Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2019, 7, 1131-1148
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Logistic and SVM Credit Score Models Based on Lasso Variable Selection

Qingqing Li

College of Science, University of Shanghai for Science and Technology, Shanghai, China
Email: L1039196980@126.com

How to cite this paper: Li, Q.Q. (2019) Logistic and SVM Credit Score Models Based on Lasso Variable Selection. *Journal of Applied Mathematics and Physics*, **7**, 1131-1148.
https://doi.org/10.4236/jamp.2019.75076

Received: April 22, 2019
Accepted: May 24, 2019
Published: May 27, 2019

Abstract

There are many factors influencing personal credit. We introduce Lasso technique to personal credit evaluation, and establish Lasso-logistic, Lasso-SVM and Group lasso-logistic models respectively. Variable selection and parameter estimation are also conducted simultaneously. Based on the personal credit data set from a certain lending platform, it can be concluded through experiments that compared with the full-variable Logistic model and the stepwise Logistic model, the variable selection ability of Group lasso-logistic model was the strongest, followed by Lasso-logistic and Lasso-SVM respectively. All three models based on Lasso variable selection have better filtering capability than stepwise selection. In the meantime, the Group lasso-logistic model can eliminate or retain relevant virtual variables as a group to facilitate model interpretation. In terms of prediction accuracy, Lasso-SVM had the highest prediction accuracy for default users in the training set, while in the test set, Group lasso-logistic had the best classification accuracy for default users. Whether in the training set or in the test set, the Lasso-logistic model has the best classification accuracy for non-default users. The model based on Lasso variable selection can also better screen out the key factors influencing personal credit risk.

Keywords

Credit Evaluation, Logistic Algorithm, SVM Algorithm, Lasso Variable Selection

1. Introduction

In the 21st century, with the rapid development of China's economy, the concept of Chinese people's consumption has undergone tremendous changes, and the credit industry has developed rapidly. Among them, the development of credit

card business is increasing day by day, and the credit risk that comes with it is not to be underestimated. Credit scoring model has been the core of credit risk management. In fact, the credit scoring model is a statistical model that analyzes a large number of customers' historical data, extracts key factors affecting credit risk, and then constructs a suitable model to evaluate the credit risk of new applicants or existing customers. Therefore, the construction of the personal credit scoring model can respond to credit risk in a timely and effective manner, which will play an important role in both banks and regulatory authorities.

In this era of information explosion, however, the emergence of big data has also led to some credit information, and the existing scoring models often cannot effectively screen out dangerous customers. At the same time, the increasing of the high customer information can lead to the complexity of the credit scoring, model bias and instability, thus variable selection becomes the key issues and difficulties in personal credit evaluation model. It is of great significance to apply the variable selection method to the development of the credit scoring model. In the credit scoring model, the subset selection such as stepwise regression is a discrete and unstable process, and the variable selection will be changed by small changes in the data set. Selection and parameter estimation also need to be carried out in two steps. Subsequent parameter estimation does not take into account the bias caused by variable selection, and accordingly it underestimates the actual variance. The calculation of subset selection is also quite complicated. In view of these defects, we adopt the Lasso method which can simultaneously perform variable selection and parameter estimation. After quantifying many explanatory variables, it is necessary to establish dummy variables as explanatory variables of the model. When using stepwise regression to select variables, only one dummy variable can be selected, which is the reason why the results are difficult to explain. However, the above problems can be well solved by Group lasso when it performs variable selection on group variables, making the dummy variables belonging to the same group be completely retained or fully eliminated in the model.

In this paper, the Logistic and SVM models of Lasso were mainly used to select and classify the influencing factors of personal credit evaluation. Then, the prediction accuracy of several models for default users is compared.

2. Literature Review

Typical credit evaluation models are: linear discriminant analysis, logistic regression, K-nearest neighbor, classification tree, neural network, genetic algorithm, support vector machine [1]-[7], etc. Among them, Logistic regression is most widely used in personal credit score, and support vector Machine (SVM) is a new artificial intelligence method developed in recent years. In 1980, Wiginton [8] first applied logistic regression to credit score analysis and analyzed the prediction accuracy of the model. Baseens and Gestel first applied the support vector machine method to the letter in 2003. In the scoring field, the support vector

machine method is obviously superior to the linear regression and neural network methods.

On the contrary, in China, the construction of the credit score system just started. Shi and Jin [9] summarize the main models and methods of personal credit score. Xiang [10] proposed to establish personal credit evaluation by using multiple discriminant analysis (MDA), decision tree, logistic regression, Bayes network (Bayes), BP neural network, RBF neural network and SVM. Shen and so on [11] did a follow-up study on support vector machines. Hu [12] believed that the most representative Logistic model are widely concerned by researchers due to its high prediction accuracy, simple calculation and strong variable explanatory ability.

There are two main methods for selecting variables: subset selection method and coefficient compression method. Subset selection method is that in linear model, all variables form a set, and each subset of the set corresponds to a model. According to certain criteria, an optimal subset fitted regression model is selected from all subsets or partial subsets.

The main research on subset selection are AIC (Akaike Information Criterion) [13] proposed by Akaike, BIC (Bayesian Information Criterion) [14] proposed by Scllwaz, CIC (covariance expansion criterion, Tibshirani and Knight) [15] and Mallows' C_P Guidelines [16]. Although these methods have strong practicability, there are many problems. For example: large algorithm complexity, high computational cost, poor interpretability of explanatory variables, etc.

With the continuing research, the variable selection method based on penalty function has been widely concerned by statistical researchers. The basic idea of this method is to add a new penalized term to the least squares or maximum likelihood function and we then minimize or maximize the augmented objective function. Thus, by compressing the regression coefficients of the insignificant variables to zero, the variables are eliminated, and the significant variables are compressed very little or it can be retained in the regression model without compression. Hence it performs the variables selection and parameters estimation simultaneously, greatly improving the speed of calculation. Regarding the penalty function, the earliest penalty function is the ridge regression method proposed by Hoerl and Kennard [17], but it cannot make variable selection. Since then, Frank and Fredman [18] have proposed the bridge regression method. The Lasso method is proposed by Tibshirani [19], which combines the advantages of ridge regression and subset selection. The least angle regression (LARS) proposed by Efron [20] gives everyone a deeper understanding of Lasso. Zou [21] overcomes the problem of excessive compression parameters of Lasso by introducing weights, and proposes an adaptive lasso model. It has the property of "Oracle properties". Yuan and Lin [22] proposed group lasso, Wang *et al.* [23] proposed group SCAD, and Huang *et al.* [24] proposed group MCP. For group variable selection, the variables in one group either all enter into the model or are all eliminated. However, in practical applications, there are cases where individual variables in some groups are not significant. Therefore, a method

which can not only select group variables but also select variables in a group is needed. That is the so-called, the so-called bi-level variable selection. After that, Huang *et al.* [25] proposed group bridge, and Simon *et al.* [26] proposed sparse group lass. All these are bi-level variable selection methods. The main contribution of this paper is to apply Logistic, Lasso-logistic, Group lasso-logistic and Lasso-SVM models to evaluate personal credit scores. Through experimental comparison, the advantages of the progressive selection, the backward selection and the Lasso method in the selection of variables are compared, and the prediction accuracy of each model is also compared.

In the third section, we present the algorithm models of Lasso-logistic, Lasso-SVM and Group lasso-logistic, and propose the method to select the parameter lambda in the model. In the fourth section, with the help of the credit data of the credit platform, SPSS software is used to preprocess the data. Section five and six use R language to compare and analyze the variable selection ability and prediction accuracy of the model through numerical experiments, so as to draw relevant conclusions.

3. Model

3.1. Logistic Model

Logistic regression is a probabilistic nonlinear model, which is a multivariable analysis method used to study the relationship between binary observation results and some influencing factors. Its basic idea is to study whether a result occurs under certain factors. For example, this paper uses some variable indicators to judge a person's credit status. Logistic regression can be expressed as:

$$P = \frac{1}{1+\mathrm{e}^{-s}},$$

$$s = \beta_0 + \sum_{i=1}^{n} \beta_i x_i.$$

where $x_i\left(i=1,2,\cdots,n\right)$ is the explanatory variable in the credit risk assessment (or the characteristic indicator of the individual), $\beta_i\left(i=1,2,\cdots,n\right)$ regression coefficient. Logistic regression value $P\in\left(0,1\right)$ is the discriminant result of credit risk.

The graph of the function in Logistic regression model has an *s* type distribution, as shown in **Figure 1**.

As you can see from **Figure 1**, *P* is a continuous increasing function of *s*, $s\in\left(-\infty,+\infty\right)$, and:

$$\lim_{s\to+\infty} P = \lim_{s\to+\infty} \frac{1}{1+\mathrm{e}^{-s}} = 1,$$

$$\lim_{s\to-\infty} P = \lim_{s\to-\infty} \frac{1}{1+\mathrm{e}^{-s}} = 0.$$

For someone $i\left(i=1,2,\cdots,n\right)$, if P_i is close to 1 (or $P_i \approx 1$), then it is judged as a "poor" credit person (or risk of default); if P_i is close to 0 (or $P_i \approx 0$),

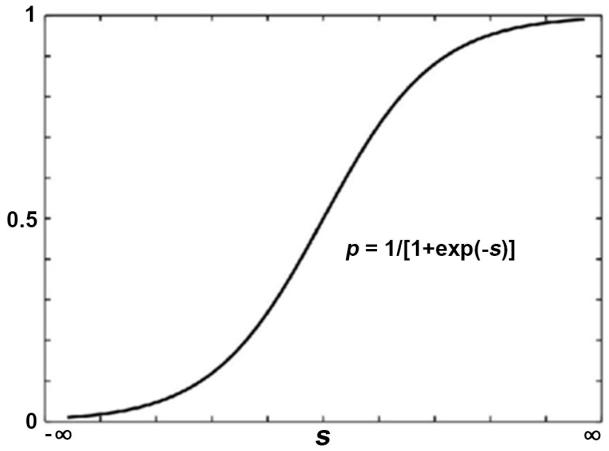

Figure 1. The graph of the logistic function.

then the person is judged to be "good". That is, the value of P_i farther away from 1 indicates that the person is less likely to fall into default set. On the contrary, it means that the risk of default is greater.

Suppose there are data variables $(x_i, y_i), i = 1, 2, \cdots, n$, where $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ which is the observed value of the explanatory variable and $y_i \in \{0,1\}$ is the observed value of the interpreted variable. In the general regression model, the observed values of the explanatory variable and the interpreted variable are often considered to be independent. In addition, assume that x_{ij} is standardized. Namely, $\frac{1}{n}\sum_i x_{ij} = 0, \frac{1}{n}\sum_i x_{ij}^2 = 1$. Let $P_i = P(y_i = 1 \mid x_i)$ be the conditional probability of $y_i = 1$ given x_i. The conditional probability under the same conditions is $P(y_i = 0 \mid x_i) = 1 - P_i$. Then, given a test sample (x_i, y_i), its probability is:

$$P(y_i) = P_i^{y_i}(1 - P_i)^{1-y_i},$$

where $P_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}}}$. Assume that each sample is independent of each other. Their joint distribution (*i.e.*, likelihood function) can be expressed as:

$$L(\beta_0, \beta_1, \cdots, \beta_m) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}.$$

The Maximum Likehood method is a good choice to estimate the parameter β. Because it can maximize the possibility that the observed value of each sample is equal to its true value. In other words, it can maximize the log likelihood function in the logistic model:

$$\begin{aligned}\ln(L(\beta_0, \beta_1, \cdots, \beta_m)) &= \ln\left(\prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} y_i\left(X_i\beta - \ln\left(1 + e^{X_i\beta}\right)\right).\end{aligned}$$

For convenience, we set $X_i = (1, x_i)$ and $\beta = (\beta_0, \beta_1, \cdots, \beta_m)^{\mathrm{T}}$. Estimating the model's parameter β by maximum likelihood estimation is equivalent to

solve the following problem:

$$\hat{\beta} = \arg\max l(\beta),$$

It is easy to know that $l(\beta)$ is concave and continuously differentiable, and therefore its local maximizer is the global maximizer. Calculate partial derivatives and make it to be zero, which leads to the likelihood equations:

$$\frac{\partial \ln\left(L\left(\beta_0, \beta_1, \cdots, \beta_m\right)\right)}{\partial \beta_0} = \sum_{i=1}^{n}\left(y_i - \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}}\right) = 0,$$

$$\frac{\partial \ln\left(L\left(\beta_0, \beta_1, \cdots, \beta_m\right)\right)}{\partial \beta_j} = \sum_{i=1}^{n}\left(y_i - \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}}\right) x_{ij} = 0.$$

But it is difficult to get an explicit solution. It needs to be solved by some iterative methods such as Newton-Raphson, EM and gradient descent algorithms. The estimated β_j obtained by the likelihood equation is called the maximum likelihood estimate, and the corresponding conditional probability P_i is estimated by $\widehat{P}_i$.

Logistic has a wide range of applications in credit scoring. The traditional Logistic method is very simple, but it is sensitive to multi-collinearity interference between individual credit variables. Therefore, some redundant variables are selected, resulting in poor prediction results. That is why we improve this method.

3.2. Lasso Model

Tishirani proposed the Lasso method which is motivated by non-negative Garrote [27].

Let $\hat{\beta} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_m\right)^{\mathrm{T}}$, the estimator $\left(\hat{\alpha}, \hat{\beta}\right)$ of the lasso method is:

$$\left(\hat{\alpha}, \hat{\beta}\right) = \arg\min_{\alpha,\beta}\left\{\sum_{i}^{n}\left(y_i - \alpha - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2\right\}, \text{ s.t.} \sum_{j=1}^{m}\left|\beta_j\right| \le t,$$

where $t \ge 0$ is the regularization parameter. For all t, one has an estimator $\hat{\alpha} = \bar{y}$ of α. Without loss of generality, we assume that $\bar{y} = 0$, Above problem can be rearranged into the following form:

$$\hat{\beta} = \arg\min_{\beta}\left\{\sum_{i}^{n}\left(y_i - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2\right\}, \text{ s.t.} \sum_{j}\left|\beta_j\right| \le t.$$

It can also be expressed in the form of the following penalty function:

$$\hat{\beta} = \arg\min_{\beta}\left\{\sum_{i}^{n}\left[\left(y_i - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{m}\left|\beta_j\right|\right]\right\}.$$

The first part of the formula represents the goodness of the model fit, and the second part represents the penalty of the parameter. The harmonic coefficient $\lambda \in [0, +\infty]$ is smaller. The smaller role of the penalty term plays the more variables is retained; the larger lambda is, the more roles of the penalty term plays, and the fewer variables are retained.

3.2.1. Logistic-Lasso Model

The Lasso method is mainly applied to linear models. The essence is to add a penalty function to the sum of squared residuals. When estimating parameters, the coefficients are compressed, and some coefficients are even compressed to 0 to achieve model variable selection. But for credit default prediction, the dependent variable is a binary value. In this case, the linear regression model cannot be used. Instead, Lasso-logistic [28] should be used. Penalized logistic regression is a modification of the logistic regression model. The negative log-likelihood function adds a non-negative penalty term to achieve good control of the coefficients.

The conditional probability of the logistic linear regression model can be expressed as:

$$\log\left\{\frac{P(y_i=1\mid x_i)}{1-P(y_i=1\mid x_i)}\right\}=\eta_\beta(x_i),$$

where $\eta_\beta(x_i)=X_i\beta$.

The coefficient estimate $\widehat{\beta_\lambda}$ in the Lasso-logistic regression model is given by the minimum value of the convex function of the following form:

$$S_\lambda(\beta)=-l(\beta)+\lambda\sum_{j=1}^{m}|\beta_j|,$$

where

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

The estimator $\hat{\beta}$ in Lasso-logistic regression model can be given as:

$$\hat{\beta}=\arg\min_\beta-\sum_{i=1}^{n}\left\{y_i\eta_\beta(x_i)-\log\left\{1+\exp\left[\eta_\beta(x_i)\right]\right\}\right\}+\lambda\sum_{j=1}^{m}|\beta_j|.$$

3.2.2. Lasso-SVM Model

The standard SVM model does not have feature selection capabilities. The specific approach of adding regularization to the SVM model is to use the regularization term with sparsity to replace the L_2 norm in the standard SVM. The L_1 norm is convex functions, with Lipschitz continuum, having properties better than other norms. L_1-SVM and its similar extensions have evolved into one of the most important tools for data analysis. The general form of Lasso-SVM is given below:

$$\min\sum_{j=1}^{m}|\beta_i|+C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.}\quad y_i\left(X_i^{\mathrm{T}}\beta\right)\geq 1-\xi_i,\ \ i=1,2,\cdots,n$$

$$\xi_i\geq 0.$$

Lasso-SVM can also be written in the following form:

$$\min \sum_{i=1}^{n}\left[1-y_i f\left(X_i\right)\right]_+ + \lambda \sum_{j=1}^{m}\left|\beta_i\right|.$$

where $\left[1-y_i f\left(X_i\right)\right]_+$ is a Hinge loss function and λ is a regularization parameter.

3.2.3. Group Lasso-Logistic Model

Group lasso was introduced by Yuan and Lin (2006), allowing pre-defined covariates to be grouped together and selected from the model. All variables in a particular group can be included or not included. It is very useful in many settings. Group lasso algorithm for logistic regression was first proposed by Kim *et al.*, and then Meier *et al.* [29] proposed a new one which can solve high dimensional problems.

Suppose there is an independent and identical distribution of observation $\left(x_i, y_i\right), i=1,2,\cdots,n$. $x_i=\left(x_{i1}, x_{i2},\cdots, x_{im}\right)$ which is an m-dimensional vector that can be divided into G groups, and the dependent variable is a binary variable $y_i \in \{0,1\}$. The independent variable can be a continuous variable or a classified variable. Assume that the degree of freedom of the group g argument is df_g, $X_i=\left(1, x_{i,1}, x_{i,2},\cdots, x_{i,G}\right), \left(g=1,2,\cdots,G\right)$. $x_{i,g}$ denotes the $x_{i,g}$ group of variables of the observation $x_{i,g}$. Similarly, β can be expressed as $\left(\beta^0;\beta^1;\beta^2;\cdots;\beta^G\right)$, β^g denotes the coefficients corresponding to group G g variables, where the labeling method is used to distinguish the β_j fraction in the case of no grouping. The probability of "default" of the dependent variable $P_{\beta(x_i)}=P_\beta\left(y=1\mid x_i\right)$ can be expressed by the following model:

$$\log\left\{\frac{P_{\beta(x_i)}}{1-P_{\beta(x_i)}}\right\}=\eta_\beta\left(x_i\right)=\beta_0+\sum_{g=1}^{G}x_{i,g}^{\mathrm{T}}\beta_g=X_i\beta,$$

where β_0 denotes intercept, β_g is the coefficient vector corresponding to group g and β is the whole coefficient vector.

The parameter $\hat{\beta}(\lambda)$ can be estimated by minimizing the convex function:

$$S_\lambda\left(\beta\right)=-l\left(\beta\right)+\lambda\sum_{g=1}^{G}s\left(df_g\right)\left\|\beta_g\right\|^2,$$

where $l(\beta)$ is a logarithmic likelihood function:

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

$s\left(df_g\right)=\left(df_g\right)^{1/2}$ and $s(\cdot)$ is used to rescale the parameter β_g vector.

3.3. The Choice of Harmonic Parameter

In the variable selection model, the key lies in the selection of the harmonic parameter lambda. That is to say, the optimal lambda determines the prediction accuracy and robustness of the model. The common methods for the optimal lambda are AIC, BIC, Cross-validation, Generalized cross-validation. Here, we use *K*-fold cross-validation to determine the optimal lambda.

The main idea of K-fold cross validation is that the data are randomly divided into K (usually 5 or 10) identical parts. Each $k=1,\cdots,K$, uses the data of the K part as the test sample, and uses the remaining K-1 parts of the data as the training sample to fit the model. Loop K times until all k are traversed. We denote the estimator by $\hat{\beta}^{-k}$. The harmonic parameter $\hat{\beta}(\lambda)$ corresponds to a classification model and the corresponding estimator $\hat{\beta}(\lambda)$. The generalization error of each model corresponding to lambda is given by the mean square prediction error. That means Cross-Validation Error (CVE) is estimated:

$$CV(\lambda)=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in C_k}\left(y_i-X_i^{\mathrm{T}}\hat{\beta}^{-k}\right)^2,$$

where C_k is the k-th partial cross-check sample, $n_k=|C_k|=\frac{n}{K}$. Minimize the above formula to find the most appropriate harmonic parameters, and the corresponding model can be considered to be the model with the best performance based on cross-check error.

4. Data

4.1. Data Source

The original data is mainly from a domestic lending institution. There are a total of 8000 records in this data set, including 25 fields. Among them, 23 fields describe the personal characteristics of the lender, including the basic personal identity information: domicile, gender, local work, education level and marital status. Also include personal economic ability: whether there is a CPF salary level. Data set also includes personal debt and debt repayment record: frequency of personal housing loan, personal commercial housing loan pen number and frequency of other loan credit card account number, number, frequency of delinquent loans, loans overdue month loan highest monthly overdue amount, maximum length, loan account number of the contract amount, loan balance has been used lines, the average individual loan maximum contract value, the average individual loans minimum contract amount, the last six months on average use. Finally, the data set also gives the total number of times of individual approval query and loan number. The result, where "0" is the performance customer and "1" is the default customer.

4.2. Data Preprocessing

There are missing and abnormal data in the original data, and the missing value filling and outlier detection are needed before analysis. The method of dealing with missing values in this paper is the average filling, and using the scatter plot to detect outliers. In the original data, such as contract amount, loan balance and used amount are continuous variables. In order to overcome the influence of the dimension, the F-score needs to be standardized and analyzed. At the same time, the ratio of the number of compliance users and default users in the sample data is about 9:2, and it is an asymmetric distribution problem, which affects the pre-

diction accuracy of the model for default customers with relatively small data capacity. Therefore, the under-sampling method is adopted for compliance users. That means some representative data are selected from the data with more samples. In order to reduce the majority of the sample, the data balance is achieved. The final data set is divided into a training set and a test set, wherein the training set has 3002 data, including 1500 compliance data and 1502 default data, and the test set has 519 data including 258 compliance data and 261 default data.

4.3. Variable Description

The classified and encoded variables in the data are shown in **Table 1**.

5. Numerical Experiment

The full-variable logistic and stepwise logistic regression models were implemented by using SPSS 22.0. The Lasso-logistic model was implemented by using the glmnet package in R language, the Lasso-SVM model was implemented by using the gcdnet package, and the Group lasso-logistic model was implemented by using grpreg. The code package uses the generalized coordinate descent method [30] to calculate the model under regularization and its generalized solution path.

5.1. Parameter Lambda Selection

Through the K-fold cross-validation, the Lasso-logistic, Lasso-SVM, and Group lasso-logistic models are changed with the value of lambda, and the model error is changed. At the top of **Figure 2**, the number of corresponding variables selected by the model is given. The value between the two dotted lines in **Figure 2** indicates the range of positive and negative standard deviation of lambda, and the dotted line on the left indicates lambda when the model error is minimized. Tibshirani contends that lambda takes a relatively small change in the model prediction bias within this interval. It is generally recommended to choose lambda which makes the model relatively simpler, namely, a large lambda within a standard deviation range. It is the best value. It can also be seen from **Figure 2** that as the value of lambda changes, the degree of compression of the model variable also changes. In other words, the number of variables to be filtered is affected by the estimate of lambda.

Figure 3 shows the filtering of variables in Lasso-logistic, Lasso-SVM, Group lasso-logistic model with the change of harmonic parameter lambda. As the value of lambda increases, the degree of the model compression increases, more variables of the model are deleted, while less variables are retained, and the function of selecting important variables is enhanced. The best value of lambda is the log(lambda), next to the right dotted line's value. In Lasso-logistic, lambda = 0.01122485; in Lasso-SVM, lambda = 0.00699683; and in Group lasso-logistic, lambda = 0.01628534.

Table 1. Variable declaration.

ID	Variable Name	The Values
y	default	$y=0$ Not default; $y=1$ Default
x_1	registered area	$x_{1,1}$ = Northeast; $x_{1,2}$ = North China Plain; $x_{1,3}$ = Central China; $x_{1,4}$ = Eastern China
x_2	gender	$x_{2,1}$ = Male; $x_{2,2}$ = Female
x_3	Whether work is local	$x_{3,1}$ = Local; $x_{3,2}$ = Not local
x_4	edu level	$x_{4,1}$ = Junior high school/Senior high school/others; $x_{4,2}$ = Junior college/Junior college and below; $x_{4,3}$ = Undergraduate; $x_{4,4}$ = Master/Doctor
x_5	marital status	$x_{5,1}$ = Maid; $x_{5,2}$ = Married; $x_{5,3}$ = Others (divorced/widowed)
x_6	Whether accumulation fund	$x_{6,1}$ = Not; $x_{6,2}$ = Yes
x_7	pay grades	$x_{7,1}$ = 0 - 3500; $x_{7,2}$ = 3501 - 8000; $x_{7,3}$ = 8000 above
x_8	The number of individual housing loans	$x_8 \in N$
x_9	The number of individual commercial housing loans	$x_9 \in N$
x_{10}	Other loans	$x_{10} \in N$
x_{11}	Debit card account number	$x_{11} \in N$
x_{12}	The number of overdue loans	$x_{12} \in N$
x_{13}	Months of overdue loans	$x_{13} \in N$
x_{14}	The maximum amount of overdue loans per month	$x_{14} \in [0,+\infty]$
x_{15}	Maximum length of loan (year)	$x_{15} \in N$
x_{16}	Total number of approval inquiries	$x_{16} \in N$
x_{17}	Loan number	$x_{17} \in N$
x_{18}	Loan account number	$x_{18} \in N$
x_{19}	contract amount	$x_{19} \in [0,+\infty]$
x_{20}	loan balance	$x_{20} \in [0,+\infty]$
x_{21}	Have used limit	$x_{21} \in [0,+\infty]$
x_{22}	Average maximum contract amount for a single lender	$x_{22} \in [0,+\infty]$
x_{23}	Average minimum contract amount for a single lender	$x_{23} \in [0,+\infty]$
x_{24}	Average usage in the last 6 months	$x_{24} \in [0,+\infty]$

5.2. Coefficient of the Models

In the logistic regression model, the dependent variable is log-occurrence ratio logit. When the log-occurrence ratio increases, the value of P also increases

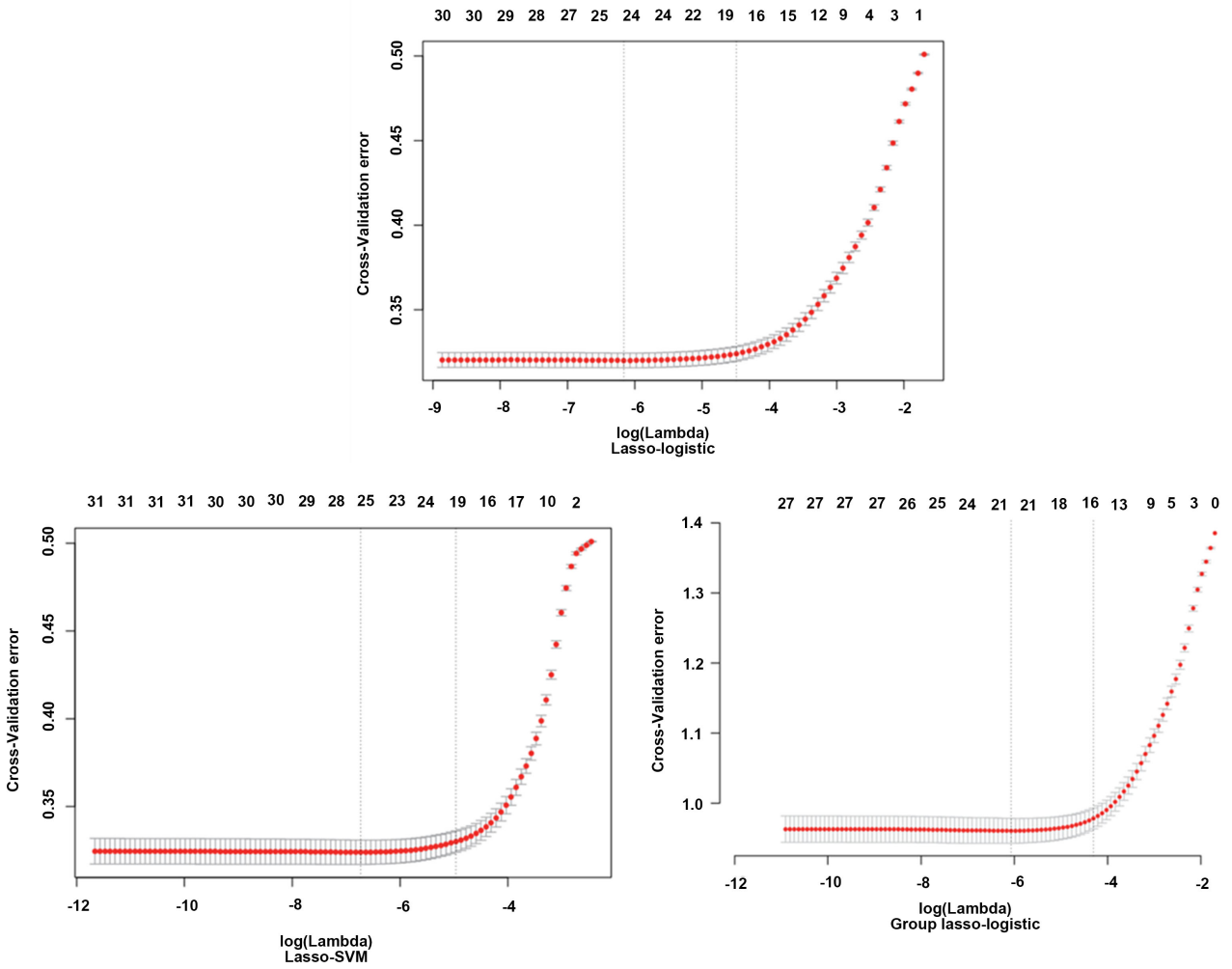

Figure 2. Lambda corresponds to the number of variables.

accordingly, which means the probability for judging credit as 1 (*i.e.*, default) increases. When the coefficient β_i is negative, it means that the variable x_i has a reverse restrictive effect on the default. When the coefficient β_i is positive, the corresponding variable x_i has a positive effect on the default, and the greater the value of β_i, the greater the promoting effect of the corresponding x_i on the customer's credit judgment as default. In full-variable logistic model, the variable x_6 (Whether accumulation fund), x_{11} (Debit card account number), x_{14} (The maximum amount of overdue loans per month), x_{22} (Average maximum contract amount for a single lender), x_{13} (Average minimum contract amount for a single lender) before the coefficients were not significant, which means the model contains too many variables, and the model is too complicated. The above non-significant variables were eliminated by stepwise regression. At the same time, x_{18} (Loan account number) and x_{21} (Have used limit) were also eliminated. Finally, 13 variables were removed for both forward and backward modes.

For Lasso-logistic model, there are 16 variables whose coefficient is compressed to 0. In other words, 18 important variables are selected to enter the

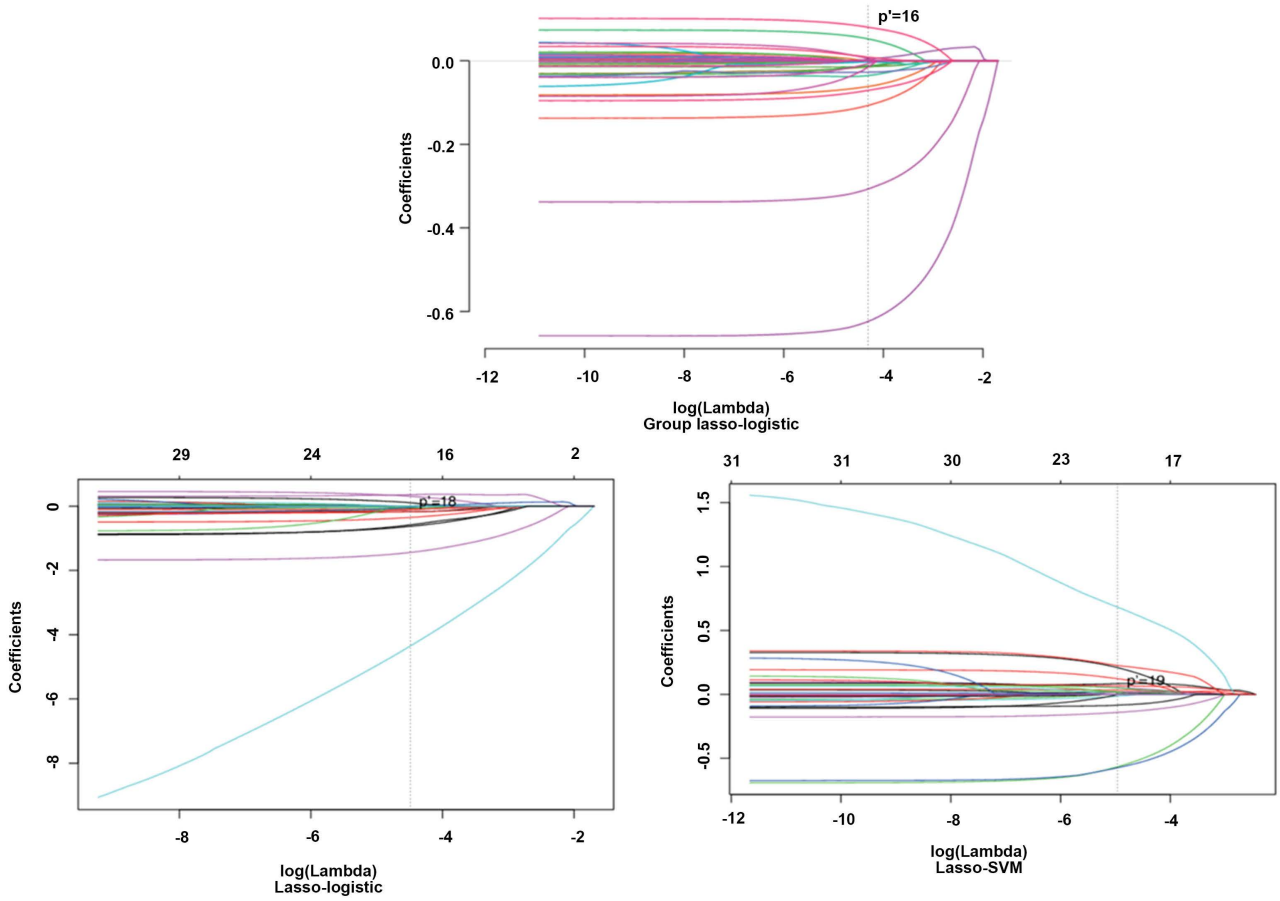

Figure 3. Lasso coefficient solution path.

model. The Lasso-SVM model eliminates 15 variables, leaving 19 variables. However, it can be seen that when using stepwise regression, Lasso-logistic and Lasso-SVM models for variable selection, the variables are excluded as classification variables, and some dummy variables in the same group are partially retained and partially eliminated, such as x_4 (edu level), which makes the result difficult to explain, showing in Table 2.

Using Group lasso, after variable selection, 18 variables were removed and 16 variables were retained. In addition, Group lasso-logistic can retain or eliminate related dummy variables of the same group as a whole, making the dummy variables have explanatory significance. We obtained from the coefficient table of the Group lasso model that in the regional variable (x_1), the north China area is the high default area, and the central China area has the lowest default risk. There was a significant gender (x_2) difference in credit risk that the default probability of male was generally higher than that of female. In the salary scale (x_7), people with low incomes were more at risk of default than those with medium and high incomes. In historical credit records, customers with overdue loans are at greater risk of default, and the number of overdue loans (x_{12}) and months of overdue loans (x_{13}) are more likely to default. More total number of approval inquiries (x_{16}) affect an individual's credit history. Variables with a

Table 2. Model coefficient table.

Variate	Full variables	Forward	Backwards	Lasso-logistic	Lasso-SVM	Group-lasso
X_{1_1}	1.715	0	1.721	0	−0.555	0.025
X_{1_2}	1.674	0	1.695	0.010	−0.567	0.001
X_{1_3}	−4.562	−4.247	−4.561	−4.354	0.667	−6.520
X_{1_4}	−1.534	−1.681	0	−1.440	0.016	−1.637
X_2	−0.868	−0.854	−0.858	−0.630	0.209	−0.824
X_3	−0.497	−0.504	−0.502	−0.352	0.122	−0.471
X_{4_1}	1.408	0.341	0.501	0.105	−0.003	0.207
X_{4_2}	0.780	0.301	0.218	0	0	−0.007
X_{4_3}	0.575	0	0	−0.111	0.015	−0.195
X_{4_4}	0.761	0	0	0	0	−0.682
X_{5_1}	−0.025	−0.002	0	0	0	0.062
X_{5_2}	−0.179	0	−0.201	−0.037	0	−0.072
X_{5_3}	−0.034	0	0	0	0	0.073
X_6	0.072	0	0	0	0	0.065
X_{7_1}	0.306	0.393	0.311	0.355	−0.085	0.570
X_{7_2}	−0.893	0.205	−1.111	−0.582	0.226	−0.554
X_{7_3}	0.934	0	0	0	0	0.251
X_8	0.123	0.101	0.098	0	0	0
X_9	0.058	0	0	0	0	0.039
X_{10}	−0.265	−0.213	−0.212	−0.104	0.053	−0.194
X_{11}	−0.078	0	0	−0.003	0.013	−0.016
X_{12}	−0.100	−0.111	−0.109	−0.084	0.033	−0.108
X_{13}	−0.190	−0.206	−0.197	−0.064	0.050	−0.170
X_{14}	−0.033	0	0	0	0	−0.021
X_{15}	0.130	0.121	0.119	0	−0.016	0.101
X_{16}	0.455	0.459	0.457	0.306	−0.142	0.431
X_{17}	0.222	−0.218	−0.209	−0.173	0.084	−0.179
X_{18}	0.164	0	0	0	0	0
X_{19}	−0.405	0	0	−0.021	0	−0.062
X_{20}	0.276	0	0	0	0	0
X_{21}	0.136	0	0	0	0	0
X_{22}	0.065	0	0	0	0	0
X_{23}	0.041	0	0	0	0	0.035
X_{24}	−0.325	−0.249	−0.263	−0.178	0.067	−0.212
Intercept term	1.026	1.275	1.836	2.575	−0.382	2.980

coefficient 0 indicate that they have been removed from the model and have little effect on credit rating.

Showing in Table 2, the number of the full-variable logistic model is the largest, and the complexity of the model is the largest. The forward and backward models excluded 13 explanatory variables, while the Lasso-logistic model excluded 16 variables, three more than the stepwise selection. The number of Lasso-SVM excluded variables was 15, one less than the Lasso-logistic model, and two more than the stepwise selection. The Group lasso-logistic model had the strongest ability to eliminate variables, with 18 variables removed. It can also be concluded that in the selection of the same group of dummy variables, the Group lasso-logistic model retains or removes the entire group of variables, making the model variables have explanatory significance.

5.3. Model Prediction Accuracy

In the actual credit risk assessment, the misclassification of default users into non-defaulting users is more of a potential loss to banks or society. Therefore, the model is more important for to correctly classify the default users than to take non-defaulting users into consideration. It is easy to see in Table 3 that in the training set, the Lasso-SVM model predicts that the number of default users will be up to 80.16%, which is 4.53% higher than the full-variable model and is higher than the stepwise forward and backward selections 6.59% and 6.39% respectively. The Lasso-logistic and Group lasso-logistic models also predicted the default users could reach 79.96% and 80.09% respectively; in the test set, the Group lasso-logistic model got the best prediction on default users, reaching 80.62%. It is higher than the full-variable, forward and backward models 8.97%, 14.34%, 13.57% respectively, and the Lasso-SVM model is the second most accurate for the default user. The Lasso-logistic model follows. Next, look at the classification of non-defaulting users. Lasso-logistic is the best rate in both the training set and the test set. The forward selection model had the worst prediction accuracy for non-defaulting users. In the overall prediction accuracy, the stepwise selection performed poorly in the test set. Lasso-logistic reached 77.21% in the training set, and Group lasso-logistic model has the highest overall prediction rate in the test set, reaching 77.26%.

6. Conclusions

In the personal credit evaluation, the Logistic model is most widely used, and the newly proposed SVM method in statistical learning also has certain application in credit evaluation. By comparing the simulation experiment analysis, the whole variable, Forward selection, Backward selection, Lasso-logistic, Lasso-SVM, and Group lasso-logistic models and empirically analyzing the personal credit data of a domestic lending platform, it can be concluded:

First, the experiment found that when all the variables were included in the full-variable Logistic mode, the coefficients before many variables could not pass

Table 3. Model prediction rate.

Model	Training set			Test set		
	Good	Bad	Total	Good	Bad	Total
Full variables	71.3	75.6	73.4	67.0	71.7	69.3
Forward selection	70.7	75.6	73.1	65.5	72.1	68.9
Backward selection	72.3	73.8	73.1	69.3	67.1	68.2
Lasso-logistic	74.5	80.0	77.2	74.7	79.5	77.1
Lasso-SVM	73.7	80.2	76.9	73.6	80.2	76.8
Group lasso	74.4	79.4	76.9	75.1	78.0	76.5

the significant level test. Thus, to some extent, the complexity of the model was increased. The interpretability of the model was reduced. The choice and Lasso overcome the multicollinearity of the full-variable model, and the coefficients of the insignificant variables in the model are compressed. Compared with the stepwise regression, the Group lasso-logistic culling variable is the strongest, followed by Lasso-logistic, Lasso-SVM model. The algorithm model based on Lasso variable selection can better select important variables, and Group lasso-logistic will retain the whole group or the entire group when the same group of dummy variables is selected, which will enhance the variables in the model to some extent.

Second, in the training set, the Lasso-SVM model has the highest prediction accuracy rate for default users; in the test set, Group lasso-logistic ranks the first in the classification accuracy of default users. Whether in the training set or in the test set, the best classification accuracy of non-defaulting users is the Lasso-logistic model. Moreover, in the training set, the overall prediction accuracy of the Lasso-logistic model is also the best. In the test set, the Group lasso-logistic model has the best overall prediction accuracy. Regardless of the prediction of defaulting users, the prediction of non-defaulting users and the overall forecasting accuracy, Lasso is better than stepwise selection. It shows that the credit scoring model based on Lasso variable selection has good extrapolation.

Therefore, based on the Logistic and SVM models established by the Lasso variable selection method, the explanatory variables can be selected more scientifically and have use value in personal credit risk assessment, which can well reduce personal credit risk.

To sum up, it is not difficult to find that in the actual rating, we often encounter some relationships between variables, thus forming a grouping structure. The traditional variable selection method cannot process the dummy variables of related groups as a whole, resulting in partial retention and partial elimination of the variables of the whole Group. In this way, the results are difficult to be explained, and Group lasso can well solve the above problems. Therefore, the Logistic and SVM models established based on the Lasso variable selection method can more scientifically select explanatory variables, which have application value

in personal credit risk assessment and can well reduce personal credit risk.

In future work, we will consider individual credit ratings for unbalanced datasets. When we use Group lasso for intra-group variable selection, the coefficient of some individual variables within the Group may not be significant. In this case, the two-layer variable selection is introduced to solve such problems.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Eisenbeis, R.A. (1978) Problems in Applying Discriminant Analysis in Credit Scoring Models. *Journal of Banking & Finance*, **2**, 205-219. https://doi.org/10.1016/0378-4266(78)90012-2

[2] Henley, W.E. (1995) Statistical Aspects of Credit Scoring. Ph.D. Thesis, Open University, Milton Keynes.

[3] Chatterjee, S. and Barcun, S. (1970) A Nonparametric Approach to Credit Screening. *Journal of the American Statistical Association*, **65**, 150-154. https://doi.org/10.1080/01621459.1970.10481068

[4] Breiman, L.I., Friedman, J.H., Stone, C.J. and Olshen, R.A. (1984) Classification and Regression Trees (CART). *Biometrics*, **40**, 874. https://doi.org/10.2307/2530946

[5] Jensen, H.L. (1992) Using Neural Networks for Credit Scoring. *Managerial Finance*, **18**, 15-26. https://doi.org/10.1108/eb013696

[6] Desai, V.S., Crook, J.N. and Overstreet Jr., G.A. (1996) A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment. *European Journal of Operational Research*, **95**, 24-37. https://doi.org/10.1016/0377-2217(95)00246-4

[7] Van Gestel, T., Baesens, B., Garcia, J. and Van Dijcke, P. (2003) A Support Vector Machine Approach to Credit Scoring. *Bank en Financiewezen*, **2**, 73-82.

[8] Wiginton, J.C. (1980) A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *The Journal of Financial Quantitative Analysis*, **15**, 757-770. https://doi.org/10.2307/2330408

[9] Shi, Q.-Y. and Jin, Y.-Y. (2004) A Comparative Study on the Application of Various Personal Credit Scoring Models in China. *Statistical Research*, **21**, 43-47.

[10] Xiang, H. and Yang, S.-G. (2011) New Developments in the Study of Key Techniques for Personal Credit Scoring. *The Theory and Practice of Finance and Economics*, **32**, 20-24.

[11] Shen, C.-H., Deng, N.-Y. and Xiao, R.-Y. (2004) Personal Credit Evaluation Based on Support Vector Machine. *Computer Engineering and Applications*, **40**, 198-199.

[12] Hu, X.-H. and Ye, W.-Y. (2012) Variable Selection in Credit Risk Analysis Model of Listed Companies. *Journal of Applied Statistics and Management*, **31**, 1117-1124.

[13] Akaike, H. (1973) Information Theory and Extension of the Maximum Likelihood Principle. In: Parzen, E., Tanabe, K. and Kitagawa, G., Eds., *Selected Papers of Hirotugu Akaike*, Springer, New York, 267-281.

[14] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*,

6, 461-464. https://doi.org/10.1214/aos/1176344136

[15] Tibshirani, R. and Knight, K. (1999) The Covariance Inflation Criterion for Adaptive Model Selection. *Journal of the Royal Statistical Society*, **61**, 529-546. https://doi.org/10.1111/1467-9868.00191

[16] Mallows, C.L. (1973) Some Comments on C_p. *Technometrics*, **15**, 661-675. https://doi.org/10.2307/1267380

[17] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12**, 69-82. https://doi.org/10.1080/00401706.1970.10488635

[18] Frank, I.E. and Friedman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135. https://doi.org/10.1080/00401706.1993.10485033

[19] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[20] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *The Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[21] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[22] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society*, **68**, 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

[23] Wang, L., Chen, G. and Li, H. (2007) Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data. *Bioinformatics*, **23**, 1486-1494. https://doi.org/10.1093/bioinformatics/btm125

[24] Huang, J., Breheny, P. and Ma, S. (2012) A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, **27**, 481-499. https://doi.org/10.1214/12-STS392

[25] Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009) A Group Bridge Approach for Variable Selection. *Biometrika*, **96**, 339-355. https://doi.org/10.1093/biomet/asp020

[26] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013) A Sparse-Group Lasso. *Journal of Computational & Graphical Statistics*, **22**, 231-245. https://doi.org/10.1080/10618600.2012.681250

[27] Breinman, L. (1995) Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, **37**, 373-384. https://doi.org/10.1080/00401706.1995.10484371

[28] Fang, K.-G., Zhang, G.-J. and Zhang, H.-Y. (2014) Personal Credit Risk Warning Method Based on Lasso-Logistic Model. *The Journal of Quantitative & Technical Economics*, **2**, 125-136.

[29] Meier, L., Van De Geer, S. and Bühlmann, P. (2008) The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society*, **70**, 53-71. https://doi.org/10.1111/j.1467-9868.2007.00627.x

[30] Friedman, J., Hastie, T. and Tibshirani, T. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1-22. https://doi.org/10.18637/jss.v033.i01