Customer Validation using Hybrid Logistic Regression and Credit Scoring Model: A Case Study

M.J. ERSHADI[1], D. OMIDZADEH[2]

[1]Corresponding author, Information Technology Department, Iranian Research Institute for Information Science and Technology (IRANDOC), Tehran, Iran; E-mail: mjershadi@gmail.com
[2]Industrial Engineering Department, Science and Research Branch, Islamic Azad University, Tehran, Iran

Abstract

In this paper a regression model is applied for validating the customers of a company. Using a Delphi method beside the expert panel the main variables which construct the regression model are extracted. A credit scoring system for validation of the customers is developed based on applied regression model. Then a Newton-Raphson method is used for determining the coefficients of regression model. Furthermore the MacFadden statistical value is calculated for approving the regression model. A case study is presented for application of proposed model. Three factors of sponsor's job, applicant's jobs and income are extracted as the main factors which are affecting the customer validation in the case. The result of the proposed model based on the case study showed that using a regression model which is empowered by Delphi system can provide a robust model for validation of the customers for deciding on granting to the customers.

Keywords: *credit scoring, regression model, Delphi method, case study.*

1. Introduction

Deciding about granting financial facilities is one of the main tasks of financial institutes and banks. Furthermore credit level and potentials to refund principle and interest of the granted facilities by the grantee shall be defined so that risk of credit level or probable no-refunding of principle and interest of the granted facilities could be reduced. One of the credit evaluation methods is a system designed to measure the credit level of the facility grantees by credit evaluation or ranking. Using such model, credit level or ranking of the applicants is measured and consequently it is decided whether to grant facilities or not.

The target of credit evaluation of the applicants in each organization is to execute equity in granting facilities. Convenience, fast identification of the applicants, and granting facilities is the main principal of such system. Moreover, avoiding facility granting to non-creditworthy clients which is the reason of increase in deferred payment of organizations, is also another advantage of credit evaluation. Integrating the information of credit applicants in each organization and specially establishing full data base of credit applicants and credit ranking are assumed as other results of credit evaluation system setup.

Nowadays, most of the financial institutes have designed ranking systems of objective validity based on scientific models and methods. Modern credit evaluation and measurement system of the applicants are based on mechanization procedures in which special advantages are allocated to some special important credit features of the applicants (such as income, job, guarantee, and capital). Therefore, if the sum of these advantages is more than expected, the client is assumed as creditworthy client otherwise the client is assumed as non-creditworthy client.

The important reasons for necessity of credit evaluation and measurement of the clients in institutes are increasing the cash flow of the institute; reducing the risk of refunding principle and interest of granted credits; ensuring the refund of granted investment by the client; convenience in deciding to grant facilities; reducing the expenses and promoting the performance of the system. (Lee et al. 2002, West 2000).

Iran Khodro Company with 550,000 unit's annual production is the biggest production factory in IRAN. One of its subsidiary companies is Iran Khodro leasing company which its main duty is funding and selling products with loans and monetary and financial leasing. This company started in 2003 and now is the subsidiary of Iran Khodro Investment Development (IKIDO) and Samand Investment.

Based on collected data in decade 2000, facilities granted by this company is equal to 7,639,799 Million Rials and differed payments equal to 1,398,861 Million Rials. Total customers in 80s were 83200 people which 53 percent have differed payments. Accumulation of differed payments has leaded to sales decline of leasing company and downtrend allocation of credit by Iran Khodro Company and banks. As a result to prevent accumulation of differed payments and to prevent losses of the company, validation of customers using the granted facilities is crucial. Validation of customers seems one of the most important challenges of Iran Khodro and its collaborating banks.

Making decision about customers' amount of credit is done by several methods. In the past Judgment based methods were very common for validating. In this way decision making is based on expert's decision which is very time consuming, expensive and subjective, in addition it does not have the necessary scientific validity. Methods based on statistical models are quite reasonable, affordable and scientific backing. Using statistical models, including memes analysis, neural networks, linear regression and logistic regression to assess customers' credit is common and useful.

The structure of this paper is as follows. In section 2 the

literature review on the main papers on the application of different models for customer validation is done. The case in which the credit scoring is applied is introduced in section 3. Our methodology in application of regression model in customer validation is presented in section 4. The main results of application of regression model and the way that our model is verified is explained in the section 5. The conclusion of the paper is proposed in section 6.

2. Literature Review

Numerous articles have been published in the field of customers' validation which in this section is introduced and briefly studied. Altman and Saunders in 1998 provided a discriminate analysis model for making decisions to assess the validity. In this model with combination and weighting financial ratios have introduced a credit rating and if the applicant individuals or companies grade doesn't be higher than the desired threshold they have no credit to get the loan. Altman)1977(also used econometric techniques, multivariate statistics, logit and probit models to measure credit risk. Doumpos and Zopounidis (2001) using hierarchy multi-criteria decision-making have addressed the credit risk. Fridson (1995) refers of the financial review process as an effective tool to evaluate credit risk. Considering some models which have been base of industries used in the past, he has separated successful and unsuccessful companies in the field of periodic credit payment. Some of these models are still being used.

Although globalization, rapid changes, fluctuations in the economic and political environment of developing countries and also an extremely high rate In terms of innovation in products and production processes make it difficult to consider such situation. In addition, the information contained in the financial statements may not be true. Even checking out this information by an independent person to verify data and checking whether this information has been deliberately entered incorrectly or not, may not yield results (may yield wrong results) (Fridson, 1995, Fraser, 1995). Bryant (2001), Matsatsinis et al., (1997), O'Leary (1995), Sangster (1995), Smith and McDuffie (1996) have evaluated the credit of companies by using their data and financial ratios. In the evaluation made by these people, only financial ratios of companies have been considered and non-financial ratios such as company history and other cases like that have not been considered. But other researchers have evaluated the credit of companies by combining financial and non-financial information. Strischek (1999), Todd et al., (1998) can be placed in this group.

Among other methods that are used for customers' credit rating, Neural networks and Genetic programming can be noted that their results are more accurate and their use has increased in recent years. Genetic programming is used to develop a genetic algorithm. Koza (1992) has introduced this method as one of the best methods available to perform credit rating. Abdo (2009) using genetic programming has been doing Egypt's rating of banks. In this article using three different approaches, genetic programming, probit analysis and weighting, he anticipate the customer's credit and also he used variables which have not been used so far in reports such as the Central Bank of Egypt. And finally by using standard accuracy rate reached the conclusion that Genetic Programming is more accurate than the other two models. Genetic Programming is used in many classification issues and predictions. For example, Etemadi et al (2009) tried to predict bankruptcy and Huang et al. (2006) have used this method in the ranking. Using neural networks is common way for prediction and according to its variety has many different types of application. Khashman (2010) compared the different neural networks and their learning types to assess the credit risk. Yu et al. (2008), considered a multi-stage model based on neural networks to assess credit risk, and one of the barriers to the use of neural networks is the necessary volume of data for the training network which covers a large part of the available data.

3. Case Study

As mentioned in the previous section the aim of this study is to evaluate the credit risk of customers of Iran Khodro Leasing Co. The company was founded on October of the year 1382. The company operates in various sectors: Including financial, administrative, financing and leasing activities. The main objective of the company is the expansion of leasing to increase profits and capture a greater share of the international market. The population of the research is applicants seeking loans from Iran Khodro leasing company in 5 years. The population is equal to 50672 people. One of the most common methods for sample selection which is introduced by using statistical simulation is that for every variable in the model at least we choose ten people who have credit, and ten people without credit (Pediosy et al. 1996). To increase the accuracy, the selected sample size of the members of society is equal to 500. The sample is chose by using simple random sampling from the community. The final sample which is used in the calculation includes 408 creditworthy people and 93 non-creditworthy people. Thus, using logistic regression and seven important variables from the point of view of experts that their selection method was described in the previous section, we chose a regression model to assess credit risk of customers that it will be explained in the following.

4. Methodology

To determine the most important variables and indicators affecting the validity of the customers, first by using the Delphi method and expert panel we have considered and reviewed seven indices of income, age, education, gender, profession, sponsor's occupation and place of residence of the applicant as the main indicators for assessing customers' credit. Delphi is one of the acceptable methods for using the knowledge of the group. Delphi is systematic approach or method for research to extract the comments of an expert group about an issue or a question. One of the purposes of Delphi is extraction, deduction and aggregation of academic and executive knowledge, and its analysis, as to guide the administrators and protect it in academic areas. The statistical population of the research is applicants seeking loans from Iran Khodro leasing company in 5 years which the number of people is 50672. Determining the suitable sample size for use in a logistic regression is a problematic issue which have not already been resolved. We use a sample with size 500. Choosing the sample has been determined by using simple random sampling from the community. After determining the sample members we calculate the creditworthy and non-creditworthy points of people. The criterion for credit-worthiness is defined as follows:

Credit-worthiness rating = number of due date installments / number of paid installments

By rounding the resulting number, we show good customers with one and bad customers with zero. This variable will be used as the dependent variable in the regression analysis and because of this binary variable the use of logistic regression is reasonable. The final sample that is used in the calculation consists of 408 creditworthy persons and 93 non-creditworthy persons. In the following logistic regression will be introduced and its features will be explained.

4.1 Logistic regression

One of the most widely used statistical models to classify and relate issues is logistic regression. The logistic regression

model, unlike other statistical tools (Such as discriminate analysis or linear regression) uses different distribution function for estimating (Press and Wilson, 1978) so it is so convenient for credit rating issues. In addition, to increase the accuracy and flexibility of this model various methods have been proposed for the development of binary logistic regression model (Agresty 1990). In fact, when the dependent variable has the discrete values using logistic regression seems to be logical. Our dependent variable in this study is Credit- worthiness rating as previously was explained the discrete values of zero and one is available. The popularity of using logistic regression is due to the shape and kind of behavior of the logistics function. This function is as follows:

$$f(z) = \frac{1}{1 + e^{-z}}$$

Domain of this function is real numbers, $\mathbb{R}$, and range of that is closed interval of 0 and 1.This feature is one of the reasons for the popularity of this model. This model is designed to describe the probability, which always have value between zero and one. So it is wise to use this model for risk assessment as always has values between zero and one, while this feature does not exist for many other models. *Figure 1* shows the S-shaped logistics function curve, which is another reason for the popularity of this model. In fact the S-shaped logistics function curve represents that the risk is low for small values of Z and after reaching the desired threshold, risk will increase (For example, for amounts larger than -4 in the figure, the risk highly increase and for amounts larger than 4, risk is almost constant) and for large values of Z, risk is around one (Kleinbaum and Klein 2010).

For fitting a logistic regression model, estimate the value of Z in the logistics function by $\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$, where independent variables are X_1 to X_t, then with estimating the existing parameters with maximum likelihood method, we estimate the probability or dependent variable risk. For example, in this study we want to classify new people applying for a loan from Iran Khodro Leasing Co. to two eligible and ineligible groups. Determine the likelihood of allotment of facility to each person by using logistic regression. In fact, the probability of granting facilities to one person is defined as follows:

$$P = P(y = 1|X_1, \ldots, X_k) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

In which X_1, ...,X_k are independent variables. In addition the model can be rewritten according to odds ratio Logarithm as follows:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k (1)$$

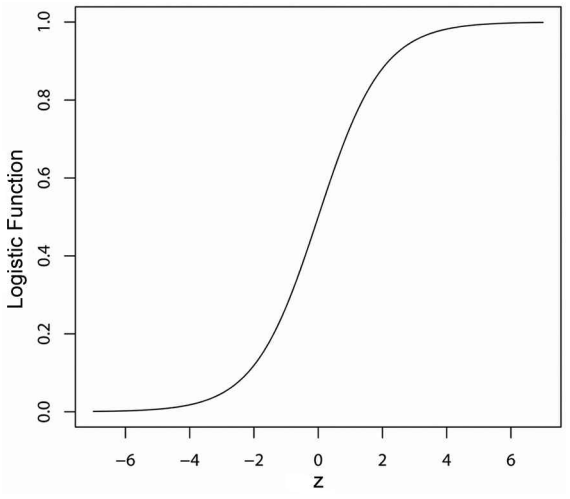

Figure 1. Logistics function

Where β_1,..., β_k are regression coefficients associated with each independent variable in the model and $\log(\frac{p}{1-p})$ is known as odd ratio logarithm. Finally, the calculated amount of probability by using logistic regression will be the decision criterion. This means that the higher the value for the applicants of these facilities, their credit to get these facilities is more and the lower the value the credit will be less.

Since the dependent variable in the model is only zero and one, the following equation can be considered for them.

$$P(Y_i = y) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

In which Yi is a binary variable. Likelihood function of the above density to estimate the unknown parameters are as follows:

$$L(y; p) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} (2)$$

With Placement of Pi from equation 1, in equation 2 we have:

$$L(y; p) = \prod_{i=1}^{n} \left(\frac{e^{\beta' x}}{1 + e^{\beta' x}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta' x}}\right)^{1-y_i}$$

Finally, to get the parameter estimation (regression coefficients) of the above function, use natural logarithm and solve the equation by derivative with respect to each parameter. Use analytical method to solve mentioned equation is impossible, so use numerical methods such as Newton-Raphson should be used to achieve the estimated value. We used R statistical software to fit the logistic regression model to mentioned data.

5. Main Results

Fitted regression model to predict customers' credit is as follows:

$y = \alpha + \beta_1 IN + \beta_2 AGE + \beta_3 EDU + \beta_4 GEN + \beta_5 JOB + \beta_6 BJOB + \beta_7 PRO$

where, IN stands for income, AGE stands for age, EDU Education, GEN Gender, JOB job applicant, BJOB sponsor's jobs, PRO applicants' Province and α is the intercept of model.

Estimation of regression coefficients of the above model is specified in *Table 1*.

Table 1. Regression coefficients

Coefficients	P-value	Regression Coefficient
Intercept	0.97	16.34
Income	0.0006	-0.69
Age	0.66	0.005
Education	0.46	0.12
Gender	0.98	-13.89
Job	0.000	-1.25
Sponsor's job	0.0009	1.25
Applicant's Province	0.084	-0.02

According to calculated P-value for each variable, it can be concluded that only factors of income, applicants' job and sponsors' job are statistically meaningful and other variables in the model can be abandoned. Therefore the regression model after removing insignificant variables is as *Table 2*.

Table 2. Regression coefficients for the significant variables

Coefficients	Regression Coefficient
Intercept	2.04
Income	-0.67
Job	-1.19
Sponsor's job	1.32

By using MacFadden statistic which is similar to coefficient of determination in linear regression we predict prediction

amount of dependent variable in logistic regression model which means the credit rating by the independent variables (MacFadden, 1973). Unlike determination coefficient amount in multiple linear regression which the more closer to one is better if the MacFadden statistical value be between 0.2 to 0.4, the fitted logistic regression model is very convenient and good and the values between 0.1 and 0.2 can also be trusted (Louviere et al. 2000). Of course, to review the accuracy of a model just using MacFadden statistic is not enough because the values of these statistic is very low in some models while the model is well fitted to the data. MacFadden statistic amount calculated for the first model (Model with all variables) equal to 0.11 and for the reduced model the value of this statistic is equal to 0.099 which represents good fitting on the data. One of the most popular and best criterion to verify the accuracy of a model is using the ROC curve. In this curve TPR rate is drawn against FPR rate. The greater the surface area under the curve it can be concluded that fitting model is more accurate. The largest area which the area under the curve gets is the amount one. *Figure 2* shows ROC curve for the models in our study.

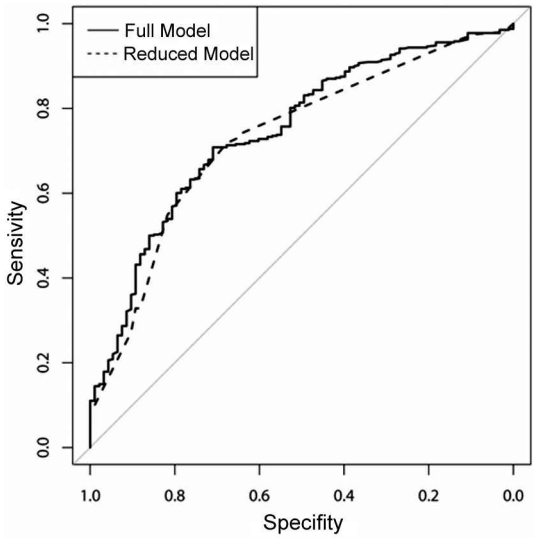

Figure 2.
ROC curve for the full model and the reduced model variables

The area under the ROC curve for the full fitted model is 0.74 and for modified models is equal to 0.72. So we can trust the results of the two models in anticipation of customers' credit. It should be noted that by removing variables that were not statistically significant by reducing the number of independent variables from seven to three, increase the accuracy and speed of calculations and the results will not have significant changes. There are several method to check the accuracy of a model which in this study, we use the most important ones to check the accuracy of the model.

6. Conclusion

Credit rating of loan applicants for banks and other financial institutions is of great importance. So today due to increase in loan applicants and because decision making about applicants in time consuming, introduction of mathematical models using traditional methods to perform this ranking is inevitable and vital. In this study, customer rating of Iran Khodro leasing company is done by using logistic regression. Based on the results of the fitted regression model, sponsor's job, applicant's jobs and income are effective factors to determine applicants' credit risk. The priority of these three variables is Sponsor's job, applicant's income and applicants' job. Based on fitted model, the more the value of calculated probability which is the response variable in the model is, the more is the credit of a person to get more facilities.

References

[1] Abdou, H. A. (2009), Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert Systems with Applications*, 36(9), 11402-11417.

[2] Agresti, A. (1990), *Categorical data analysis*. New York: Wiley.

[3] Altman, E.I., Haldeman, R.G., Narayanan, P. (1977), Zeta analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, 1, 29-54.

[4] Altman, E.I., Saunders, A. (1998), Credit risk measurement: Developments over the 20 years. *Journal of Banking and Finance*, 21, 1721-1744.

[5] Bryant, K. (2001), ALEES: An agricultural loan evaluation expert system. *Expert Systems with Applications*, 21, 75-85.

[6] Doumpos, M., Zopounidis, C. (2001), Assessing financial risks using a multi criteria sorting procedure: The case of country risk assessment. *Omega*, 29, 97-109.

[7] Etemadi, H., Rostamy, A.A.A., & Dehkordi, H.F. (2009), A genetic programming model for bankruptcy prediction: Empirical evidence from Iran. *Expert Systems with Applications*, 36(2), 3199-3207.

[8] Fraser, L.M. (1995), *Understanding Financial Statements*, 4th Edition. Prentice Hall, Englewood Cliffs, NJ.

[9] Fridson, M.S. (1995), *Financial Statement Analysis*, 2nd Edition. Wiley, New York, NY.

[10] Huang, J.J., Tzeng, G.H., & Ong, C.S. (2006), Two-stage genetic programming (2SGP) for the credit scoring model. *Applied Mathematics and Computation*, 174(2), 1039-1053.

[11] Khashman, A. (2010), Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9), 6233-6239.

[12] Kleinbaum, D.G., & Klein, M. (2010), *Logistic regression: a self-learning text*. Springer Science & Business Media.

[13] Koza, J.R. (1992), *Genetic programming: on the programming of computers by means of natural selection* (Vol. 1). MIT press

[14] Lee, T.S., Chiu, C.C., Lu, C.J., & Chen, I.F. (2002), Credit scoring using the hybrid neural discriminant technique. *Expert Systems with applications*, 3(3), 245-254.

[15] Louviere, J.J., Hensher, D.A., & Swait, J.D. (2000), *Stated choice methods: analysis and applications*. Cambridge University Press.

[16] Matsatsinis, N.F., Doumpos, M., Zopounidis, C. (1997), Knowledge acquisition and representation for expert systems in the field of financial analysis. *Expert Systems with Applications*, 12, 247-262.

[17] McFadden, D. (1973), Conditional logit analysis of qualitative choice behavior. In: *Frontiers in Econometrics*, ed. by P. Zarembka, Wiley, New York.

[18] O'Leary, D.E. (1995), AI in accounting, finance and management. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 4, 149-153.

[19] Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., & Feinstein, A.R. (1996), A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12), 1373-1379.

[20] Press, S.J., & Wilson, S. (1978), Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(4), 699-705.

[21] Sangster, A. (1995), The bank of Scotland's COMPASS – the future of bank lending? *Expert System with Applications*, 9, 457-468.

[22] Smith, L.M., McDuffie, R.S. (1996), Using an expert system to teach accounting for business combinations. *Expert System with Applications*, 10, 181-191.

[23] Strischek, D. (1999), A written policy for lending to contractors. The *Journal of Lending & Credit Risk Management*, 81, 32-42.

[24] Tam, M.C.Y., Tummala, V.M.R. (2001), An application in vendor selection of a telecommunications system. *Omega*, 29, 171-182.

[25] Todd, M., Kennedy, R., Fried, C. (1998), Ten steps to better credit-scoring. *The Journal of Lending & Credit Risk Management*, 81, 54-59.

[26] West, D. (2000), Neural network credit scoring models. *Computers & Operations Research*, 27(11), 1131-1152.

[27] Yu, L., Wang, S., & Lai, K.K. (2008), Credit risk assessment with a multistage neural network ensemble learning approach. *Expert systems with applications*, 34(2), 1434-1444.