Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 1 (2012) 2463–2468

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Science, ICCS 2010

Credit scorecard based on logistic regression with random coefficients

Gang Dong[a], Kin Keung Lai[b]*, Jerome Yen[c]

[a]*Department of Management Sciences, City University of Hong Kong, Hong Kong*
[b]*School of Business Administration, North China ElectricPower University, China*
[c]*Department of Finance and Economics, Tung Wah College, Hong Kong*

Abstract

Many credit scoring techniques have been used to build credit scorecards. Among them, logistic regression model is the most commonly used in the banking industry due to its desirable features (e.g., robustness and transparency). Although some new techniques (e.g., support vector machine) have been applied to credit scoring and shown superior prediction accuracy, they have problems with the results interpretability. Therefore, these advanced techniques have not been widely applied in practice. To improve the prediction accuracy of logistic regression, logistic regression with random coefficients is proposed. The proposed model can improve prediction accuracy of logistic regression without sacrificing desirable features. It is expected that the proposed credit scorecard building method can contribute to effective management of credit risk in practice.

Keywords: Credit scorecard; Logistic regression; Random coefficients; Bayesian procedures;

1. Introduction

Credit scorecard systems are widely used in banking industry nowadays, especially after Basel Accord II was implemented in 2007. Scores earned by applicants for new loans or existing borrowers seeking new loans are used to evaluate their credit status. Credit scores are awarded on the basis of different techniques designed by individual lenders. However, irrespective of the varying nature of techniques used, credit scoring is invariably used to answer one key question - what is the probability of default within a fixed period, usually 12 months. Credit scoring can be divided into application scoring and behavior scoring, based on the information used when modeling. Application scoring uses only the information provided in application, while behavior scoring uses both the application information, and (past) behavior information.

Basically, there are three kinds of methods that have been studied in credit scoring; classification techniques, Markov Chain and Survival analysis. Among them, classification technique is the one that has been studied most

* Corresponding author. Tel.: +852-2788-8563.
E-mail address: mskklai@cityu.edu.hk.

1877-0509
doi:10.1016/j.procs.2010.04.278

extensively. A large number of classification techniques for credit scoring can be found in literature. These techniques can be roughly categorized into five groups: (1) statistical models; (2) operational research methods; (3) artificial intelligence techniques; (4) hybrid approaches; and (5) ensemble models. Statistical models mainly comprise logistic regression techniques [1], linear discriminant analysis [2], *k*-nearest neighbor [3] and classification tree [4]. Operational research methods include linear programming [5] and quadratic programming [6]. Artificial intelligence techniques include neural networks [7], support vector machine [8], genetic algorithm [9] and genetic programming [10]. Hybrid approaches primarily include fuzzy systems and neural networks [11], fuzzy systems and support vector machines [12] and neural networks and multivariate adaptive regression splines [13]. In case of ensemble models, the neural network ensemble is a typical example. Interested readers can refer to [14] for more details. Application of Markov Chain and survival analysis on credit scoring can be found in [15] and [16] respectively.

Although these techniques have been tested and compared in the context of credit scoring, many of them have not been widely used in developing operational credit scorecards. The reasons are twofold – robustness and transparency. Some methods like neural networks and support vector machines may lead to slightly better classifiers on a set of data, but the interactions they use make them more vulnerable as the population characteristics change. More importantly, regulators now require that banks give reasons for rejecting an applicant for credit [17]. Support vector machine and neural networks are always described as "black box" because they do not require any information about the functional relationships between variables. Their results can not be easily interpreted and banks can not give rejecting reasons according to the results of these methods.

Other techniques mentioned above suffer one or both of these two problems. In banking industry, logistic regression, linear regression, linear programming and classification tree have been used to develop credit scorecard systems. Among them, logistic regression is the most commonly used one due to its distinctive features which can be found in [17]. Although logistic regression does not involve the above two issues, its prediction ability is inferior to some methods like neural networks and support vector machines. Therefore, the logistic regression model with random coefficients is proposed to improve the prediction accuracy of logistic regression without sacrificing its desirable features. The main contribution of this paper is to propose a new model to improve the prediction accuracy of logistic regression without sacrificing its desirable features rather than comparisons of prediction accuracy of various methods.

2. Logistic Regression with Random Coefficients

A logistic regression model with random coefficients is applied, where the coefficients follow multivariate normal distribution. In the model, the probability of individual being "good" is expressed as follows:

$$P_{n,G} = \frac{\exp(\sum_{k=1}^{K} \beta_{n,k} x_{n,k})}{1+\exp(\sum_{k=1}^{K} \beta_{n,k} x_{n,k})} \tag{1}$$

where

$\beta_{n,k}$ = the coefficient of the k th attribute of individual n

$x_{n,k}$ = the value of the k th attribute of individual n

Under a random coefficients specification, the parameters are assumed to be randomly distributed across individuals, that is, for individual n, the vector of parameters $\theta_n = \{\beta_{n,1}, \beta_{n,2}, ...\beta_{n,K}\}$ follows a multivariate normal distribution $N(\mu, \Omega)$, where μ is the mean and Ω is the covariance matrix.

3. Parameter Estimation

The parameters to be estimated include μ, Ω and $\theta_n \ \forall n$. Bayesian procedures are used to estimate these parameters. The posterior distribution of μ, Ω and $\theta_n \ \forall n$, by definition, can be written as:

$$K(\mu, \Omega, \theta_n \ \forall n \mid \ ) = \frac{1}{M} \prod_{n}^{N} L(y_n \mid \theta_n) \phi(\theta_n \mid \mu, \Omega) k(\mu, \Omega) \tag{2}$$

where $Y = \{y_1, y_2 ... y_N\}$ and y_n denotes whether individual n is good, i.e. y_n equals 1 if individual n is "good", otherwise y_n equals 0. M is the normalizing constant, which is difficult to calculate, since it involves integration. However, the parameters can be estimated without knowing or calculating M of the posterior distribution. $L(y_n \mid \theta_n)$ can be expressed as:

$$L(y_n \mid \theta_n) = \left(\frac{\exp(\sum_{k=1}^{K} \beta_{n,k} x_{n,k})}{1+\exp(\sum_{k=1}^{K} \beta_{n,k} x_{n,k})}\right)^{y_n} * \left(\frac{1}{1+\exp(\sum_{k=1}^{K} \beta_{n,k} x_{n,k})}\right)^{1-y_n} \quad (3)$$

$k(\mu, \Omega)$ is the prior on μ and Ω.

Draws from above posterior are obtained through Gibbs sampling. A draw of each parameter is taken, conditional on other parameters:

(1) Take a draw of μ conditional on values of Ω and θ_n $\forall n$. Given a diffuse prior on the posterior of μ, the posterior follows $N(\bar{\theta}, \Omega / N)$.

(2) Take a draw of Ω conditional on values of μ and θ_n $\forall n$. Given a diffuse prior on posterior of Ω, the posterior is Inverted Wishart with $K+N$ degrees of freedom and scale matrix $(KI + NS_1)/(K+N)$, where

$$S_1 = (1/N)\sum_{n=1}^{N}(\theta_n - \mu)(\theta_n - \mu)' \quad (4)$$

The posterior for each individual's θ_n, conditional on their results and the population mean and variance, is difficult to draw from, and the Metropolis-Hasiting (MH) algorithm is used.

Gibbs sampling starts with any initial values μ^0, Ω^0 and θ_n^0 $\forall n$. The t th iteration of the Gibbs sampler consists of these steps:

Step1. Draw μ^t from $N(\bar{\theta}^{t-1}, \Omega^{t-1}/N)$, where $\bar{\theta}^{t-1}$ is the mean of the $\bar{\theta}_n^{t-1}$'s.

Step2. Draw Ω^t from $IW(K+N, (KI + NS^{t-1})/(K+N))$, where

$$S^{t-1} = (1/N)\sum_{n=1}^{N}(\theta_n^{t-1} - \mu^t)(\theta_n^{t-1} - \mu^t)' \quad (5)$$

Step3. For each n, draw θ_n^t using one iteration of MH algorithm, starting from θ_n^{t-1} and using the normal density $\phi(\theta_n \mid \mu^t, \Omega^t)$.

These three steps are repeated for many iterations. The resulting values converge to draws from the joint posterior distribution of μ, Ω and θ_n $\forall n$.

4. Empirical Experiment

4.1. Data Analysis

To evaluate the performance of our proposed algorithm, a German Credit Data Set from University of California at Irvine (UCI) Machine Learning Repository is applied. The dataset can be found at http://archive.ics.uci.edu/ml/datasets.html. The dataset includes 20 characteristics and classification results. Among them, 7 characteristics are numerical and 13 characteristics are qualitative. The dataset includes 1000 samples, where 700 samples are "good" and 300 samples are "bad".

In the dataset, there are 3 continuous characteristics including "Duration", "Credit Amount" and "Age". Each of these 3 continuous characteristics is classified into several bins. The detailed classification is shown in Table 1.

The dataset is randomly divided into 10 subsets of the same size. Each subset includes 100 samples, where 70 samples are "good" and 30 samples are "bad". Each time, 9 subsets are used to build a logistic regression model with fixed coefficients using SPSS. Hence, 10 logistic regression models with fixed coefficients are obtained. For each model, a set of characteristics with significant coefficients is constructed. The shared characteristics of these 10 sets are used as the characteristics for our proposed model. There are five characteristics shared by the 10 sets, including "Status of existing checking account (CA)", "Duration in months (Duration)", "Credit history (CH)", "Savings account/bonds (Savings)" and "Installment rate in percentage of disposable income (IR)". Furthermore, these five characteristics are recoded into 22 dummy variables.

Table 1: Classification of Continuous Characteristics

Characteristics	Bins	Values
Duration in Month (Dr)	Dr < 1 year	1
	1 year $\leq$ Dr < 2 years	2
	2 years $\leq$ Dr < 3 years	3
	3 years $\leq$ Dr	4
Credit Amount (CA)	CA < 2000	1
	2000 $\leq$ CA < 5000	2
	5000 $\leq$ CA < 10000	3
	10000 $\leq$ CA	4
Age	Age < 28	1
	28 $\leq$ Age < 41	2
	41 $\leq$ Age < 51	3
	51 $\leq$ Age	4

4.2. Experiment Settings and Results

We randomly selected 9 subsets as the training set and 1 set as testing set. Based on the selected dataset, the logistic regression model with fixed coefficients (called LRF) is trained. The coefficients of LRF are estimated using SPSS, and the results are shown in Table 2.

For the logistic regression model with random coefficients (called LRR), coefficients are estimated under the initial assumption that they are independently normally distributed in the population. That is, $\theta_n \sim N(\mu, \Omega)$ with diagonal Ω. For the Bayesian procedure, 10000 iterations of the Gibbs sampling are performed. For Gibbs sampling, it starts with the estimated coefficients of LRF. The coefficients of 900 samples are randomly generated by the normal distribution with mean $\overline{\theta}$ and covariance matrix W, where $\overline{\theta}$ is the estimated coefficients of LRF and W is the matrix with diagonal elements equalling 0.25. Since the Gibbs sampling starts with $\overline{\theta}$, it converges very soon. Therefore, the first 500 iterations are used as the burn-in process and the last 9500 iterations are used for the estimates. The estimates of parameters μ and Ω are shown in Table 3.

Table 2: Coefficients estimates of LRF

Coefficients	Estimates	Coefficients	Estimates
CA1	-1.387	CH4	0.094
CA2	-0.967	CH5	0
CA3	-0.413	Savings1	-0.036
CA4	0	Saving2	0.007
Duration1	2.234	Saving3	0.625
Duration2	1.58	Saving4	0.763
Duration3	1.294	Savings5	0
Duration4	0	IR1	0.751
CH1	-1.413	IR2	0.714
CH2	-1.146	IR3	0.530
CH3	-0.183	IR4	0

Table 3: Coefficients estimates of LRR

Coefficients		Bayesian Estimates	Coefficients		Bayesian Estimates
CA1	Mean	-1.449804 (0.03597475)	CH4	Mean	0.1411739 (0.04077080)
	St.dev	0.03277543 (0.001573993)		St.dev	0.03281068 (0.001510439)
CA2	Mean	-0.9059897 (0.05470632)	CH5	Mean	0.2220893 (0.08302383)

	St.dev	0.03269786 (0.001534337)		St.dev	0.03279539 (0.001539268)
CA3	Mean	-0.5661041 (0.07748404)	Savings1	Mean	-0.04378045 (0.0370465)
	St.dev	0.03273943 (0.001502446)		St.dev	0.03271315 (0.001520649)
CA4	Mean	0.1390510 (0.05313521)	Savings2	Mean	-0.06965887 (0.04913048)
	St.dev	0.03273189 (0.001495568)		St.dev	0.03282662 (0.001544255)
Duration1	Mean	2.257448 (0.03561973)	Savings3	Mean	0.6833126 (0.03866932)
	St.dev	0.03282765 (0.001533944)		St.dev	0.03277337 (0.001534817)
Duration2	Mean	1.612257 (0.04343867)	Savings4	Mean	0.8225605 (0.03704827)
	St.dev	0.03276230 (0.001556254)		St.dev	0.03284451 (0.001522668)
Duration3	Mean	1.241366 (0.06809208)	Savings5	Mean	-0.04087436 (0.07015044)
	St.dev	0.03280230 (0.001535490)		St.dev	0.03280604 (0.001556045)
Duration4	Mean	-0.03959597 (0.0521015)	IR1	Mean	0.632054 (0.08061883)
	St.dev	0.03277398 (0.001540658)		St.dev	0.03280673 (0.001553577)
CH1	Mean	-1.453215 (0.02883353)	IR2	Mean	0.7681777 (0.03660409)
	St.dev	0.03278287 (0.001550613)		St.dev	0.03278261 (0.001546148)
CH2	Mean	-1.070917 (0.04255506)	IR3	Mean	0.520004 (0.0426243)
	St.dev	0.03274843 (0.001541318)		St.dev	0.03275891 (0.001523483)
CH3	Mean	-0.1460130 (0.02718792)	IR4	Mean	0.1264567 (0.06775553)
	St.dev	0.03284983 (0.001519736)		St.dev	0.03289945 (0.001564571)

4.3. Comparisons of Prediction Accuracy

There are many criteria that can measure the quality of credit scorecards. Among them, prediction accuracy is the most important one. In this paper, Percentage Correctly Classified (PCC) is used as the criterion of measuring prediction accuracy. PCC represents the percentage of observations that are correctly classified. Prediction accuracy results of LRF and LRR are shown in Tables 4 and 5 respectively.

Table 4: Prediction accuracy of LRF

Observed	Predicted		
	0	1	PCC
0	13	17	43.33
1	12	58	82.86
Overall Percentage			71

Table 5: Prediction accuracy of LRR

Observed	Predicted		
	0	1	PCC
0	13	17	43.33
1	9	61	87.14
Overall Percentage			74

5. Conclusions

In this paper we propose a logistic regression model with random coefficients for building credit scorecards. The empirical results indicate the proposed model can improve prediction accuracy of the logistic regression with fixed coefficients without sacrificing its desirable features. However, the proposed model needs much more time to estimate parameters.

Acknowledgements

The work described in this paper is supported by CityU strategic research grant (project no. 7008023).

References

[1] J.C. Wiginton, A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis 15 (1980) 757-770.
[2] E. Rosenberg and A. Gleit, Quantitative Methods in Credit Management: A Survery. Operations research 42 (4) (1994) 589-613.
[3] S. Chatterjee and S. Barcun, A Nonparametric Approach to Credit Screening. Journal of the American Statistical Association 65 (329) (1970) 150-154.
[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees, Wadsworth International Group, 1984.
[5] O. L. Mangasarian, Linear and Nonlinear Separation of Patterns by Linear Programming, Operations research 13 (1965) 444-452.
[6] B. Vladimir, K. Hiroshi, U. Stanislav, Credit Cards Scoring with Quadratic Utility Functions, Journal of Multicriteria Decision Analysis 11 (4-5) (2002) 197.
[7] H. L. Jensen, Using Neural Networks for Credit Scoring. Managerial Finance 18 (6) (1992) 15.
[8] I. T. V. Gestel, B. Baesens, I. J. Garcia, P. V. A. Dijcke, Support Vector Machine Approach to Credit Scoring. Bank- en Financiewezen 2 (2003) 73-82.
[9] V. S. Desai, D. G. Conway, J. N. Crook, G. A. J. Overstreet, Credit-Scoring Models in the Credit-Union Environment Using Neural Networks and Genetic Algorithms. IMA Journal of Management Mathematics 8 (4) (1997) 323-346.
[10] J. J. Huang, G. H. Tzeng, C. S. Ong, Two-Stage Genetic Programming (2sgp) for the Credit Scoring Model. Applied Mathematics and Computation 174 (2) (2006) 1039-1053.
[11] R. Malhotra, D. K. Malhotra, Differentiating between Good Credits and Bad Credits Using Neuro-Fuzzy Systems. European Journal of Operational Research 136 (1) (2002) 190-211.
[12] Y. Wang, S. Wang, K. K. Lai, A New Fuzzy Support Vector Machine to Evaluate Credit Risk. Fuzzy Systems, IEEE Transactions on Fuzzy Systems 13 (6) (2005) 820-831.
[13] T. S. Lee, I. F. Chen, A Two-Stage Hybrid Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines. Expert Systems with Applications 28 (4) (2005) 743-752.
[14] L. Yu, S. Y. Wang, K. K. Lai, L. G. Zhou, Bio-Inspired Credit Risk Analysis-Computational Intelligence With Support Vector Machines, Springer Berlin Heidelberg, Berlin Heidelberg, 2008.
[15] H. Frydman, J. G. Kallberg, D. L. Kao, Testing the adequacy of Markov chains and Mover–Stayer models as representations of credit behaviour. Operations Research 33 (1985) 1203-1214.
[16] M. Stepanova and L. Thomas, Survival Analysis Methods for Personal Loan Data. Operations Research 50 (2002) 277-289.
[17] L. Thomas, Consumer Credit Models – Pricing, Profit and Portfolios, OXFORD University Press, UK, 2009.