Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 862132, 9 pages
http://dx.doi.org/10.1155/2014/862132

Research Article

A Model for Recognizing Key Factors and Applications Thereof to Engineering

Baofeng Shi and Guotai Chi

School of Business Management, Dalian University of Technology, Dalian 116024, China

Correspondence should be addressed to Guotai Chi; chigt@dlut.edu.cn

Received 9 September 2013; Revised 6 November 2013; Accepted 1 December 2013; Published 8 January 2014

Academic Editor: Gelan Yang

This paper presents an approach to recognize key factors in data classification. Using collinearity diagnostics to delete the factors of repeated information and Logistic regression significant discriminant to select the factors which can effectively distinguish the two kinds of samples, this paper creates a model for recognizing key factors. The proposed model is demonstrated by using the 2044 observations in finical engineering. The experimental results demonstrate that the 13 indicators such as "marital status," "net income of borrower," and "Engel's coefficient" are the key factors to distinguish the good customers from the bad customers. By analyzing the experimental results, the performance of the proposed model is verified. Moreover, the proposed method is simple and easy to be implemented.

1. Introduction

With the advent of the era of big data, the data classifications exist in fingerprint recognition, facial recognition, customer classification, DNA identification, product category, and so forth. So it has become more and more important for researchers to find the key factors which are capable of effectively distinguishing the data. For this purpose, many mathematical models are explored as the decision support methods to classify the data.

In the literature, there are two main data classification methods. Artificial intelligence method is one of the classification methods for recognizing key factors. Hu et al. proposed an adaptive multilevel kernel machine method for scene classification and experimented on two popular benchmark datasets, which demonstrated that the proposed model outperformed the original spatial PACT [1]. In order to obtain the performance of customer classification models, Finlay compared the performance of several multiple classifiers and found that Error Trimmed Boosting outperformed all other multiple classifiers on UK credit data [2]. Akkoc proposed a three stage hybrid Adaptive Neuro Fuzzy Inference System client classification model, which is based on statistical techniques and Neuro Fuzzy. The proposed model performs better than the Linear Discriminant Analysis, Logistic Regression Analysis, and Artificial Neural Network (ANN) approaches [3]. By combining the biometric fractal pattern and particle swarm optimization (PSO)-based classifier, a fingerprint recognition model was established [4]. Twala explored the predicted behavior of five classifiers for different types of noise in terms of credit risk prediction accuracy and how such accuracy could be improved by using classifier ensembles. The experimental evaluation showed that the ensemble of classifiers technique has the potential to improve prediction accuracy [5]. Chen studied the classification problem of default customers and nondefault customers by using Support Vector Machines. Experiment demonstrated that the proposed model can effectively recognize the key factors [6].

Statistics and measurement method is another classification tool to solve this problem. Căleanu et al. studied the problems of feature extraction and classifier design in facial recognition by combining a feature extraction technique and a k-NN statistical classifier method. Experimental results showed that the approach enables them to achieve both higher classification accuracy and faster processing time [7]. Compared with conventional models such as multiple

discriminant analysis, logistic regression analysis, and neural networks for the classification problems of bankrupt firms and nonbankrupt firms, Min and Lee proposed the DEA classification model [8]. Shi and Chi studied the customer classification problem by combining correlation analysis and Probit regression. Experiment demonstrated that the proposed model can recognize the key factors which can effectively distinguish the default customers from the nondefault ones [9]. In order to distinguish good customers and bad customers, Hwang et al. established ordered semiparametric Probit customers classification model by substituting ordered semiparametric function for linear regression function [10]. Sun et al. presented a classification method for distinguishing distressed enterprises and nondistressed enterprises based on gray forecasting and pattern recognition. Then the calculating result was classified to judge state of enterprise with the pattern recognition model [11]. Because attribute interactions toward classification were not considered in the classification methods, a new nonlinear classification method with nonadditive measures was proposed. Experimental results showed that applying nonadditive measures on the classic optimization-based models could improve the classification robustness and accuracy by comparing with some popular classification methods [12]. Because the existing model-based approaches are often conceptually and numerically instable for large and complex data sets, Corander et al. considered a Bayesian model-based method for unsupervised classification of discrete valued vectors, which have certain advantages over standard solutions based on latent class models [13].

Although the existing researches have made great progress, there are still some drawbacks. Firstly, the collinearity between factors cannot be excluded in the existing classification researches. Secondly, the existing models include the factors which are unable to effectively distinguish the two types of samples.

The purpose of this paper is to set up a model for recognition key factors, which is based on collinearity diagnostics and Logistic regression significant discriminant. Using a Chinese state-owned commercial bank's 2044 petty loans for farmers, the proposed model is tested by screening the key factors which can effectively distinguish the good customers from the bad ones.

The rest of the paper is structured as follows. We will give the constructing principle of the model for recognition key factors in Section 2. The third part is the construction steps of this model. The fourth part presents the data and the empirical results. Conclusions are given in Section 5.

2. The Constructing Principle of the Model for Recognizing Key Factors

(1) The Principle of Screening Key Factors. It is obtained by the bilateral probability P_j of regression coefficient c_j for every factor x_j^i by constructing the Logistic regression model among factors x_j^i and default state y_i of customers (wherein, y_i is equal to 0 denoting the ith customer is a good customer and y_i is equal to 1 denoting the ith customer is a bad customer). Comparing the bilateral probability P_j with the given critical probability P_0, it can distinguish whether the factor x_j^i has an obvious effect on default state of customers. By deleting these factors that have no obvious effect on default status, it ensures that the reserved factors can effectively distinguish the bad customers from the good ones.

(2) The Principle of Eliminating Redundant Information between Factors. The more redundant factors data system includes, the more disorder the data classification results will be. This paper eliminates the repeated information of the factors by using collinearity diagnostics.

Flowchart of the research methodology is shown in Figure 1.

3. The Recognition Key Factors Model Based on Collinearity Diagnostics and Logistic Regression Significant Discriminant

3.1. Data Standardization. There are two kinds of factors in practice. One is called quantitative factors (namely, quantitative indicators) and the other is called qualitative factors (namely, qualitative indicators).

(1) The Data Standardized of Quantitative Factors. The quantitative factors include positive factors, negative factors, and interval factors. The positive factors are the factors whose values are the bigger, the better; the negative factors are the factors whose values are the smaller, the better. And the interval factors are the factors whose values are reasonable only when they lie in certain intervals.

Let x_j^i denote the standardization score of the ith observed value of the jth indicator. Let v_{ij} denote the factor data of the ith observed value of the jth indicator. Let n denote the number of observations. The standardization equations of the positive factors and the negative factors are shown as (1) and (2), respectively [12],

$$x_j^i = \frac{v_{ij} - \min_{1\le i\le n}\left(v_{ij}\right)}{\max_{1\le i\le n}\left(v_{ij}\right) - \min_{1\le i\le n}\left(v_{ij}\right)}, \tag{1}$$

$$x_j^i = \frac{\max_{1\le i\le n}\left(v_{ij}\right) - v_{ij}}{\max_{1\le i\le n}\left(v_{ij}\right) - \min_{1\le i\le n}\left(v_{ij}\right)}. \tag{2}$$

Let q_1 denote the left boundary of the ideal interval. Let q_2 denote the right boundary of the ideal interval. The standardization of the interval factors is shown as follows [12]:

$$x_j^i = \begin{cases} 1 - \dfrac{q_1 - v_{ij}}{\max\left(q_1 - \min_{1\le i\le n}\left(v_{ij}\right), \max_{1\le i\le n}\left(v_{ij}\right) - q_2\right)}, \\ \qquad\qquad v_{ij} < q_1 \ (a), \\ 1 - \dfrac{v_{ij} - q_2}{\max\left(q_1 - \min_{1\le i\le n}\left(v_{ij}\right), \max_{1\le i\le n}\left(v_{ij}\right) - q_2\right)}, \\ \qquad\qquad v_{ij} > q_2 \ (b), \\ 1, \quad q_1 \le v_{ij} \le q_2 \ (c). \end{cases} \tag{3}$$

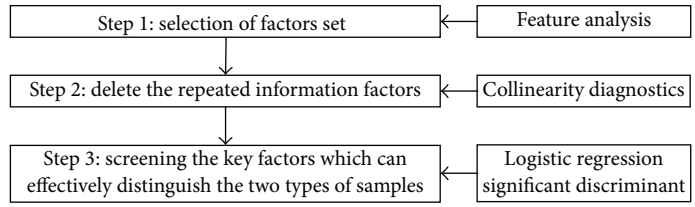

Figure 1: Flowchart of the research methodology.

The meanings of the rest of letters in (3) are the same as the letters in (1).

(2) The Data Standardization of Qualitative Factors. By rational analysis and expert investigation for qualitative factors, the scoring standard of qualitative factors can be obtained.

3.2. Deleting the Repeated Information Factors Based on Collinearity Diagnostics

(1) The Steps of Collinearity Diagnostics

Step 1 (building regression equation). Let x_j^i denote the standardization score of the ith observed value of the jth factor ($i = 1, \ldots n,\ j = 1, \ldots, m$); then the regression equation of this factor with the rest of the factors is as follows:

$$x_j^i = a_0 + a_1 x_1^i + \cdots + a_{j-1} x_{j-1}^i + a_{j+1} x_{j+1}^i + \cdots + a_m x_m^i. \tag{4}$$

The estimated values a_i can be obtained by using least squares estimation in (4). Substituting these parameters a_i into (4), the estimated value $\hat{x}_j^i$ of factor x_j^i can be obtained.

Step 2 (calculating the determination coefficient R_j^2). Let $\overline{x}_j$ denote the mean value of the jth indicator. Then

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}. \tag{5}$$

Let R_j^2 denote the determination coefficient of the jth indicator. Then

$$R_j^2 = \frac{\sum_{i=1}^{n} \left(\hat{x}_{ij} - \overline{x}_j\right)^2}{\sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2}. \tag{6}$$

The economic meanings of (6) are as follows. The bigger determination coefficient R_j^2 is, the stronger the correlation between the jth factor and the rest of factors will be. That is to say, the rest of factors can reflect the jth factor information effectively, and the jth factor should be deleted.

Step 3 (calculating the variance inflation factor VIF). Let VIF_j denote the variance inflation factor of the jth indicator. Then

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}. \tag{7}$$

The economic meanings of (7) are as follows. The variance inflation factor VIF_j reflects the correlation between the jth factor and the rest of factors in the same feature layer. If the variance inflation factor VIF_j is greater than 10 [14], it indicates that there is a multicollinearity between the jth factor and the rest of factors, and the jth factor should be deleted.

(2) The Standard of Collinearity Diagnostics Screening. Factors, reflecting repeated information, constitute a set, and the factors only whose variance inflation factors are smaller than 10 are reserved [14].

3.3. Screening the Key Factors Based on Logistic Regression Significant Discriminant

(1) The Establishment of Logistic Regression Function. Let y_i denote the data status of the ith observed value; let y_i equal to 0 denote that the ith observed value belongs to the first kind of sample, for example, good sample; let y_i equal to 1 denote that the ith observed value belongs to the other kind of sample, for example, bad sample. Let a and c_j denote regression coefficients. Let m denote the number of factors. Let x_j^i denote the standardization of the ith observed value of the jth indicator. Let ε_i denote random error. The Logistic multiple linear regression function between data status y_i and factors x_j^i is as follows [14]:

$$\log it\left(y_i\right) = a + c_1 x_1^i + c_2 x_2^i + \cdots + c_m x_m^i. \tag{8}$$

The function of (8): it is obtained by the bilateral probability P_j of regression coefficient c_j for every factor x_j^i by constructing the Logistic regression function between evaluation factors x_j^i and data state y_i of farmers. Comparing the bilateral probability P_j with the given critical probability P_0, it can distinguish whether the factors x_j^i have an obvious effect on data state y_i. Deleting these factors that do not have an obvious effect on data status, it ensures that the reserved factors can effectively distinguish bad samples from good ones.

(2) The Standard of Logistic Regression Significant Discriminant Screening. As a matter of experience, the threshold probability P_0 equals 0.05 [14].

If $P_j \geq P_0 = 0.05$ [14], accept the assumption that the true value of regression coefficient c_j corresponds to the factor

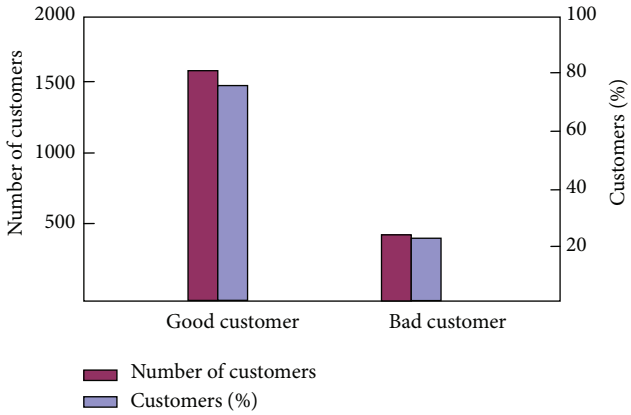

FIGURE 2: The distribution of customers.

FIGURE 3: The classification demo figure of good customers and bad customers.

x_j^i. It indicates that probability of the true value of regression coefficient c_j being zero is at least 95%. In other words, the factor x_j^i cannot significantly distinguish the data status and it should be deleted.

Conversely, if $P_j < P_0 = 0.05$ [14], refuse the assumption that the true value of regression coefficient c_j corresponds to the factor x_j^i. It indicates that probability of the true value of regression coefficient c_j not being zero is at least 95%. In other words, the factor x_j^i can significantly distinguish the data status and it should be reserved.

4. Empirical Study

4.1. Samples and Data Source

(1) Samples. In order to verify the effectiveness of the proposed model, this paper recognizes the key factors which can effectively distinguish the good customers from the bad ones by using a Chinese state-owned commercial bank's 2044 petty loans for farmers [15]. The sample includes 1589 nondefault customers (i.e., good customers) and 455 default customers (i.e., bad customers). The distribution of customers is shown in Figure 2.

(2) The Establishment of Extensive Factors Set. According to the available factors from a Chinese national commercial bank [15], this paper selects 68 factors of petty loans for farmers, which includes five feature layers, that is, "basic information," "repayment ability," "repayment willing," "guarantee and joint guarantee," and "macro environment," as shown in Column 1, 2, 5 of Table 1.

At the beginning of screening factors, we removed 18 unavailable factors, such as "credit status of joint guarantor" and "technical support efforts." Other 50 factors are left. The deleted factors are marked with "unavailability delete" in Column 7 of Table 1.

(3) Data Source. The data in the first to 50th Row and the first to 2044th Column of Table 2 are from the farmers petty credit loan system of a Chinese national commercial bank headquarter [15]. Since there exists no missing data, we used all data without any adjustment. The default status of each customer is shown in the corresponding Column, the 51st Row of Table 2. The number "1" denotes default customers, and the number "0" denotes nondefault customers.

Next, the key factors which can distinguish the two types of customers effectively will be selected. The classification demo result of good customers and bad customers is shown in Figure 3.

4.2. The Establishment of Recognizing Key Factors Model

4.2.1. The Standardization of Factors Data

(1) Scoring the Quantify Factors. It should be pointed out that there are two interval factors in this credit system, that is, "Consumer price index" and "Age". The ideal interval of "Consumer price index" is [101, 105] [12]. Inflation or deflation is nonexistent within this interval. The ideal interval of "Age" is [31, 45] [12]. The repayment ability and repayment willingness of these customers are strong in the interval.

According to the factors type in Column d of Table 2, take the original data of positive factors v_{ij} from Column 1 to 2044 of Table 2 into (1), the original data of negative factors v_{ij} into (2), and the original data of interval factors v_{ij} into (3), and then the standardized data of factors x_j^i are obtained. The results are shown in Column 2045 to 4088 of Table 2.

(2) Scoring the Qualitative Factors. The scoring standard of qualitative factors can be obtained by rational analysis, as shown in Column 2 to 6 of Table 3.

According to the factor type in Column d of Table 2, standardized scores of qualitative factors can be obtained in Table 2 based on the scoring criteria of qualitative factors in Table 3. The results are shown in Column 2045 to 4088 of Table 2.

Table 1: Extensive factors set.

(1) Feature layers	(2) Factors	(3) References	(4) Screening result	(5) Factors	(6) References	(7) Screening result
Basic information	Loan purpose	[9, 16, 17]	Deleted by collinearity diagnostics	Supporting population	[16, 18]	Deleted by collinearity diagnostics
	Age	[3, 9, 16–25]	Reserved	Family number/labor force	[16]	Deleted by significance
	Value of house owing	[9, 16, 18, 24, 25]		Number of members	[18, 24]	
	Marital status	[3, 9, 16, 17]		Number of labor force	[18, 22]	
	Education background	[3, 9, 16, 17]	Deleted by significance	...	...	...
	Household expenses	[9, 17, 18]		Area ratio of disposable assets	[9, 16]	Unavailability delete
Repayment ability	Expenses/incomes	[16, 19, 20, 22, 23]	Reserved	Agricultural production incomes	[16, 18, 24]	Deleted by collinearity diagnostics
	Nonagricultural incomes/total incomes	[16, 18, 24]		Net agricultural incomes	[16, 18, 24]	Deleted by significance
	Net income of borrower	[16, 18, 24]		Total expenses	[9, 25]	
	Education cost of children each year	[16, 18, 24]	Deleted by significance	...	...	...
	Agricultural production expenses	[9, 16, 17]		Total property	[16]	Unavailability delete
Repayment willingness	Private loans	[9, 16, 17]	Deleted by significance	Loaning records of borrower	[9, 16–18, 22, 24, 25]	Deleted by significance
	Residential stability	[3, 16, 17]	Reserved	...	...	...
	Residential status	[3, 9]		Social reputation status	[16, 17]	Unavailability delete
Guarantee and joint guarantee	Strength of guarantor	[9, 16, 18]	Reserved	Age of guarantor	[9, 16, 18, 25]	Deleted by significance
	Marital status of guarantor	[9, 16–18, 25]		...	...	...
	Gender of guarantor	[9, 16, 24]	Deleted by significance	Credit status of joint guarantor	[16]	Unavailability delete
Macro environment	Engel's coefficient	[9, 16, 26]	Reserved	Regional government policy	[9, 18, 25]	Deleted by collinearity diagnostics
	Increasing rate of regional GDP	[9, 16, 26]		...	...	...
	Per capita agricultural output value	[17, 23]		Technical support efforts	[9, 18, 24]	Unavailability delete

4.2.2. Deleting the Repeated Information Factors. By substituting the standardized data of factors x_j^i in the 2045th and the 4088th Column of Table 2 into (4)–(6), the determination coefficient R_j^2 of all factors is obtained, as shown in the fourth Column of Table 4. By taking the determination coefficient R_j^2 in the fourth Column of Table 4 into (7), the variance inflation factor VIF_i of all factors is obtained, as shown in the fifth Column of Table 4.

According to the standard of collinearity diagnostics screening shown in Section 3.2. (2), if the variance inflation factor VIF of an indicator is greater than 10, the indicator or factor should be deleted. In this progress, eight factors which reflect repeated information are deleted and 42 factors are reserved. The deleted factors include "Loan purpose," "House value," and "Regional government policy," as shown in the sixth Column of Table 4 marked as "Deleted."

4.2.3. Screening the Key Factors

(1) The Establishment of Logistic Regression Model. By substituting the ultimate 42 factors reserved in Table 4 into (8), the

Table 2: The original data and standardized data of factors.

(a) No.	(b) Feature layers	(c) Factors	(d) Factor type	Original data of factors v_{ij}			Standardized data of factors x_j^i					
							Nondefault customers			Default customers		
				(1) M. Song	⋯	(2044) M. Xu	(2045) M. Song	⋯	(3920) D. Liu	(3921) F. Chen	⋯	(4088) M. Xu
1	Basic information	Loan purpose	Qualitative	3	⋯	3	0.600	⋯	0.600	0.600	⋯	0.600
⋯		⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
14		House value	Positive	0.130	⋯	0.000	0.39	⋯	0.33	0.31	⋯	0.29
15	Repayment ability	Expenses/incomes	Positive	18.989	⋯	8.124	1.000	⋯	0.887	0.000	⋯	0.123
⋯		⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
31		Expense of family's daily life	Negative	120	⋯	8	0.737	⋯	0.534	1.000	⋯	0.001
32	Repayment willingness	Private loans	Qualitative	1.12	⋯	0.51	0.014	⋯	0.025	0.019	⋯	0.006
⋯		⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
39		Repayment to net income ratio	Negative	35.28	⋯	64.84	0.688	⋯	1.000	0.991	⋯	0.426
40	Guarantee and joint guarantee	Strength of guarantor	Positive	0.066	⋯	0.009	0.041	⋯	0.157	0.347	⋯	0.005
⋯		⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
45		Age of guarantor	Interval	45	⋯	36	1.000	⋯	0.059	0.094	⋯	1.000
46	Macro environment	Engel's coefficient	Negative	0.373	⋯	0.399	0.892	⋯	0.135	0.246	⋯	0.773
⋯		⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
50		Regional government policy	Qualitative	1	⋯	3	1.000	⋯	0.600	0.000	⋯	0.000
51	—	Default or not y_i	—	0	⋯	1	0	⋯	0	1	⋯	1

Table 3: The scoring criteria of qualitative factors.

(1) No.	(2) Feature layers	(3) Factors	(4) Options number	(5) Options	(6) Scoring
1	Basic information	Education background	1	Undergraduate and above	1.00
2			2	Junior college	0.90
3			3	High school and technical secondary school	0.60
4			4	Junior high school	0.40
5			5	Primary school	0.20
6			6	Other	0.00
...	...	...	...	...	...
70	Guarantee and joint guarantee	Group membership of coguarantee	1	Friendly relations, associating frequently, very familiar, business partners or neighbors	1.00
71			2	Ordinary relations, a little familiar	0.80
72			3	Unknown	0.50

Table 4: Collinearity diagnostics of factors.

(1) No.	(2) Feature layers	(3) Factors	(4) Determination coefficient R_j^2	(5) Variance inflation factor (VIF_i)	(6) Screening result of collinearity diagnostics
1	Basic information	Loan purpose	0.942	17.241	Deleted
2		Age	0.091	1.100	Reserved
...		...	...	...	...
14		House value	0.928	13.889	Deleted
15	Repayment ability	Expenses/incomes	0.731	3.717	Reserved
...		...	...	...	...
31		Expense of family's daily life	0.199	1.248	Reserved
32	Repayment willingness	Private loans	0.063	1.067	Reserved
...		...	...	...	...
39		Repayment to net income ratio	0.497	1.988	Reserved
40	Guarantee and joint guarantee	Strength of guarantor	0.456	1.838	Reserved
...		...	...	...	...
45		Age of guarantor	0.320	1.471	Reserved
46	Macro environment	Engel's coefficient	0.265	1.361	Reserved
...		...	...	...	...
50		Regional government policy	0.968	31.250	Deleted

Logistic regression model between data status (i.e., default status) y_i and factors x_j^i is obtained as follows:

$$\log it\left(y_i\right) = a + c_1 x_1^i + c_2 x_2^i + \cdots + c_{42} x_{42}^i. \quad (9)$$

The parameter i equals $1, 2, \ldots, 2044$, respectively, in (9). By taking the standardized data of factors x_j^i in Column 2045 to 4088 of Table 2 into (9), the regression coefficients c_j of 42 factors and the corresponding bilateral probability P_j are obtained, as shown in Column four and five of Table 5.

(2) Recognizing the Key Factors. Based on the key factors screening standard shown in Section 3.3. (2), if $P_j \geq P_0 = 0.05$, the factor x_j^i cannot significantly distinguish the default status and should be deleted. On the contrary, the factor x_j^i should be retained. And the threshold probability P_0 equals 0.05 in this paper, as shown in Column 6 of Table 5.

Comparing the critical probability $P_0 = 0.05$ with data in the first Row and the fifth Column of Table 5, the bilateral probability P_1 corresponding to the regression coefficient c_1 of the first factor "Age" is less than the critical probability 0.05; that is, $P_1 = 0.007 < P_0 = 0.05$. It indicates that the factor "Age" can significantly distinguish the default status and should be reserved. The result is marked with "Reserved" in the first Row and the seventh Column of Table 5.

Similarly, comparing the critical probability 0.05 with the other data in the fifth Column of Table 5, the screening results were listed in the corresponding row in the seventh Column of Table 5. The Logistic regression significant discriminant screening deleted 29 factors, such as "Private loans" and "Repayment to net income ratio."

Table 5: Screening the key factors based on Logistic regression significant discriminant.

(1) No.	(2) Feature layers	(3) Factors	(4) Regression coefficients c_j	(5) Bilateral probability P_j	(6) Critical probability P_0	(7) Screening result
1		Age	−1.827	0.007		Reserved
⋯	Basic information	⋯	⋯	⋯	0.05	⋯
11		Number of labor force	2.017	0.405		Deleted
12		Expenses/incomes	0.079	0.044		Reserved
⋯	Repayment ability	⋯	⋯	⋯	0.05	⋯
26		Expense of family's daily life	−0.328	0.834		Deleted
27		Private loans	6.220	0.104		Deleted
⋯	Repayment willingness	⋯	⋯	⋯	0.05	⋯
33		Repayment to net income ratio	2.962	0.762		Deleted
34	Guarantee and joint guarantee	Strength of guarantor	−1.973	0.000		Reserved
⋯		⋯	⋯	⋯	0.05	⋯
38		Age of guarantor	0.264	0.772		Deleted
39		Engel's coefficient	−0.152	0.018		Reserved
⋯	Macro environment	⋯	⋯	⋯	0.05	⋯
42		CPI	3.211	0.336		Deleted

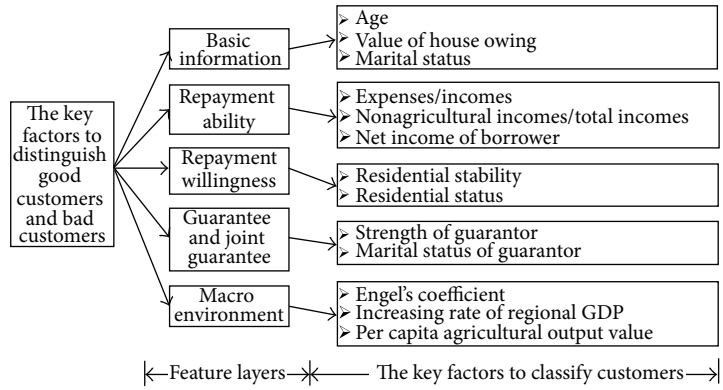

Figure 4: The key factors to distinguish good customers and bad customers.

In conclusion, this paper extracts thirteen factors which can effectively distinguish good customers from bad ones, as shown in Figure 4.

5. Conclusion

With the advent of the era of big data, data classification puzzles have emerged in DNA identification, fingerprint recognition, customer classification, facial recognition, and so forth. Recently, recognizing key factor methods and classifier models have been proposed for solving this problem. So it has become more and more important for researchers to find key factors which are capable of effectively distinguishing the data. To do that, many mathematical models are explored as the decision support methods to classify the data.

We propose a model for recognition key factors, which is based on the combination of collinearity diagnostics and logistic regression significant discriminant. To demonstrate the performance of the proposed model, factors screening tasks were performed by an empirical study of the 2044 observations in finical engineering. Our empirical results show that the proposed model can accurately screen the key factors, which can effectively distinguish the good customers and bad customers. Moreover, the proposed method is simple and easy to be implemented.

The main contribution of this study is as follows: deleting the factors that reflect repeated information by using collinearity diagnostics and recognizing the factors which can effectively distinguish the two kinds of samples by using Logistic regression significant discriminant; this paper established a recognition key factors model for data classification.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of the paper.

Acknowledgments

The research is supported by the National Natural Science Foundation of China (Grant no. 71171031), the Ministry of Education of China as Science and Technology Research Project (Grant no. 2011-10), the China Banking Regulatory Commission as a Risk Management Project of Banking Information Technology (Grant no. 2012-4-005), and the Bank of Dalian as credit rating and loan pricing systems for small business (Grant no. 2012-01). The authors thank the organizations mentioned above.

References

[1] J. Hu, L. Wang, F. Duan, and P. Guo, "Adaptive multilevel kernel machine for scene classification," *Mathematical Problems in Engineering*, vol. 2013, Article ID 324945, 9 pages, 2013.

[2] S. Finlay, "Multiple classifier architectures and their application to credit risk assessment," *European Journal of Operational Research*, vol. 210, no. 2, pp. 368–378, 2011.

[3] S. Akkoc, "An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: the case of Turkish credit card data," *European Journal of Operational Research*, vol. 222, no. 1, pp. 168–178, 2012.

[4] C. Lin, J. Chen, Z. Gaing, and M. Younis, "Combining biometric fractal pattern and particle swarm optimization-based classifier for fingerprint recognition," *Mathematical Problems in Engineering*, vol. 2010, Article ID 328676, 14 pages, 2010.

[5] B. Twala, "Multiple classifier application to credit risk assessment," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3326–3336, 2010.

[6] Y. Chen, "Research on evaluation and decision system of small amount loans for farmers based on support vector machines," *Dalian Univisity of Technology*, pp. 22–58, 2011.

[7] C.-D. Căleanu, X. Mao, G. Pradel, S. Moga, and Y. Xue, "Combined pattern search optimization of feature extraction and classification parameters in facial recognition," *Pattern Recognition Letters*, vol. 32, no. 9, pp. 1250–1255, 2011.

[8] J. H. Min and Y.-C. Lee, "A practical approach to credit scoring," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1762–1770, 2008.

[9] B. Shi and G. Chi, "A credit risk evaluation index screening model of petty loans for small private business and its application," *Advances in Information Sciences and Service Sciences*, vol. 5, no. 7, pp. 1116–1124, 2013.

[10] R.-C. Hwang, H. Chung, and C. K. Chu, "Predicting issuer credit ratings using a semiparametric method," *Journal of Empirical Finance*, vol. 17, no. 1, pp. 120–137, 2010.

[11] X. Sun, W.-H. Qiu, and B.-J. Tang, "Research about model of early-warning of enterprise crisis based on gray forecasting and pattern recognition," *System Engineering Theory and Practice*, vol. 25, no. 5, pp. 36–42, 2005.

[12] N. Yan, Z. Chen, Y. Shi, Z. Wang, and G. Huang, "Using non-additive measure for optimization-based nonlinear classification," *American Journal of Operations Research*, vol. 2, no. 3, pp. 364–373, 2012.

[13] J. Corander, M. Gyllenberg, and T. Koski, "Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy," *Advances in Data Analysis and Classification*, vol. 3, no. 1, pp. 3–24, 2009.

[14] A. Rencher and G. Schaalje, *Linear Models in Statistics*, John Wiley & Sons, Hoboken, NJ, USA, 2008.

[15] Chi Guotai Research Group of Dalian University of Technology, "Credit risks decision and evaluation system of microcredit for farmers for postal savings bank of China," *Dalian University of Technology*, 2011.

[16] Postal Savings Bank of China, "Merchant credit rating table of Postal Savings Bank of China," *Postal Savings Bank of China*, 2009.

[17] Construction Bank of China, "Small business customers evaluation approaches of China Construction Bank," *Construction Bank of China*, pp. 1–8, 2007.

[18] Emei credit cooperatives, *The operating rules and regulationsof micro-credit loans to farmers for Rural Credit Cooperatives of Sichuan Province*, 2013, http://www.emxh.com/article/2007/0803/file_142.html.

[19] Standard and Poor's Ratings Services, "S&P's study of China's top corporates highlights their significant financial risks," *Standard & Poor's Ratings Services*, pp. 175–199, 2012.

[20] Moody's Investors Service, "Global credit research," *Moody's Investors Service*, pp. 136–147, 2009.

[21] Fitch Ratings, "Fitch ratings global corporate finance 2012 transition and default study," *Credit Market Research—Fitch Ratings*, pp. 2–27, 2013.

[22] Fair Isaac Corporation, "Free FICO Credit Score," Fair Isaac Corporation, 2013, http://www.myfico.com/Default.aspx.

[23] T.-C. Wang and Y.-H. Chen, "Applying rough sets theory to corporate credit ratings," in *Proceedings of the IEEE International Conference on Service Operations and Logistics, and Informatics*, pp. 132–136, June 2006.

[24] Agricultural Bank of China, "Management of farmers creditrating of Agricultural Bank of China," *Agricultural Bank of China*, 2008.

[25] Industrial and Commercial Bank of China, "The notice about printing the evaluation method of the small business corporate clients credit rating of China Industrial and Commercial Bank means," *Industrial and Commercial Bank of China*, no. 78, 2005.

[26] K. Carling, T. Jacobson, J. Lindé, and K. Roszbach, "Corporate credit risk modeling and the macroeconomy," *Journal of Banking and Finance*, vol. 31, no. 3, pp. 845–868, 2007.