecasting private sector bank deposits in Greece: Determinants for trend and shock effects. International Journal of Banking Accounting and Finance, 9(2): 141-169.

[2] Mitat U. (2017). Comparison of ARIMA and RBFN Models to Predict the Bank Transactions. Information Technology Journal, 6(3): 475-477.

[3] Hosmer D W, Hosmer T, Le C S, Lemeshow S. (2015). A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine, 16(9): 965-980.

[4] Keerthi S S, Shevade S K, Bhattacharyya C, Murthy K. (2014). Improvements to platt's smo algorithm for svm classifier design. Neural Computation, 13(3): 637-649.

[5] Stroie L. (2014). Predicting consumer behavior with artificial neural networks. Procedia Economics and Finance, 15: 238-246.

[6] Kumar U, Simaiya S, Prasad D. (2021). Hybrid weighted random forests method for prediction & classification of online buying customers. Journal of Information Technology, 13(2): 245-259.

[7] Krawczyk B, Galar M, Jelen L, Herrera F. (2016). Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Applied Soft Computing, 38: 714-726.

[8] Abellan J, Mantas C J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, 41(8): 3825–3830.

[9] Kontsewaya Y, Antonov E, Artamonov A. (2021). Evaluating the effectiveness of machine learning methods for spam detection. Procedia Computer Science, 190: 479-486.

[10] Raghuwanshi B S, Shukla S. (2018). Class-specific extreme learning machine for handling binary class imbalance problem. Neural Networks, 105: 206-217.

[11] Zhang D, Wei L, Gong X, Hui J. (2011). A novel improved smote resampling algorithm based on fractal. Journal of Computational Information Systems, 7(6): 2204-2211.

[12] Hastie T, Rosset S, Zhu J, Zou H. (2009). Multi-class adaboost. Statistics and its Interface, 2(3): 349-360.

[13] Wang K J, Makond B, Chen K H, Wang K M. (2014). A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients. Applied Soft Computing Journal, 20: 15-24.

[14] Luthfiarta A, Zeniarja J, Faisal E, Wicaksono W. (2020). Prediction on deposit subscription of customer based on bank telemarketing using decision tree with entropy comparison. Journal of Applied Intelligent System, 4(2): 57-66.

[15] Wang W, Sun D. (2021). The improved adaboost algorithms for imbalanced data classification. Information Sciences, 563: 358-374.