Credit Risk Model Based on Logistic Regression and Weight of Evidence

Xiang YANG[1, a, *], YongbinZHU[1, b],Li YAN[1,c], Xin WANG[1, d]

[1]School of Engineering Technology, Honghe University, Mengzi, 661100, China

[a]email:806464637@qq.com, [b]email:410239127@qq.com, [c]email: 904382815@qq.com, [d]email: 2535563360@qq.com, [*]corresponding author

Keywords: Credit Risk, Logistic Regression, Weight of Evidence, Scorecard

Abstract. Many techniques have been used to build credit risk model. Among them, logistic regression is a more appropriate technique due to its desirable features (e.g., interpretability and prediction accuracy). In this paper, toimplement credit risk assessment quickly, a method for constructing credit risk model (in the form of a scorecard) based on logistic and weight of evidence is proposed.

Introduction

Effective management of credit risk is important for a loans institution. A loans institution always willing to lend to creditworthy customers (called good customers), and reject the loan application of customers with poor credit (called bad customers). To distinguish good customers and bad customers using application information, many techniques have been used to build credit risk model: linear discriminant analysis [1], artificial neural network [1] [3], *k*-nearest neighbor [2], classification tree [3], logistic regression [3] [4] and so on.

Although these techniques can provide good discrimination, logistic regression is a more appropriate technique to build credit risk model due to its distinctive features (interpretability, prediction accuracy and so on [5]). Especially, regulators require that loans institutions give reasons for rejecting a loan application; in this regard, logistic regression is satisfactory because of its interpretability.

Simple logistic regression model is not convenient in practical applications because of a lot of computing. Therefore, the credit risk model (in the form of a scorecard) based on logistic regression and weight of evidence is proposed to distinguish good customers and bad customers quickly.

Logistic regression on credit data

Logistic regression used to divide multidimensional vectors into two categories. On credit data, logistic regression form is given by

$$p_n=\frac{\exp(\beta_0+\beta_1 x_{n,1}+\beta_2 x_{n,2}+\ldots+\beta_k x_{n,k})}{1+\exp(\beta_0+\beta_1 x_{n,1}+\beta_2 x_{n,2}+\ldots+\beta_k x_{n,k})} \quad (1)$$

where

p_n= the probability of costumer n's class label is bad (true or 1)

$x_{n,i}$= the value of costumer n on characteristic i

β_i = the coefficient of the model

In logistic regression, the value of class label for a costumer is discrete; the values ofβcan be estimated by the method of maximum likelihood estimation. The likelihood function form is given by

$$L=\prod_{i=1}^{N}\frac{\exp(\beta_0+\beta_1 x_{n,1}+\beta_2 x_{n,2}+\ldots+\beta_k x_{n,k})}{1+\exp(\beta_0+\beta_1 x_{n,1}+\beta_2 x_{n,2}+\ldots+\beta_k x_{n,k})} \quad (2)$$

where N is the number of costumers. Then, the values ofβcan be obtained by solving the following equations

$$\left\{\frac{\partial L}{\partial \beta_j} = 0 \ (\ j = 0, 1, 2 \ \ldots, k\ ) \right. \qquad (3)$$

A good logistic regression model should have significant linear relationship between the tuple of (x_1, $x_2 \ldots x_k$) and$logit(p)$ (i.e.,$\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k$). Therefore, set the hypotheses as

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_1: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ is false

The test statistic is adopted is the log likelihood ratio, and the form is given by

$$G^2=2 \ln \frac{L}{L_0} \qquad (4)$$

where L_0 = the value of L whileH_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$. The$G^2$ obeys chi-square distribution with df =k. After calculation, null hypothesis will be rejected if the p-value is less than significance level, and the linear relationship between the tuple of (x_1, x_2, …, x_k) and $logit(p)$ is considered to be significant.

Analogously, a good logistic regression model should have significant linear relationship between x_i and $logit(p)$. Therefore, set the hypotheses:H_0: $\beta_i = 0$, andH_1: $\beta_i \neq 0$. The test statistic is the *Wald*that obeys chi-square distribution with df=1 and the form is given by

$$Wald = \left(\frac{\beta_i}{S_{\beta_i}}\right)^2 \qquad (5)$$

where S_{β_i} is the standard error ofβ_i. The linear relationship betweenx_i and $logit(p)$ is considered to be significant if the p-value is less than significance level.

Weight of evidence on credit data

The value of weight of evidence (woe) can maximize the predictive ability of variable [6]. For example, the values of characteristicxare divided into three bins: bin 1, bin 2 and bin 3; then, bin1's woe is given by

$$\ln\left(\frac{gr}{br}\right) \times 100 \qquad (6)$$

where gr= the number of good costumers (while the value on x belongs to bin 1) is divided by the number of all good costumers; br= the number of bad costumers (while the value on x belongs to bin 1) is divided by the number of all bad costumers. Need to be explained, the reason of binning is to meet scorecard's needs; a costumer's value on a characteristic must belong to a bin, and the credit is measured by the bin's score.

Empirical experiment

For modeling, a credit dataset (to protect privacy, the dataset was processed specially) from University of California at Irvine Machine Learning Repository is applied. The dataset include 690 costumers' information; 15 characteristics (A1~A15; A1, A4, A5, A6, A7, A9, A10, A12 and A13 are discrete; A2, A3, A8, A11, A14 and A15 are continuous) and a class label (A16, "+": good costumers or "-": bad costumers).In this section, the software of Clementine was used for modeling process and analysis.

In 690 samples, there are 37 samples (5%) with incomplete information; therefor, the 37 samples were discarded; in the remaining 653 samples, 296 are good, and 357 are bad.

The characteristics that have little effect on the prediction for class label must be discarded. Therefore, the adopted strategies is: (1) a discrete characteristic will be discarded if the number of samples with a special value on this characteristic divided by the number of all samples is greater than 90%; (2) a discrete characteristic will be discarded if the number of arbitrary two sample values on this characteristic are not the same divided by the number of all samples is greater than 95%; (3) a continuous characteristic will be discarded if its variance coefficient is less than 0.1; (4) a characteristic will be discarded if the p-value (based on the hypothesis testing: H_0: the

characteristic is not related to the class label, H_1: the characteristic is related to the class label) is greater than 0.05. So, A13 is discarded according to strategy no. 1; A1 and A12 were discarded according to strategy no. 4.

To reduce the computational intensity, the characteristics that are redundant for predicting class label must be discarded. Therefore, the adopted strategy is to retain one of the two if the two continuous characteristics are related (the p-value is less than 0.05). So, A2, A3 and A11 are discarded. Note: to measure the correlation between continuous characteristic 1 and 2, the hypothesis is: H_0: characteristic 1 and 2 are not related, H_1: characteristic 1 and 2 are related.

According to the features of a scorecard, a costumer's value on a characteristic must belong to a bin. Therefore, the adopted strategy is: (1) the number of bins should be appropriate; (2) each bin of a characteristic has special effect on class label distribution. The weight of each bin is represented by woe that can be calculated by Equation 6. After processing, the rest of characteristics, their bins and woes of bins are shown in Tab. 1.

Tab. 1. Characteristics' bins and bins' woes

	Bins	Woes		Bins	Woes		Bins	Woes
A6	ff	-162.791	A8	<1.21	-79.608	A4	l	148.462
	x	179.681		>=1.21	85.595		u	18.337
	aa, m, c, w	-2.319	A9	f	-257.824		y	-67.879
	e, q, r, cc	85.947		t	155.232	A5	g	18.337
	i, d, k, j	-82.668	A10	f	-88.950		gg	148.462
A7	ff	-156.182		t	106.977		p	-67.879
	z	128.599	A14	<100	68.270			
	v, bb, o, n	-8.706		>=100	-31.987			
	h	74.126	A15	<500	-45.203			
	dd, j	-40.041		>=500	157.367			

It should be noted: (1) for discrete characteristics with small number of possible values, each possible value regards as a bin; (2) for discrete characteristics with big number of possible values (e.g., A6 and A7), first, each possible value regards as a bin, and calculate the woe of each bin; second, merge bins if the bins' woe are similar (by the method of *K* mean clustering); third, recalculate the woe of each bin.

Now, a costumer's information on a characteristic is replaced by woe. Credit dataset is divided into a training dataset and test dataset randomly; each one includes 50% of samples. The logistic regression after training is shown in Tab. 2.

Tab. 2.The coefficients of the logistic regression after training

β_0	β_4	β_6	β_8	β_9	β_{10}	β_{15}
0.1475	-0.01673	-0.007306	-0.005736	-0.009532	-0.006775	-0.008818

From Tab. 2, obviously, 3 characteristics (A5, A7, and A14) are discarded by the model.

By the analysis, the statistics ofG^2 is 282.903, and the p-value is very close to 0; the *Wald* of A4, A6, A8, A9, A10 and A15 are 11.661, 7.655, 5.308, 64.049, 10.487 and 10.816 respectively, and the p-values are all very close to 0.

Percentage Correctly Classified (PCC) represents the percentage of observations that are correctly classified. PCC is also proposed to measure the logistic regression model goodness of fit. PCC based on the training dataset is shown in Tab.3, and Tab.4 for PCC based on the test dataset.

Tab. 3. PCC based on the training dataset

Predicted	Observed		
	0	1	PCC
0	140	17	89.2
1	15	158	91.3
Overall Percentage			90.3

Tab. 4. PCC based on the test dataset

Predicted	Observed		
	0	1	PCC
0	120	29	80.5
1	21	153	87.9
Overall Percentage			84.5

Based on the above analysis, it can be seen that the logistic regression is good.

To implement scorecard, the score of each bin must be determined. In banking industry, Equation 7 is often used to calculate a costumer's credit score.

$$Score=Offset+Factor\times \ln(odds) \quad (7)$$

Where*odds* = the number of good costumers is divided to the number of bad costumers on*Score*. Usually, the*odds* should have doubled if the*Score* adds some. Let*odds* is 30/1 while*Score* is 500, and*odds* will have doubled while*Score* for each additional 50; then, a costumer's*Score* is shown by Equation 8.

$$Score=254.6553+72.1348\times \ln(odds) \quad (8)$$

Because of$logit(p) = -\ln(odds)$, Equation 9 is obtained by Equation 8 and the logistic regression.

$$Score = \sum\left[\frac{254.6553}{6}-72.1384\times\left(\frac{0.1475}{6}+\beta_i\times woe_i\right)\right] \quad (9)$$

According to Equation 9, the score of a bin on A*i* is

$$\frac{254.6553}{6}-72.1384\times\left(\frac{0.1475}{6}+\beta_i\times woe\right) \quad (10)$$

where, the "*woe*" is the bin's woe. According to Equation 10, each bin's score is shown in Tab.5.

Tab.5. Bin's scores

	Bins	Scores		Bins	Scores		Bins	Scores
A6	ff	-45	A8	<1.21	8	A4	l	220
	x	135		>=1.21	76		u	63
	aa, m, c, w	39	A9	f	-137		y	-41
	e, q, r, cc	86		t	147	A15	<500	12
	i, d, k, j	-3	A10	f	-3		>=50	141
				t	93			

The scorecard was analyzed using K-S index method on credit dataset, and the analysis result on credit dataset is given by Fig. 1.

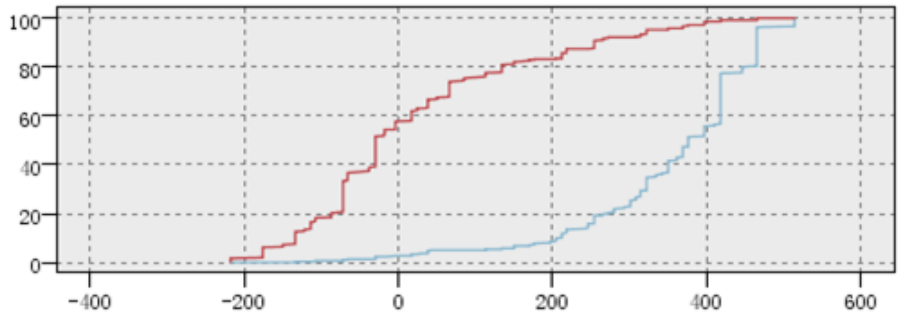

Fig. 1. Analysis of the scorecard using K-S

In Fig. 1, the horizontal axis represents*Score*, and the vertical axis represents cumulative ratio. The under curve is to good customer, and the up curve is to bad customer; the two curves are far apart, and it indicates that the scorecard model is good. After calculating, the distance between the two curves is furthest on the*Score* of 170;on*Score* of 170, the cumulative ratio of bad customers is 82.633%, and 6.757% for good customers; it means that, if accept a customer with over 170 and reject a customer with less 179, will reject 82.633% of bad customers and 6.757% of good customers. The result shows that the scorecard model (Tab. 3) is good, and the critical value of*Score* is 170 (note: the*Score* range is [-206, 812]).

Conclusion

In this paper, a model construction methodfor evaluating credit risk is proposed.

For credit risk evaluation, logistic regression is a more appropriate technique due to its desirable features. A scorecard model can evaluate a costumer's credit risk quickly at front-end of business. Because of the features of a scorecard (i.e., a costumer's value on a characteristic must belong to a bin), characteristics in credit dataset was binned;replace the value with bin's woe. The logistic regression model was proposed based on the training dataset; the analysis result shows that the model is good. Based on the logistic regression model, the scores of bins were determined; the score range and the critical value of score were determined too; the analysis result shows that the scorecard model is good.

To protect privacy, the credit dataset was processed specially; in practical applications, a data analyst can do more works using semantic analysis; for example, obtaining more interesting conclusion, understanding dataset better, guiding business, improving model.

Acknowledgement

This work was financially supported by Scientific Research Foundation Project of Honghe University of China (Project No. XJ15Y21).

References

[1] Bhattacharya S, Kumar K. Artificial neural network vs linear discriminant analysis in credit ratings forecast: A comparative study of prediction performances [J]. Review of Accounting & Finance, 2006, 5(3):216-227.

[2] Abdelmoula A K. Bank Credit Risk Analysis with K-Nearest-Neighbor Classifier: Case of Tunisian Banks [J]. Journal of Accounting & Management Information Systems, 2015, 14.

[3] Arminger G, Enache D, Bonne T. Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks [J]. Computational Statistics, 1997, 12(2):293-310.

[4] David Andrich. A General Form of Rasch's Extended Logistic Model for Partial Credit Scoring [J]. Applied Measurement in Education, 1988, 1(4):363-378.

[5] Thomas L C. Consumer credit models: pricing, profit and portfolios [M]// Oxford University Press, 2009.

[6] WEED D L. Weight of evidence: a review of concept and methods [J]. Risk Analysis, 2005, 25(6): 1545-57.