Imbalanced Data Credit Scoring Model Based on Group-Lasso Method

WEI Yongfeng, XIANG Yibo

Abstract

In view of the complexity of the customers' credit risk faced by commercial banks at the present, how to manage customers' credit risk is very important. Customers' credit risk modeling is a key step. We use the credit card data of a commercial bank to construct a credit scoring model and predict the default probability. We construct a credit scoring model based on Logistic regression, using the group-Lasso (AUC criterion) method to select variables, and using the ROSE (Random Over Sampling Examples) method to deal with the imbalanced categories. The results are compared and analyzed, and the new model constructed in this work has certain advantages in discriminating ability and predictive ability. It can play a guiding role for banks and other financial institutions in evaluating customer credit risk and can be used as an effective basis for customer credit evaluation decisions. In practice, it also has good operability.

Keywords: Credit Scoring; Logistic Regression; Group-Lasso Method; ROSE

With the rapid development of the financial industry, the changes in the public's consumption concept, the continuous enrichment of personal credit products, and the substantial increase in the scale of personal credit business, the problem of personal credit risk has become increasingly prominent. Personal credit risk is the most important and complex risk faced by commercial banks at present, so how to carry out effective personal credit risk management to reduce default risk has become the core issue of research. "Basel New Capital Accord" pointed out that qualified banks should implement the internal rating method and calculate the default probability of customers by building models on historical data [1]. As one of the key factors affecting credit risk, default probability is the basis of credit risk measurement [2]. In credit risk management, credit scoring models play an important role.

At present, the personal credit scoring models based on machine learning include linear discriminant analysis [3], Logistic regression [4], neural network [5], support vector machine [6], genetic algorithm [7], cluster analysis [8], the K- nearest-neighbor model [9], decision tree [10] and Bayesian method [11], etc. Among them, logistic regression is the most widely used in personal credit scoring. The logistic model has the advantages of simple calculation, strong interpretability, and high prediction accuracy, and has been verified in China's mortgage credit risk assessment [12]. However, with the continuous development of society and the economy, the factors affecting personal credit risk are increasing, traditional logistic regression cannot effectively deal with the problem of multicollinearity of independent variables in the model, and too many variables increase the complexity of the model, which reduces the prediction accuracy of the model. Therefore, scholars continue to conduct in-depth research on this basis. Aiming at the shortcomings of logistic regression, Lee and Zhang [13] improved the prediction performance of the Logistic regression model by optimizing the non-uniform sampling of samples in the personal credit scoring model. Jongh et al. [14] proposed to eliminate the influence of variable multicollinearity in Logistic regression on personal credit scores through large sample data. Wei Qiuping and Zhang Jingxiao [15] constructed a credit scoring model based on the partial least squares method, Shi Xiaokang and He Xiaoqun [16] based on the biased Logistic regression method for the application of the personal credit rating model, both of which can effectively reduce multicollinearity in the independent variables in the model. However, the above-mentioned models also have certain limitations. They do not fully consider the imbalance and lack of sample data themselves and do not make necessary selections of variables. Tibshirani[17] proposed that the Lasso method can realize variable selection and parameter estimation at the same time, and can compress the coefficients of some independent variables in the model to make them tend to zero, so as to achieve the purpose of variable selection. Zhang Tingting and Jing Yingchuan [18] directly introduced the improved adaptive Lasso-Logistic regression model to personal credit scoring, and their method has better interpretability and higher prediction accuracy than Lasso-Logistic regression. However, when there are categorical variables in the data, the Lasso method usually cannot get satisfactory results, because the Lasso method can only select a single dummy variable, not the entire categorical variable, and the Group-Lasso method solves this problem well. Zhang Juan and Zhang Beibei [19] used the generalized semi-parametric additive model based on the Group-Lasso method for the application research of the credit scoring model, although the model

considers the Group-Lasso method for variable selection, and the dummy variable is used as a group for the overall selection, but the necessary processing of the sample data is lacking.

To solve the above problems, this paper tries to use the ROSE (Random Over Sampling Examples) method to deal with the unbalanced credit card data and then uses the Group-Lasso method to select variables in the Logistic regression model to build a personal credit scoring model.

1 Model and Method

1.1 Group-Lasso Method

In the linear regression model, the continuous response variable $Y \in R^n$, $n \times p$ matrix is $\boldsymbol{X}$, coefficient vector $\beta \in R^{\mathrm{p}}$. The Lasso estimate is defined as follows：

$$\hat{\beta}_\lambda = argmin(\|Y - \boldsymbol{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|) \quad (1)$$

Where $u \in R^n$, $\|u\|_2^2 = \sum_{i=1}^{n} u_i^2$. When the value of the penalty parameter λ is large, some components of $\hat{\beta}_\lambda$ are set to 0 to achieve the purpose of selecting variables. Lasso's L_1 penalty term can also be used for other models, such as the Cox model, the Logistic regression, etc. The method is to replace the residual sum of squares term in the above estimation with the corresponding negative log-likelihood function. In the linear regression model, when the independent variables are not only continuous variables but also include categorical variables, the Lasso method often cannot obtain satisfactory results. Traditional Lasso can only select a single dummy variable instead of the entire categorical variable. Group-Lasso [20] solves this problem based on Lasso. Its estimate is defined as follows:

$$\hat{\beta}_\lambda = argmin(\|Y - \boldsymbol{X}\beta\|_2^2 + \lambda \sum_{g=1}^{G} \left\|\beta_{\mathfrak{T}_g}\right\|_2) \quad (2)$$

Where $\mathfrak{T}_g$is the subscript set of the variables in the gth group, $\beta_{\mathfrak{T}_g}$is the coefficient vector of the variables in the gth group, $(g = 1,2,\cdots,G)$. The penalty term of Group-Lasso can be regarded as an intermediate state between L_1 penalty and L_2penalty [21], and Group-Lasso selects variables at the group level, that is, selects variables in groups. For example, consider a categorical variable with K levels. In the process of modeling a linear model, the categorical variable is transformed into $K-1$ variables of 0-1, which can be regarded as a group. Group-Lasso can select these $K-1$ dummy variables at the same time, while Lasso may only select a part of these $K-1$ dummy variables, which is meaningless.

1.2 Group-Lasso Variable Selection in Logistic Regression Model

Suppose there are independent and identically distributed samples (x_i, y_i), $i = 1,2,\cdots,n$, where $x_i \in R^p$corresponds to the G group of independent variables, the binary response variable $y_i \in \{0,1,\}$, the independent variable can be continuous or categorical. Denote df_g as the degrees of freedom of the g-th independent variable [22], that is, the number of independent variables in the g-th group, so it can be written as $x_i = (x_{i,1}^T, \cdots, x_{i,G}^T)^T$, each group of variables $x_{i,g} \in R^{df_g}$, $g = 1,2,\cdots,G$. A linear logistic regression model models the conditional probability$p_\beta(x_i) = P_\beta(y = 1|x_i)$ as follows:

$$og\frac{p_\beta(x_i)}{1-p_\beta(x_i)} = \eta_\beta(x_i) \quad (3)$$

$$\eta_\beta(x_i) = \beta_0 + \sum_{g=1}^{G} x_{i,g}^T \beta_g \quad (4)$$

$\beta_0 \in R$ is the intercept term, $\beta \in R^{df_g}$ is the coefficient vector of the gth group of independent variables, $\beta = (\beta_0, \beta_1^T, \cdots, \beta_G^T)^T \in R^{p+1}$ is the coefficient vector for all independent variables. Logistic regression Group-Lasso estimates $\hat{\beta}_\lambda$ can be obtained by minimizing the following convex function:

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{g=1}^{G} s(df_g)\|\beta_g\|_2 \quad (5)$$

where $l(\cdot)$ is the log-likelihood function:

$$l(\beta) = \sum_{i=1}^{n} y_i \eta_\beta(x_i) - log[1 + exp\{\eta_\beta(x_i)\}] \quad (6)$$

The parameter λ controls the penalty intensity and the function $s(\cdot)$readjusts the penalty intensity according to the dimension of the coefficient vector. Unless otherwise specified, use $s(df_g) = df_g^{1/2}$[23], the parameter estimation adopts the Block Coordinate Descent algorithm [24].

1.3 ROSE (Random Over Sampling Examples)

In practical binary classification problems, there are many situations where there are very few samples in one class, and it is usually the class we are more interested in. Numerous studies have shown that the model pays more attention to large classes and ignores the influence of small classes, and this class imbalance problem will seriously affect the fitting and prediction effects of machine learning algorithms [25]. To deal with the above problems, the general method is to directly adjust

the sample size of the original data set, so that the data samples of different categories reach a balanced state. There are two main methods to solve the non-equilibrium problem based on this direct generation of new data samples: one is based on the data level, and the other is based on the algorithm level.

The methods based on the data level are mainly random undersampling and random oversampling. (1) The random undersampling method is to directly reduce the sample size of the large-class samples to balance the two types of samples but using this method will also directly cause the large-class to lose a lot of important information, resulting in inaccurate results. (2) The random oversampling method is to directly increase the sample size of small-class samples to balance the two types of samples, but there is also a problem with this method, that is, repeating the addition of small-class sample data will make the information repeat and increase the computational burden is likely to lead to overfitting. There are two main methods based on the algorithm level: one is the cost-sensitive learning method, and the other is the artificial sample data synthesis method. (1) The cost-sensitive learning method does not directly generate a balanced data set but generates a cost matrix to deal with the unbalanced problem by adjusting the cost of misclassification. However, this method has great limitations in unbalanced data, the uneven distribution of response variables will reduce the accuracy of the algorithm, and the prediction accuracy for small classes will be very low. In unbalanced data, neither algorithm can obtain enough information from the class with a small sample size to make accurate predictions [26]. (2) The synthetic method of artificial sample data is to use the generated artificial sample data instead of repeating the original sample data to deal with the unbalanced problem and solve the problem of overlapping generated samples. Compared with the random undersampling method or random oversampling method, there is no duplication. With sample data, there is also no reduction in sample information.

In some cases, the artificial sample data synthesis method has certain advantages over other methods of dealing with non-equilibrium, among which the most effective and commonly used are SMOTE algorithm and ROSE algorithm. The SMOTE algorithm is to generate new data similar to small-class observations, specifically, randomly selecting points on the connection line between the sample point and its neighbors as the generated artificial samples [27]; the ROSE algorithm is based on the automatic Conditional kernel density estimation of variables, producing

class-balanced artificial samples [28]. After using two methods to deal with the data imbalance, the results show that the empirical results after processing by the ROSE algorithm are more effective. Therefore, the following will focus on the ROSE algorithm [29].

Consider the training set $T_n = \{(x_i, y_i), i = 1,2,\cdots,n\}$, the class labels are $y_i \in \{\mathcal{Y}_0, \mathcal{Y}_1\}$, x_i is a realization of some population $\in R^d$ (a sample from the population x), the probability density $f(x)$ of the population is unknown. $n_j < n$ represents the number of categories $\mathcal{Y}_j$. The ROSE algorithm generates a new artificial sample including the following steps:

(1) Select $y = \mathcal{Y}_j \in \mathcal{Y}$ with probability 1/2;

(2) Select samples (x_i, y_i) in the training set T_n with probability $p_i = \frac{1}{n_j}$ so that $y_i = y$;

(3) Sampling from $K_{H_j}(\cdot, x_i)$, where K_{H_j} is a probability distribution centered at x_i, and $H_j$is the scale parameter matrix.

First, select a sample from the training set, and then generate a new sample in the neighborhood of this sample, and the width of this neighborhood is determined by x_i. In general, $K_{H_j}$is a unimodal and symmetric probability distribution. When a class label $\mathcal{Y}_j$ is given, generating a new sample is equivalent to sampling by a kernel density estimate of $f(x|\mathcal{Y}_j)$, where the kernel function is K_{H_j}. The choice of the kernel function K and the "window width" H_j is a pure kernel density estimation problem. When given a label, the conditional density is as follows:

$$\hat{f}(x|y = \mathcal{Y}_j) = \sum_{i=1}^{n_j} p_i \, Pr(x|x_i) = \sum_{i=1}^{n_j} \frac{1}{n_j} Pr(x|x_i) \; = \sum_{i=1}^{n_j} \frac{1}{n_j} K_{H_j}(x - x_i) \quad (7)$$

2 Empirical Analysis

2.1 Data Collection and Preprocessing

The data used in this paper comes from the credit card user data of a commercial bank's credit center. The total number of samples in this dataset is 67,773, and each sample represents the information corresponding to a customer. Each customer information contains 45 attributes, which means that the number of independent variables selected in this paper is 45, including the basic information of customers, personal financial target assets, personal consumption, and personal loans, etc. (See Appendix Table A1). A label attribute classifies customer categories into two categories, in which "good" customers are labeled "0", that is, customers without a default; "bad"

customers are labeled "1", that is, customers who have defaulted. Among the 67,773 total samples, 66,051 customers were defined as "good", accounting for 97.5% of the total sample, and 1,722 customers were defined as "bad", accounting for 97.5% of the total sample. 2.5%.

In actual situations, the category labels of credit card user data collected by commercial banks are often seriously unbalanced, and the data is also seriously missing because only a small number of customers actually default or have complete data information. Since the quality of the sample data set is directly determined by the data analysis results, it is necessary to preprocess the sample data set. When preprocessing missing data, three processing methods are mainly used: deletion method, filling method, and no processing. At present, there is still a lack of a method that is generally applicable to dealing with missing values in practical applications. In this paper, the missing numerical value is filled with the mean of the variable, and the missing categorical variable is filled with "null".

2.2 Evaluation Criteria and Model Establishment

This paper takes discriminative ability and prediction accuracy as the evaluation criteria and conducts an empirical analysis of the personal credit data. ROC curve, AUC, and KS values were used for verification. Considering that the nature of credit scoring is a binary classification problem, we also use two types of errors to explain the reliability and accuracy of the model prediction results.

The ROC curve [30] is a common method for evaluating the effect of binary classifiers and an effective tool to assist in determining probabilistic segmentation values. Generally, the x-axis of the ROC curve is the false positive rate (FPR), and the y-axis is the true rate (TPR). In the confusion matrix of binary prediction, the row direction is the actual category value of the observation, and the column item is the predicted category value. Generally given a two-class model and its threshold, we can calculate a coordinate point from the actual value and predicted value of these sample data. If the coordinate point is closer to the upper left corner, it means that the prediction accuracy is higher. The closer the lower right corner is, the lower the prediction accuracy. As a supplement to the ROC curve, the AUC value represents the area under the ROC curve, so a classifier with a larger AUC value has a higher classification accuracy. The KS value here only represents the ability of the model to segment samples and cannot fully indicate whether the segmentation is accurate. If the KS value is greater than 0.2, it can be considered that the model

has good prediction accuracy. Even if the good and bad customers are completely wrong, the KS value can still be very high.

From the perspective of two types of errors, the first type of error is about predicting a "good" customer to become a "bad" customer, and the second type of error is about predicting a "bad" customer into a "good" customer. Although both types of errors are what we want to avoid, in practice, the cost of making the second type of error is several times that of the first type of error. Therefore, our main purpose is to minimize the type 2 error while reducing the overall error rate to minimize the loss.

2.2.1 Group-Lasso Logistic Regression Model

The Group-Lasso Logistic regression method was used to establish a personal credit scoring model. Since the data contains a large number of categorical variables, Lasso cannot be used directly for variable selection. Therefore, here we use the Group Lasso method for variable selection and perform simultaneous variable selection on dummy variables generated by categorical variables. At this time, the deviation of the sub-model relative to the saturated model is selected as the model selection criterion and the model selection criterion based on the AUC value is selected for comparative analysis.

1) Group-Lasso Logistic Regression Model (Deviation)

First, the deviation of the sub-model relative to the saturated model is used as the criterion for model selection, 5-fold cross-validation is performed, and the model corresponding to the minimum cross-validation deviation (CV Deviance) is selected. where the deviation is defined as follows:

$$D = -2(log\ (p(y|\widehat{\boldsymbol{\beta}}_{\mathbf{0}})) - log\ (p(y|\widehat{\boldsymbol{\beta}}_{\boldsymbol{s}}))) \quad (8)$$

$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$ is the parameter estimated by the sub-model and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{s}}$ is the parameter estimated by the saturated model. The parameter λ is related to the cross-validation bias as Figure 1. The optimal parameter $\lambda = 0.000415$ is obtained, 40 variables are selected, and 5 variables are eliminated. Use the selected 40 variables to train the Logistic regression model, and the sample numbers corresponding to the training set and test set are the same as the previous one (implemented by a fixed random

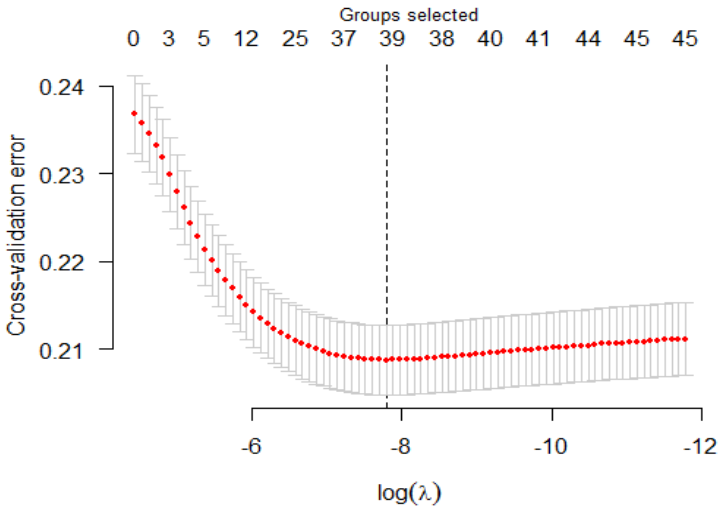

Figure 1 Path of Parameter λ

seed). Obtain the ROC curve on the test set (Figure 2). It can be obtained from the ROC curve: the AUC value is 0.777, and the calculated KS value is 0.4343.

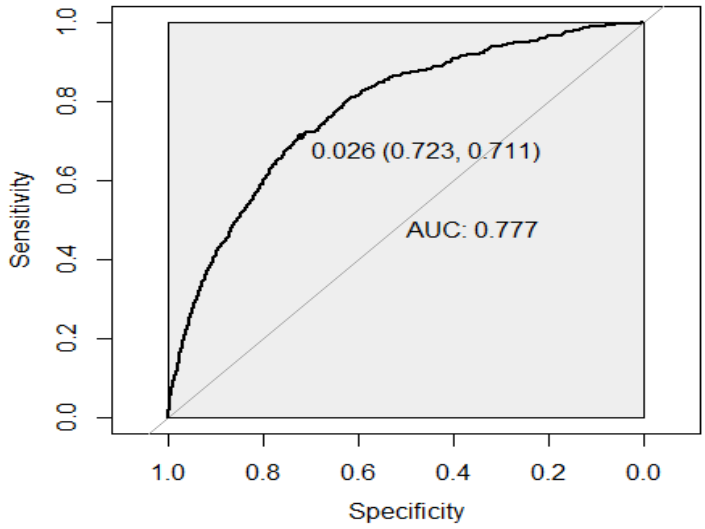

Figure 2 ROC Curve under Group-Lasso Logistic Regression Model (Deviation)

2) Group-Lasso Logistic Regression Model (AUC)

Secondly, the model is selected based on the AUC value. Cross-validation is not used here, but the model corresponding to the maximum AUC value is directly selected according to the AUC values of different sub-models on the test set. At this point, the model selected 39 variables and eliminated 6 variables. Use the selected 39 variables to train the logistic regression model. The sample

numbers corresponding to the training set and the test set are the same as the previous ones (implemented by a fixed random seed), and the ROC curve on the test set is obtained (Figure 3). From the ROC curve, it can be obtained that the maximum AUC value is 0.778, and the KS value corresponding to the model with the maximum AUC is 0.4341.

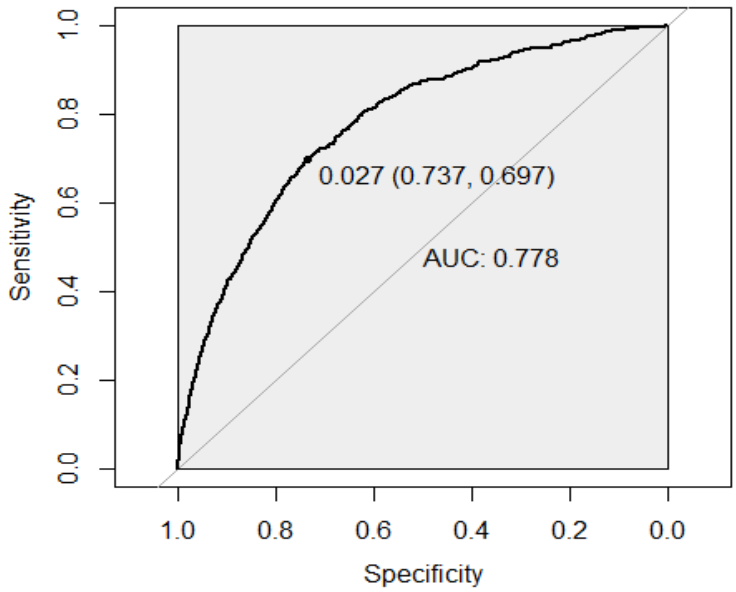

Figure 3 ROC Curve under Group-Lasso Logistic Regression Model (AUC)

The empirical results show that using the bias as the model selection criterion, although the accuracy of "0" is lower than that of the other model, the accuracy of "1" is higher than that of the other model, that is, it reduces the accuracy of the second model. The probability of class error and the KS value are also slightly improved, indicating that the results obtained by using bias as the model selection criterion are more effective.

2.2.2 ROSE Group-Lasso Logistic Regression Model

1) ROSE Group-Lasso Logistic Regression Model (Deviation)

First, 40 variables were modeled using bias as the criterion for model selection. At this time, the training set and the test set remain unchanged. The sample capacity of the training set is 40664, there are 39631 samples with the label "0", and there are 1033 samples with the label "1". By using the ROSE method to generate a balanced data set, the samples of the balanced data set labeled "0" and "1" are 20399 and 20265, respectively, and the model is trained with the balanced data to

obtain the ROC curve on the test set (Figure 4). Obtained from the ROC curve: the AUC value is 0.78, and the KS value is 0.4345 after calculation.

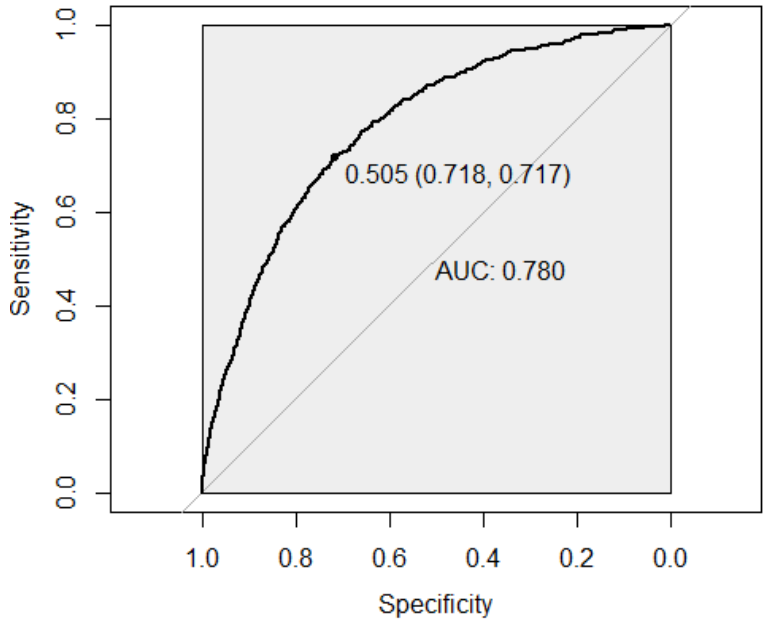

Figure 4 ROC Curve under ROSE Group-Lasso Logistic Regression Model (Deviation)

2) ROSE Group-Lasso Logistic Regression Model (AUC)

Secondly, 39 variables selected with AUC value as the criterion for model selection were used for

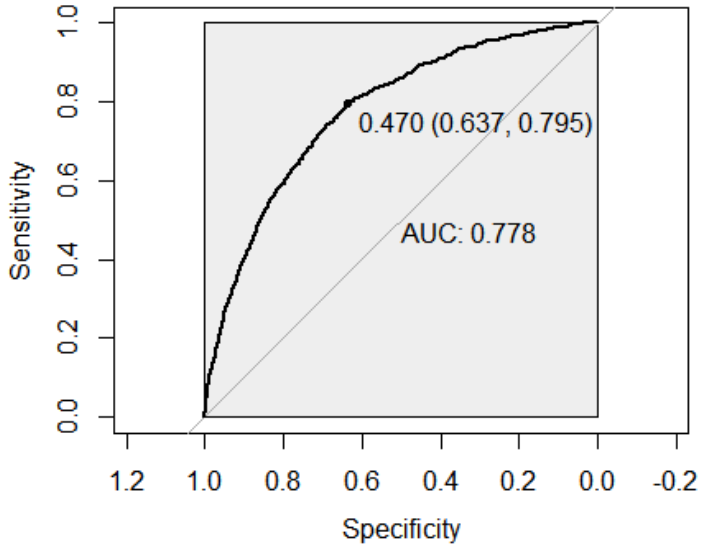

Figure 5 ROC Curve under ROSE Group-Lasso Logistic Regression Model (AUC)

the training set is 40664, there are 39631 samples with the label "0", and there are 1033 samples with the label "1". Similarly, the balanced data set is generated by using the ROSE method. The samples with the labels of "0" and "1" in the balanced data set are 20399 and 20265, respectively. The model is trained with the balanced data, and the ROC curve on the test set is obtained (Figure 5). Obtained from the ROC curve: the AUC value is 0.778, and the KS value is 0.4347 after calculation.

2.2.3 SMOTE Group-Lasso Logistic Regression Model

1) SMOTE Group-Lasso Logistic regression model (Deviation)

To further verify the validity of the personal credit scoring model established in this paper, the SMOTE algorithm is selected to perform unbalanced processing on the sample data. First, 40 variables were modeled with deviation as the criterion for model selection, and other conditions remained unchanged. The ROC curve on the test set was obtained (Figure 6). Obtained from the ROC curve: the AUC value is 0.742, and the KS value is 0.4167 after calculation.

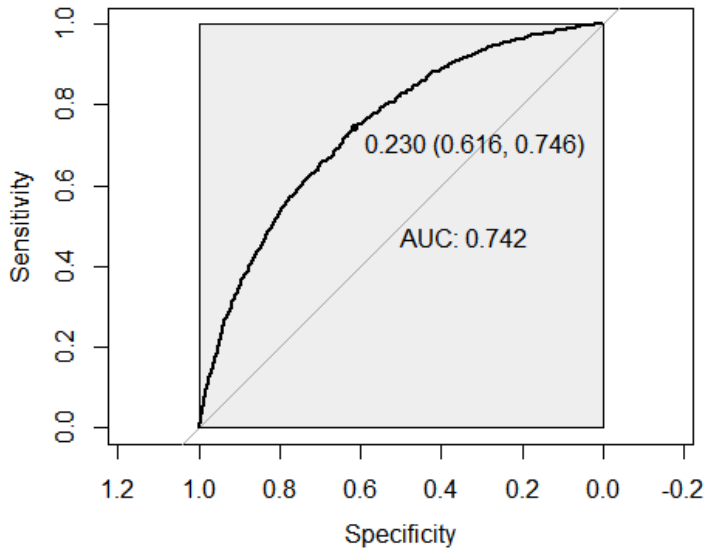

Figure 6 ROC Curve under SMOTE Group-Lasso Logistic Regression Model (Deviation)

2) SMOTE Group-Lasso Logistic Regression Model (AUC)

Secondly, 39 variables were selected with AUC as the model selection criterion, and other conditions remained unchanged. The ROC curve on the test set was obtained (Figure 7). Here is obtained from the ROC curve: the AUC value is 0.748, and the KS value is 0.4209 after calculation.

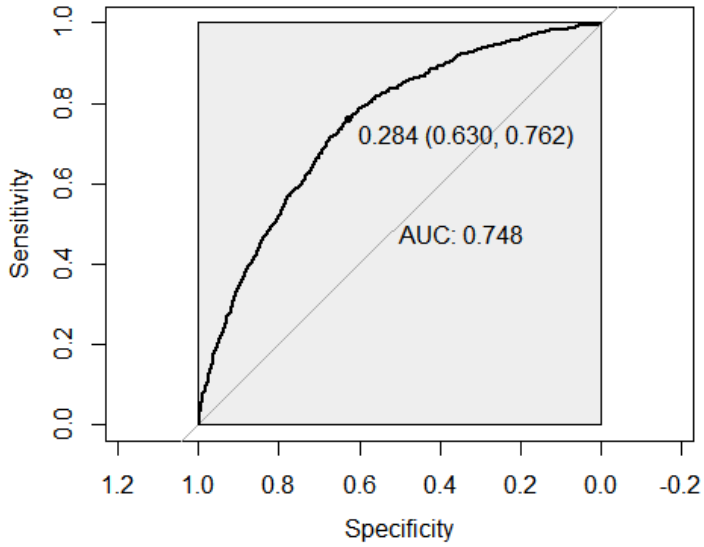

Figure 7 ROC Curve under SMOTE Group-Lasso Logistic Regression Model (AUC)

The empirical results show that the results obtained in this paper are very unsatisfactory after using the SMOTE algorithm to process the unbalanced sample data. Although both belong to the artificial sample data synthesis method, it is obvious that the SMOTE algorithm does not apply here.

2.3 Comparison and Analysis of the Results of each Model

In this paper, the discriminant ability and prediction accuracy are used as the evaluation criteria to analyze the personal credit data and use the ROC curve, AUC, and KS value to verify. At the same time, considering that the nature of credit scoring is a binary classification problem, the reliability of the model result prediction is further explained from the perspective of two types of errors. By comparing the results of the above models (as shown in Table 1), it can be seen that compared with other models, the AUC and KS values of the results obtained by the ROSE Group-Lasso Logistic regression model (AUC criterion) constructed in this paper are higher, indicating that the model The overall prediction accuracy is improved, and the accuracy rate of "1" is greatly improved compared to other models, that is, the probability of making Type II errors is reduced. Therefore, the personal credit scoring model based on the ROSE Group-Lasso Logistic regression model (AUC criterion) is more effective.

Table1 Comparison of Results of each Model

Model	Criterion			
	AUC	KS	0 Accuracy	1 Accuracy
Group-Lasso Logistic （Deviation）	0.777	0.4343	0.723	0.711
Group-Lasso Logistic （AUC）	0.778	0.4341	0.737	0.697
ROSE Group-Lasso Logistic （（Deviation）	0.780	0.4345	0.718	0.717
ROSE Group-Lasso Logistic （AUC）	0.778	0.4347	0.637	0.795
SMOTE Group-Lasso Logistic （Deviation）	0.742	0.4167	0.616	0.746
SMOTE Group-Lasso Logistic （AUC）	0.748	0.4209	0.630	0.762

3 Conclusions

This paper introduces the Group-Lasso Logistic method into the personal credit scoring model. The main innovations are as follows: (1) The sample data comes from a domestic commercial bank credit card center, and a large number of real sample data make the empirical results more practical reference value; (2)) In view of the serious imbalance between defaulting customers and non-defaulting customers in the sample data, the ROSE algorithm is creatively used to perform unbalanced processing on the sample data, which improves the data quality. Finally, the empirical results show that the ROSE Group-Lasso Logistic (AUC criterion) method is more effective than other models in terms of discriminative ability and predictive ability. Therefore, the credit scoring model constructed based on the method in this paper can play a certain guiding role for banks and other financial institutions in assessing customer credit risk and can be used as an effective basis for customer credit evaluation decision-making, and has good practical application. However, the research in this paper is only an empirical discussion of the personal credit scoring model from a statistical method, which has limitations. Future research can turn to build dynamic scoring models

suitable for personal credit scoring. The dynamic credit scoring model can not only further optimize the credit risk management of commercial banks but also estimate the default probability more accurately and timely.

References

[1] Ba Shusong. Research on the New Basel Capital Accord [M]. Beijing: China Financial Press, 2003.

[2] Wang Ying, Nie Guangli, Shi Yong. Research on customer default probability of commercial banks in my country based on credit scoring model [J]. Management Review, 2012, 24: 78-87.

[3] Katarzyna S, Tomasz S, Piotr F. Heteroscedastic Discriminant Analysis Combined with Feature Selection for Credit Scoring[J]. Statistics in Transition New Series, 2016, 17(2): 265-280.

[4] Kordas G. Credit Scoring Using Binary Quantile Regression[C]// Dodge Y. Statistical Data Analysis Based on the L1-Norm and Related Methods. Basel: Birkhäuser Press, 2002: 125-137.

[5] Baesens B, Setiono R, Mues C. Using Neural Network Rule Extraction and Decision Tables for Credit-risk Evaluation[J]. Management Science, 2003, 49(3): 312-329.

[6] Marcano C A, Marin B A. Artificial Metaplasticity Neural Network Applied to Credit Scoring[J]. International Journal of Neural Systems, 2011, 21(4):

[7] Desai V S, Conway D G. Credit-scoring Models in the Credit-union Environment Using Neural Networks and Genetic Algorithms[J]. IMA Journal of Management Mathematics, 1997, 8(4): 323-346.

[8] Lundy M. Cluster Analysis in Credit Scoring: Credit Scoring and Credit Control[M]. New York: Oxford University Press, 1993.

[9] Henley W E, Hand D A. K-nearest-neighbour Classifier for Assessing Consumer Credit Risk[J]. The Statistician, 1996, 45(1): 77-95.

[10] Khanbabaei M, Alborzi M. The Use of Genetic Algorithm: Clustering and Feature Selection Techniques in Construction of Decision Tree Models for Credit Scoring[J]. International Journal of Managing Information Technology, 2013, 5(4): 13.

[11] Bunn D W. An Empirical Bayes Procedure for the Credit Granting Decision[J]. Operations Research Letters, 1981, 1(1): 10-12.

[12] Liu Xihe, Guo Na. Analysis of credit risk factors of housing mortgage loans in my country [J]. Shandong Social Sciences, 2012, 3: 105-108.

[13] Lee T H, Zhang M. Bias Correction and Statistical Test for Developing Credit Scoring Model through Logistic Regression Approach[J]. International Journal of Information Technology & Decision Making, 2003, 2(2): 299-311.
[14] De Jongh P J, De Jongh E, Santana L, et al. The Impact of Pre-selected Variance Inflation Factor Thresholds on the Stability and Predictive Power of Logistic Regression Models in Credit Scoring[J]. African Journals Online, 2015, 31(1): 17-37.
15] Wei Qiuping, Zhang Jingxiao. Credit scoring model based on partial least squares method [J]. Statistics and Decision, 2012, 10: 4-6.
[16] Shi Xiaokang, He Xiaoqun. The biased logistic regression model and its application in personal credit rating [J]. Quantitative Statistics and Management, 2015, 34: 1048-1056.
[17] Tibshirani R. Regression Shrinkage and Selection Via the Lasso[J]. Royal Statist Soc B, 1996, 58(1): 267-288.
[18] Zhang Tingting, Jing Yingchuan. Adaptive Lasso-Logistic Regression Analysis of Personal Credit Scores" [J]. Practice and Understanding of Mathematics, 2016, 46: 92-99.
[19] Zhang Juan, Zhang Beibei. Research on the application of generalized semi-parameter additive credit scoring model based on Group-LASSO method [J]. Mathematical Statistics and Management, 2016, 35: 517-524.
[20] Lukas M, Geer S V, Peter B. The Group Lasso for Logistic Regression[J]. Journal of the Royal Statistical Society, 2008, 70(1): 53-71.
[21] Zou H, Hastie T, Tibshirani R. On the Degrees of Freedom of the Lasso[J]. The Annals of Statistics, 2007, 35(5): 2173-2192.
[22] Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables[J]. Journal of the Royal Statistical Society, 2006, 68(1): 49-67.
[23] P. Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization[J]. Journal of Optimization Theory and Applications, 2001, 109(3): 475-494.
[24] Saraswat M. Practical Guide to deal with Imbalanced Classification Problems in R. USA: Analytics Vidhya Content Team, (2016-03-28)
[25] Chawla NV, Japkowicz N, Kotcz A. Special Issue on Learning from Imbalanced Data Sets[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 1-6.
[26] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[27] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning [J]. Journal of the Royal Statistical Society, 2009, 167(1): 192-192.
[28] Menardi G, Torelli N. Training and Assessing Classification Rules with Imbalanced Data[J]. Data Mining & Knowledge Discovery, 2014, 28(1): 92-122.
[29] Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning[J]. R Journal, 2014, 6(1): 79-89.
[30] Mamitsuka H. Selecting Features in Microarray Classification Using ROC Curves[J]. Pattern Recognition, 2006, 39(12): 2393-2404.

文章编号:2095-6134(2021)02-0181-08

基于 Group-Lasso 方法的非均衡数据信用评分模型*

韦勇凤,向一波†

(中国科学技术大学管理学院, 合肥 230026)

(2019 年 5 月 17 日收稿; 2019 年 7 月 8 日收修改稿)

Wei Y F, Xiang Y B. Imbalanced data credit scoring model based on Group-Lasso method[J]. Journal of University of Chinese Academy of Sciences, 2021,38(2):181-188.

摘 要 目前商业银行面临的个人信用风险问题极其复杂,如何对个人信用风险进行管理非常重要。个人信用风险建模是其中很关键的一步。利用某商业银行信用卡数据,构建信用评分模型,预测客户的违约概率。通过采用 ROSE (random over sampling examples)方法处理类别不均衡的问题,利用 Group-Lasso(AUC 准则)方法进行变量选择,构建基于 Logistic 回归的信用评分模型。实证结果表明,该方法对样本数据进行类别不均衡处理的结果比其他模型在判别能力和预测能力上更为有效。采用该方法所构建的模型能够作为客户信用评价决策的有效依据,指导银行及其他金融机构评估顾客个人信用风险,在实际运用中具有良好的可操作性。

关键词 信用评分;Logistic 回归;Group-Lasso 方法;ROSE

中图分类号:F832.479 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2021.02.004

Imbalanced data credit scoring model based on Group-Lasso method

WEI Yongfeng, XIANG Yibo

(*School of Management*, *University of Science and Technology of China*, *Heifei* 230026, *China*)

Abstract In view of the complexity of the customers' credit risk faced by commercial banks at the present, how to manage customers' credit risk is very important. Customers' credit risk modeling is a key step. We use the credit card data of a commercial bank to construct a credit scoring model and predict the default probability. We construct a credit scoring model on the basis of Logistic regression, using the group-Lasso (AUC criterion) method to select variables and using the ROSE (random over sampling examples) method to deal with the unbalanced categories. The results are compared and analyzed, and the new model constructed in this work has certain advantages in discriminating ability and predictive ability. It can play a guiding role for banks and other financial institutions in evaluating customer credit risk and can be used as an effective basis for customer credit evaluation decision. In practice, it also has good operability.

Keywords credit scoring; Logistic regression; Group-Lasso method; ROSE

* 安徽省自然科学基金(1808085MG222)资助

† 通信作者,E-mail:xybll@mail.ustc.edu.cn

随着金融行业快速发展、大众消费观念极速改变,个人信贷产品不断丰富,以及个人信贷业务规模大幅提升,个人信用风险问题日益突出。个人信用风险是目前商业银行面临的风险中最为重要和复杂的,因此如何进行有效的个人信用风险管理来降低违约风险成为研究的核心问题。《巴塞尔新资本协议》指出有条件的银行要实施内部评级法,通过对历史数据构建模型测算客户的违约概率[1]。违约概率作为影响信用风险的关键因素之一,准确地评估客户的违约概率是信用风险计量的基础[2]。而在信用风险管理方面,信用评分模型发挥着重要的作用。

目前,基于机器学习的个人信用评分模型有:线性判别分析[3]、Logistic 回归[4]、神经网络[5]、支持向量机[6]、遗传算法[7]、聚类分析[8]、最近邻模型[9]、决策树[10]和贝叶斯方法[11]等。其中,Logistic 回归在个人信用评分中应用最为广泛。Logistic 模型具有计算简单、解释性强与预测精度较高的优点,并且在中国房贷信用风险评估中得到验证[12]。然而,随着社会和经济的不断发展,影响个人信用风险的因素在不断增加,传统的 Logistic 回归不能有效地处理模型中自变量存在多重共线性的问题,且过多变量增加了模型的复杂度,进而降低了模型预测的精准度。因此,不断有学者在此基础上进行深入研究。针对 Logistic 回归存在的缺陷,Lee 和 Zhang[13]通过对个人信用评分模型中样本的非均匀抽样进行优化,提高了 Logistic 回归模型的预测性能。Jongh 等[14]提出通过大样本数据消除 Logistic 回归中的变量多重共线性对个人信用评分中的影响。魏秋萍和张景肖[15]基于偏最小二乘方法构建信用评分模型,史小康和何晓群[16]基于有偏的 Logistic 回归方法进行个人信用评级模型的应用研究,这两种方法都能有效地缓解评分模型中自变量存在多重共线性的问题。然而,上述模型也存在一定的局限性,没有充分考虑样本数据本身存在的非均衡性和缺失性,也没有对研究变量进行必要的选择。Tibshirani[17]提出 Lasso 方法能够同时实现变量选择与参数估计,可以将模型中部分自变量的系数压缩使之趋于零,从而达到变量选择的目的。张婷婷和景英川[18]直接将改进的 adaptive Lasso-Logistic 回归模型引入个人信用评分,他们的方法相比 Lasso-Logistic 回归具有更好的解释性与更高的预测精准度。然而,当数据中有分类变量时,Lasso 方法通常不能得到满意的结果,因为 Lasso 方法只能选择单个哑变量,而不是整个分类变量,Group-Lasso 方法很好地解决了这个问题。张娟和张贝贝[19]采用基于 Group-Lasso 方法的广义半参数可加模型进行信用评分模型的应用研究,虽然该模型考虑利用 Group-Lasso 方法进行变量选择,将哑变量作为组进行整体的选择,但是缺乏对样本数据进行必要的处理。

为解决上述问题,本文尝试采用 ROSE(random over sampling examples)方法处理类别不均衡的信用卡数据之后,再使用 Group-Lasso 方法在 Logistic 回归模型中进行变量选择,构建个人信用评分模型。

1 模型与方法

1.1 Group-Lasso 方法介绍

在线性回归模型中,记连续型响应变量为 $\boldsymbol{Y} \in \mathbb{R}^n$,$n \times p$ 维设计矩阵为 $\boldsymbol{X}$,系数向量为 $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$。Lasso 估计[17]定义如下

$$\hat{\boldsymbol{\beta}}_\lambda = \operatorname{argmin}\left(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{1}$$

其中 $\boldsymbol{u} \in \mathbb{R}^n$,$\|\boldsymbol{u}\|_2^2 = \sum_{i=1}^{n} u_i^2$。当惩罚参数 λ 取值较大时,$\hat{\boldsymbol{\beta}}_\lambda$ 的一些分量取 0,从而达到选择变量的目的。Lasso 方法的 L_1 惩罚项也可以用于其他模型,比如 Cox 模型、Logistic 回归等,做法是将上述估计中的残差平方和项换为相应的负对数似然函数。

在线性回归模型中,当自变量除连续型变量,还含有分类变量时,Lasso 方法通常不能得到满意的结果。传统 Lasso 方法只能选择单个的哑变量,而不是整个分类变量,Group-Lasso 方法[20]在 Lasso 方法的基础上解决了这个问题。其估计定义如下

$$\hat{\boldsymbol{\beta}}_\lambda = \operatorname{argmin}\left(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_{\mathfrak{T}_g}\|_2 \right), \tag{2}$$

其中 $\mathfrak{T}_g$ 是第 g 组变量的下标集,$\boldsymbol{\beta}_{\mathfrak{T}_g}$ 是第 g 组变量的系数向量,$(g = 1, 2, \cdots, G)$。Group-Lasso 方法的惩罚项可以看作是 L_1 惩罚和 L_2 惩罚的中间状态[21],Group-Lasso 方法在组的水平上选择变量,即成组地选择变量。例如,考虑一个具有 K 个水平的分类变量,在线性模型建模过程中,该分类变量是被转化为 $K-1$ 个 0-1 变量,就可以看作

一个组。Group-Lasso 方法可以对这 $K-1$ 个哑变量同时选择,而 Lasso 方法只能选出这 $K-1$ 个哑变量中的一部分,这是没有实际意义的。

1.2 Logistic 回归模型中的 Group-Lasso 变量选择

假设有独立同分布的样本 $(\boldsymbol{x}_i, y_i), i = 1,2,\cdots,n$,其中 $\boldsymbol{x}_i \in \mathbb{R}^p$ 对应于 G 组自变量,二元响应变量 $y_i \in \{0,1\}$,自变量可以为连续型变量或分类变量。记 df_g 为第 g 个自变量的自由度(degrees of freedom)[22],即第 g 组自变量的个数,于是可记 $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}^{\mathrm{T}}, \cdots, \boldsymbol{x}_{i,G}^{\mathrm{T}})^{\mathrm{T}}$,每一组变量 $\boldsymbol{x}_{i,g} \in \mathbb{R}^{\mathrm{df}_g}, g = 1,2,\cdots,G$。

线性 logistic 回归模型对条件概率 $p_\beta(\boldsymbol{x}_i) = P_\beta(y = 1 \mid \boldsymbol{x}_i)$ 建模:

$$\log \frac{p_\beta(\boldsymbol{x}_i)}{1 - p_\beta(\boldsymbol{x}_i)} = \eta_\beta(\boldsymbol{x}_i), \tag{3}$$

其中,

$$\eta_\beta(\boldsymbol{x}_i) = \boldsymbol{\beta}_0 + \sum_{g=1}^{G} \boldsymbol{x}_{i,g}^{\mathrm{T}} \boldsymbol{\beta}_g, \tag{4}$$

$\boldsymbol{\beta}_0 \in \mathbb{R}$ 是截距项,$\boldsymbol{\beta} \in \mathbb{R}^{\mathrm{df}_g}$ 是第 g 组自变量的系数向量,$\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1^{\mathrm{T}}, \cdots, \boldsymbol{\beta}_G^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+1}$ 是所有自变量的系数向量。Logistic 回归的 Group-Lasso 估计 $\hat{\boldsymbol{\beta}}_\lambda$ 可通过最小化如下凸函数得到:

$$S_\lambda(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} s(\mathrm{df}_g) \|\boldsymbol{\beta}_g\|_2, \tag{5}$$

其中 $l(\cdot)$ 是对数似然函数:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \eta_\beta(\boldsymbol{x}_i) - \log[1 + \exp\{\eta_\beta(\boldsymbol{x}_i)\}]. \tag{6}$$

参数 λ 控制惩罚力度,函数 $s(\cdot)$ 根据系数向量的维数重新调整惩罚力度,如无特别说明,采用 $s(\mathrm{df}_g) = \mathrm{df}_g^{1/2}$[23],参数估计采用 Block Coordinate Descent 算法[23]。

1.3 ROSE(random over sampling examples)

在实际的二分类问题中,有许多情形是其中有一类样本非常少,而且通常是我们比较感兴趣的那一类。然而大部分模型更关注大类而忽视了小类的影响,这类类别不均衡问题会严重影响机器学习算法的拟合和预测效果[24]。为处理上述问题,一般采取的方法是直接调整原始数据集的样本量,使得不同类别的数据样本之间达到均衡状态。基于这种直接产生新的数据样本来解决非均衡问题的主要方法有两种:一是基于数据层面,二是基于算法层面。

基于数据层面的方法主要是随机欠抽样法和随机过抽样法。1)随机欠抽样法是直接减少大类样本的样本量使得两类样本均衡,但采用这种方法会使得大类损失不少重要信息,导致结果不准确。2)随机过抽样法是增加小类样本的样本量使得两类样本均衡,但采用这种方法会重复增加小类的样本数据,增大计算负担,还有可能导致过拟合。基于算法层面的方法主要也有两种:一是代价敏感学习法,二是人造样本数据合成法。1)代价敏感学习法不直接生成均衡数据集,而是通过调节错分代价的方式,生成代价矩阵处理非均衡问题。该方法在非均衡数据的处理中具有较大的局限性,响应变量的不均匀分布使得算法精度下降,对于小类的预测精度会很低。而在非均衡的数据中,任一算法都没法从样本量少的类中获取足够的信息进行精确预测[25]。2)人造样本数据合成法利用人造样本数据而不是重复原始样本数据处理非均衡问题,解决了生成样本重叠的问题。相对于随机欠抽样法或随机过抽样法而言,该方法没有重复利用样本数据,也没有减少样本信息。

在某些场合下,人造样本数据合成法相对于其他处理非均衡的方法而言具有一定的优势,其中最为有效和常用的是 SMOTE(synthetic minority over-sampling technique)算法和 ROSE 算法。SMOTE 算法是生成与小类观测相似的新数据,具体地说是在样本点和它近邻点的连线上随机投点作为生成的人造样本[26];ROSE 算法则是基于各类别对应的自变量的条件核密度估计,产生类别均衡的人造样本[27]。本文运用 2 种方法对数据进行非均衡处理之后,发现通过 ROSE 算法进行处理后的实证结果更加有效。所以,本文最终采用 ROSE 算法[28]对数据进行非均衡处理。

考虑训练集 $T_n = \{(\boldsymbol{x}_i, y_i), i = 1,2,\cdots,n\}$,这里 $y_i \in \{y_0, y_1\}$ 是类别标签,$\boldsymbol{x}_i$ 是某个总体 $\boldsymbol{x} \in \mathbb{R}^d$ 的一次实现(来自总体 $\boldsymbol{x}$ 的一个样本),总体的概率密度 $f(\boldsymbol{x})$ 未知。$n_j < n$ 表示类别 y_j 的个数,ROSE 算法通过以下步骤产生一个新的人造样本:

1)以概率 1/2 选择 $y = Y_j$, $j = 1,2$,

2)以概率 $p_i = 1/n_j$ 在训练集 T_n 中选择样本 $(\boldsymbol{x}_i, y_i)$ 使得 $y_i = y$,

3)从 $K_{\boldsymbol{H}_j}(\cdot, \boldsymbol{x}_i)$ 中采样,其中 $K_{\boldsymbol{H}_j}$ 是一个概率分布,中心在 $\boldsymbol{x}_i$,$\boldsymbol{H}_j$ 是刻度参数矩阵。

先从训练集中选择一个样本,然后在这个样本的邻域中产生一个新的样本,这个邻域的宽度由 $\boldsymbol{H}_j$ 决定。通常,$K_{\boldsymbol{H}_j}$ 是一个单峰和对称的概率分布。在给定类别标签 Y_j 时,产生新的样本等价于由 $f(\boldsymbol{x} \mid Y_j)$ 的核密度估计来采样,其中核函数为 $K_{\boldsymbol{H}_j}$。核函数 K 和"窗宽" $\boldsymbol{H}_j$ 的选择是纯粹的核密度估计问题。给定标签时,条件密度如下

$$\begin{aligned}\hat{f}(\boldsymbol{x} \mid y = Y_j) &= \sum_{i=1}^{n_j} p_i \Pr(\boldsymbol{x} \mid \boldsymbol{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j} \Pr(\boldsymbol{x} \mid \boldsymbol{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j} K_{\boldsymbol{H}_j}(\boldsymbol{x} - \boldsymbol{x}_i). \quad (7)\end{aligned}$$

2 实证分析

2.1 数据采集与预处理

本文数据来源于某商业银行信用中心的信用卡用户数据,该数据集样本总量为 67 773 个,其中每一个样本代表一个客户对应的信息。每个客户信息含有 45 个属性,表示本文选取的自变量为 45 个,包含客户的基本情况、个人金融标的资产、个人消费和个人贷款等方面(见附录表 A1)。用一个标签属性对客户类别进行二分类,其中表示"好"的客户标签为"0",即没有违约的客户;"坏"的客户标签为"1",即有违约的客户。在这 67 773 个样本总量中,被定义为"好"的客户样本量有 66 051 个,占总样本量的 97.5%,被定义为"坏"的客户样本量有 1 722 个,占总样本量的 2.5%。

商业银行收集到的信用卡用户数据类别标签通常是严重非均衡的,同时数据也是严重缺失的,因为真正违约或数据信息完整的客户只有很少的一部分。由于样本数据集的质量直接决定了数据分析结果,因此,对样本数据集进行预处理是十分必要的。针对缺失数据进行预处理时,主要采用 3 种处理方法:删除法、填充法和不处理。本文采用常见的填充法来处理,对于数值型的缺失用该变量的均值填补,对于属性变量的缺失用"空"填补。

2.2 评价标准与模型建立

本文以判别能力和预测精度作为评价标准,对个人信用数据进行实证分析。采用 ROC 曲线(receiver operating characteristic curve)、AUC(area under the curve of ROC)值和 KS(Kolmogorov-Smirnov)值进行验证。考虑到信用评分的本质是一个二分类问题,本文采用两类错误对模型预测结果的可靠性和精度进行解释。

ROC 曲线[29]是评估二元分类器效果的常用方法,也是辅助确定概率分割值的有效工具。一般 ROC 曲线的 x 轴为假正率(FPR),y 轴为真正率(TPR),二分类预测的混淆矩阵中,行项为观测的实际类别值,列项为预测类别值。一般给定一个二分类模型和它的阈值,就能从这些样本数据的真实值和预测值计算出一个坐标点。坐标点离左上角越近,表示其预测准确率越高;离右下角越近,表示其预测准确率越低。作为 ROC 曲线的补充,AUC 值表示 ROC 曲线下方的面积。AUC 值越大的分类器,其分类正确率越高;这里的 KS 值仅代表模型分割样本的能力,并不能完全表示分割是否准确。在极端情况下,即便好坏客户完全分错,KS 值也可以很高。但通常情况下,KS 值大于 0.2 就可认为模型有比较好的预测准确性。

从两类错误角度考虑,第 1 类错误即将"好"的客户预测成为"坏"的客户,第 2 类错误即将"坏"的客户预测成为"好"的客户。虽然这两类错误都是我们所要避免的,但是在实际情况中,犯第 2 类错误所付出的代价是第 1 类错误的好几倍。因此,我们的主要目的是将整体错误率降低的同时,将第 2 类错误降到最低,以便将损失减少到最小。

2.2.1 Group-Lasso Logistic 回归模型

采用 Group-Lasso Logistic 回归方法建立个人信用评分模型。由于该数据中含有大量的分类变量,所以不能直接用 Lasso 方法进行变量选择。因此,在这里采用 Group-Lasso 方法进行变量选择,对由分类变量产生的哑变量做同时变量选择。分别以子模型相对于饱和模型的偏差和 AUC 值作为模型选择的准则进行比较分析。

1)Group-Lasso Logistic 回归模型(偏差准则)

以子模型相对于饱和模型的偏差作为模型选择的准则,做 5 折交叉验证,选择交叉验证偏差(CV deviance)最小时对应的模型。偏差定义如下

$$D = -2(\log(p(y \mid \hat{\boldsymbol{\beta}}_0)) - \log(p(y \mid \hat{\boldsymbol{\beta}}_s))), \quad (8)$$

其中:$\hat{\boldsymbol{\beta}}_0$ 是由子模型估计出的参数,$\hat{\boldsymbol{\beta}}_s$ 是由饱和模型估计出的参数。参数 λ 与交叉验证偏差的关系见图 1。

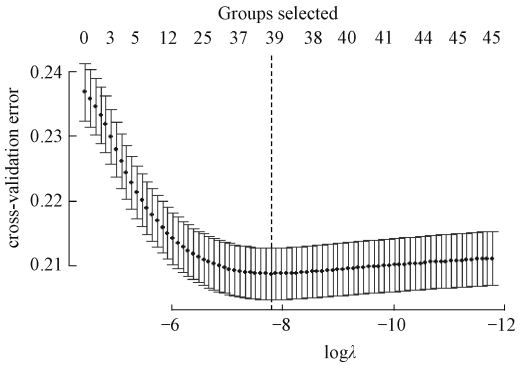

图 1　参数 λ 的路径

Fig. 1　Path of parameter λ

得到最优参数 $\lambda = 0.000\,415$, 选出 40 个变量,剔除 5 个变量。用选出的 40 个变量训练 Logistic 回归模型,训练集和测试集对应的样本编号都与前文相同(通过固定随机种子实现),得到在测试集上的 ROC 曲线(图 2)。由 ROC 曲线可得 AUC 值为 0.777,经计算可得 KS 值为 0.434 3。

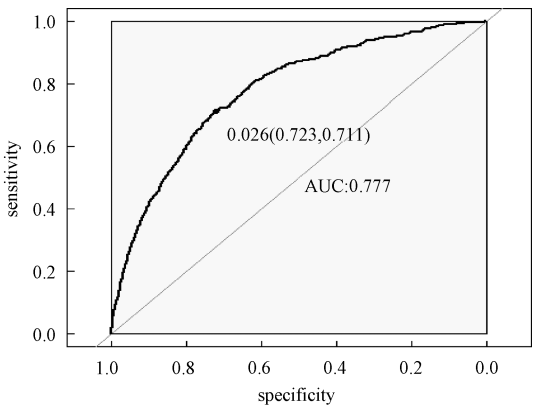

图 2　Group-Lasso Logistic 回归模型(偏差准则)下的 ROC 曲线

Fig. 2　ROC curve under Group-Lasso Logistic regression model (deviation criterion)

2) Group-Lasso Logistic 回归模型(AUC 准则)

以 AUC 值作为模型选择的准则。在这里没有用到交叉验证,而是直接根据不同子模型在测试集上的 AUC 值选择 AUC 值最大时对应的模型。此时,模型选出 39 个变量,剔除 6 个变量。用选出的 39 个变量训练 Logistic 回归模型,训练集和测试集对应的样本编号都与前文相同(通过固定随机种子实现),得到在测试集上的 ROC 曲线(图 3)。由 ROC 曲线可得:AUC 值最大为 0.778,经计算可得 AUC 值最大的模型对应的 KS 值为 0.434 1。

实证结果表明,以偏差作为模型选择准则得到的结果,虽然"0"的准确率相对另一个模型有所下降,但是"1"的准确率比另一个模型有所提升,即降低了犯第 2 类错误的概率,同时 KS 值也略有提升,表明以偏差作为模型选择准则得到的结果更为有效。

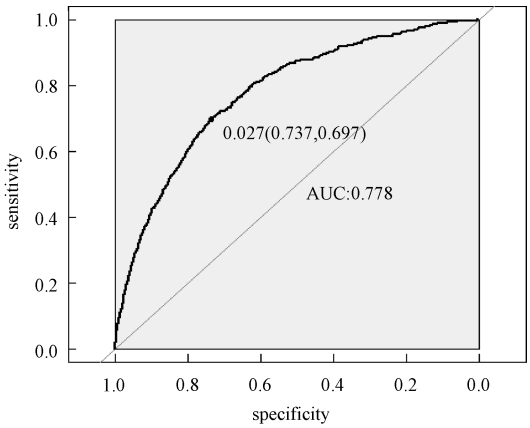

图 3　Group-Lasso Logistic 回归模型(AUC 准则)下的 ROC 曲线

Fig. 3　ROC curve under Group-Lasso Logistic regression model (AUC criterion)

2.2.2　ROSE Group-Lasso Logistic 回归模型

1) ROSE Group-Lasso Logistic 回归模型(偏差准则)

采用以偏差作为模型选择的准则选出的 40 个变量来建模。此时,训练集和测试集不变,其中训练集样本容量为 40 664,标签为"0"的样本有 39 631 个,标签为"1"的样本有 1 033 个。通过采用 ROSE 方法产生均衡数据集,均衡数据集标签为"0"和"1"的样本分别为 20 399 个和 20 265 个,用均衡数据训练模型,得到在测试集上的 ROC 曲线(图 4)。由 ROC 曲线得到 AUC 值为 0.78,经过计算得到 KS 值为 0.434 5。

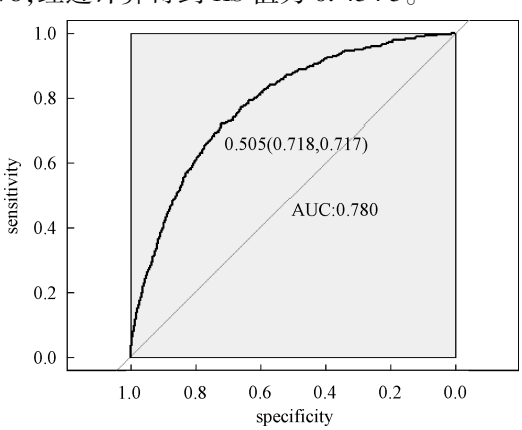

图 4　ROSE Group-Lasso Logistic 回归模型(偏差准则)下的 ROC 曲线

Fig. 4　ROC curve under ROSE Group-Lasso Logistic regression model (deviation criterion)

2) ROSE Group-Lasso Logistic 回归模型(AUC 准则)

采用以 AUC 值作为模型选择的准则选出的 39 个变量来建模。此时,训练集和测试集不变,其中训练集样本容量为 40 664,标签为“0”的样本有 39 631 个,标签为“1”的样本有 1 033 个。同样通过采用 ROSE 方法产生均衡数据集,均衡数据集标签为“0”和“1”的样本分别为 20 399 个和 20 265 个,用均衡数据训练模型,得到在测试集上的 ROC 曲线(图 5)。由 ROC 曲线得到 AUC 值为 0.778,经过计算得到 KS 值为 0.434 7。

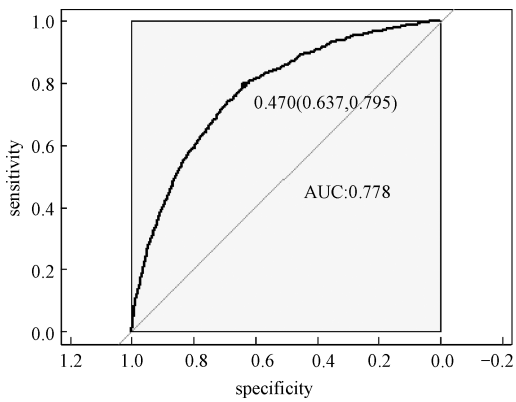

图 5 ROSE Group-Lasso Logistic 回归模型(AUC 准则)下的 ROC 曲线

Fig. 5 ROC curve under ROSE Group-Lasso Logistic regression model (AUC criterion)

实证结果表明,采用 ROSE 算法对样本数据进行非均衡处理之后,以 AUC 值作为模型选择准则得到的结果,虽然“0”的准确率相对于偏差作为模型选择准则有所下降,但是“1”的准确率比另一个模型大幅提升,即降低了犯第 2 类错误的概率,同时 KS 值也略有提升,因此,本文认为以 AUC 值作为模型选择准则得到的结果更为有效。

2.2.3 SMOTE Group-Lasso Logistic 回归模型

1) SMOTE Group-Lasso Logistic 回归模型(偏差准则)

采用以偏差作为模型选择的准则选出的 40 个变量来建模,其他条件不变。得到在测试集上的 ROC 曲线(图 6)。由 ROC 曲线得到 AUC 值为 0.742,经过计算得到 KS 值为 0.416 7。

2) SMOTE Group-Lasso Logistic 回归模型(AUC 准则)

采用以 AUC 值作为模型选择的准则选出的 39 个变量来建模,其他条件不变。得到在测试集

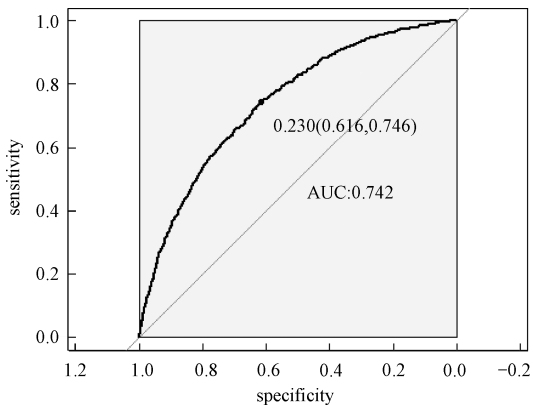

图 6 SMOTE Group-Lasso Logistic 回归模型(偏差准则)下的 ROC 曲线

Fig. 6 ROC curve under SMOTE Group-Lasso Logistic regression model (deviation criterion)

上的 ROC 曲线(图 7)。这里由 ROC 曲线得到 AUC 值为 0.748,经过计算得到 KS 值为 0.420 9。

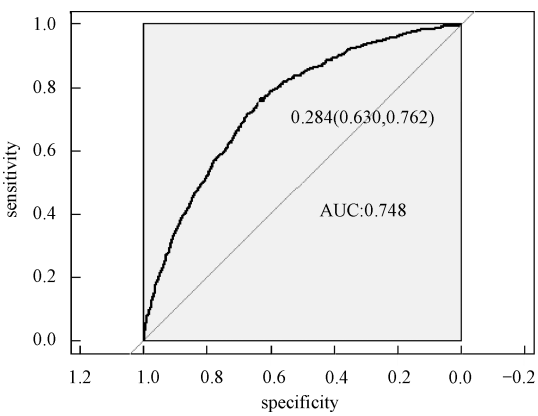

图 7 SMOTE Group-Lasso Logistic 回归模型(AUC 准则)下的 ROC 曲线

Fig. 7 ROC curve under SMOTE Group-Lasso Logistic regression model (AUC criterion)

实证结果表明,采用 SMOTE 算法进行样本数据非均衡处理,所得结果非常不理想。虽然两者都属于人造样本数据合成法,但显然 SMOTE 算法在这里并不适用。

2.3 各模型结果比较分析

本文以判别能力和预测精度作为评价标准,对个人信用数据进行分析,采用 ROC 曲线、AUC 值和 KS 值进行验证。同时考虑到信用评分的本质是一个二分类问题,所以从两类错误角度进一步对模型结果预测的可靠性进行解释。通过比较上述几种模型结果(表 1)可以看出,相对于其他模型,本文构建的 ROSE Group-Lasso Logistic 回归模型(AUC 准则)所得结果的 AUC 值和 KS 值都较高,说明该模型整体的预测精度有所提高,并且

"1"的准确率大幅提升,即降低了犯第 2 类错误的概率。因此,基于 ROSE Group-Lasso Logistic 回归模型(AUC 准则)构建的个人信用评分模型更为有效。

表 1　各模型的结果比较

Table 1　Comparison of results among different models

模型	指标值			
	AUC	KS	0 准确率	1 准确率
Group-Lasso Logistic(偏差准则)	0.777	0.434 3	0.723	0.711
Group-Lasso Logistic(AUC 准则)	0.778	0.434 1	0.737	0.697
ROSE Group-Lasso Logistic(偏差准则)	0.780	0.434 5	0.718	0.717
ROSE Group-Lasso Logistic(AUC 准则)	0.778	0.434 7	0.637	0.795
SMOTE Group-Lasso Logistic(偏差准则)	0.742	0.416 7	0.616	0.746
SMOTE Group-Lasso Logistic(AUC 准则)	0.748	0.420 9	0.630	0.762

3　结论与展望

本文将 Group-Lasso Logistic 方法引入个人信用评分模型,主要创新点如下:1)样本数据来源于国内某商业银行信用卡中心,大量真实的样本数据使得实证结果更加具有实际参考价值;2)针对样本数据中违约客户与未违约客户的严重不均衡状况,创造性地采用 ROSE 算法对样本数据进行非均衡处理,提升了数据质量。最后,实证结果表明,ROSE Group-Lasso Logistic(AUC 准则)方法在判别能力和预测能力上相对其他模型更为有效。因此,本文构建的信用评分模型,能够作为客户信用评价决策的有效依据,指导银行及其他金融机构评估顾客个人信用风险,并且在实际运用中也具有良好的可操作性。

本文的研究只是从统计方法上对个人信用评分模型进行实证探讨,具有局限性。未来的研究可以转向构建适合个人信用评分的动态评分模型。动态信用评分模型不仅可以进一步优化商业银行的信用风险管理,还能更加及时和精确地估计违约损失率。

参考文献

[1] 巴曙松. 巴塞尔新资本协议研究[M]. 北京:中国金融出版社,2003.

[2] 王颖,聂广礼,石勇. 基于信用评分模型的我国商业银行客户违约概率研究[J]. 管理评论,2012,24(2):78-87.

[3] Stapor K, Smolarczyk T, Fabian P. Heteroscedastic discriminant analysis combined with feature selection for credit scoring[J]. Statistics in Transition New Series, 2016, 17(2): 265-280.

[4] Kordas G. Credit scoring using binary quantile regression[C]// Dodge Y. Statistical Data Analysis Based on the L1-Norm and Related Methods. Basel: Birkhäuser Press, 2002: 125-137.

[5] Baesens B, Setiono R, Mues C, et al. Using neural network rule extraction and decision tables for credit-risk evaluation [J]. Management Science, 2003, 49(3): 312-329.

[6] Marcano-Cedeño A, Marin-De-La-Barcena A, Jimenez-Trillo J, et al. Artificial metaplasticity neural network applied to credit scoring[J]. International Journal of Neural Systems, 2011, 21(4): 311-317.

[7] Desai V S, Conway D G, Crook J N, et al. Credit-scoring models in the credit-union environment using neural networks and genetic algorithms[J]. IMA Journal of Management Mathematics, 1997, 8(4): 323-346.

[8] Lundy M. Cluster analysis in credit scoring: credit scoring and credit control[M]. New York: Oxford University Press, 1993.

[9] Henley W E, Hand D J. A k-nearest-neighbour classifier for assessing consumer credit risk[J]. Journal of the Royal Statistical Society, 1996, 45(1): 77-95.

[10] Khanbabaei M, Alborzi M. The use of genetic algorithm: clustering and feature selection techniques in construction of decision tree models for credit scoring[J]. International Journal of Managing Information Technology, 2013, 5(4): 13-32.

[11] Bunn D W. An empirical Bayes procedure for the credit granting decision[J]. Operations Research Letters, 1981, 1(1): 10-12.

[12] 刘喜和,郭娜. 我国住房抵押贷款信用风险因素分析[J]. 山东社会科学,2012(3):105-108.

[13] Lee T H, Zhang M. Bias correction and statistical test for developing credit scoring model through logistic regression approach[J]. International Journal of Information Technology & Decision Making, 2003, 2(2): 299-311.

[14] De Jongh P J, De Jongh E, Pienaar M, et al. The impact of pre-selected variance inflation factor thresholds on the stability and predictive power of logistic regression models in credit scoring[J]. ORiON, 2015, 31(1): 17-37.

[15] 魏秋萍,张景肖. 基于偏最小二乘方法的信用评分模型[J]. 统计与决策,2012(10):4-6.

[16] 史小康,何晓群. 有偏 Logistic 回归模型及其在个人信用评级中的应用研究[J]. 数理统计与管理,2015,34(6):1048-1056.

[17] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective[J]. Journal of the Royal Statistical Society B, 2011, 73(3): 273-282.

[18] 张婷婷,景英川. 个人信用评分的 Adaptive Lasso-Logistic 回归分析[J]. 数学的实践与认识,2016,46(18):92-99.

[19] 张娟,张贝贝. 基于 Group-Lasso 方法的广义半参数可加信用评分模型应用研究[J]. 数理统计与管理,2016,35(3):517-524.

[20] Meier L, Van de Geer S, Bühlmanm P. The group lasso for logistic regression[J]. Journal of the Royal Statistical Society B, 2008, 70(1): 53-71.

[21] Zou H, Hastie T, Tibshirani R. On the "degrees of freedom" of the lasso[J]. The Annals of Statistics, 2007, 35(5): 2173-2192.

[22] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society B, 2006, 68(1): 49-67.

[23] Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization[J]. Journal of Optimization Theory and Applications, 2001, 109(3): 475-494.

[24] Chawla N V, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1-6.

[25] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[26] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning[M]. 2nd ed. New York: Springer, 2009.

[27] Menardi G, Torelli N. Training and assessing classification rules with imbalanced data[J]. Data Mining and Knowledge Discovery, 2014, 28(1): 92-122.

[28] Lunardon N, Menardi G, Torelli N. ROSE: a package for binary imbalanced learning[J]. The R Journal, 2014, 6(1): 79-89.

[29] Mamitsuka H. Selecting features in microarray classification using ROC curves[J]. Pattern Recognition, 2006, 39(12): 2393-2404.

附录 A

表 A1 个人信用评分模型指标体系主要指标变量解释说明

变量模块	变量信息			
	变量名称	变量名	变量类型	变量含义
标签类别	违约情况	target	二分类	0:未违约;1:违约
个人信息特征	户籍类型	ADDR_TYPE	二分类	H:非本地;B:本地
	居住城市	org	分类变量	如:南京、合肥等
	年龄	PAI_AGE	连续变量	分值越大,年龄越大
	单位性质	emp_sta	分类变量	如:国有企业、私营企业
	学历	education	分类变量	如:本科、硕士等
	性别	PAI_SEX	二分类	M:男性;F:女性
	婚姻状况	PAI_MARITAL_STATUS	二分类	M:已婚;S:未婚
	是否有小孩	child_flag	二分类	有/无
	公司行业	emp_type	分类变量	如:教育、制造业
	工作属性	work_state	分类变量	全职、其他
	大学录取批次	PY_Edu_sch_enroll	分类变量	如:本科第一批
	学校是否为 211	PY_Edu_sch_211	二分类	是/否
财务信息特征	住房类型	home_type	分类变量	如:自购、单位分配
	公司职位	PAI_EMP_POSITION	连续变量	其分值越高,职位越高
	工龄	PAI_EMP_WORKYEARS	连续变量	分值越大,工龄越长
	月收入	PAI_MONTH_INCOME	连续变量	分值越大,收入越高
	信用支付	PAI_CREDIT_PAY_OTHER	连续变量	分值越大,消费能力越高
历史信用特征	车辆登记与身份是否一致	PY_Veh_Per_ID_flag	分类变量	一致/不一致/无法核查/空缺
	车辆登记信息	PY_Veh_ID_flag	分类变量	一致/不一致/无法核查/空缺
	车辆是否有违法	PY_Veh_Sta_flag	分类变量	如:违法未处理、正常
	按揭租金	PAI_MORTAGAGE_rent	连续变量	分值越大,租金越多
	手机与身份是否一致	PY_TelPho_Per_ID_flag	分类变量	一致/不一致/无法核查/空缺
	按揭贷款	PAI_MORTAGAGE_loan	连续变量	分值越大,贷款越多