International Journal of Disaster Recovery and Business Continuity
Vol.11, No. 1 (2020), pp. 302-310

Credit Risk assessment Model For Maskan Bank Legal Customers Using Logistic Regression

Sobhan Hoosein Beigy [1], Yasanolah Poor Ashraf [2*], Karam Khalili [3]
1- Ph.D. Candidate, Department of Financial Management, Ilam Branch, Islamic Azad University, Ilam, Iran, Email: sobhanho@gmail.com
2- Associate Professor, Department of Planning & Management, Ilam, Iran, Email: Yasan_ashraf@yahoo.com
3- Assistant Professor, Ilam Branch, Islamic Azad University, Ilam, Iran, Email: karam.khalili@yahoo.com

Abstract

Providing credit facilities to clients can be regarded as one of the most important tasks of banks. Banks in each country, after collecting financial resources, allocate these resources to different economic sectors. In fact, this action by the banks will strengthen the various economic sectors in each country to perform better their duties, and ultimately provide the necessary background for the country's economic growth and development. If the banks can do this, it is important to properly allocate financial resources to eligible customers. Proper allocation of financial resources, while achieving the above objective, will provide the necessary ground for the continued life of the banks. In this case, it is important to correctly identify the risk-averse customers before granting them facilities in order to enhance the effectiveness of the decisions taken. Estimating that a company will go bankrupt in the future is very important for facilitators and creditors, so finding the model that best fits the companies has always been a concern.

Method: *In this study, credit risk assessment of legal clients of Maskan Bank was investigated by using logistic regression and feature selection by genetic algorithm. It is noteworthy that for the dataset, a normalization method was used (ratio of distance from mean to data standard deviation) and the results were obtained based on both normalized and abnormal data sets to determine the impact of data normalization on the data set to get the right prediction percentage from customers.*

Results: *In this study, the regression coefficients on the Maskan Bank dataset are calculated based on the logistic regression model using IVIOS software, and then the prediction of correct results based on the logistic regression will be obtained in the MATLAB mathematical software. In addition, based on feature selection with genetic algorithm, the results of logistic regression are optimized. Most of the work done in this field by logistic regression did not have a prediction percentage above%80 and this prediction method was lower than other prediction methods in the lower classes. In this study, we have obtained the correct prediction percentage of %94.8 based on the use of appropriate features collected from customers in the Maskan Bankk dataset by using feature selection through genetic algorithm.*

Keywords: *Credite Classification of Bank Clents, Logistic Regression, Feature Selection, Genetic Algorithm.*

ISSN: 2005-4289 IJDRBC

1. Introduction

The banking system is considered as one of the main pillars of any economic system. Banks and credit and financial institutions play an important role in monetary policy implementation. Therefore, their proper and principled functioning can significantly contribute to the economic growth and prosperity of society (Rajabzadeh Moghani et al., 2015). Hence, a healthy, profitable system can play a stronger role in the stability of the financial system and resist economic shocks.

Most experts believe that banking resources are mainly used for granting credit and the bank main interests are pursued in this way. Banking operations deal with various risks due to the characteristics of banking activities. Moreover, the bank is the main lending (credit-granting) institution that has exposed banking activities to credit risk. Thus, banks pay particular attention to the discussion of dynamic risk management (DRM) and the design of internal models for risk management, as well as to tailor various structures and organizations for optimal risk management in banks. Credit risk is one of the main risks to which banks are exposed. Lack of proper management and control of this risk causes the bank to go through crisis and bankruptcy. These crises disrupt the entire economic and social system as the bank is an influential institution in the economic system of any country (Mirghafouri & Ashuri, 2015).

Credit risk is the risk, based on which the loan is repaid with a delay or not repaid at all, which disrupts banks' cash flow and has a negative impact on the liquidity and return on investment (ROI) of the bank. To control and mitigate credit risk, the bank must properly identify its credit facility applicants and differentiate between low-risk applicants capable of timely loan repayment and high-risk applicants, which can be managed through effective credit risk management (Chen et al. 2012).

Therefore, an ounce of prevention is worth a pound of cure in granting banking facilities because the costs of prevention are always lower than those of treatment. Prior to granting facilities, if banks and financial institutions properly accredit and rate the facility applicants and also grant the facilities correctly, the facility granting costs would definitely be lower than the costs involved in deferred claims and debt collection. Resources are wasted if clients are not accredited and there is no supervision and control. One of the causes of inflation is that resources have a double-edged function rather than providing the working capital of a production unit. That is, they cause an increase in demand and ultimately an increase in prices rather than an increase in supply and ultimately a decrease in prices. Thus, due to their limited financial resources, banks should strive to allocate these limited resources optimally for greater profitability to tap into the manufacturing and service sectors in society (Jalili et al., 2010).

Given the above, there seems to be no coherent and organized move towards establishing credit risk models in Iran, despite the importance of credit risk in banking and financial institutions. For example, on the one hand, the absence of credit risk indicators and the institutions that rate them is clearly felt in the Iranian financial markets. On the other hand, a consistent and orderly trend has not been observed in determining credit risk and ratings and credit ceilings based on risk indices in granting credit facilities to clients. Experts and the Credit Committee are currently appointing them. In this case, not only does the existence of an efficient risk model facilitate the decision-making process for granting credits but it also provides an

ISSN: 2005-4289 IJDRBC

efficient paradigm for allocating capital to different economic sectors for the banking system and subsequently the country itself. In this regard, this paper mainly aims to present a model to provide cover for credit risk in Bank Maskan by applying the logistic regression prediction method. This research will be conducted to meet Bank Maskan's research needs and its results will be used to enhance the efficiency of its credit risk coverage system.

2. Theoretical Foundations and Research Background

The history of the linear regression model goes back to when Fitzpatrick (1932) conducted studies on predicting a client's inability to repay a loan as one of the most important and controversial issues in finance. This issue has become an important area of theoretical and empirical research in the field of financial economics over the last 80 years.

In recent years, neural network models have gained a prominent place in comparison with some classical models such as auditory analysis in estimation (Cooper, 1999). These models can be considered as a generalization of nonlinear regression models. Hence, they are widely used by researchers to identify trends and patterns in data as well as generate knowledge from data.

Salchenberger et al. (1992) used multilayer perceptron (MLP) neural networks to predict the financial health of savings and loans. They compared this method with the logit regression model. They studied S&L data from January 1986 to December 1987 and demonstrated the superiority of the neural network over the logit model (Salchenberger et al., 1992).

Other methods used to solve classification and credit risk problems are the decision tree method and the support vector machine (SVM). Liu et al. (2010) used GA-optimized SVM for credit scoring (Liu et al., 2010).

In his thesis entitled "Designing a Credit Rating Model Based on Inter-Company and Industry Variables Using Artificial Neural Networks (ANNs)", Sahandi (2010) has measured the accuracy of the neural network model in identifying creditworthy and uncreditworthy clients by a neural network. According to the results, the neural network performed better than logistic regression, with the classification of about 89% of uncreditworthy clients and 85% of creditworthy clients and about 83% of uncreditworthy clients and 82% of creditworthy clients, respectively (Sahandi, 2010).

(Table 1) lists some of the work done in this area related to our work.

ISSN: 2005-4289 IJDRBC

Table 1. Some credit risk studies

Name	Year	Title	Result
Akhbari and Rafi'i	2010	Providing a fuzzy neural model for the credit rating of banks' legal clients	Their model considers debt ratio, activity ratio, and equity-to-total assets ratio as input variables and the client's probability of default (PD) as output variables. After training and testing the model based on Bank Keshavarzi data from 2001 to 2006, the proposed model predicted the client's credit status with an accuracy of 69.36%.
Dehmardeh et al.	2012	Bank customer validation using credit scoring approach at Bank Sepah branches in Zahedan	According to estimation results, based on statistical indices, logistic regression is significant with respect to coefficients, with high resolution and credit in bank credit risk management.
Gong Dong et al.	2010	A logit regression model with random coefficients to generate credit scorecard	According to the experimental results, the proposed model can improve the prediction accuracy of the logit regression model with constant coefficients without eliminating its desirable features.
Tabagari	2015	Credit rating using logistic regression	In this model, the values of the parameters are not absolute and are usually measured in relative terms. In this work, 16 independent variables are considered. Important variables include age, debt repayment-loan, life span in the current location, type of job, amount of credit, other debt.

3. Logistic Regression Model

This model is a special case of regression models with discrete dependent variables. The simplest of these models are those in which the dependent variable *yi* chooses binary values 0 and 1 without loss of generality.

ISSN: 2005-4289 IJDRBC

Logistic regression can be considered as a special case of the general linear model and linear regression. The logistic regression model is based on a series of assumptions (about the relationship between the dependent and independent variables) completely different from linear regression. The important difference between these two models lies in the two features of logistic regression. First, the conditional distribution Y|x is a Bernoulli distribution rather than a Gaussian (normal) distribution because the dependent variable is binary. Second, the prediction values are probabilistic whose ranges are obtained in the interval [0, 1] by the logit distribution function. Logistic regression predicts the output probability.

Logit Distribution Function

There are different statistical methods for estimating credit rating models. The general form of the proposed model is given by Equation 1:

$$Y = F(X1, X2, X3, \ldots Xn) \quad (1)$$

where Y is the response variable that determines the credit applicant's status which is discrete because the bank's clients are divided into two categories: creditworthy (0) and uncreditworthy (1). Now consider Equation 2:

$$E(Y = 0 \mid X_i) = \frac{1}{1+e^{-z}} = p_i \quad (2)$$

where *e* is the base of the natural logarithm and $Z_i = B_1 + B_2 X_i$. The above equation represents what is called the "cumulative logistic distribution function." P_i is the probability of timely loan repayment by the client. In this case, the ratio of P_i to (1- P_i), i.e., the probability of non-repayment of the loan, is Equation 3:

$$\frac{P_i}{1-P_i} = e^z \quad (3)$$

By taking the natural logarithm of Equation 3, we obtain Equation 4:

$$Ln(\frac{P_i}{1-P_i}) = Z_i = B_1 + B_2 X_i \quad (4)$$

The logarithm of the "odds ratio" is linear with respect to the parameters obtained by estimating the coefficients using the maximum likelihood estimation (MLE) method. Thus, the coefficient estimation yields a model each of the coefficients of which indicates the value of the variable (i.e., the logarithm of the odds to the benefit of default) for a unit of change in the independent variable. Bank clients can be classified by calculating the PD using Equation 2.

4. Bank Maskan Dataset Details

This dataset covers 6014 Bank Maskan legal clients with the most bank claims in 2018, consisting of 21 features. A "Good Client" or "Bad Client" label was assigned to each client based on past due, deferred,

ISSN: 2005-4289 IJDRBC

and doubtful claims related to that client. A label with a value of 0 was used for a good client and a label with a value of 1 was used for a bad client. A data normalization operation is performed for this data so that it is placed within the same range. A code was written in MATLAB to apply data normalization to the Bank Maskan dataset. The code written to normalize the data first calculates the average data for each feature and then obtains the standard deviation of the data for each. The normalized value for a feature is obtained from a particular client as the ratio of the distance between data and the average of that feature to the standard deviation of that feature. This normalization method is used when population parameters are known.

Bank Maskan client data features include:

- Lender Branch Management Code (a natural number within the range 65-264)
- Lender Branch Code (a four-digit natural number)
- Loan Amount (in million Tomans)
- The time from the contract expiry date to debt maturity (in months)
- Borrower's activity history (in months)
- The history of client interaction with the bank (in months)
- Average client bank account balance (in million USD)
- Client financial ratios (a number within the range 1-200)
- Obligation Fulfillment History (a number within the range 1-200)
- Vision (a number within the range 1-180)
- Guarantee amount (Guarantee amount/loan ratio, a number within the range 1-180)
- Capital to Amount Ratio (a number within the range 1-180)
- Bank Credit Unit Comment (a number within the range 1-150)
- Branch Score (a number within the range 1-1000)
- Branch Rate (a natural number between 1 and 6)
- Client Rate (number within the range 1-1000)
- Client Rate (a natural number between 1 and 6)
- Loan renewal frequency (a natural number between 1 and 11)
- Past Due Amount (in million USD)
- Deferred Claims (in million USD)
- Doubtful Claims (in million USD)

5. Data Analysis and Findings

EViews is used to obtain logistic regression coefficients. A non-time work file with 6014 observations was created in EViews for the Bank Maskan client dataset. Then, the logistic regression can be calculated from the “Equation Estimation” menu. In the Equation Estimation menu, we select the binary estimation method; then, in the window that opens, we select the “Logit” option. (Figure 1) shows the logistic regression results in EViews for abnormal data. The coefficients matrix is obtained once again based on the normalized data.

Wald statistic is used to test the significance of logistic regression coefficients. The null hypothesis means that the variable in question has no effect on the dependent variable. As shown in Figure 1, the values of the “Prob” column represent the value or probability to accept or reject the H0. According to Figure 1, the

ISSN: 2005-4289 IJDRBC

null hypothesis would be rejected for variables whose value is less than 0.05, and the coefficients obtained will be significant.

Dependent Variable: Y
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 05/25/19 Time: 20:51
Sample: 1 6014
Included observations: 6014
Convergence achieved after 7 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-6.568360	2.315066	-2.837224	0.0046
S1	0.015684	0.003886	4.035497	0.0001
S2	4.10E-05	6.99E-05	0.587232	0.5570
S3	3.48E-05	1.63E-05	2.130471	0.0331
S4	0.047581	0.009690	4.910076	0.0000
S5	0.010096	0.005500	1.835622	0.0664
S6	0.015205	0.005646	2.693164	0.0071
S7	-0.006770	0.003440	-1.968259	0.0490
S8	0.007350	0.003971	1.850824	0.0642
S9	0.001300	0.002374	0.547504	0.5840
S10	-0.002495	0.005407	-0.461485	0.6445
S11	0.001142	0.002871	0.397849	0.6907
S12	0.002796	0.002964	0.943244	0.3456
S13	0.001037	0.003251	0.318936	0.7498
S14	-0.003918	0.001694	-2.312119	0.0208
S15	0.048140	0.113852	0.422832	0.6724
S16	-0.001062	0.001930	-0.550343	0.5821
S17	-0.135918	0.189387	-0.717672	0.4730
S18	0.743970	0.036644	20.30241	0.0000

McFadden R-squared	0.221346	Mean dependent var	0.051380
S.D. dependent var	0.220790	S.E. of regression	0.204906
Akaike info criterion	0.321765	Sum squared resid	251.7097
Schwarz criterion	0.342938	Log likelihood	-948.5481
Hannan-Quinn criter.	0.329117	Deviance	1897.096
Restr. deviance	2436.380	Restr. log likelihood	-1218.190
LR statistic	539.2837	Avg. log likelihood	-0.157723
Prob(LR statistic)	0.000000		

Obs with Dep=0	5705	Total obs	6014
Obs with Dep=1	309		

Figure 1. Results of the logistic regression model for the Bank Maskan client dataset

For classification based on the obtained coefficients matrix, clients are classified from abnormal and normal data modes and modes of total features and omitted features.

For classification by logistic regression method, classification can be done based on Equation 4 as well as coefficients obtained from logistic regression in EViews. To classify using MATLAB, a code has been written to execute the classification process that takes the coefficients matrix from EViews and holds the bank client dataset as a matrix. After calculating P_i for each client and comparing the prediction made with the actual label of that client, it is determined what percentage of clients is correctly classified. If P_i is less than 0.5, the client falls into the category labeled 0, meaning he/she is a creditworthy client, and if it is greater than 0.5, the client falls into the category labeled 1, meaning he/she is an uncreditworthy client.

ISSN: 2005-4289 IJDRBC

(Table 2) presents the results predicted by the logistic regression model.

Table 2. Client Classification (Categorization) Results for the Logistic Regression Model

Percentage of correct predictions	Abnormal data	Normal data
Total 18 Bank Maskan dataset features	94.1	94.7

Next, the prediction results will be optimized using a feature selection method. In this method, the prediction is used by logistic regression as a fitness function and GA as the optimal solution search algorithm. To do so, the proposed feature selection scheme will be coded in MATLAB.

The proposed feature selection algorithm runtime for the Bank Maskan client dataset with 100 runs of the algorithm is about 40 seconds.

(Table 3) presents the results of the feature selection implementation.

Table 3. Client Classification Results for Logistic Regression Model Using Feature Selection

Percentage of correct predictions	Abnormal data	Normal data
Prediction percentage by applying feature selection	94.8	94.8
Number of selected features	3, 8, 10, 11, 13, 15	1, 2, 6, 10, 11, 13, 15, 16, 17

6. Discussion and Conclusion

This study assessed the credit risk of Bank Maskan's legal clients by applying logistic regression and feature selection using a genetic algorithm (GA). It is worth noting that a normalization method (i.e., the ratio of the distance from the mean to the standard deviation of the data) was used for the dataset. Then, the results were obtained based on both datasets, i.e., normalized and abnormal, to obtain the effect of data normalization on the percentage of correct prediction of clients. In this study, first, the regression coefficients on the bank dataset are obtained based on the logistic regression model using EViews. Then, the percentage of correct predictions results are obtained based on logistic regression in MATLAB based on the relations described in Section 3 of this paper. Then, the results of logistic regression are optimized based on feature selection with GA. According to the results, a very high prediction rate is obtained for this dataset. The prediction percentage of most of the work done in this area was not more than 80% using logistic regression. Hence, this prediction method is ranked lower than other prediction methods. In this study, a correct prediction percentage 94.8 was obtained as a result of using appropriate features collected from clients in the Bank Maskan dataset as well as feature selection using GA.

ISSN: 2005-4289 IJDRBC

References

[1] Ahmadian Yazdi, Farzaneh, Ebrahimi Salari, Taghi, Jandaghi, Fereshteh, and Rajabzadeh Moghani, Nahid (2015). "Investigating Factors Influencing Human Capital Accumulation in Iran During 1971-2012," *Journal of Applied Economics Studies in Iran (AES)*, Vol. 4, No. 15, Fall 2015, pp. 201-228.

[2] Akhbari, Mahdieh, and Mokhatab Rafi'i, Farimah (2010). "Application of Neuro-Fuzzy Reasoning Systems to Credit Ratings of Banks' Legal Clients," *Journal of Economic Research*, Vol. 45, No. 3.

[3] Jalili, Mohammad, Khodaei Valeh Zaghred, Mohammad, and Kaneshloo, Mahdieh (2010). "Validating Real Clients in the Country's Banking System," *Journal of Quantitative Studies in Management*, No. 3, pp. 127-148.

[4] Dehmardeh, Mahboubeh, Yaghoobi, Nour Mohammad, and Shokri, Ali (2012). "Studying the Structural Empowerment of Organizational Agility in the Banking System," *Journal of Thought on Strategic Management*, Vol. 6, No. 1, pp. 133-158.

[5] Mirghafouri, Seyed Habibullah, and Amin Ashuri, Zohreh (2015). "Credit Risk Assessment of Banks Clients," *Journal of Business Management Research*, Vol. 7, No. 13, pp. 147-166.

[6] Chen, C. and Lu, H. andSougiannis, T. (2012). The Agency Problem, Corporate Governance, and the Asymmetrical Behavior of Selling, General, and Administrative Costs. Contemporary Accounting Research, 29 (1).

[7] Cooper, M.T., Bray, S.J. (1999). Frizzled regulation of Notch signalling polarizes cell fate in the Drosophila eye. Nature 397(6719): 526--530.

[8] Dong, G. etal. (2010). Credit scorecard based on logistic regression with random coefficients, Journal of Procedia Computer Science, 1: 2463-2468.

[9] Liu, J.; Siu, O. & Shi, K. (2010). "Transformational leadership and employee well being: The mediating role of trust in the leader and self-efficacy", APPlied Psychology, 59(3), PP: 454-479.

[10] Salchenberger LM, Cinar EM, Lash NA, (1992), "Neural networks: a new tool for predicting thrift failures". Decision Sciences, Volume 23, Issue 4, pages 899–916.

[11] Tabagari, S. (2015), "Credit Scoring by Logistic Regression", (MS), Uuniversity of Tartu.

ISSN: 2005-4289 IJDRBC