academicJournals

Vol. 7(16), pp. 1414-1421, 28 April, 2013
DOI: 10.5897/AJBM11.394
ISSN 1993-8233 © 2013 Academic Journals
http://www.academicjournals.org/AJBM

African Journal of Business Management

Full Length Research Paper

Calculating the best cut off point using logistic regression and neural network on credit scoring problem- A case study of a commercial bank

Mehrnaz Heidari Soureshjani and Ali Mohammad Kimiagari*

Department of Industrial Engineering, Amirkabir University of Technology, Tehran, Iran.

Accepted 1 June, 2012

Credit scoring is a method used to estimate the probability of default or becoming delinquent of a loan applicant or existing borrower. There are several methods used for scoring, such as traditional statistics models like probit and logistic regression, data mining approaches and also artificial intelligence algorithms. In this paper, two high-usage methods on real data of legal customers of a commercial were used, and also their performance have been compared. It was found that logistic regression as a statistic model can estimate a good econometrics model which is able to calculate the probability of defaulting, and also neural networks is a very high performance black box method which can be used in credit scoring problems. Also the best cut off point in both logistic regression and neural network is calculated by these methods which have minimum errors on the available data.

Key words: Credit scoring, logistic regression, goodness of fitness, cut off point, neural network.

INTRODUCTION

Devoting capital and asset to economic activities is accomplished through financial market, in which banking is the main part. This is done through lending to the bank's customers. Credit risk means the probability of default or becoming delinquent of a loan applicant or existing borrower. Reducing and controlling this risk is one of the main parameters which resulted to better lending approaches and performance improvement of the banks.

There are several methods used to solve the problem of credits scoring. These different useful techniques, known as the credit scoring models, have been developed by the banks and researchers in order to solve the problems involved during the evaluation process. The objective of credit scoring models is to assign credit applicants to either a "good credit" group that is likely to repay financial obligation or a "bad credit" group who has high possibility of defaulting on the financial obligation. Therefore, credit scoring problems are basically in the scope of the more general and widely discussed classification problems (Lu and Chen, 2009).

As a legal customer applied to obtain a loan, the bank should assess the applicant and decide whether to approve the loan request or not. This is done through human analysis on the main characteristics of the customer. But accepting the characteristics remarked by customer itself may mislead the decision maker. So the bank must make the decision based on real, trustworthy and documented information. In order to handle these information altogether, the bank need to automate the credit evaluation process.

As a result, customers with high probability of default accounts can be monitored and necessary proceedings can be taken in order to prevent the account defaulting. In response, the statistical methods, non-parametric statistical methods, and artificial intelligence approaches have been proposed to support the credit approval decision process.

*Corresponding author. E-mail: kimiagar@aut.ac.ir Tel: +982164542228.

Credit scoring models are commonly structured along the lines of Altman's (1968) Z-score model using historical loan and borrower data to identify which borrower characteristics are able to distinguish between defaulted and non defaulted loans. Other general introductions to credit scoring are proposed by Mays (1998), Hand and Henley (1997), Mester (1997), Viganò (1993), and Lewis (1990). A *credit-scoring* model is a formula that puts weight on different characteristics of a borrower, lender, and loan (Nanni and Lumini, 2009).

Among several methods used for credit scoring, discriminant analysis and logistic regression are two most commonly used data mining techniques to construct credit scoring models. However, linear discriminant analysis (LDA) has often been criticized because of its assumption about the categorical nature of the data and the fact that the covariance matrices of different classes are unlikely to be equal. In addition to the LDA approach, logistic regression is an alternative to an alternative way to set down credit scoring. Basically, the logistic regression model emerged as the technique in predicting dichotomous outcomes. A number of logistic regression models for credit scoring applications have been reported in the literature. Harrell and Lee (1985) found out that logistic regression is as efficient as LDA (Lee et al., 2006).

In addition to LDA and logistic regression, credit scoring also lends itself to a recent development of neural networks approach. Neural networks provide an alternative way for LDA and logistic regression, particularly in situations where the dependent and independent variables exhibit complex nonlinear relationships. Even though neural networks have been reported to have better credit scoring capability than LDA and logistic regression (Desai et al., 1996; Jensen, 1992; Piramuthu, 1999), they are, however, also being criticized for their long training process in designing the optimal network's topology and inability to identify the relative importance of potential input variables, as a result of which they have limited its applicability in handling credit scoring problems (Piramuthu, 1999).

Beside these methods, there are several data mining methodologies used in previous years, such as classification and decision trees (CART), multi-adaptive regression splines (MARS) and bootstrap aggregation (Bagging) (Hothorn and Lausen, 2003).

Data mining approaches are becoming a common alternative for making credit scoring models due to their associated memory characteristic, generalization capability, and outstanding credit scoring capability, but these approaches are also being criticized for their long training process, inability to identify the relative importance of potential input variables, and certain interpretative difficulties (Lee et al., 2006).

There are two main scoring types: External scoring and internal scoring. Some international committees have published standard scoring methods based on probability of default, such as Basel committee, Fitch, S&P and other committees (Basel, 2000). Scoring based on these standards usually is named as external scoring. But each bank in its own country, area of activity and capability can have a special scoring model, which will be its scoring model. This is usually called internal scoring.

Purpose of this study

In this article we tried to find a special model for calculating the probability of default of legal customers in a commercial bank using two approaches which have generally been used for credit risk problems, logistic regression and neural network.

Every modeling method may have errors, and especially in credit risk modeling, there are two types of errors: Error type 1 which means reporting a bad credit as a good one, and Error type 2 which means reporting a good credit as a bad one. Error type 1 may cause losing the loan, losing the loan's profit and also pursuit costs but error type 2 may cause losing a good opportunity for bank. Finding a model which can minimize both error type 1 and 2, is the goal of this article.

METHODOLOGY

As mentioned earlier, two main approaches to calculate probability of default in this article are logistic regression and neural network. The case study contains 127 legal customer's data of a commercial bank. These customers have borrowed a loan and 21 of them did not repay their loan. These historical data have been gathered and some variables have been calculated based on their balance sheet information at the moment of applying for the loan.

There are 4 main financial ratios which are liquidity ratios, leveraging ratios, activity ratios, and profitability ratios. Liquidity ratios measure the availability of cash to pay debt. Leveraging ratios or Debt ratios measure the firm's ability to repay long-term debt. Activity ratios measure the effectiveness of the firm's use of resources. Profitability ratios measure the company's use of its assets and control of its expenses to generate an acceptable rate of return (Williams et al., 2008).

These ratios are the main financial information which a bank can used to calculate the probability of default of a customer. Other information such as management measures, Characteristics and perspectives of the products and competitions are also important (Jiao et al., 2007). But this information can be kept in view as this information was not in the scope of this article.

Logistic regression

Distribution of the credit information data is usually non-normal and in this case a suitable extension is a generalized linear model known as logistic regression or logit model. Given a vector of application characteristics x, the probability of default p is related to vector x by the following equation:

$$logit(p) = \ln\frac{p}{1-p} = w_0 + w_1x_1 + w_2x_2 + \cdots + w_kx_k$$

Logistic regression provides a method for modeling a binary response variable, which takes values 1 and 0 by mapping the data on a logit curve (Figure 1).

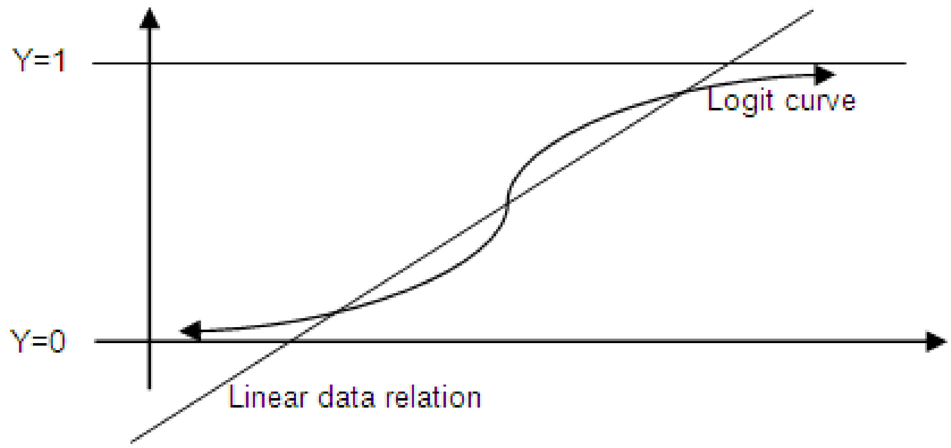

Figure 1. Logit curve.

The response variable here is 1 for those customers who have defaulted, and 0 for those repaid their loan at regular time. The vector x is the vector of characteristics which are actually the financial ratios for each customer in the moment of applying for debt. The vector w is calculated through maximum likelihood estimation. In this method, a function is defined based on the probability and w, named likelihood function. Maximizing the logarithm of the likelihood function will maximize the prediction rate of the model.

To evaluate the illustrated logit model, there are four main tests, (a) overall model evaluation, (b) statistical tests of individual predictors, (c) goodness-of-fit statistics, and (d) validations of predicted probabilities (Ying et al., 2002).

A logistic model is said to provide a better fit to the data if it demonstrates an improvement over the intercept-only model (also called the null model). An improvement over this baseline is examined using the LR or -2lnL(null)- 2lnL(model), which lnL(model) is maximum likelihood as the estimated variables are meaningful in the model, and lnL(null) is likelihood with assuming zero for all variables.

The statistical significance of individual regression coefficients is tested using the Wald chi-square statistic. If the statistic is less than 0.05, then the variable merely should be included in the model.

Goodness-of-fit statistics assess the fit of a logistic model against actual outcomes. The inferential goodness-of-fit test is the Hosmer–Lemeshow (H–L) test. This statistic tests H_0 hypothesis of $H_0 : E[Y] = \frac{exp(x'w)}{1+exp(x'w)}$ using chi-square, and if becomes more than 0.05, shows that the model fits well to data.

Logistic regression predicts the logit of an event outcome from a set of predictors, and it can be transformed back to the probability scale:

$$p = \frac{exp(x'w)}{1 + exp(x'w)} = \frac{e^{(w_0 + w_1 x_1 + \cdots + w_k x_k)}}{1 + e^{(w_0 + w_1 x_1 + \cdots + w_k x_k)}}$$

The resultant predicted probabilities can then be revalidated with the actual outcome to determine if high probabilities are indeed associated with events and low probabilities with nonevents. The degree to which predicted probabilities agree with actual outcomes is expressed as either a measure of association or a classification table.

As the study use classification table for assessing the logit model, the cut-off point should be considered. In a classification table, if the predicted probability of default for a customer becomes more than cut off point, we can report the customer as a "bad" customer, and if this probability becomes less than the cut-off point, we can report the customer as a "good" one. The pre-defined cut off point value in statistical and econometrics software such as *Eviews* and *SPSS*, is usually 0.5. But what is the best cut off point?

As mentioned earlier, the best model for a bank is the one that minimizes both type 1 and 2 errors. Selecting the best cut off point is done by minimizing these errors.

Neural network

Neural networks are artificial intelligence algorithms that allow for some learning through experience to discern the relationship between borrower characteristics and the probability of default and to determine which characteristics are most important in predicting default (Mester, 1997).

Research into neural networks began in 1943 with the publication written by MCCULLOCH and PITT. According to the proclaimed principle, a mathematical model had to be developed that could simulate the natural operation of the neuron. For us, the important parts of a neuron are the dendrites, through which the neuron receives signals, and the axons, which help to forward processed information to other neurons. Synapses play a significant part in processing information. It is through them that axons connect to the dendrites of other neurons (Ferenc, 2003). The operation of the mathematical neuron model is similar to human's brain. Using a given function, they process the information received from dendrites, and if the incoming signal exceeds a so-called stimulus threshold, they forward the information via axons. The most important property of the neuron is that it is continuously changing its operation (that is, its internal function) based on the data received – it is 'learning'. Synapses play an important role in this learning process, as they are able to amplify or subdue the signals coming from other neurons. In the learning process, signal amplification factors change on synapses (referred to as 'weights' in the model in light of their function). In the neuron model, the change or modification of these weights means learning.

Characteristics of the customers are input variables in neural network for credit scoring problem. The output variable is the actual condition of the customers, either they have defaulted or not.

To assess a neural network, there are 3 main indicators that show whether the network can estimate the outputs or not (Ravi et

Table 1. Variables in the survey to find effective variables.

Category	Name	Formula	Variable name
Liquidity ratios	Current ratio (working capital ratio)	Current asset/ current liabilities	X_1
	Quick ratio	Current asset – (Inventories+ prepayments)/ current liabilities	X_2
Leveraging ratios	Debt ratio	Total liabilities/ total assets	X_3
	Debt to equity ratio	Long term debt+ value of leases/ average shareholder equity	X_4
Activity ratios	Asset turnover	Net sales/ total assets	X_5
	Receivable turnover ratio	Net sales/ average net receivables	X_6
	Stock turnover ratio	Cost of goods sold/ average inventory	X_7
Profitability ratios	Return on assets	Net income/ total assets	X_8
	Return on equity	Net income/ average shareholder equity	X_9
	Profit margin	Net profit/ net sales	X_{10}

al., 2002). These indicators are PR, RMSE and MAPE. PR means the amount of outputs predicted correctly (F) on the amount of total predicted outputs (N):

$$PR = \frac{F}{N}$$

root mean square (RMS) is calculated from the following equation, which O_m^i is the predicted output from neural network and O_a^i is the actual amount of output.

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}\left(O_a^i - O_m^i\right)^2}{n}}$$

MAPE means the average of absolute error:

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\frac{|actual_i - forecast_i|}{actual_i}$$

The best learned neural network is that which can predict outputs more correctly. More PR and less RMS and less MAPE can show this. But the meaning of PR is closely related to the cut-off point we define in calculating error, or what we say "correct prediction". So in this article, besides calculating RMS and choosing the best neural network, the study also check different cut off points to find the best for calculating credit risk. The amount of MAPE can't be calculated, because there are zeros as denominator.

RESULTS

The two main approaches used in this article have been briefly discussed before. In this article, the study have used the approaches and illustrated model to specify variables influencing default of the customers. Almost 1000 data of legal customers of a commercial bank gathered and because the data base has had several missing data, filtered. As mentioned, the data of 127 legal customers were finally selected and taken into account for illustrating the logit model.

The main variables which were in the survey were financial ratios which are listed in Table 1. The variables entered into the model illustration process one by one, and the best model which has got the characteristics mentioned earlier selected as final logit model. The Eviews output window for illustrating the model is shown in Table 2. The final equation which models the default of these 127 customers is:

$$\begin{aligned} logit(p) &= ln\frac{p}{1-p} \\ &= -7.233126173 - 4.814803512 * X_1 + 13.63992793 * X_3 \\ &\quad - 22.78927478 * X_8 \end{aligned}$$

This result shows that 3 main variables affect the default of a customer which are current ratio, debt ratio and return on assets. As it is shown in Table 2, the model's LR amount is 81.35 and the statistic is almost 0<0.05, so the hypothesis of being zero for all variables is not true, so the logit model is truly illustrated. Also the chi-square statistic for each variable is less than 0.05 and shows that these variables merely should be included in the model.

As shown in Table 3, the probability of chi-square statistic in Hosmer-Lemeshow goodness of fitness test shows that the model fits data very well.

To calculate the default probability of customers using logit model, we can use the following equation:

$$p = \frac{e^{(-7.233126173 - 4.814803512 * X_1 + 13.63992793 * X_3 - 22.78927478 * X_8)}}{1 + e^{(-7.233126173 - 4.814803512 * X_1 + 13.63992793 * X_3 - 22.78927478 * X_8)}}$$

Calculating the probability for each customer and comparing the estimated probability with actual position

Table 2. Output window for final illustrated model.

Equation: FINALEQUATION Workfile: LOGIT

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: ML - Binary Logit (Quadratic hill climbing)
Date: 02/02/11 Time: 15:22
Sample: 1 127
Included observations: 127
Convergence achieved after 8 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-7.233126	2.847219	-2.540418	0.0111
X1	-4.814804	1.615577	-2.980238	0.0029
X3	13.63993	3.920269	3.479335	0.0005
X8	-22.78927	10.06199	2.264888	0.0235

Mean dependent var	0.165354	S.D. dependent var	0.372971
S.E. of regression	0.204930	Akaike info criterion	0.319283
Sum squared resid	5.165524	Schwarz criterion	0.408864
Log likelihood	-16.27448	Hannan-Quinn criter.	0.355679
Restr. log likelihood	-56.95224	Avg. log likelihood	-0.128146
LR statistic (3 df)	81.35553	McFadden R-squared	0.714243
Probability(LR stat)	0.000000		

Obs with Dep=0	106	Total obs	127
Obs with Dep=1	21		

Table 3. Hosmer-Lemeshow goodness of fitness test.

Equation: UNTITLED Workfile: LOGIT

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: ML - Binary Logit (Quadratic hill climbing)
Date: 02/02/11 Time: 16:22
Sample: 1 127
Included observations: 127
Andrews and Hosmer-Lemeshow Goodness-of-Fit Tests
Grouping based upon predicted risk (randomize ties)

	Quantile of Risk		Dep=0		Dep=1		Total	H-L
	Low	High	Actual	Expect	Actual	Expect	Obs	Value
1	2.E-10	0.0001	12	11.9998	0	0.00020	12	0.00020
2	0.0001	0.0004	13	12.9967	0	0.00328	13	0.00328
3	0.0004	0.0009	13	12.9926	0	0.00738	13	0.00738
4	0.0010	0.0024	12	11.9825	0	0.01751	12	0.01754
5	0.0026	0.0042	13	12.9555	0	0.04451	13	0.04466
6	0.0043	0.0147	13	12.8870	0	0.11300	13	0.11399
7	0.0150	0.0285	12	11.7795	0	0.22050	12	0.22462
8	0.0337	0.1533	12	11.9756	1	1.02437	13	0.00063
9	0.2917	0.8668	6	5.50343	7	7.49657	13	0.07770
10	0.8690	0.9865	0	0.92731	13	12.0727	13	0.99854
		Total	106	106.000	21	21.0000	127	1.48854

H-L Statistic:	1.4885	Prob. Chi-Sq(8)	0.9929
Andrews Statistic:	89.5675	Prob. Chi-Sq(10)	0.0000

Table 4. Classification table, with cut off point equal to 0.5.

Observed			Predicted		
			y		Percentage
			.00	1.00	correct
Step 1	y	0.00	103	3	97.2
		1.00	5	16	76.2
	Overall percentage				93.7

Table 5. Sensitivity and specifity of model with different cut off points.

Cut off point	Specifity	Sensitivity	Sum
0.01	69.7	100	169.7
0.1	90.6	95.2	185.8
0.2	94.3	95.2	189.5
0.3	95.3	95.2	190.5
0.4	97.2	90.5	187.7
0.5	97.2	76.2	173.4
0.6	97.2	71.4	168.6
0.7	97.2	71.4	168.6
0.8	97.2	71.4	168.6
0.9	100	33.3	133.3
0.99	100	0	100

of customers, either have defaulted or not, the study have classified customers in Table 4, with cut off point equal to 0.5.

The two error types are clear in the classification table. For cut-off point equal to 0.5, the Type 1 error is 23.8% and the Type 2 error is 2.8.

To minimize the overall error, Korsholm proved that if one of 4 main situations occurs, the amount of overall error will be minimized (Korsholm, 2004). If the proportion of defaulted customer's output variable was named (y=1) correctly predicted as "sensitivity of the model" and the proportion of non-defaulted customer's output variable (y=0) correctly predicted as "specifity of the model", these 4 situations are:

1) Sum of sensitivity and specifity degree of the model becomes maximal
2) If the sensitivity becomes more than 80%, then sum of sensitivity and specifity degree of the model becomes maximal
3) The minimum amount between sensitivity and specifity degree of the model, becomes maximal
4) If the type1 error is x multifold important than type 2 error for the bank, then the amount of (x* sensitivity + specifity) becomes maximal.

It is notable that sensitivity degree equals to 1- error type 1 and specifity degree equals to 1- error type 2. Using different cut off points and calculation the classification tables, error type 1 and error type 2 and sensitivity and specifity degree of the model for each cut off point, the best cut off point will be 0.3, as is summarized in Table 5 and is shown in Figure 2.

Another point that can confirm the result of logit modeling is relative operating characteristic (ROC) curve, which shows a receiver operating characteristics and used for evaluating the logit model as well. The ROC plot is merely the graph of points defined by sensitivity and (1 – specificity). Customarily, sensitivity takes the y axis and (1 – specificity) takes the x axis. If the area under the curve becomes maximum amount, then the model fits data well. The curve is shown in Figure 3.

To evaluate the overall illustrated model, the study can compare the results of logit model with the results of neural network. For this purpose, the study take 112 data as learning data and 15 data as test data. And the 3 variables that finally illustrate the logit model used as input data and the actual position of customers as output data. It used a feed-forward network-error back propagation algorithm with 3 neurons in hidden layer. The RMS of test data was 0.151 and the study calculated the prediction rate (PR), with two amounts of 0.5 as a default probability cut off point and then 0.3 which is the best cut off point in logit modeling. The results are shown in Table 6.

Conclusion

In this article we calculated the best cut off point which

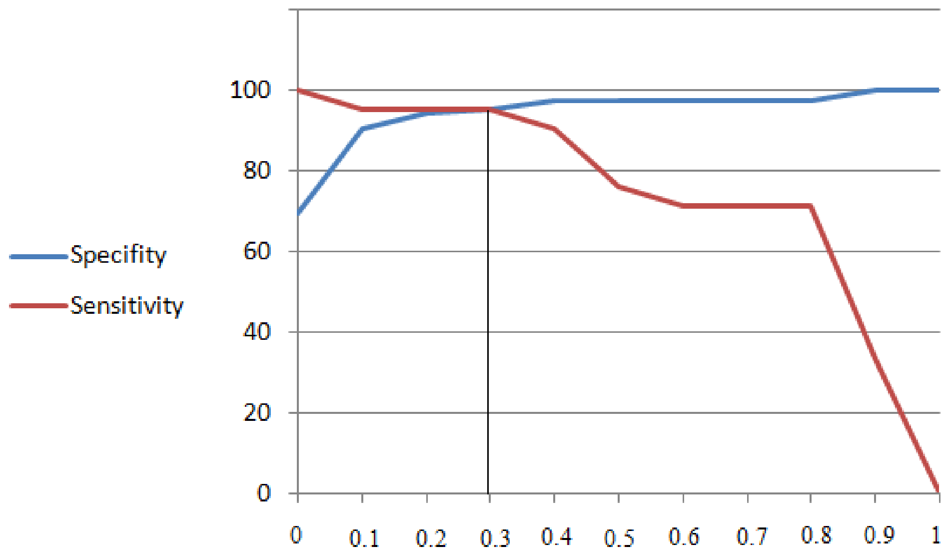

Figure 2. Best find cut off point that maximizes the minimum of specifity and sensitivity degree.

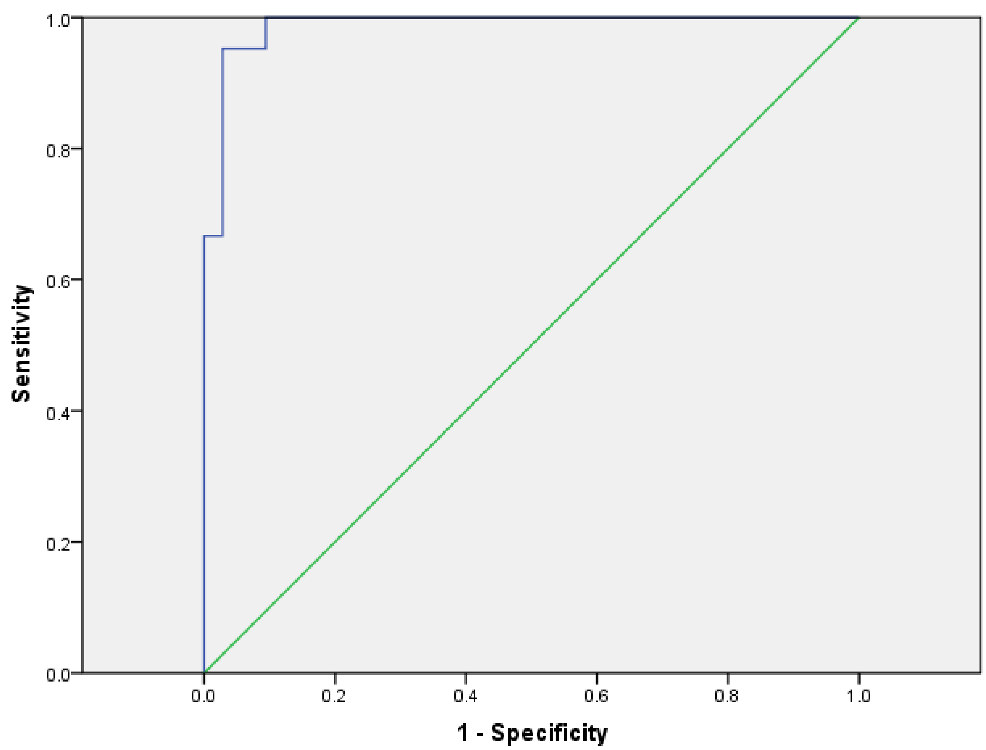

Figure 3. ROC curve- The area under curve is 0.987.

Table 6. RMS and PR using neural network.

Input data: 3 variables (current ratio, debt ratio, return on assets)			
RMS	PR- cut off point equal to 0.3	PR- cut off point equal to 0.5	Neurons of hidden layer
0.151	124/127=0.97	124/127=0.97	3

minimizes the overall error of modeling credit risk, both with logistic regression and neural network methods which are two most used methodologies in credit risk analysis.

The best amount of cut off point for every bank is the amount that minimizes the overall error of illustrated model. In logistic regression modeling, the cut-off point is the point that the decision maker decides whether to accept the loan application or not. If the probability becomes more than the cut-off point, the customer will be in the class of "bad customers", otherwise will be in the class of "good customers".

In neural network, for calculating the prediction rate of the network, we should use a threshold of correct prediction on incorrect one which is similar to the cut-off point in logit modeling. Using the neural network shows that the calculated best cut off point via logit modeling is also good for neural network.

REFERENCES

Basel (2000). Committee on Banking Supervision, Principal for Management of Credit Risk, September.

Desai VS, Crook JN, Overstreet Jr. GA (1996). “A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment”. Eur. J. Oper. Res. 95:24-37

Ferenc (2003). Credit scoring process from a knowledge management perspective. Periodical Polytechnic Ser. Soc.Manag.Sci. (11/1):95-110.

Harrell FE, Lee KL (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. In P. K. Sen (Ed.): Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences. North-Holland: Elsevier Science Publishers pp.333-343.

Hothorn T, Lausen B (2003). Double-bagging: combining classifiers by bootstrap aggregation. Pattern Recognit. 36:1303-1309.

Jensen HL (1992). “Using Neural Networks for Credit Scoring”, Manag. Financ. 18:15-26

Jiao Y, Syau YR, Lee ES (2007). Modeling credit rating by fuzzy adaptive network. Math. Comput. Model. 45:717-731.

Korsholm L (2004). Analysis Of Diagnostic Studies, Sensitivity and specificity positive and negative predicted values ROC curves tests based on logistic regression, Department Of Statistics And Demography, University Of Southern Denmark.

Lee TS, Chiu CC, Chou YC, Lu CJ (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Comput. Stat. Data Anal. 50:1113-1130.

Lu CL, Chen TC (2009). A study of applying data mining approach to the information disclosure for Taiwan’s stock market investors. Expert Syst. Appl. 36:3536-3542

Mester L (1997). What’s the point of the credit scoring? Business Review-Federal Reserve Bank of Philadelphia

Nanni L, Lumini A (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Syst. Appl. 36:1-4.

Piramuthu S (1999). Financial Credit-Risk Evaluation with Neural and Neuro-fuzzy Systems. Eur. J. Oper. Res. 112:310-321.

Ravi BS, Warren FW, Jos LGAM (2002). Modelling and Evaluating Quality Measurement using Neural Networks. Int. J. Oper. Prod. Manage. 22(10):1162-11855.

Williams JR, Haka FS, Bettner MS, Carcello JV (2008). Financial and Managerial Accounting. McGraw-Hill Irwin.

Ying CH, Peng J, Lee KL, Ingersoll GM (2002). An Introduction to Logistic Regression Analysis and Reporting. J. Educ. Res. 96:1.