DYNAMIC PREDICTION MODEL FOR FINANCIAL DISTRESS IN CONSTRUCTION INDUSTRY USING DATA MINING

S. Ping Ho
Nation Taiwan University
spingho@ntu.edu.tw

Fang Ying Lin
Nation Taiwan University
r91521705@ntu.edu.tw

Abstract

The early awareness of a potential financial distress is crucial to firm's managers for understanding their clients, suppliers and their own firms, and crucial to fund suppliers for assessing the construction firm's credit worthiness. The purpose of this paper is to develop a dynamic prediction model for financial distress in construction industry using Data Mining. This research expects to provide construction firm managers and creditors an effective index for evaluating the credit risk a construction firm. Results show that the proposed model has higher accuracy and stability for distress prediction and can provide a more effective quantitative framework for evaluating the financial standing of a construction firm.

Keyword: Financial Distress, Distress Prediction, Data Mining, CART (Classification and Regression Tree), Construction Industry.

1. INTRODUCTION

The early awareness of a potential financial distress is crucial to firm's managers for understanding their clients, suppliers and their own firms, and crucial to fund suppliers for assessing the construction firm's credit worthiness. Researchers have been trying to build effective financial distress prediction models by applying various approaches. The purpose of this study is to develop a dynamic prediction model for financial distress in construction industry using Data Mining.

Data Mining is one of the decision support technique for knowledge discovery. One of the classification algorithms is Decision Tree technique. This study expects to use the CART algorithm (Classification and Regression Tree) to build the financial distress predict model, and build the rules that can identify those companies that are highly possible to encounter financial distress.

2. LITERATURE REVIEW

Traditionally, the prediction of corporate failure relied on financial ratio analysis and there was no theory on how or when financial distress will occur. Bankruptcy prediction models were pioneered by Beaver's (1966) univariate test and Altman's (1968) multivariable discriminant analysis. Both studies show that financial variables can be used to predict bankruptcy. Since then, the prediction of corporate failure has been a topic of much interest. Recent works have extended this line of research into four directions: 1. utilizing different techniques, 2. examining the relationship between various definitions of bankruptcy and prediction models, 3. exploring a greater variety of explanatory variables, and 4. studying financial distress in particular industry. For example, in first direction, Tam and Kiang (1992) use one of the data mining classification techniques -C4.5 to estimate the probability of bankruptcy. In the second direction, Poston, Harmon, and Gramlich (1994) assign firms into one of three groups in financial distress according to each firm's financial condition: turnarounds, business failures, and survivors. They find that financial ratios are not so useful in distinguishing between financially distressed firms that are able to turn around and those that are unable to avoid failure. In the third direction, Rose and Giroux (1982) examine 28 business cycle indicators and find that economic conditions affect the failure process.

In the fourth direction, as suggested by Altman (1993), the characteristics of different industries are considered in the distress prediction models. Keasey, K., and McGuinness, P. (1990), Kangari, R., Farid, F., and Elgharib, H. M. (1992), Langford, Iyagba and Koma (1993), and Abidali and Harris（1995）are all focus in construction industry to build the financial distress prediction models.

3. METHODOLOGY AND DATA

3.1Method

Data mining has no particular assumptions on sample data, such as the normal distribution assumption in discriminated analysis. Decision tree is a typical classification tool for data mining. Whereas most decision trees algorithms have the same structure, various algorithms such as CHAID, C4.5/C5.0 and CART produce trees that differ from one another in the number of splits allowed at each level of the tree, how the splits are chosen when the trees are built, and how the tree growth is limited to prevent over-fitting. The fact that software packages often allow the user to choose a splitting criterion reflects that there is no single best choice for all problems. Data mining must experiment so as to determine which one gives the best results for the data set in hand.

CART (Classification And Regression Tree) algorithm is one of the main types of decision trees. Each branch of a decision tree is a test on a single variable that cuts the space into two or more pieces, and each split in this algorithm is constrained to be binary. The process for building decision tree is called recursive partitioning, an iterative process of splitting the data up into partitions. The algorithm chooses the split that partitions the data into two parts that are "purer" than the original. This splitting or partitioning procedure is then applied to each of the new boxes/categories. The process continues until no more useful splits can be found. So, the heart of the algorithm is the rule that determines the initial split.

The measure used to evaluate a potential splitter is the increase in purity. Purity represents the correct rate in the node, measured by Gini index (Berry, 2000). It can be interpreted as the probability that any two random elements of the population will belong to different classes. Since the probability index is simply one minus the sum of the all the Pi^2. The formula for the diversity index for binary targets is 2 P_1 (1-P_1), where P_1 is the probability of class one. It has been shown that the Gini criterion tends to favor splits that isolate the largest class in one branch of the tree.

3.2Sample and Data

The sample is from firms listed in the Taiwan Stock Exchange Corporation (TSEC) during 1985 to 2004. Financial distress is defined in this study as firms that either (1) were closed down by government authorities (all of which were banks and finance companies) or (2) were required by the Bank of Taiwan or TSEC to submit restructuring plans.

There were 76 firms listed in the TSEC, of which 31 were financially distressed by this definition. Firms' seasonal financial data were collected and each set of data from seasonal report is considered as one sample. As a result, total sample database consists of 1576 non-financially distressed samples and 495 financially distressed samples.

We choose the split points are 1995 to 2002, that partitions the data into two parts. Forward part data is training data set, and test data set is behind. The number of total database is 144. Several studies, for example, Beaver and Merwin (1966), indicate that firms may begin to exhibit the tendency toward failure as much as five years prior to the actual failure. Little is mentioned, however, of the actual significance of these early signs of distress. As Altman (1968) states: "Is it enough to show that a firm's position is deteriorating or is it more important to examine when in the life of a firm's does its eventual failure?" Here we will consider the financial data as much as five years prior to the actual failure. The definition of class show is in Table 1.

Table 1. The delineation of Class

Model	Types of Classes
D1	Normal, Distress (0 to 1 year before distress)
D2	Normal, Distress (0 to 2 years before distress)
D3	Normal, Distress (0 to 3 years before distress)
D4	Normal, Distress (0 to 4 years before distress)
D5	Normal, Distress (0 to 5 years before distress)

3.3Research design and Procedures

Considering the sample Choice-Based Biases and the sample Selection Biases (Zmijewski, 1984), we use all of the usable data. To detect financial distress, we set the null hypothesis, H_0, as "Distress," and we develop rules for identifying Normal firms. If a firm is identified as Normal, then H_0 is rejected and the firm is considered as Normal, otherwise, the firm is considered as Distress. However, when the firm is considered as Normal when H_0 is rejected, there is the type I error that H_0 is mistakenly rejected. As a result, type I error also means that a Distressed firm is considered as a Normal firm. Thus, the type I error (α) can be measured by the ratio of misjudgment through the rules developed for identifying the Normal firms. The procedures in this study for developing the financial distress models are as follows:

Step1. Data search and collect. We use the financial data from the TSEC, period ranging from 1985 to 2004. We also use the corporate basic data to build the class in Taiwan Economic Journal Data Bank (TEJ Data Bank).

Table2. Financial Variable

ROA-EBIT (%)	YOY-Total Assets% (Year-on-year growth)
ROE-NI (%)	YOY-Total Equity%
ROE -NI Exclude Disposal%	YOY-Fixed Assets%
Gross Margin (%)	YOY-ROA%

Realized Gross Profit%	Retention Ratio
Operating Income (%)	Current (%)
Pre-Tax Income (%)	Acid Test (%)
Net Income (%)	Interest Expense (%)
Non- Operating Inc./Revenue (%)	Debt/Equity (%)
Net Income%	Liabilities (%)
Operating Expense (%): Operating expense/ Revenue	Equity/TA (%)
CF-Operating CF/Current Liability (%)	(Long-term Liability+ Equity)/FA (%)
Interest Exp/Debt (%)	Interest-Bearing (%): debt/equity
PS(per share)-equity	Times Interest Earner
PS-Cash flow	Operation Income/Capital
PS-Sales	Pre-Tax Income/Capital
PS-Operating Income	Inv.& A-R/Equity
PS-Pre_ Tax Income	Total Asset Turnover
YOY-Sales%	A/R&N/R Turnover
YOY-Gross Margin%	Days A/R
YOY-Realized GM%	Inventory Turnover
YOY- Operating Income%	Days Inventory
YOY-Pre-Tax Income%	Fixed Asset Turnover
YOY-Net Income %e	Equity Turnover
YOY-Ordinary Income %	Days-A/P Turnover
YOY- NI Ex Disposal%	Net Operating Cycle

Step2. Data preparation：This step includes cleaning noises and handling missing values. The tasks of cleaning noises include finding repeated data or wrong property of database, and reconditioning data. In this study, after we delete missing values and repeated data, we have 76 companies, and 2071 seasonal samples.

Step3. Data converting and data warehousing: This step includes the determination and converting of data type, and the building Data Warehouse using IBM DB2.

Step4. Model developing: This step is to build the prediction Model using the CART (Classification and Regression Tree) algorithm used by IBM Intelligent Miner. Since this model will be revised later, we shall call the model before revision the "Original Model." According to the argument test, the best prediction power occurs when the minimum sample size in a node is 5, and the maximum node correct ratio is 100%.

Step5. Model evaluation and revision: After considering the cost of misclassification, we build the Original model is revised according to the criterion of pruning tree. Then consider the problem for over-fitting, we build the Revised Model when the maximum node correct ratio is 80% in the model.

4. THE DYNAMIC PREDICTION MODEL

In the research design, we find that a split of 70-30 percent for training set and test set works well in our test. This result is same to Berry (2000). So, we use the year from 1985 to 2000 to be training data set, and year 2000 to 2004 to be test data set.

The prediction ability, (total correct rate) is about 89.64% in Original Model. But the depth of decision trees in the Original Model is too much. So, we consider the problem for over-fitting to build the Revised Model.

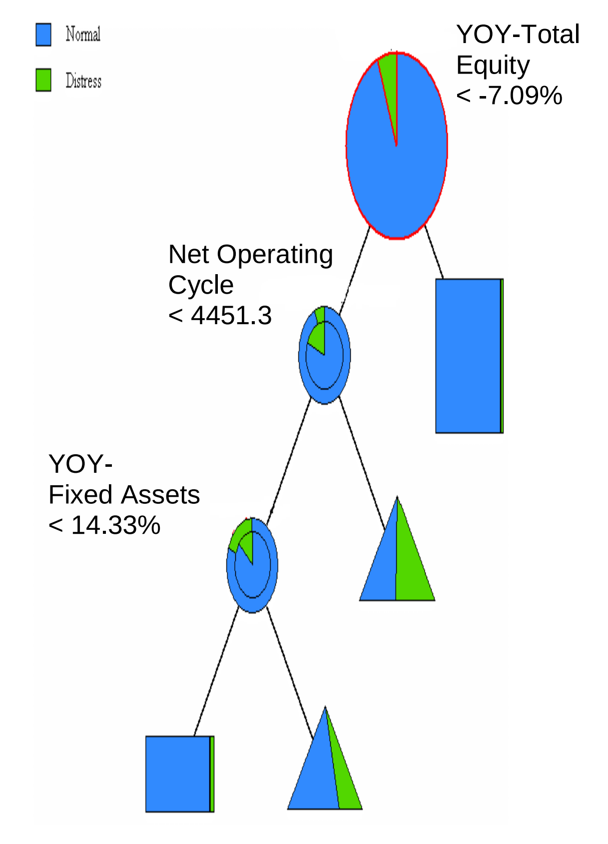
The prediction ability, (total correct rate) is about 73.93% in Revised Model. The performance of type I error in the four model (Original Training & Test Model, Revised Training & Test Model) is lower than 10%, and the type Π error rate is 53.78. Even we consider the long tern prediction result, the cost of misclassification is better than the literature (Altman, 1968). The performance of total minimum error rate, is the lowest in D1 model, the secondary is in D2 model. The error rate for Original Model and Revised Model is shown in table 3. The Revised Model rule is show in table 4.

Table 3.The performance for the Original Model and the Revised Model

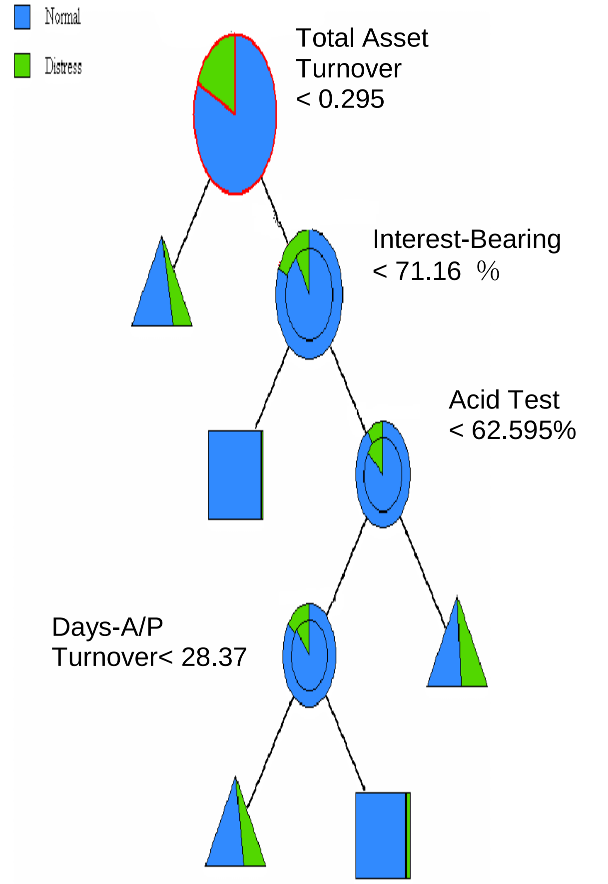
Model	Revised Model			Original Model		
	α	β	Total	α	B	Total
The error rate in Training Model (%)						
D1	3.3	62	6.2	4	17	4.9
D2	6	63	18.9	7	21	8
D3	8	69	35	7	20	9
D4	7	62	36	8	13	9
D5	7	65	44	8	16	10
The error rate in Test Model (%)						
D1	5.6	87	7.2	2	100	3.7
D2	2.5	82	18.4	4	69	7
D3	3	85	38	5	73	14
D4	9	65	20	7	72	18
D5	8	48	37	9	65	20

P.S. (1) Type I error: α (%) (2) Type Π error: β (%)

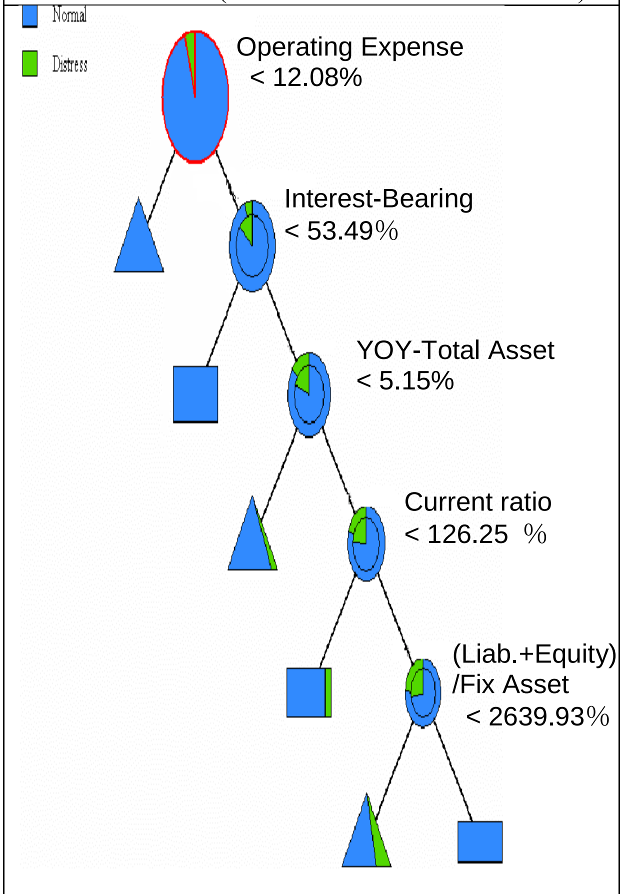
Table5. The Revised Model rule



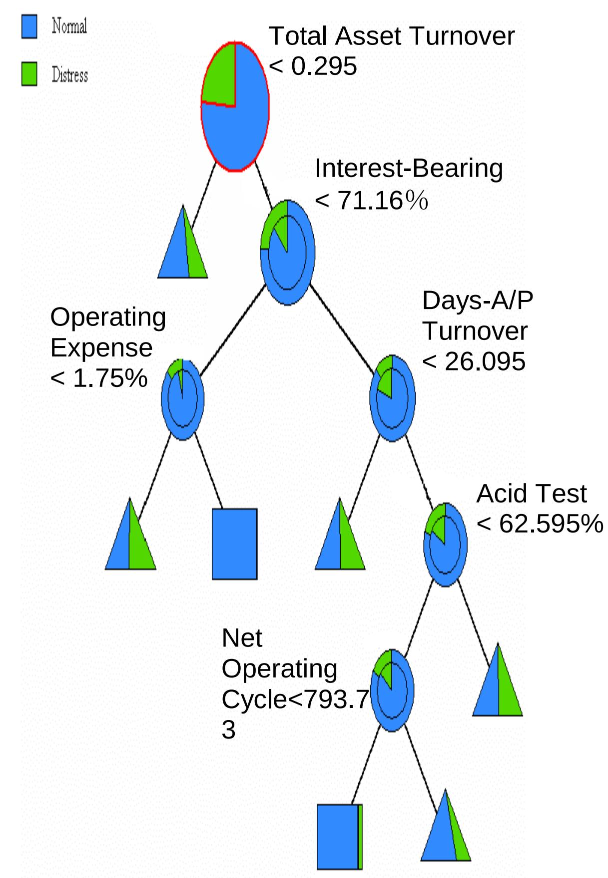
Model Name: D1(train:1985~2000/ test:200~2004)



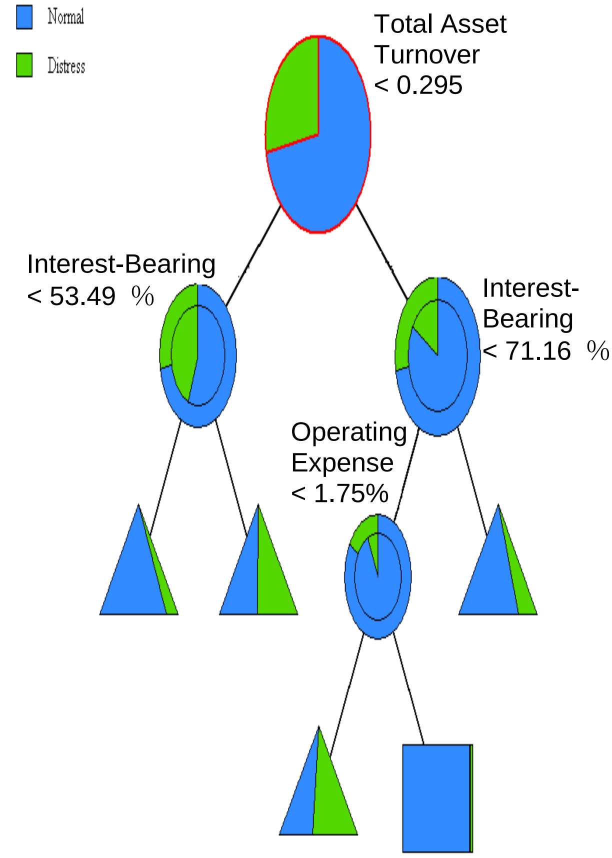
Model Name: D3(train:1985~2000/ test:200~2004)



Model Name: D2(train:1985~2000/ test:200~2004)



Model Name: D4(train:1985~2000/ test:200~2004)



Model Name: D5(train:1985~2000/ test:200~2004)

According to table 3, the most importance critica l predicting variables are YOY-Total Equity%, Net Opera t i n g Cycle , YOY-Fixed Assets , Operating Expense , Interest-Bearing , YOY-Total Asset, Current ratio, (Liability +Equity) / Fix Asset , Total Asset Turnover , Days-A/P Turnover and Acid Test. Same of the variabl es same to the results by Abidali & Harris (1995) and Langford, Iyagba & Koma (1993).

Our models show that the Interest-Bearing(%) is the most important variables in differentiating bet ween normal company and financial distress com pany. The suggested range of Interest-Bearing(%) in the short-term for a normal company is 53.49 % or so, and 71.16% in the long-term.And in lo ng tern prediction ability, (3 to 5 years) Total Asset Turnover <0.295 occur to the financia l distress company frequently.

Since current assets include inventory, if the curr ent ratio is greater than 126.25%, the company may have too much inventory in hand, and as a result, the company may have higher probability of having financial distress.

5. CONCLUSIONS AND DISCUSSION

In this paper, we use all usable data as samples for minimum sample choice bias and apply CART algorithm to build the Original Model first and then revise it to the Revised Model. In the Revised Model, Interest-Bearing ratio and Total Asset Turnover ratio have great capability to differentiate between normal company and financial distress company. It shows that the construction company's profit is deeply affected by the problem for liabilities frequently. For creditor, we suggest you consider four years financial statement and use Revised Model–D1and D2 to predict financial distress company. If you have the data enough, you can use the Revised Model –D4 to predict the long tern performance of the company.

After consider the bias in literature, the overall predict correct rate is 89.64%. Then, we consider the problem for over-fitting to build the Revised Model. The prediction ability, (total correct rate) is about 73.93% in Revised Model. The performance of type I error in the four model (Original Training & Test Model, Revised Training & Test Model) is lower than 10%, better than the literature (Altman, 1968) and so on]. But the type Π error rate is 53.78. Perhaps the model performance is not very satisfied, but the predict result (1)is to be close to the fact ,(2) less uncertain than other model that didn't consider sample bias and(3)few cost of misclassification at least

In the research design, we have found a split of 70-30 percent for training set and test set works well in our practice. This result is same to Berry's research. [Berry, 2000]

For the follow-up researchers, we suggest you can use C4.5 algorithms to build the model, or use new variables, about industry characteristic, accede to the new models, to improve the model predict performance.

REFERENCES

[1] Abidali, A.F., and Harris, F., "A methodology for predicting company failure in the construction industry." *Construction Management and Economics,* Vol. 13, No. 3, pp.189-196, 1995.
[2] Altman, E.I., "Financial ratios, Discriminated analysis and the prediction of corporate bankruptcy." *Journal of Finance*, Vol. 23, No. 4, pp. 589-610, 1968.
[3] Altman, E.I., *Corporate Financial Distress and Bankruptcy*, John Wiley & Sons, 1993.
[4] Beaver, W.H., "Financial ratios as predictors of failure." *Journal of Accounting Research*, Vol. 4, No. 3, pp.71-112, 1966.
[5] Berry, M., and Linoff, G., Master Data Mining: The Art and Science of Customer Relationship Management, John Wiley & Sons, 2000.
[6] Kangari, R., Farid, F., and Elgharib, H. M.,"Financial performance analysis for construction industry." *Journal of Construction*

Engineering and Management, ASCE, Vol. 118, No. 2, pp. 349-361, 1992.

[7] Keasey, K., and McGuinness, P., “The Failure of UK Industrial Firms for the Period 1976-1984, Logistic Analysis and Entropy Measures” *Journal of Business Finance and Accounting,* Vol. 17, No. 1, pp. 119-135, 1990.

[8] Langford, D., Iyagba, R., and Komba, D.M., “Prediction of solvency in construction companies.” *Construction Management and Economics,* Vol. 11, No. 5, pp. 317-325, 1993.

[9] Poston, K. M., Harmon, W. K., and Gramlich, J. D. , “A test of financial ratios as predictors of turnaround versus failure among financially distressed firms” *Journal of Applied Business Research*, Vol. 10, pp. 41-56, 1994.

[10] Rose, R. S.; Andrews, W. T.; and Giroux, G., “A Predicting business failure: A macroeconomic perspective” *Journal of Accounting, Auditing*, and Finance, No. 3, pp. 20-31, 1982.

[11] Tam, K.Y., and Kiang, M.Y., “Managerial applications of neural network: the case of failure predictions.” *Management Science,* Vol. 38, No. 7, pp. 926-947, 1992.

[12] Zmijewski, M.E., “Methodological Issues Related to the Estimation of Financial Distress Prediction Models.” *Journal of Accounting Research*, Vol. 22, pp. 59-82., 1984.