Summer 7-26-2022

Prediction of Credit Card Defaulters Based on SMOTE-XGBoost Model

Lingmin Jin
College of Economics and Management, China Jiliang University, China

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

Jiali Zhao
College of Economics and Management, China Jiliang University, China

Full Research Paper

Prediction of Credit Card Defaulters Based on SMOTE-XGBoost Model

Lingmin Jin[1], *Zengyuan Wu*[1*], *Jiali Zhao*[1]

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

Abstract: Credit card defaulters are on the rise year by year, which would lead commercial banks into a serious business crisis. It is important for commercial banks to control the default rate of credit cards. According to the low percentage of defaulters, it is challenging to predict them using a traditional machine learning algorithm. To address this problem, an improved ensemble learning model is proposed, where the Synthetic Minority Oversampling Technique (SMOTE) is used to oversample the data set, and the Extreme Gradient Boosting algorithm (XGBoost) is introduced to construct the predicting model. For clarity, this model is called a SMOTE-XGBoost model. Customer default data from the UCI machine learning dataset is used to empirically test the effectiveness. In terms of Recall, ACC, and AUC values, ten-fold cross-validation is carried out to evaluate and compare the performance between the SMOTE-XGBoost model and other models, including the general XGBoost model and Random Forest. The empirical results show that the SMOTE-XGBoost model performs well and outperforms other models.

Keywords: credit card default, unbalanced data, the Extreme Gradient Boosting algorithm(XGBoost), the Synthetic Minority Oversampling Technique (SMOTE)

1. INTRODUCTION AND LITERATURE REVIEW

With the continuous advancement of technology and the improvement of people’s living standards, the application scenarios of credit card payment are spreading in all areas of life [1]. According to credit card industry statistics, China's credit card loan balance has rapidly increased from 0.16 trillion yuan in 2009 to 7.59 trillion yuan in 2019. At the end of 2018, there are approximately 530 million natural persons with credit collection records, and the number of credit card holders is expected to reach 500 million. As of the end of 2019, the proportion of credit card loans overdue for more than six months in China's banking industry is 1.15 percent of total credit card loan receivables. Simultaneously, credit card repayable balances and credit utilization rates are rising. It causes an increase in spending and overdraft amounts, which inevitably creates some risk. The ability of borrowers to repay the loan on time will have a significant impact on the commercial bank's daily operations [2]. As a result, it is critical to find ways to reduce default risk through data mining by effectively analyzing and utilizing the data generated by credit card users [3].

As the total amount of current loans grows, so does the likelihood of customer default continues to rise, and the risk of customer loans becomes more apparent [4]. Scholars have currently conducted extensive research in the field of credit card default. Li et al. [5] investigated the impact of diversity, independence and social factors on credit card defaults in China, and discovered that credit card defaults were not related to the income of credit card customers, but significantly to the stability of their income. Bursztyn et al. [6] investigated the role of ethical factors in credit card debt repayment. They discovered that ethical claims significantly reduced personal credit default rates and reduced defaults by customers with the highest prior credit risk. The preceding study introduces the factors of credit card default and the analysis of the forecast scheme. Aside from these aspects, of course, the field of credit card default research is vast.

Recent research has focused on the development of a credit card default prediction model and how to improve the model's accuracy. Ogundimu [7] reviewed some methods for dealing with class imbalanced data, and his results showed that the log-F prior and ridge regression methods were preferred among the above models, and the

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

SMOTE improved predictive accuracy than random oversampling technology. Leow and Crook [8] estimated the exposure at the default of the obligor by estimating the outstanding balance of an account. The results showed that their prediction was more accurate than other models at any time over the entire default loan period.

The preceding study mainly uses a single classification algorithm to study the risk prediction of default. A single classification algorithm has the advantages of simple modeling and fast model training. But some defects are prone to local optimality. In the current mainstream machine learning algorithm, ensemble learning is a new and effective algorithm to enhance the effectiveness of the model. By cascading several base classifiers to generate strong classifiers, an ensemble learning algorithm can improve the accuracy and stability of the model. Recently, several scholars have proposed an ensemble learning model to assess default risk. Yu et al. [9] used a multi-view ensemble learning method based on model distance and adaptive clustering to predict default risk in P2P lending. Hayashi et al. [10] first used a one-dimensional fully connected layer CNN with recursive rule extraction algorithm with decision tree and this method was very effective in extracting highly concise rules for heterogeneous credit scoring datasets.

The preceding literature uses an ensemble learning algorithm to predict the risk of customer default and obtain better prediction results than the single classification algorithm. However, there are still some limitations in applying ensemble learning algorithms in default risk prediction directly. This is due to the unbalanced nature of customer credit data, as evidenced by the extremely unbalanced ratio of defaulting to non-defaulting customers. However, few studies have taken into account the unbalanced data and used data under-sampling methods to keep data balanced. As the popular under-sampling technology, SMOTE is employed by some scholars during the data pre-processing stage, which can effectively improve performance. Byeon [11] used machine learning algorithms to build models for predicting depression in older people living in the community and confirmed that the SMOTE-based random forest algorithm showed the highest accuracy and best predictive performance in random forest, GBM and logistic regression analyses. Wang et al. [12] proposed a new near-infrared spectroscopy identification model for diesel brands that combined tree-based feature selection, SMOTE, and XGBoost ensemble learning to achieve high accuracy and speed.

According to the existing literature in the field of credit card default prediction, it is not difficult to find that they mainly focus on high-dimensional and nonlinear data[13]. However, few studies have focused on the imbalanced data, which may not be conducive to the effective identification of credit card credit defaulters for the following reasons. First, in binary classification, the strong classifier generated in an ensemble learning model are generally superior to one single base classifiers. When a single classifier is applied in a dataset, data overfitting often occurs [14]. Data overfitting means that the final model works well on the training set, but it works on the testing set badly. That is, the generalization ability of the model is weak, and the prediction accuracy is reduced. Second, given binary classification, class imbalance occurs when minority samples are far less than majority samples. When the categories of non-defaulters and defaulters of credit card loans are unevenly distributed, it is difficult to effectively distinguish minority class samples from majority samples, which can easily lead to poor performance in predicting score card defaulters.

Therefore, this paper addresses the following question: according to the unbalanced characteristics of credit card credit data, how can we use machine learning to improve the accuracy of predicting credit card default? Currently, there are two solutions to the problem of unbalanced data classification in machine learning. The first is data-level research. The second is algorithm-level research, which aims to improve machine learning algorithms. In the first data-level research, resampling methods are mainly used to adjust the distribution of training data sets to keep them balanced. For example, Ogundimu [7] applied random oversampling techniques and synthetic minority over-sampling techniques to predict credit-card defaults and concluded that synthesizing a few oversampling techniques could improve predictive accuracy. In the second algorithm-level research, Chen et al.

[15] proposed a novel sparse data perception algorithm for sparse data, where a weighted quantile sketch was used to approximate the learning of implementing trees. More importantly, they provided insights into cache access patterns, and data compression to build a malleable lift tree system. By combining these insights, the XGBoost can handle billions of dollars of data with far fewer resources than other models. This model supports column sampling and row sampling, which can reduce both the risk of overfitting and the calculation. Furthermore, it can help improve the accuracy of the prediction results.

Therefore, in response to the above-mentioned unbalanced distribution problem of credit card customer loan data, an ensemble learning model is proposed in this paper. It is named SMOTE-XGBoost, where SMOTE is used to address the problem of unbalanced distribution by oversampling, and the XGBoost algorithm is introduced to construct the predicting model. Furthermore, the data from the UCI machine learning dataset is used to test the model's performance. Finally, the model's performance is compared with the general XGBoost model and Random Forest. Recall, ACC and AUC values are used as the evaluation index for classification to evaluate the validity of prediction results.

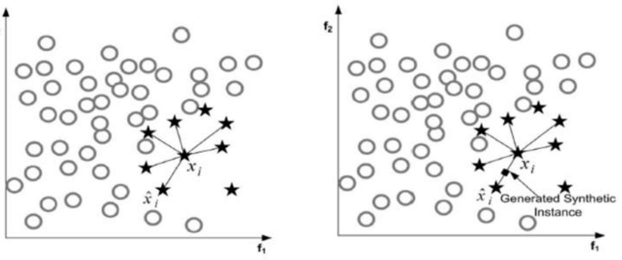
The rest sections of this paper are organized as follows. In Section 2, the details of SMOTE and XGBoost algorithm are provided. In Section 3, the SMOTE-XGBoost model is implemented to predict credit card defaults. In section 4, several conclusions and future research are drawn.

2. SMOTE-XGBOOST MODEL

2.1 The Synthetic Minority Oversampling Technique (SMOTE).

The SMOTE is mainly used to solve the classification problem of the unbalanced data set. From the perspective of training models, the large size of negative samples with low-value density tends to have a negative effect on the classification of positive samples with high-value density. Such models are trained with less value. SMOTE is an improved scheme based on the random oversampling technique. The basic idea is to randomly select n minority class sample points from the original sample point' K nearest neighbors, and a new sample point is generated by randomly selecting a point on the line between the original sample point and one of the above n sample points. The core of SMOTE is that the features of the neighboring points on the feature space are similar. It does not sample on the data space, but in the feature space, so it will be more accurate than the traditional sampling method.

As shown in Figure 1, the oversampling process is as follows:



(a) (b)

Figure 1. Example of SMOTE

- For each sample *x* in the minority class, the Euclidean distance is used as a criterion to calculate its distance from this sample to all the remaining samples in the minority class sample set to obtain its *k*-nearest neighbors.
- A sampling ratio is set according to the sample imbalance ratio to determine the sampling multiplicity *N*. For each minority class sample *x*, a number of samples are randomly selected from its *k*-nearest neighbors, assuming the selected nearest neighbors are x_n.
- For each randomly selected nearest neighbor, a new sample is constructed separately from the original

sample according to the following equation (1).

$$X_{new} = x + rand(0,1) * |x - x_n| \tag{1}$$

2.2 Subsection.

XGBoost is an ensemble tree model whose sparse-aware algorithm has an inherent advantage for handling sparse data. Meanwhile, a regularization term is added to its objective function, which can effectively avoid overfitting.

XGBoost belongs to Gradient Boosting Decision Tree (GBDT) models. The basic idea of GBDT is to let the new base model (GBDT uses categorical regression tree as the base model) fit the deviation of the previous model, thus continuously reducing the deviation of the additive model. Compared to the classic GBDT, XGBoost has made some improvements, resulting in a significant improvement in effectiveness and performance. First, GBDT expands the objective function to the first order based on Taylor expansion, while XGBoost expands the objective function to the second order. More information about the objective function is retained, which is helpful for the boosting effect. Second, GBDT is to find a new fit label for the new base model (negative gradient of the previous additive model), while XGBoost is to find a new objective function for the new base model (second-order Taylor expansion of the objective function concerning the new base model). Third, XGBoost incorporates L2 regularization terms for the number of leaf nodes and leaf node weights, thus facilitating the model to obtain a lower variance. Fourth, XGBoost adds a strategy to automatically handle missing value features. By dividing samples with missing values into left subtree or right subtree respectively and comparing the advantages and disadvantages of the objective functions under the two schemes, the samples with missing values are automatically divided without the need to pre-process the missing features for filling. In addition, XGBoost supports candidate quantile cuts, feature parallelism, etc., which can improve the performance.

The process of the XGBoost algorithm is described as follow:

The input is the training set samples $I = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, maximum number of iterations is *T*, loss function is *L*, regularization factors are λ and γ. The output is the strong learner *f(x)*. For the number of iterative rounds *t = 1, 2, ...T*, we have as follows:

- Step 1: Compute loss function *L* of the *i*-th sample (*i=1, 2, ... m*) in the current round based on the first order derivative of $f_{t-1}(x_i)$ which called g_{ti}, and second order derivative called h_{ti}. Calculate the sum of the first order derivatives of all samples $G_t = \sum_{i=1}^{m} g_{ti}$ and the sum of second order derivatives of all samples $H_t = \sum_{i=1}^{m} h_{ti}$.
- Step 2: Try to split the decision tree based on the current node, default score is 0, *G* and *H* are the sum of the first-order derivatives and second-order derivatives of the nodes that currently need to be split.

For feature serial number *k=1,2...K.*

（a）G_L=0, H_L=0.

（b）Arrange the samples by feature k from smallest to largest, take out the i-th sample in turn, after calculating the current sample into the left subtree in turn, the sum of the first- and second-order derivatives of the left and right subtrees is

$G_L=G_L+g_{ti}$, $G_R=G-G_L$,

$H_L=H_L+h_{ti}$, $H_R=H-H_L$,

（c）Try to update the maximum *score.*

$$score = \max(score, \frac{1}{2}\frac{G_L^2}{H_L+\lambda} + \frac{1}{2}\frac{G_R^2}{H_R+\lambda} - \frac{1}{2}\frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma) \tag{2}$$

- Step 3: Split subtree based on the division features and eigenvalues corresponding to the maximum *score.*
- Step 4: If the maximum *score* is 0, then the current decision tree is built. Calculate w_{tj} for all leaf regions, get the weak learner $h_t(x)$ and update strong learner $f_t(x)$, then go to the next round of weak learner iterations. If the maximum *score* is not 0, then go to step 2 and continue to try to split the decision tree.

3. EMPIRICAL ANALYSIS

3.1 Data description and pre-processing.

The data set used in this study is taken from UCI Machine Learning Repository "default of credit card clients Data Set". It provides a total of 30,000 pieces of data covering credit amount, age, gender, education level, marital status, account status, payment status, payment amount, etc.The names and meanings of the variables are shown in Table 1.

Table 1. Names and meanings of variables

Variables	Name of variables	Meaning of variables
Y	DEFAULT PAYMENT NEXT MONTH	default payment (Yes = 1, No = 0)
X1	LIMIT_BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2	SEX	Gender (1 = male; 2 = female).
X3	EDUCATION	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4	MARRIAGE	Marital status (1 = married; 2 = single; 3 = others).
X5	AGE	Age (year).

X6-X11	PAY_0, PAY_2,
	PAY_3, PAY_4,
	PAY_5, PAY_6
		History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:
		X6 = repayment status in Sept. 2005; X7 = repayment status in Aug. 2005;
		X8= repayment status in Jul. 2005; X9= repayment status in Jun. 2005;
		X10=repayment status in May 2005; X11=repayment status in Apr. 2005.
		The measurement scale for the repayment status is:
		-1 = pay duly;
		1 = payment delay for 1 month; 2 = payment delay for 2 months;
		3 = payment delay for 3 months; 4 = payment delay for 4 months;
		5 = payment delay for 5 months; 6 = payment delay for 6 months;
		7 = payment delay for 7 months; 8 = payment delay for 8 months;
		9 = payment delay for 9 months and above.

X12-X17	BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6	Amount of bill statement (NT dollar). X12 = amount of bill statement in Sept. 2005; X13 = amount of bill statement in August, 2005; X14 = amount of bill statement in July, 2005; X15 = amount of bill statement in June, 2005; X16 = amount of bill statement in May, 2005; X17 = amount of bill statement in April, 2005.

X18-X23	PAY_AMT1,
	PAY_AMT2,
	PAY_AMT3,
	PAY_AMT4,
	PAY_AMT5,
	PAY_AMT6
		Amount of previous payment (NT dollar).
		X18 = amount paid in September, 2005;
		X19 = amount paid in August, 2005;
		X20 = amount paid in July, 2005;
		X21 = amount paid in June, 2005;
		X22 = amount paid in May, 2005;
		X23 = amount paid in April, 2005.

The "X3" education level variable in this dataset contains 14 missing values. The "X4" marital status variable contains 54 missing values. In "X12", the negative amount of September billing indicates that there is a deposit balance without repayment, so there is no possibility of default in this case, but the sample shows that the number of defaulters is 109, and such data is invalid and should be discarded. The actual number of samples obtained after pre-processing is 29,823.

According to the data of the credit card defaulters in this paper, the non-default sample size was 23,301 and the default sample size was 6,522, accounting for 21.81% of the total. The ratio of non-default sample to default sample is 3.57:1. Generally, data ratios of more than 3:1 are considered to be a significant imbalance in the sample categories. The sample-set is now partitioned in a ratio of 9: 1 and the feature variables are separated from the target variables.

3.2 Evaluation indicators.

The prediction of loan risk in this experiment is a dichotomous model, so the Recall, ACC, and AUC values are selected as evaluation indicators, and the risk prediction confusion matrix is shown in Table 2, defining positive samples as loan non-default and negative samples as loan default. Where TP implies that observation is positive, and is predicted to be positive. FP implies that observation is negative, but is predicted positive. TN implies that observation is negative, and is predicted to be negative. FN implies that observation is positive, but is predicted negative.

Table 2. Risk forecasting confusion matrix

Confusion Matrix		Predicted Class	
		Positive Sample	Negative Sample
Actual Class	Positive Sample	TP	FN
	Negative Sample	FP	TN

Recall is a metric of coverage, and the metric has multiple positive cases classified as positive cases, which is calculated as shown in equation (3).

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Accuracy (ACC) reflects the rate at which the classifier accurately identifies true positives and false negatives, which is calculated as shown in equation (4).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

AUC (Area Under Curve) is defined as the area under the ROC curve. We tend to use the AUC value as the evaluation criterion of the model because the ROC curve usually does not indicate which classifier is better, and as a value, the classifier corresponding to a larger AUC is better.

The ROC curve, known as the receiver operating characteristic curve, is a curve based on a series of different dichotomies (cut-off values or decision thresholds), with the true positive rate (TPR) as the vertical coordinate and the false positive rate (FPR) as the horizontal coordinate, reflecting the relationship between sensitivity and specificity. where TPR indicates the percentage of samples that were correctly judged as positive among all samples that were actually positive, and FPR indicates the percentage of samples that were incorrectly judged as positive among all samples that were actually negative. The formulae for each of the two are shown in equations (5) and (6).

$$TPR = \frac{TP}{TP+FN} \tag{5}$$

$$FPR = \frac{FP}{FP+TN} \tag{6}$$

AUC is a performance metric that measures the merit of a learner. From the definition, AUC can be obtained by summing the area of each part under the ROC curve. The better the classification, and thus the higher the ROC curve, the greater the value of AUC. The AUC value is generally floating between 0.5 and 1. If the AUC is less than 0.5, it means that the model is even less accurate than the random results. Similarly, if the AUC is equal to 1, it indicates that the classification accuracy is 100%.

3.3 Model training and evaluation.

3.3.1 Feature selection

Since there are many input variables, not every variable contributes to the training performance of the model. On the contrary, it may reduce the effectiveness of the model due to correlation. It is necessary to perform feature selection first and eliminate variables of low importance [16].

In this paper, the importance of 23 input variables was rated using F-scores, and the top 10 variables with high importance were selected: “X5 (AGE)”, “X19 (PAY_AMT2)”, “X6 (PAY_0)”, “X2 (SEX)”, “X21 (PAY_AMT4)”, “X20 (PAY_AMT3)”, “X1 (LIMIT_BAL)”, “X18 (PAY_AMT1)”, “X23 (PAY_AMT6)”, “X4 (MARRIAGE)”, as shown in Figure 2.

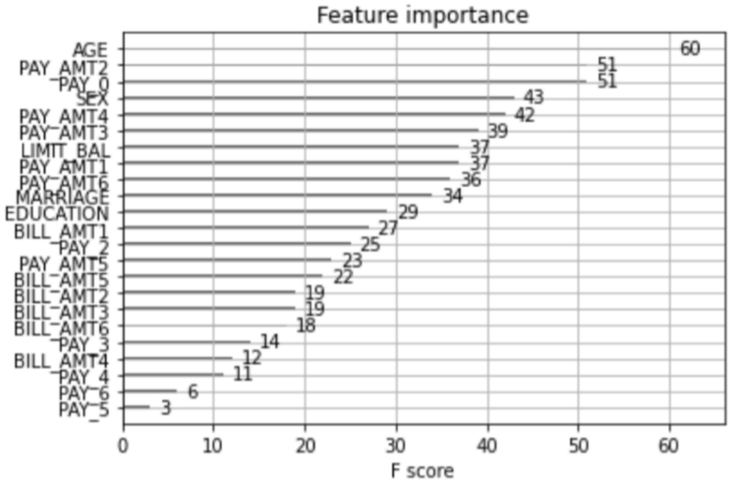


Figure 2. Importance ranking of feature selection variables

3.3.2 Descriptive analysis

The age distribution of credit card customers in the sample is mainly concentrated between 20 and 40 years old, where the youngest credit card customer is 21 years old and the oldest is 79 years old. The age distribution of credit card customers is shown in Table 3.

Table 3. Age distribution of credit card customers

Age range	(20,30]	(30,40]	(40,50]	(50,60]	(60,70]	(70,80]
Account	10974	10639	5963	1977	255	15

The distribution of credit card customer loan default rates in the sample data with respect to age and gender is shown in Figure 3 and Figure 4. From their bar and line graphs, the credit card customer loan default sample is mainly concentrated between the ages of 20 and 30, followed by the age group of 30 to 40 years old, with a default rate of nearly 33%. The age group between 60 and 80 years old has the lowest default rate, accounting for only 1% of these loans. Compared with men, the default rate for female loans is much higher than that of male credit card customers. This result indicates that credit card customers in their 20s and 30s have higher loan demand than credit card customers in other age groups. Similarly, credit card customers in this age group have higher default rates than credit card customers in other age groups. With the rapid changes of the times, young people's demand for life is growing. They are quick to accept new concepts of consumption and financial management. Young people have a strong sense of overspending, and their demand for borrowing is growing fast. However, due to their young age, low capital accumulation and weak risk tolerance, they are less able to deal with external emergencies and are prone to shortage of liquidity, resulting in late repayment. Commercial banks need to strengthen the eligibility of credit card customers in this age group when lending money and strictly control the default of credit card customers in this age group.

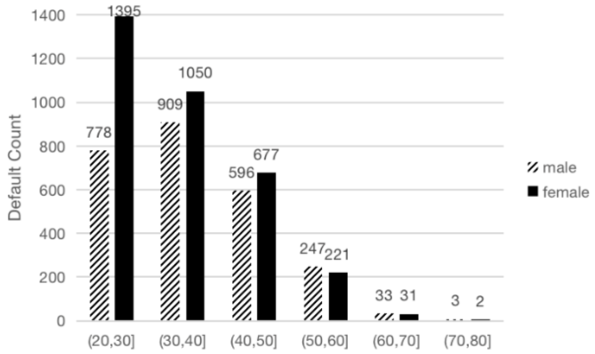


Figure 3. Credit card customer gender, age and number of defaults bar chart

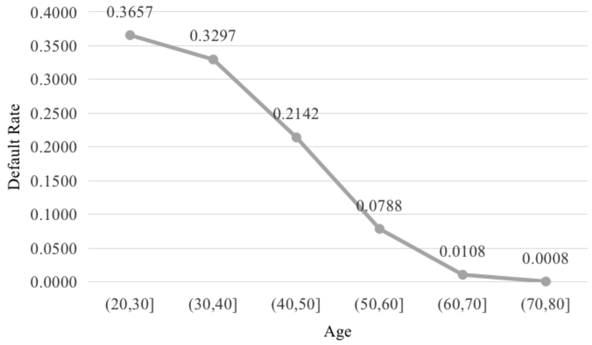


Figure 4. Line graph of the relationship between credit card customer age and default rate

3.3.3 Model evaluation

The XGBoost algorithm and the random forest algorithm are constructed for training on the training set, and also tested on the test set. Since the ratio of positive and negative classes in the original training set is severely imbalanced, in order to solve this problem, it is necessary to ensure that SMOTE oversampling is performed after the training and test sets are split, and only the training set is oversampled. Therefore, the total sample size of 29,823 is divided into a 9:1 ratio to construct a new training set with a sample size of 26,841 and a new test set with a sample size of 2,982. The specific findings are shown in Table 4.

Table 4. The evaluation results of model application

	Random Forest	XGBoost	SMOTE-XGBoost
Recall	0.558	0.646	0.724
ACC	0.781	0.769	0.783
AUC	0.765	0.796	0.856

From Table 4, the Recall of random forest is only 0.558 and the AUC value is also the lowest among the three algorithms when predicting credit card customer default risk by all three algorithms. However, the performance is better in ACC value. The XGBoost algorithm outperforms the random forest algorithm in terms of Recall and AUC values, which are 0.646 and 0.796, respectively. However, it performs poorly in terms of ACC values. XGBoost algorithm is a highly efficient machine learning algorithm. Compared with other methods, it can overcome the problem that traditional methods are not efficient enough when the data size is large. The model training speed is better than other algorithms, and the accuracy can be improved effectively by using large-scale data. The SMOTE-XGBoost model performed the best performance in terms of Recall, ACC and AUC values with 0.724, 0.783, and 0.856, respectively, by oversampling with the SMOTE algorithm and increasing the number of minority class samples. It can be seen that the SMOTE algorithm does make better use of the majority class sample to obtain more information in the category imbalanced dataset, thus improving its accuracy.

In the existing literature on credit card default, some scholars also use the XGBoost model in combination with other algorithms to build models and predict default behaviors. For example, Zhang et al. [17] used random forests to extract features from the perspective of Filter and Wrapper, and sampled training set samples using the SMOTE algorithm. In the model training stage, the particle swarm optimization algorithm is used to improve the classification accuracy of the XGboost model. Finally, the data provided by an open-source bank dataset is used for instance validation. The final results show that the AUC value of the model that has not been processed by SMOTE is 0.689, the value of the model AUC that has not been extracted by feature is 0.658, and the AUC value of the model used in this study is 0.786. Since the AUC value is used as the evaluation index in their study, we compare the AUC value in this paper with our results. It is clear that the default prediction accuracy of the SMOTE-XGBoost model used in our study is higher, and the AUC value reaches 0.856. It shows that the performance of SMOTE-XGBoost model is better when predicting credit card defaults.

The combined comparison of the three metrics, Recall, ACC, and AUC value, shows that ensemble learning has better prediction results than the single classifier model. In the unbalanced dataset, the SMOTE-XGBoost model predicts better after applying the SMOTE algorithm to balance the data, indicating that the combined use of the SMOTE algorithm and ensemble learning is better for predicting default risk.

4. CONCLUSION AND PROSPECT

With the rapid growth of the credit card business, it is necessary for banks and other financial institutions to establish a systematic monitoring and prediction model in terms of credit risk. On the one hand, it guides and informs the actual business expansion. On the other hand, it achieves the purpose of early warning and prevention and control of default risk, so that it comes from the business but ultimately serves the business. The SMOTE technique and a XGBoost algorithm are combined for predicting credit card defaulters. Firstly, the SMOTE algorithm is used to oversample the training set samples in the training set and to pre-process the input data. Secondly, the SMOTE-XGBoost model is proposed. Furthermore, the model is tested by using the data from the UCI machine learning dataset. The performance of the model is evaluated in terms of Recall, ACC, and AUC values. Ten-fold cross-validation is carried out to evaluate the performance among the SMOTE-XGBoost model, the general XGBoost model, and Random Forest.

The results demonstrate that the SMOTE-XGBoost model performs well than other models. First, the SMOTE-XGBoost model comprehensively adopts the resampling technique and ensemble learning algorithm. It is effective to solve the problem of data imbalance. Second, compared with the other two models, the SMOTE-XGBoost model has higher overall classification accuracy and better performance in predicting credit card defaulters. This confirms that our model is suitable for working with unbalanced data. Finally, with ten-fold cross-validation, the model proposed in this paper is more stable than the other two methods.

Of course, there are some shortcomings associated with SMOTE-XGBoost model. First, more personal data should be examined, such as income, housing, vehicle ownership et al. Second, an XGBoost algorithm with the SMOTE oversampling technique is used in various fields of research, such as orphan genes [18], brand recognition of diesel fuel [12]. However, whether the sampling results of SMOTE algorithm are the most representative? whether the parameter adjustment of the XGBoost algorithm has reached the optimal degree? In the future, it is a good attempt to introduce other algorithms for feature engineering or use other algorithms to achieve better performance. In addition, other credit card debit datasets could be used to further validate the SMOTE-XGBoost model.

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] He H, Fan Y A. (2021). A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. Expert Systems with Applications, 176(4): 114899

[2] Butaru F, Chen Q P, Clark B, Das S, Lo A W, Siddique A. (2016). Risk and risk management in the credit card industry. Journal of Banking & Finance, 72: 317-331

[3] Moradi S, Rafiei F M. (2019). A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. Financial Innovation, 5(1): 15

[4] Martins C, Oliveira T, Popovic A. (2014). Understanding the Internet banking adoption: A unified theory of acceptance and use of technology and perceived risk application. International Journal of Information Management, 34(1): 1-13

[5] Li Y Y, Li Y, Li Y. (2019). What factors are influencing credit card customer's default behavior in China? A study based on survival analysis. Physica A-Statistical Mechanics and its Applications, 526: 120861

[6] Bursztyn L, Fiorin S, Gottlieb D, Kanz M. (2019). Moral Incentives in Credit Card Debt Repayment: Evidence from a Field Experiment. Journal of Political Economy, 127(4): 1641-1683

[7] Ogundimu E O. (2019). Prediction of default probability by using statistical models for rare events. Journal of the Royal Statistical Society Series A, 182(4): 1143-1162

[8] Leow M, Crook J. (2016). A new mixture model for the estimation of credit card exposure at default. European Journal of Operational Research, 249(2): 487-497

[9] Yu S A, Yw A, Xin Y A, Dw B, Yy C, Yw A. (2020). Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in p2p lending. Information Sciences, 525: 182-204

[10] Hayashi Y, Takano N. (2020). One-Dimensional Convolutional Neural Networks with Feature Selection for Highly Concise Rule Extraction from Credit Scoring Datasets with Heterogeneous Attributes. Electronics, 9(8): 1318

[11] Byeon H. (2021). Predicting the Depression of the South Korean Elderly using SMOTE and an Imbalanced Binary Dataset. International Journal of Advanced Computer Science and Applications, 12(1): 74-79

[12] Wang S, Liu S, Zhang J, Che X, Yuan Y, Wang Z. (2020). A new method of diesel fuel brands identification: smote oversampling combined with xgboost ensemble learning. Fuel, 282: 118848

[13] Carcillo F, Borgne Y, Caelen O, Kessaci Y, Bontempi G. (2021). Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection. Information Sciences, 557: 317-331

[14] Wu Z, Zhou C, Xu F, Lou W. (2020). A CS-AdaBoost-BP model for product quality inspection. Annals of Operations Research. https://doi.org/10.1007/s10479-020-03798-z

[15] Chen T Q, Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. In: Assoc Comp Machinery, eds. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, 2016. New York: ASSOC COMPUTING MACHINERY, 785-794

[16] Xie Y, Liu G, Cao R, Li Z, Yan C, Jiang C. (2019). A Feature Extraction Method for Credit Card Fraud Detection. In: IEEE, eds. 2nd International Conference on Intelligent Autonomous Systems (ICoIAS), Nanyang Technol Univ, SINGAPORE, 2019. New York: IEEE, 70-75

[17] Zhang Lei, Wang Jiaqi, Fei Zhiyou, Luo Shuai, Sui Jingqi. (2020). Bank users' private credit risk assessment model based on RF-SMOTE-XGboost. Modern Electronics Technique, 43(16): 76-81(in Chinese)

[18] Gao Q, Jin X, Xia E, Wu X, Li S. (2020). Identification of orphan genes in unbalanced datasets based on ensemble learning. Frontiers in Genetics, 11: 820