Data Mining Approach to Prediction of Going Concern using Classification and Regression Tree (CART)

Dr. Mahdi Salehi[1] and Dr. Mahdi Salehi[2]

[1] Ferdowsi University of Mashhad

Received: 10 December 2012 Accepted: 2 January 2013 Published: 15 January 2013

Abstract

This paper has employed a data mining approach for Going Concern Prediction (GCP) for one year ahead and has applied Classification and Regression Tree (CART) and Naïve Bayes Bayesian Network (NBBN) based on feature selection method in Iranian firms listed in Tehran Stock Exchange (TSE). For this purpose, at the first step, using the Stepwise Discriminant Analysis (SDA) has opted the final variables from among of 42 variables and in the next stage, has applied 10-fold cross-validation to figure out the optimal model. McNemar test signifies that there is a significant difference between the two models in terms of prediction accuracy and CART model is able to predict going concern more accurately. The CART model reached 99.92 and 98.62 percent accuracy rates so as to training and holdout data.

Index terms— data mining, going concern prediction, classification and regression tree, naïve bayes bayesian network, financial ratios, iran.

1 Introduction

oing Concern Prediction (GCP) is an important element in investor's decision-making. Rapid advances in technology, vast environmental changes and increasing competition has affected the security of investment. On the other hand, based on the requirements of Statement on Auditing Standards (SAS) No.59 on every audit the auditor should evaluate whether substantial doubt exists about the firm's ability to continue as a going concern **??**AICPA, 1988). However, SAS 59 contained the relevant criticized guidelines because of deeply subjective, general, ambiguous (Koh & Killough 1988) and, consequently, assessment of GCP sometimes is a tough process and the complexity of GCP has led the development of several models by employing a multiple financial and non-financial variables that might be signifying going concern opinion for auditor (Martens et al, 2008). Early studies of GCP developed by applying statistical techniques such as multiple discriminant analysis and Logit, probit (McKee, 1976;Kida, 1980;Koh, 1987;Menon & Schwartz, 1987;Koh & Brown, 1991). In recent years, data mining has established, developed and began to appear and grow promptly in the financial area and constructed a new approach for the deep research. Data mining technique via utilizing a large number financial data can be extracting, valuable and unknown knowledge dynamically. Using data mining techniques several research have been conducted in GCP area and the findings indicate that these techniques are able to predict the going concern status of firms and accounting data are useful in GCP (Brabazon & Keenan, 2004;Koh & Kee Low, 2004;Martens et al, 2008;Mokhatab et al., 2011). Nowadays these methods because of the restrictive assumptions of statistical techniques (such as normality, linearity and independence of variables) are used less. This research has applied Classification and Regression Tree (CART) and Naïve Bayes Bayesian Network for GCP. Results from this study will help a manager to keep track of company's performance and to identify significant problems and take efficient measure to reduce the coincidence of failure. In addition, this model helps lenders and other stakeholders to have a clear and comprehensive picture of the firm's prospective status. In addition, auditor can use the survey results in the final stages of the audit engagement as a qualitycontrol device or as a benchmark in auditor judgment. Particularly, the GCP model in this paper can be applied for auditors to assess potential clients and as a means to identify non-going concern firms that might require further consideration.

2 II.

3 Research Development

The data set is composed of 146 Iranian manufacturing companies including 73 matched companies in bankrupt firms and firms with going concern status that all of them were or still are listed in the Tehran Stock Exchange (TSE) from 2001-2011. As you can see in Table 1, the 42 proposed variables used in this study are shown. After data collection, this paper applied process of future selection by T-test and Stepwise Discriminant Analysis (SDA) at a significant level of 0.05 and selected final variables. The potential advantages of feature selection are facilitating data visualization and understandable data, reducing the measurement and storage requirements (Ashoori & Mohammadi, 2011). Another purpose of these tests is to determine the financial ratios that can distinguish between the two companies (going concern and nongoing concern status). The result of SDA process is shown in Table 2. The ratios that are entered in the model are total liabilities to total assets (?? 9), Retained earnings to total assets (?? 31), Operational income to sales (?? 36) and Net income to total assets (?? 34). After extraction of financial ratios, a model was constructed that explained as a discriminant model in below: Z= -0.374 X9+ 0.293 X31+ 0.359 X36+ 0.384 X34

(1) CART, methodology was popularized in 80s by Breiman et al. (1984). In the area of GCP, the goal of the analysis via CART is to obtain a set of if-then rules with acceptable accuracy that determine what companies will have going concern or not in the future. Furthermore, reasons for selecting CART are that is nonparametric and can easily handle outliers. It is flexible and has an ability to adjust in time **??**Timofeev 2004). In order to obtain the best predictive accuracy, CART is built to minimize the misclassification cost, which takes both variance, and misclassification rates into consideration. It is a significant step to choose the splits on the features that are employed to predict membership in corresponding class of firms. CART computational detail includes itself in finding the best split rules in order to make an uncomplicated, informative and accurate tree. The CART regards all variables as independent in the calculations of split with the training data set. The ??th samples is expressed as (?? 1 ?? , ?? 2 ?? , ?,?? ?? ?? , ?, ?? ??), where ?? ?? ?? is the value of the ??th sample firm on the ??th feature and the label value of the sample is ?? ?? . Since CART is a binary recursive partitioning method that every leaf of the data splits to two sub-leaves, for classification problem the values of ?? ?? are binary, e.g., -1 or 1. In the process of splitting, if a feature value ?? ?? ?? ? ?? ? is met, CART follows the rule that a sample goes right, otherwise it goes left. Split at each node will occur only when the split can go to greatest improvement in accuracy of prediction. Specific types of node impurity measure that Breiman et al. (1984) proposed to apply Gini index as the criteria used in order to reduce the impurity in splitting for classification, since it can be estimated more rapidly and be readily extended to include symmetries costs can measure this. In the classification problem of GCP, the Gini index of impurity of a node can be signified as follows (Breiman et al., 1984):( ) 2 1 ? ? = j j gini c p I

Where ??(?? ??) indicates the relative frequency of the first class in the node. The Gini index reaches a value of zero when only one class is obtained at a node.It means that if all cases in a node belong to the same class, the Gini index will be zero (Li, Sun & Wu, 2010). CART applied backward pruning algorithms. Pruning will be necessary to build smaller tree models that perform better on new data and not just on the training data. CART uses pruning and selecting in each node in the tree when the tree is fit (Soni, 2010). As the classification or regression tree is constructed, it can be used for classification of new data. The output of this stage is an assigned class or response value to each of the new observations. By set of questions in the tree, each of the new observations will get to one of the terminal nodes of the tree. A new observation is assigned with the dominating class/ response value of b) The Method of Naïve Bayes Bayesian Network (NBBN)

Bayes networks are a powerful tool for relationships between a set of variables and they are a suitable tool for dealing with uncertainty conditions in expert systems (Markov, 2007). The purpose of Bayes network is to establish a model that can classify companies correctly using financial ratios. A NBBN is based on Bayes' rule that is expressed as follows: In problem solving of going concern, P(A)??(??/??)= ??(??/??) ??(??) ??(??)**(2)**

shows the percentage of companies with going concern status and P(B) indicates the share of each of the independent variables are used for GCP and P(A/B) is probability of going concern status during one year ahead. An example of a NBBN can be seen in Figure 1. In this figure A is dependent variable and ?? 1 , ?? 2 , ?? 3 , and ?? 4 are independent variables (Sun & Shenoy, 2007).

4 Experimental Results

The proposed CART and NBBN models are implemented by using MATLAB 7.6.They are results from the 10 testing data sets by using 10-fold cross validation (See Table 3 As shown in Table 5, the result of McNemar test at 5% level indicates that there are significant differences between the two models in GCP. According to Table 6, Type I error is the probability that a company with non going concern status to be classified as a company with going concern status and Type II error is the probability that a company with going concern status to be classified as a company with non going concern status.

Costs related to these two types of errors are very different. Costs resulting from incorrectly classifying a company with non-going concern as a company with going concern status (Type I error) is much larger than the Type II error (incorrectly classifying a company with going concern as a company with non-going concern status). In holdout data type I and II error are also equal to 2.5 and 0 percent in CART model and 22.64 and 22.65 percent for obtained model by NBBN.

5 Conclusion

The current study demonstrated feasibility of applying CART and NBBN to predict going concern status with data collected from Iran. This paper considered a set of features that include 42 variables proposed in prior literature dealing with financial status prediction models in Iran and applied SDA to identify potential variables for GCP model and finally four financial ratios were selected and constructed CART and NBBN GCP models based on selected features. Based on the conclusions, the empirical tests show that CART and NBBN models have achieved 98.62 and 75.55 percent accuracy rates for training and holdout data, respectively. Moreover, McNemar's test results indicate that there are significant differences between the two models in predicting of going concern. In summary, obtained results from this research from 146 companies of Iran signify that: CART model has appropriate ability for GCP of firms. Further, this research empirically tested future selection using statistical technique that data mining algorithms can be used for future research. [1]

Figure 1: D

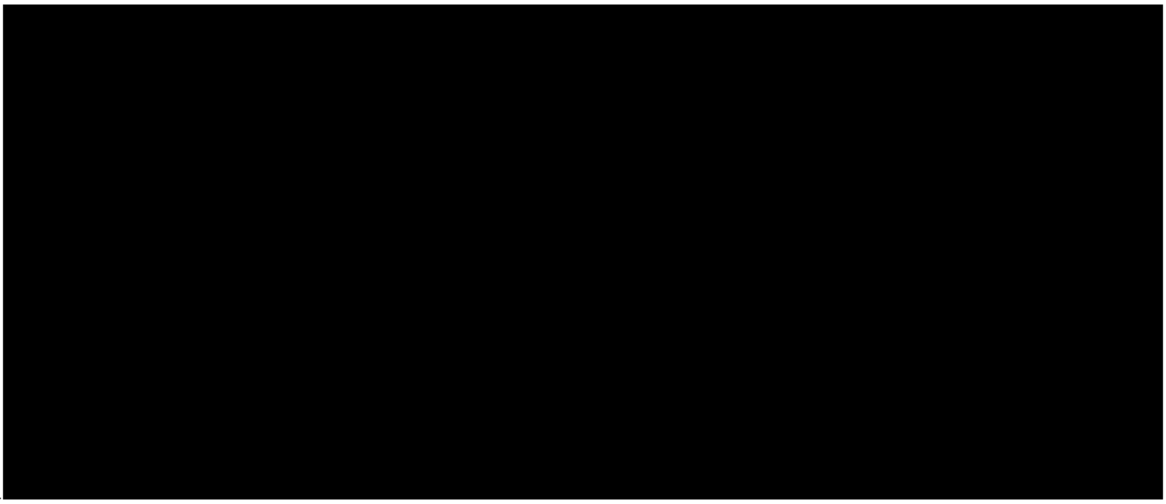
1 

Figure 2: Figure 1 :

1

2013
ear
Y
Volume XIII Issue III Version I
( )
Global Journal of Management and Business Research

[Note: DData Mining Approach to Prediction of going Concern using Classification and Regression Tree (CART)]

Figure 3: Table 1 :

2

#	Definition of variables	Means of Group 1	Means of Group 0	Sig level	#	Definition of variables	Means of Group 1	Means of Group 0	S l
1	EBIT/TA	0.18	0.05	0.00	2	LTD/SE	0.20	0.56	0
3	RE/SC	0.65	0.02	0.00	4	MVE/TL	1.40	0.66	0
5	MVE/SE	2.42	2.57	0.22	6	MVE/TA	0.77	0.48	0
7	Ca/TA	0.05	0.03	0.00	8	Size(logTA)	5.25	5.23	0
9	TL/TA*	0.67	0.80	0.00	10	CL/SE	2.27	4.76	0
11	CL/TL	0.86	0.85	0.94	12	(Ca+STI)/CL	0.11	0.05	0
13	(R+Inv)/TA	0.57	0.57	0.88	14	R/S	0.53	0.40	0
15	R/Inv	1.18	1.00	0.93	16	SE/TL	0.63	0.32	0
17	SE/TA	0.35	0.22	0.00	18	CA/CL	1.31	1.07	0
19	QA/CL	0.70	0.57	0.00	20	QA/TA	0.37	0.36	0
21	FA/(SE+LTD)	0.60	0.91	0.01	22	FA/TA	0.22	0.24	0
23	CA/TA	0.70	0.68	0.66	24	Ca/CL	0.09	0.04	0
25	IE/GP	-0.02	-1.21	0.48	26	S/Ca	35.30	44.80	0
27	S/TA	0.93	0.70	0.00	28	WC/TA	0.13	0.00	0
29	PIC/SE	0.53	0.86	0.00	30	S/WC	2.87	1.73	0
31	RE/TA*	0.08	-0.03	0.00	32	NI/SE	0.42	-0.03	0
33	NI/S	0.16	-0.02	0.00	34	NI/TA*	0.13	0.00	0
35	S/CA	1.34	1.07	0.00	36	OI/S*	0.20	0.06	0
37	OI/TA	0.17	0.03	0.00	38	EBIT/IE	-5.21	-0.45	0
39	EBIT/S	0.52	0.10	0.00	40	GP/S	0.27	0.15	0
41	S/SE	3.32	4.68	0.05	42	S/FA	6.29	6.44	0

Group 1: going concern firms and Group 0: non-going concern firms
* : Final variables selected by SDA

CA: Current assets
Ca: Cash
CL: Current liabilities
PIC: Paid in capital
EBIT: Earnings before interest & taxes
FA: Fixed assets
GP: Gross profit
IE: Interest expenses
Inv: Inventory
LA : Liquid assets
LTD: Long term debt
MVE: Marked value of equity
NI: Net income
OI: Operational income
QA: Quick assets
R: Receivables
RE: Retained earnings
S: Sales
SC: Stock capital
SE: Shareholders' equity
STI: Short term investments
TA: Total assets
TL: Total liabilities
WC: Working capital

Step		Tolerance	F to Remove	Wilks' Lambda
1	Net income to total assets	1.00	100.77	
2	Net income to total assets	0.94	56.24	0.75
	Total liabilities to total assets	0.94	9.07	0.55
3	Net income to total assets	0.51	8.62	0.52
	Total liabilities to total assets	0.91	11.10	0.53
	Operational income to sales	0.55	6.11	0.51
4	Net income to total assets	0.48	4.75	0.49
	Total liabilities to total assets	0.90	8.55	0.50
	Operational income to sales	0.54	4.57	0.49
	Retained earnings to total assets	0.77	4.37	0.49

3

Fold	and NBBN model CART Training data	Hold-out data	NBBN Training data	Hold-out data
1	100.00	100.00	100.00	80.00
2	100.00	100.00	100.00	80.00
3	100.00	100.00	100.00	66.67
4	93.33	99.23	100.00	66.67
5	100.00	100.00	100.00	80.00
6	92.86	100.00	100.00	85.71
7	100.00	100.00	100.00	64.29
8	100.00	100.00	100.00	78.57
9	100.00	100.00	100.00	82.21
10	100.00	100.00	100.00	71.43
Min	92.86	99.23	100.00	64.29
Max	100.00	100.00	100.00	85.71
Median Variance	100.00 9.28	100.00 0.07	100.00 0.00	85.71 61.99
Mean	98.62	99.92	100.00	75.55

Figure 5: Table 3 :

4

Fold	Cont Rule	Height Tree
1	3	2
2	3	2
3	3	2
4	2	1
5	3	2
6	2	1
7	3	2
8	3	2
9	3	2
10	3	2

Figure 6: Table 4 :

5

Methods NBBN	
CART	-3.536 (0.011)

[Note: D]

Figure 7: Table 5 :

6

	Real status	
Prediction	Non going concern status	Going concern status
	1-P 22 (Type I error)	P 11
	P 22	1-P 11 (Type II error)

Figure 8: Table 6 :

[Brabazon and Keenan ()] 'A hybrid genetic model for the prediction of corporate failure'. A Brabazon , B Keenan . *Computational Management Science* 2004. Springer-Verlag. p. .

[Menon and Schwartz ()] 'An empirical investigation of audit qualification decisions in the presence of going concern uncertainties'. K Menon , K Schwartz . *Contemporary Accounting Research* 1987. 3 (2) p. .

[Kida ()] 'An investigation into auditors, continuity and related qualification judgments'. T Kida . *Journal of Accounting Research* 1980. 18 (2) p. .

[Breiman et al. ()] *Classification and regression trees*, L Breiman , J Friedman , R Olshen , C Stone . 1984. (International Group)

[Ashoori and Mohammadi ()] 'Compare failure prediction models based on feature selection technique: empirical case from Iran'. S Ashoori , S Mohammadi . *Procedia Computer Science* 2011. 3 p. .

[Mckee ()] 'Discriminant prediction of going concern status: A model for auditors'. T Mckee . *Selected Papers of the AAA Annual Meeting*, 1976.

[Mokhatab Rafiei et al. ()] 'Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence'. F Mokhatab Rafiei , S M Manzari , S Bostanian . *Expert Systems with Applications* 2011. 38 p. .

[Koh and Kee Low ()] 'Going concern prediction using data mining techniques'. H C Koh , C Kee Low . *Managerial Auditing Journal* 2004. 19 (3) p. .

[Soni ()] 'Implementation of multivariate data set by CART algorithm'. S Soni . *Journal of Information Technology and Knowledge Management* 2010. 2 (2) p. .

[Li et al. ()] 'Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods'. H Li , J Sun , J Wu . *Expert Systems with Applications* 2010. 37 p. .

[Martens et al. ()] 'Predicting going concern opinion with data mining'. D Martens , L Bruyneseels , B Baesens , M Willekens , J Vanthienen . *Decision Support Systems* 2008. 45 p. .

[Koh ()] *Prediction of going-concern status: A probit model for the auditors*, H Koh . 1987. Virginia Polytechnic Institute and State University (Ph.D. dissertation)

[Markov ()] *Probabilistic reasoning with naïve bayes and Bayesian networks, PhD dissertation*, Z Markov . 2007. Central Connecticut State University

[Koh and Brown ()] 'Probit prediction of going and non-going concerns'. H Koh , R Brown . *Managerial Auditing Journal* 1991. 6 (3) p. .

[Koh and Killough ()] 'Proposed statement on auditing standards: the auditor's consideration of an entity's ability to continue existence'. H C Koh , L N Killough . *Virginia Accountant Quarterly* 1988. 40 (2) p. .

[Aicpa ()] 'The auditor's consideration of an entity's ability to continue in existence'. Aicpa . *Statement on auditing standards*, 1998.

[Sun and Shenoy ()] *Using Bayesian Networks for Bankruptcy Prediction: Some*, L Sun , P Shenoy . 2007.