Research Journal of Applied Sciences, Engineering and Technology 9(10): 856-861, 2015
DOI:10.19026/rjaset.9.2635
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: November 19, 2014 Accepted: January 8, 2015 Published: April 05, 2015

Research Article

A Multi-industry Default Prediction Model using Logistic Regression and Decision Tree

[1]Suresh Ramakrishnan, [1]Maryam Mirzaei and [2]Mahmoud Bekri
[1]Faculty of Management, Universiti Teknologi Malaysia, Malaysia
[2]Economic and Statistic Institute, Karlsruhe Institute of Technology, Germany

Abstract: The accurate prediction of corporate bankruptcy for the firms in different industries is of a great concern to investors and creditors, as the reduction of creditors' risk and a considerable amount of saving for an industry economy can be possible. Financial statements vary between industries. Therefore, economic intuition suggests that industry effects should be an important component in bankruptcy prediction. This study attempts to detail the characteristics of each industry using sector indicators. The results show significant relationship between probability of default and sector indicators. The results of this study may improve the default prediction models performance and reduce the costs of risk management.

Keywords: Decision tree, default prediction, industry effects, logistic regression

INTRODUCTION

Prediction of corporate default is of a great concern to investors/creditors, borrowing firms and governments. During last two decades, the world has experienced a large number of financial crisis in emerging market economies of Latin America and Asia during 1994-1998 and the recent crisis in USA due to the sub-prime mortgage during 2008. These financial crisis were not confined individual economy, but affected directly or indirectly almost all the countries of the world. As a result, many voices have called for a revolution of existing default warning system to detect or prevent default problems in real time. Any organization may face default due to the competition and uncertainty which is observed increasingly in the business environment. An improvement in model accuracy in the default likelihood assessment leads to enormous future savings for the credit industry. Therefore, various profits such as cost decline in credit analysis, an increased debt collection rate and better monitoring attain as of accurate default prediction.

Review of literature on the subject confirmed hand full of studies conducted in the last four decades. Despite of these studies, the recent credit crisis indicated that yet there are areas of the study that needs researchers' attention. Moreover, emerging of the regulatory changes such as Basel III accord and the need for more precise and comprehensive risk management procedures justifies need of research in area of credit risk modelling and banking supervision. This requirement like these pushes companies especially banks and insurance companies to have a very robust and transparent risk management system.

Since the study of Fitz (1932), default prediction becomes a challenging issue in corporate finance. A number of default prediction models have developed extremely due to the emergent accessibility of data and the improvement of econometrical methods during the 1980s and 1990s. Most of this study has been persuasively directed by a small number of early studies (Beaver, 1966; Altman, 1968; Ohlson, 1980; Zavgren, 1985) on US extracted companies. Earlier, most of the studies on default risk focused on firm-specific indicators as a predictor of firms default across United States including (Courtis, 1978; Deakin, 1972; Jackendoff, 1962; Merwin, 1942; Meyer and Pifer, 1970; Smith and Winakor, 1935). Although, majority of the studies used the firm-specific variables, some researchers tried to use some other indicators such as interest rate, stock index return and GDP that affects default prediction. As a result of relationship between general economic and bankruptcy rates, some attempts have been made to predict default based on macroeconomic variables.

Over the years, a large strand of research on default prediction remained restricted to firm-level factors. Based on the surveys of the literature, not much attention has been paid to industry effects. Yet, there are some reasons which represent the importance of industry effects on default prediction. It is plausible that probability of default can differ for firms in different industries due to different levels of competition amongst various industries. Different industries may have different accounting principles, involving that the probability of default can vary for firms in different industries with otherwise the same balance sheets. Keeping in view the importance of external

Corresponding Author: Maryam Mirzaei, Faculty of Management, Universiti Teknologi Malaysia, Malaysia

environmental factors, little attention has so far been paid to sectors and industry factors. Recent developments in the literature of default prediction have highlighted the importance of the effects of industry factor. In this regard, Lang and Stulz (1992) argued that sectors have distinctive nature and need to be intensively explored. Accordingly, related researches stress the need to examine the industry's behavioral effect on firm's default and support the importance of industry effects on default (Opler and Titman, 1994).

These studies on default prediction employ dummy variable to control the industry effect on default. We attempt to detail the characteristics of each industry, following the Kayo and Kimura's (2011) approach that justifies the characteristics influencing leverage. According to Gianneth (2003), firms in sectors with highly volatile returns are more likely to default due to temporary illiquidity; longer debt can help reducing inefficiencies in these sectors. In order to capture the more realistic effect of sector or industry on default prediction, this study employs munificence, dynamism and firm's concentration of an industry.

This study contributes to the default prediction literature by outlining a procedure to be used by banks to assess the likelihood for borrower default. Rather than focusing on financial measures which may be backward looking, this study investigates three industry factors including: munificence, dynamism and HHIndex as part of mechanism for selecting potentially distressed firms. Thus, this study intends to fill this gap comparing different methods including logistic regression, decision trees.

Variable selection: In default prediction, the most important concern of interest among researchers is to construct the prediction model which characterizes the association between the default and financial ratios and then deploy the model to identify the high risk of default in the future. A large number of characteristics are usually incorporated so that the training data is not enough to cover the decision space, which is represented as the curse of dimensionality. Feature selection represents the problem by excluding unimportant, redundant and correlated features in order to increase the accuracy and simplicity of classification model, reducing the computational effort and enhancing the use of models. The representative features for default prediction can be presented as follows:

Profitability: Profit before interest and tax/Total assets.

Size: Natural logarithm of sales.

Tangibility: Ratio of fixed assets to total assets.

In order to capture the more realistic effect of sector or industry on corporate default prediction, this study employed three variables at industry level including: munificence, dynamism and Herfindhal-Hirschman Index (HHIndex). The first two variables are derived from the model of Dess and Beard (1984), known as multidimensional model of environment. This model so far has been used in the context of corporate strategies. Consequently, effects of industry specific properties on bankruptcy prediction of firm have been analysed by Kayo and Kimura (2011).

Munificence: The ability of an atmosphere to preserve a constant expansion is called munificence (Dess and Beard, 1984). The sectors/ industries operating in normal environment with high munificence tend to have larger level of opportunities as compared to industries with low munificence (Almazan and Molina, 2005).

Dynamism: Generally, the environmental dynamism describes the rate and instability of changes in firm's external environment (Dess and Beard, 1984; Simerly and Li, 2000).

HHIndex: On the basis of industry concentration, it can be divided as high and low concentrated industries. The level of industry concentration affects the firm leverage differently.

METHODOLOGY

Logistic regression: Logistic regression is a type of regression methods (Allison, 2001; Hosmer and Lemeshow, 2000) where the dependent variable is discrete or categorical, for instance, default (1) and non-default (0). Logistic regression examines the effect of multiple independent variables to forecast the association between them and dependent variable categories. According to Morris (1997) and Martin (1977) was the first researcher who used logistic technique in corporate default perspective. He employed this technique to examine failures in the U.S. banking sector. Subsequently, Ohlson (1980) applied logistic regression more generally to a sample of 105 bankrupt firm and 2,000 non-bankrupt companies. His model did not discriminate between failed and non-failed companies as well as the Multiple Discriminant Analysis (MDA) models reported in previous studies. According to Dimitras *et al*. (1996), logistic regression is in the second place, after MDA, in default prediction models. This method creates a score for every observation's dependent variable based on its independent pointers' weights. This score demonstrates the likelihood of membership in the objective category. For instance, the following equation can be used for default prediction model:

$$Probability\ (Default|X_1, \ldots X_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_t x_t)}} \quad (1)$$

where, Probability (Default $|X_1, \ldots, X_i$) is the probability of default, X_i (i = 1,…, n) are independent variables such as firm-specific variables and β_1 to β_i are coefficients which have estimated by the model. This model can be explained as the probability of default based on firm's given characteristics. In this model, maximum probability function is applied. In this regard, the weights are employed to make best use of the probability of default for the identified failed companies and the probability of non-default for non-failed companies. Thus, based on this technique, using a broken-off point, a firm is classified as failed or non-failed. Logistic regression is also able to verify the significance of individual variables in the model (Allison, 2001; Hosmer and Lemeshow, 2000).

Decision tree: Decision trees are the most popular and powerful techniques for classification and prediction. The foremost cause behind their recognition is their simplicity and transparency and consequently relative improvement in terms of interpretability. Decision tree is a non-parametric and introductory technique, which is capable to learn from examples by a procedure of simplification. Frydman *et al*. (1985) first time employed decision trees to forecast default. Soon after, some researchers applied this technique to predict default and bankruptcy including (Carter and Catlett, 1987; Gepp *et al*., 2010; Messier and Hansen, 1988; Pompe and Feelders, 1997).

Decision trees allocate data to predefined classification groups. For instance, in terms of business default prediction, this technique assigns each firm to a failed or non-failed group. Decision tree is a non-parametric and introductory technique, which is capable to learn from examples by a procedure of simplification. Generally, decision trees are binary trees include a set of branches (paths from roots to leaf nodes), leaf nodes (objects classes) and nodes (decision rules) which classifies objects according to their attributes (Dimitras *et al*., 1996). Therefore, the decision tree takes the form of top-down term structure, which divides the data to generate leaves. Under the structure, one target class is central and each record flows through the tree along a path determined by a series of tests until it obtains a terminal node (Quinlan, 1986).

There are two types of decision tree models, regression trees when the response variable is continuous and classification trees when the response variable is quantitative discrete or categorical. There are various algorithms to make decision trees which the most popular are C4.5 and CART. The main advantage of decision tree is that there is no restrict statistical requirement such as normality for dataset as this technique is a non-parametric method. Also due to the simplicity of the model, this technique became so popular and easy to use for the purpose of classification.

EMPIRICAL RESULTS

Data description: The dataset was used to classify a set of firms into those that would default and those that would not default on loan payments. It consists of 285 observations of Malaysian companies during 2007-2012 from four different sectors including: trading and services, manufacturing sectors (consumer product and industrial product) and Construction and Property Sector. Of the 147 cases for training, 67 belong to the default case under the requirements of PN4, PN17 and Amended PN17 respectively and the other 201 to non-default case. Consulting an extensive review of existing literature on corporate default models, the most common financial ratios that are examined by various studies were identified. The variable selection procedure should be largely based on the existing theory. The field of default prediction, however, suffers from a lack of agreement as for which variables should be used. The first step in this empirical search for the best model is therefore the correlation analysis. If high correlation is detected, the most commonly used and best performing ratios in the literature are prioritized. Therefore, the choice of variables entering the models is made by looking at the significance of ratios.

The components of the financial ratios which are estimated from data are explained below and Table 1 shows the summary statistics for selected variables for default and non-default firms. The most significant variables based on two methods were identified. These variables selected from the significant indicators for the model which could best discriminate the default firms from the non-default firms.

Logistic regression to model default prediction: As shown in Table 2, five independent variables made statistically significant contribution to the model. The independent variables are size, profitability, liquidity, munificence and HHIndex. This is based on the Wald test that shows the contribution of each of the predictor or independent variables to a model. Its interpretation is similar to the F or t values for the significance testing of regression coefficients (Hair *et al*., 2006). Variables that contribute significantly to the models should have significance value of less than 0.05 (Pallant, 2007). A remarkable result specifies a predictor that is faithfully associated with the outcome (Tabachnick and Fidell, 2007). The B coefficient value is shown in Table 2 for each significant determinant.

Therefore, based on Table 2, the equation for the different sectors using financial ratios and sector indicators is:

Table 1: Summary statistics for selected variables

		Non-default firms		Defaulted firms	
Variable	Formulation	Mean	S.D.	Mean	S.D.
Liquidity	Current assets/current liabilities				
Profitability	Profit before interest and tax/total assets	7.61	6.65	5.07	10.06
Tangibility	Ratio of fixed assets to total assets	8.72	2.15	6.87	0.56
Size	Natural logarithm of sale	5.56	3.32	5.98	1.65
Munificence	• Regressing time against sales of an industry over the period of study • Taking the ratio of the regression slope coefficient to the mean value of sales over the same period	0.82	0.93	0.81	0.92
Dynamism	Standard error of the munificence regression slope co-efficient divided by the mean value of sales over the study period	1.82	1.71	1.31	1.73
HHI	The HHI is calculated by the summing of squares of each firm's market share within the industry	0.67	0.46	0.44	0.32

S.D.: Standard deviation

Table 2: Estimation results of logistic regression

IV	B	S.E.	Wald	df	p-value
Constant	-3.930	0.294	7.2530	1	0.020
Size	0.001	0.000	15.234	1	0.000
Profitability	0.003	0.001	11.159	1	0.001
Liquidity	0.001	0.002	10.399	1	0.001
Munificence	0.001	0.000	4.272	1	0.039
HHI	-0.279	0.089	9.888	1	0.002

S.E.: Standard error

Table 3: Detailed accuracy by class

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.822	0.121	0.796	0.822	0.809	0.896	Y
0.879	0.178	0.896	0.879	0.888	0.895	N
Weighted avg.						
0.858	0.157	0.859	0.858	0.859	0.895	

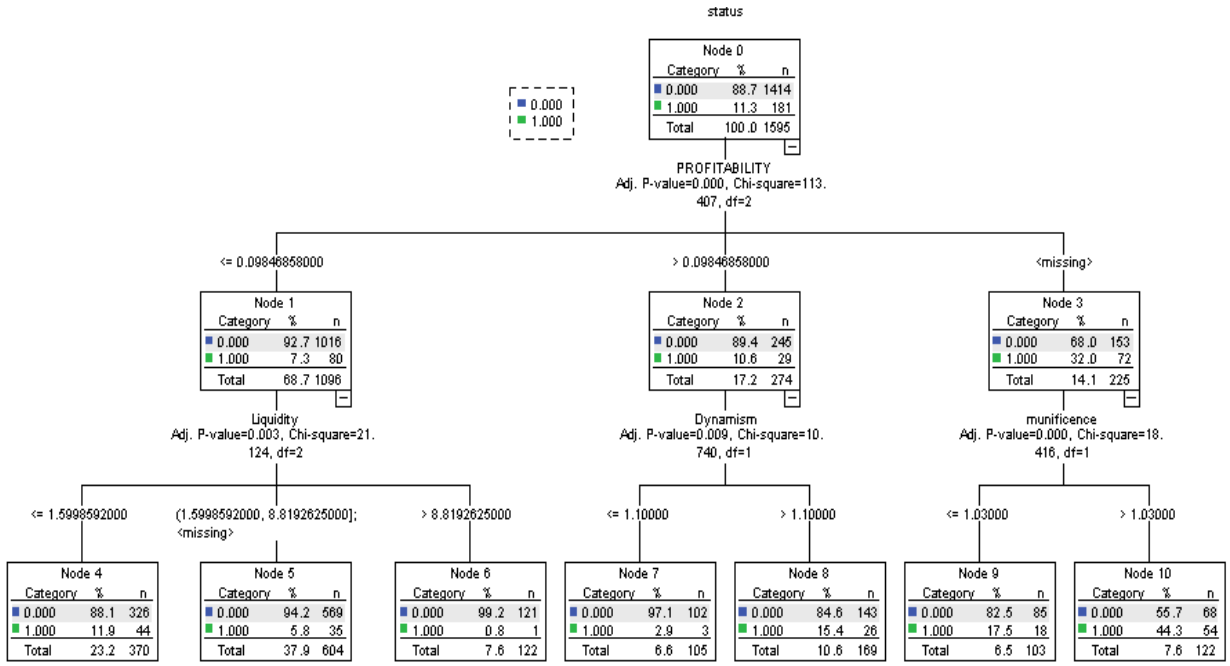

Fig. 1: Decision tree to model default prediction

$$P = \frac{1}{1+e^{-(-3.930+0.001X_1+0.003X_2+0.001X_3+0.001X_4-0.279X_5)}} \quad (2)$$

where,

X_1 = Size
X_2 = Profitability
X_3 = Liquidity
X_4 = Munificence
X_5 = HHI

The findings show that sector variables in corporate with financial ratios can be used to predict corporate default among different industries in Malaysia.

Decision tree to model default prediction: The tree diagram is a graphic representation of the tree model. This tree diagram shows that:

- Using the CHAID method, profitability is the best predictor of firm's default.
- For the low profitability category, the next best predictor is liquidity. Of the firms in this category, only 7.3% have defaulted on loans.
- For the high profitability category, the model includes one more predictor: dynamism. About 15% of those firms with the value more than 1.1 of dynamism have defaulted on loans (Fig. 1).

The decision tree model shows an accuracy about 85.83% and mean absolute error about 0.16 (Table 3):

Correctly Classified Instances	85.8268%
Incorrectly Classified Instances	14.1732%
Kappa statistic	0.6961
Mean absolute error	0.1583
Root mean squared error	0.3565

DISCUSSION AND CONCLUSION

Default prediction takes an important role in the prevention of corporate default, which makes the accuracy of default prediction model be widely concerned by researchers. Appropriate identification of firms 'approaching default is undeniably required. There is a large volume of published studies describing the role of firm- specific factors in default prediction models and during the past 40 years, the use of firm-specific variables in default prediction models has been subject of many studies. It is evoked (implied) by researches that there exists significant relationship between default prediction and firm specific variables. The results of this study supports the literature on default prediction. According to the results, financial ratios such as liquidity and profitability affects the probability of default significantly.

Although the main part of default prediction literature across developed and developing economies focused on firm specific and macroeconomic indicators. However, a number of studies on default prediction highlighted the importance of industry on default prediction. These studies on default prediction employ dummy variable to control the industry effect on default. We attempt to detail the characteristics of each industry, following the Kayo and Kimura's (2011) approach that justify the characteristics influencing leverage. The results show a significant relationship between industry indicators including munificence and HHI on default prediction. As compared to developing countries, the business environment in developed markets is more competitive, therefore, the munificence tends to be insignificant in developed countries. Since the nature of every industry is different in developing countries and every industry is subject to different level of competitiveness. Therefore, it is plausible to find significant relationship between probability of default and munificence. The results of this study may improve the default prediction models performance and reduce the costs of risk management. However, this study also has the limitation that the experimental data sets are only collected from Malaysian listed companies and further investigation can be done based on other countries' real world data sets in future study.

REFERENCES

Allison, P.D., 2001. Logistic Regression Using the SAS System: Theory and Application. SAS Publishing, BBU Press, Cary, NC.

Almazan, A. and C.A. Molina, 2005. Intra-industry capital structure dispersion. J. Econ. Manage. Strat., 14: 263-297.

Altman, E.I., 1968. Financial ratios, discriminant analaysis and the prediction of corporate bankruptcy. J. Finance, 23(4): 589-609.

Beaver, W.H., 1966. Financial ratios as predictors of failure. J. Accounting Res., 3: 71-111.

Carter, C. and J. Catlett, 1987. Assessing credit card applications using machine learning. IEEE Expert, 2: 71-79.

Courtis, J., 1978. Modeling a financial ratios categories framework. J. Bus. Finance Account., 5(4): 371-386.

Deakin, E., 1972. A discriminant analysis of predictors of business failure. J. Account. Res., 10(1): 167-179.

Dess, G.G. and D.W. Beard, 1984. Dimensions of organizational task environments. Adminis. Sci. Quarterly, 29: 52-73.

Dimitras, A.I., S.H. Zanakis and C. Zopounidis, 1996. A survey ofbusiness failure with an emphasis on prediction methods and industrialapplication. Eur. J. Operat. Res., 90: 487-513.

Fitz, P.P., 1932. A comparison of ratios of successful industrial enterprises with those of failed companies. The Certified Public Accountant (October, November, December): 598-605, 656-662 and 727-731, Respectively.

Frydman, H., E. Altman and D. Kao, 1985. Introducing recursive partitioning for financial classification: The case of financial distress. J. Finance, pp: 269-291.

Gepp, A., K. Kumar and S. Bhattacharya, 2010. Business failure prediction using decision trees. J. Forecasting, 29(6): 536-555.

Gianneth, M., 2003. Do better institutions mitigate agency problems? Evidence from corporate finance choices. J. Rnancial Quanftttatlve Anal., 38(1).

Hair, J., W. Black, B. Babin, R. Anderson and R. Tatham, 2006. Multivariate Data Analysis. 6th Edn., Pearson Prentice Hall, Uppersaddle River, N.J.

Hosmer, D.W. and S. Lemeshow, 2000. Applied Logistic Regression. Wiley, New York.

Jackendoff, N., 1962. A Study of Published Industry Financial and Operating Ratios. Temple University, Bureau of Economic and Business Research, Philadelphia.

Kayo, E.K. and H. Kimura, 2011. Hierarchical determinants of capital structure. J. Banking Finance, 35: 358-371.

Lang, L.H.P. and R.M. Stulz, 1992. Contagion and competitive intra-industry effects of bankruptcy announcements. J. Financial Econ., 32: 45-60.

Martin, D., 1977. Early warnings of bank failure: A logit regression approach. J. Banking Finance, 1: 249-276.

Merwin, C., 1942. Financing small corporations in five manufacturing industries, 1926-1936. National Bureau of Economic Research, New York.

Messier, Jr., W. and J. Hansen, 1988. Inducing rules for expert system development: An example using default and bankruptcy data. Manage. Sci., 34(12): 1403-1415.

Meyer, P. and H. Pifer, 1970. Prediction of bank failures. J. Finance, 25(4): 853- 868.

Morris, R., 1997. Early warning indicators of business failure. Ashgate Publishing, Aldershot.

Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res., 18(1).

Opler, T.C. and S. Titman, 1994. Financial distress and corporate performance. J. Finance, 49: 1015-1040.

Pallant, J., 2007. Survival Manual: A step by step guide to data analysis using SPSS for Windows. 3rd Edn., McGraw Hill, New York.

Pompe, P. and A. Feelders, 1997. Using machine learning, neural networks and statistics to predict corporate bankruptcy. Microcomput. Civil Eng., 12: 267-276.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1: 81-106.

Simerly, R.L. and M. Li, 2000. Environmental dynamism, capital structure and performance: A theoretical integration and an empirical test. Strategic Manage. J., 21: 31-50.

Smith, R. and A. Winakor, 1935. Changes in financial structure of unsuccessful industrial corporations. Bureau of Business Research. Bulletin Urbna University of Illinois Press, 51.

Tabachnick, B.G. and L.S. Fidell, 2007. Using Multivariate Statistics. 5th Edn., Allyn and Bacon, Boston.

Zavgren, C., 1985. Assessing the vulnerability to failure of American industrial firms: A logistic analysis. J. Bus. Finance Account., 12(1): 19-45.