Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Predicting Financial Distress: A Comparison of Survival Analysis and Decision Tree Techniques

Adrian Gepp and Kuldeep Kumar*

Faculty of Business, Bond University, Gold Coast, Australia

Abstract

Financial distress and then the consequent failure of a business is usually an extremely costly and disruptive event. Statistical financial distress prediction models attempt to predict whether a business will experience financial distress in the future. Discriminant analysis and logistic regression have been the most popular approaches, but there is also a large number of alternative cutting – edge data mining techniques that can be used. In this paper, a semi-parametric Cox survival analysis model and non-parametric CART decision trees have been applied to financial distress prediction and compared with each other as well as the most popular approaches. This analysis is done over a variety of cost ratios (Type I Error cost: Type II Error cost) and prediction intervals as these differ depending on the situation. The results show that decision trees and survival analysis models have good prediction accuracy that justifies their use and supports further investigation.

1. Introduction

History reveals that a goal of no businesses experiencing financial distress and then becoming bankrupt is unrealistic. For example, the Belgian car company Minerva failed in 1956 and the US – based computer hardware company Commodore International declared bankruptcy in 1994. More recently, during and after the Global Financial Crisis a vast number of businesses failed such as Waterford Wedgewood in 2009. However, we can improve our ability to identify businesses that are at risk of experiencing financial distress in the future. This prior warning might assist some businesses to make appropriate changes to avoid financial distress and future bankruptcy, but more commonly this advanced warning could be used to mitigate the costs of financial distress and business failure, for example,

- Finance institutions could better control their risk exposure and future number of bad debts,
- Investors could more accurately control the risk profile of their investments and potentially improve their performance by not investing in future failures and,
- Other stakeholders such as suppliers and customers would have better information on which to make decisions such as long – term exclusive arrangements.

*Corresponding author. Tel.: +61-7-5595-3305; Fax: +61-7-5595-1160.
E-mail address: kkumar@bond.edu.au

Accurate prior warnings would promote economic stability and reduce costly contagious effects similar to those which have been seen with the recent Global Financial Crisis when a business failing results in another business failing and so on. The field of financial distress prediction, also known as business (or firm) failure prediction and bankruptcy prediction involves developing statistical models and data – mining approaches based on publicly available information such as financial ratios that provide these prior warnings. It should come as no surprise that the Global Financial Crisis has created renewed interest in this field.

The remainder of this paper is structured as follows. A brief review of background and related work is presented first. Some important issues with models for predicting financial distress are then discussed including an overview of the two primary modelling techniques used in this paper. The data and methodology used are then presented, followed by the results and conclusions.

2. Background and Related Work

Many different statistical and mathematical models have been used to predict financial distress since the first publications in the 1960s. Beaver[1] presented a univariate model in 1966 and two years later Altman[2] pioneered the use of discriminant analysis in the field. Altman's work has been the subject of much later research, including that of Deakin[3] who increased the number of explanatory variables and Edminster[4] who focused on small businesses. Agarwal and Taffler[5] analyse an Altman – style model using UK data over a period of 25 years. From this extensive empirical test, they concluded that the discriminant analysis approach has useful real – world predictive ability. It is also interesting to note that Agarwal and Taffler recommend that once models become out-of-date, which can take a long time, it is preferable to build a completely new model rather than simply re-estimate coefficients of existing models.

In 1980, Ohlson's[6] pioneering work used logistic regression as a way of overcoming restrictive assumptions of discriminant analysis, such as normality and equal covariances. In 1993, Theodossiou[7] introduced sequential cumulative sum (CUSUM) procedures to predict financial distress with excellent empirical results. The soft computing methods known as artificial neural networks have also been used for financial distress prediction – Tan[8] provides a summary. More recently in 2011, du Jardin and Severin[9] outlined an approach using a Kohonen map that has shown promise in making predictions over longer periods. There are also many other techniques that have been used to predict financial distress, including support vector machines and nearest neighbour, survival analysis and decision trees.

In 1985, Frydman *et al.*[10] were the first to use decision trees to predict business failure. They found their decision tree to be a superior predictor of business failure compared with discriminant analysis. This success, however, failed to stimulate a lot of research on the use of decision trees in this field. Joos *et al.*[11] compared logit analysis and See5 decision trees without an overall superior technique emerging, while Huarng *et al.*[12] found CART to be better than See5 at predicting business failure on a very small data set. On a larger data set, Gepp *et al.*[13] more recently found CART to be superior to See 5, as well as being superior to discriminant analysis. Gepp *et al.* also discuss other studies that use decision trees. These studies mostly support decision trees as being superior to discriminant and logit analyses in this field. While the work of Chen[14] did not find that CART outperformed See5, it found that decision – tree techniques in general outperformed logit analysis, particularly over shorter prediction intervals.

Lane *et al.*[15] pioneered the use of survival analysis for financial distress prediction in 1986. Their use of the Cox model was empirically comparable to discriminant analysis but with fewer Type I Errors. Similar encouraging results were also found by Crapp and Stevenson[16]. Laitinen and Luoma[17] also used the Cox model, but found it slightly inferior to both discriminant and logit analysis. Shumway[18] used an accelerated failure time survival analysis model that outperformed the traditional techniques at predicting financial distress. More recently in 2008, Gepp and Kumar[19] found the Cox model to be comparable at equal misclassification costs but inferior in adapting to higher Type I Error costs when compared with discriminant analysis and logistic regression. It is important to note that some research has raised questions about whether the proportional hazards assumption of the Cox model is appropriate for financial distress prediction[20]. Survival analysis techniques have also been used to study the influential factors in the survival of specific classes of businesses, such as Internet – based businesses[21] and a particular type of Italian credit bank[22].

Laitinen and Kankaanpää[23] compared decision trees with the Cox model and did not find a statistically significant difference. Apart from this paper, there is a lack of comparisons between decision trees and survival analysis in the financial distress prediction literature.
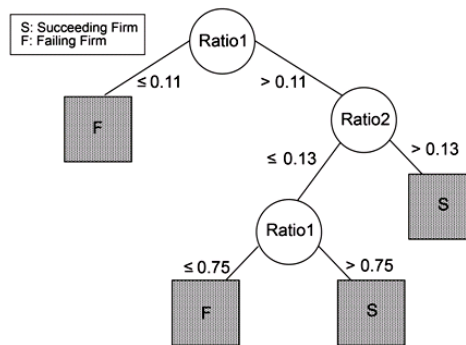
Fig. 1. An example decision tree for classifying businesses.

3. Modelling Financial Distress and Business Failure Prediction

Because of the real – world applicability of these models, Agarwal and Taffler[5] are critical of focusing on theory and instead advise that models be first subjected to thorough empirical testing. While a model's ability to correctly classify the data from which it is developed is important it also needs to be tested for predictive accuracy on data separate from the model building stage as this is a much better guide for future performance.

It is clearly important for financial distress models to minimise both types of misclassification errors:

- Missing financially distressed businesses (Type I Error) will result in financial losses, including those to debtors, investors, suppliers and customers. It will also damage economic stability; and,
- Falsely classifying businesses without financial distress (Type II Error) produces opportunity costs such as missed gains from an investment or business association. It can also result in difficulties for the misclassified business tasks such as raising capital, purchasing on credit and receiving payment before delivery.

While it is fairly safe to assume that a Type I Error is more critical than a Type II Error, a quantifiable difference in misclassification costs has not been agreed upon in the literature as it seems to be subjective and will vary depending on the situation and point of view of the user. Consequently, testing a model over a range of misclassification costs is beneficial.

In addition to accuracy, automated systems need to produce predictions early enough to be useful. In some cases one year ahead might be long enough, but in other cases an earlier prediction might be needed such as when considering long – term business decisions. Hence, as with misclassification costs, testing models over various prediction intervals would be beneficial.

3.1 An overview of decision tree models

Decision trees, or classification trees, are a non-parametric data – mining technique. The trees are built by a recursive process of splitting data when moving from higher-to-lower levels. Figure 1 depicts a tree for predicting business failure that classifies each business as either a succeeding or failing firm. It also illustrates that every non-classification node contains a splitting rule (usually univariate) that describes how data are split. If Ratio1 in Fig. 1 was Current Assets over Total Assets, then the first splitting rule would be to classify each business into the

- Left sub-tree if Current Assets over Total Assets ≤ 0.11, or
- Right sub-tree if Current Assets over Total Assets > 0.11.

The example in Fig. 1 also illustrates that decision trees are simple to understand and interpret. They are also a powerful multivariate approach that can easily model interactions and handle missing data, as well as being simple to develop into automated systems. Unlike parametric models, decision trees do not have assumptions about the underlying distribution of the data and there is no need to consider monotonic transformations such as logarithms.

There are different algorithms that can be used to build decision trees, and the choice of algorithm can make a difference to performance. Decision tree software called THAID was developed for classification tasks by Morgan and Messenger[24] in the 1970s. A decade later, Breiman *et al.*[25] developed a new, sophisticated tree – based software called CART (Classification and Regression Trees), which is now sold by Salford Systems in a popular expanded commercial form. Other algorithms include the popular See5/C5.0 (based on See4.5 and ID3), CHAID (based on THAID) and those found in the MATLAB Statistics Toolbox, IBM SPSS software and S-Plus packages.

3.2 An overview of survival analysis models

Survival analysis techniques analyse the time until a certain event occurs and have been extensively used in the medical sciences. The use of this approach to business failure is fundamentally different from other approaches because it models a timeline instead of a classification problem. This timeline is most commonly described by the survival or hazard function (each is derivable from the other). The survival function $S(t)$ indicates the probability that an individual survives until time t. When applied to financial distress prediction, an individual can be a business and survival represents the absence of financial distress. Contrastingly, the hazard function $h(t)$ indicates the instantaneous rate of death or financial distress at a certain time t.

There are many different survival analysis techniques including regression – based approaches that are well suited for making predictions. The most common is the semi-parametric proportional hazards (PH) model proposed by Cox[26] in 1972, but there are alternatives such as accelerated failure time (AFT) models and Aalen's additive model. Cox's PH model is defined as

$$h(t) = h_0(t)e^{X'\beta+c}, \text{ where}$$

- $h_0(t)$ is the non-parametric baseline function that describes the change in the hazard function over time; and
- $e^{X'\beta+c}$ describes how the hazard function relates to the explanatory variables (X) and is the parametric part of the model, where β is a vector of variable coefficients and c a constant estimated by a method very similar to the maximum likelihood method.

The survival function is then computed as $S(t) = e^{-H(t)}$, where $H(t)$ is the cumulative hazard function from time 0 to t. Survival probabilities can then be compared with cut–off values as is performed when using discriminant analysis and logistic regression.

4. Data and Methodology

This study compares a decision – tree technique and a survival – analysis technique with benchmark techniques. The Cox technique was chosen to represent survival analysis as it is the best known and most commonly applied. CART was chosen to represent decision trees because of its promising empirical results in previous studies[12,13]. Additionally, discriminant analysis (DA) and logistic regression (LR) have been used as benchmarks for this comparison, because of their longstanding and wide – spread use in the field of financial distress prediction.

4.1 Data

A large panel data set kindly provided by Panayiotis Theodossiou has been used for this research. These data have previously been used by Gepp and Kumar[19] to assess a Cox model and Kahya and Theodossiou[27] to test their CUSUM procedure. The sampling methodology has been detailed by Kahya and Theodossiou[27]. Financially distressed firms are deemed such based on debt default criteria to minimise contaminated data and avoid many of the problems with using a legal definition of bankruptcy. Some financially distressed companies never file for bankruptcy because of an acquisition. Furthermore, US law at that time was such that many businesses filed for bankruptcy for legal reasons other than financial distress[27].

The main properties of this data set are described in Table 1. 27 financial variables are used as explanatory variables, comprising 7 liquidity ratios, 8 profitability ratios, 3 management efficiency ratios, 1 activity ratio, 3 leverage ratios, 1 market structure ratio and 4 size variables. These variables were chosen because of their significance in previous

Table 1. Description of data used.

Property	Value
Businesses (Type)	Manufacturing and Retail
Selection Procedure	Random selection
Businesses (Number)	189: 117 successful and 72 failed
Duration	18 years (1974-1991)
Number of Business – Years (Instances)	2,954: 1,923 training sample + 1,031 hold – out test sample

Table 2. Misclassification costs used.

Property	Values Used						
Ratio of Type I to Type II Error Cost (ECR)	1	1.5	2	2.5	3	4	5
Corresponding Cut-off Values [ECR=cut-off / (1-cut-off)]	0.5	0.6	0.67	0.71	0.75	0.8	0.83

financial distress research – more detailed reasoning is provided by Kahya and Theodossiou[27]. The differences between the current and previous year values for each of the 27 financial variables were also used to total 54 explanatory variables. Kahya and Theodossiou[27] clearly stated that these differences provided useful information to financial distress models.

A separate hold – out data set was created to test each model's ability to predict the failure or success of businesses not used in the model training process, which gives a much more realistic view of potential real – world accuracy. This hold – out data set was created as a stratified random sample approximately 35% of the size of the initial data set with 47 successful and 25 failed businesses.

4.2 Methodology

The Cox, DA and LR models were estimated by IBM SPSS Statistics using a forward stepwise procedure (termed 'Forward:LR' in the software for Cox and LR). The significance – level boundaries for entry and removal were set to 5% and 10% respectively for all three models. This is the same method used in Gepp and Kumar[19]. The prior probabilities for the DA models were set to be computed from the proportions in the initial data set, as this has previously been shown to increase the predictive ability of DA models.

Models were assessed on prediction intervals of 1, 2, 3, 5 and 10 years to compare the adaptability of the models to changing lengths of prediction. An x year prediction interval indicates whether a business experiences financial distress within the next x years. Separate CART, DA and LR models were developed for each prediction interval. However, as the Cox model incorporates time, only one Cox model is needed. For example, the one and three – year prediction intervals with the Cox model are obtained using $S(1)$ and $S(3)$ respectively.

Comparisons over various misclassification costs were also conducted because the real world costs are unknown and vary case by case. The misclassification costs were provided as inputs to CART and consequently a separate CART model was developed for each variation. On the other hand with Cox, DA and LR, one model was used for the entire variety of misclassification costs by varying the cut-off values as shown in Table 2.

As the study includes analyses over different misclassification costs the two techniques were compared based on a weighted – error cost measure described in equation (1), for which lower values indicate superior performance. This method was chosen in preference to the area under the Receiver Operating Characteristic (ROC) curve, because of the problems David Hand[28] identified with using the ROC curve.

$$\text{Weighted Error Cost} = \frac{ECR + n_f + n_s}{N} \tag{1}$$

where

- n_f is the number of truly financially distressed businesses incorrectly classified (Type I Error);
- n_s is the number of businesses without financial distress incorrectly classified (Type II Error); and,
- N is the number of instances in the relevant data set (in-sample or hold-out). The division by N is performed to reduce the range of measures in order to improve the ease of interpretation.
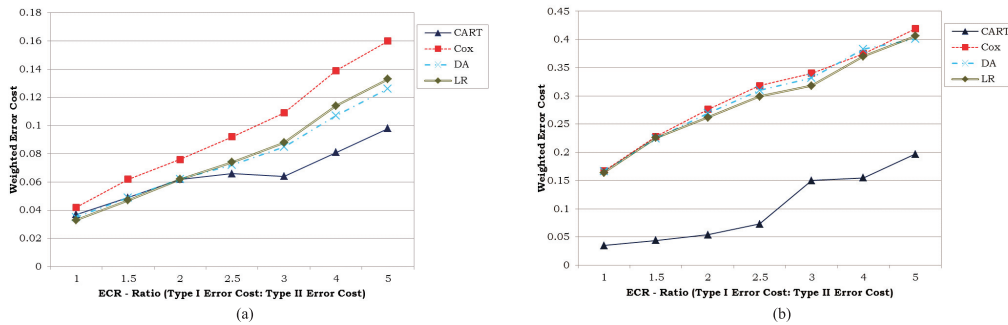
Fig. 2.   (a) The cost of 1-year classifications; (b) The cost of 10-year classifications.

5. Results and Discussion

The models will be compared for classification accuracy and then prediction accuracy following some general findings. Lessons for model development in financial distress prediction revealed in this research include that:

- Profitability ratios are better than other ratios in distinguishing between financial distress and financial health;
- Prediction accuracy decreases for longer predictions and more variables are important in models that predict further into the future, both of which are likely a consequence of the intuitive idea that longer predictions are more complicated. This finding supports using more variables when predicting further into the future;
- In-sample classifications were more accurate than hold – out predictions. In-sample classifications had lower weighted – error cost across all models, ECRs and intervals except for the Cox model with a two – year interval and an ECR of 1.5. This demonstrates the importance of using hold – out samples to test the accuracy of models; and,
- It is important to include both the value for specific variables as well as the one-year change, because they offer different yet important information. This finding is demonstrated by the inclusion of both specific variables and their corresponding one-year change in the most important variable list.

5.1 Comparative performance analysis for classification accuracy

These results represent each model's ability to classify the data that was used to develop the models. This means that the models are classifying data they have already seen in the model development stage. The CART model had the lowest cost of classification for almost all misclassification costs and prediction intervals, except for four cases where it either tied with or was within 13% of the LR model with the lowest cost. In addition, CART improved its classification superiority relative to the other models as the prediction interval increased. The Cox model was the worst classifier in all but five cases, but was comparable to DA and LR for longer prediction intervals. The DA and LR models had similar classification ability. Figure 2 illustrates these results.

5.2 Comparative performance analysis for prediction accuracy

These results represent each model's accuracy when applied to new, unseen data and so are a much better indicator of real – world performance where all the data are new. Figure 3 graphically illustrates the prediction accuracy of each model. The most obvious result is the poor performance of the LR model, which will be discussed at the end of this section. Aside from this model, the overall prediction accuracy of the CART, Cox and DA models is very comparable. The Cox model has the lowest weighted – error cost in 40% of the cases, while the DA and CART models share approximately the other 60%.

While an overall best predictor is not apparent, there are performance differences in specific cases. Some separation in prediction accuracy does occur with higher relative Type I Error costs (ECRs). DA is clearly the best two-year predictor when the ECR is above 2. It could be argued that the Cox model is the best one-year predictor and CART
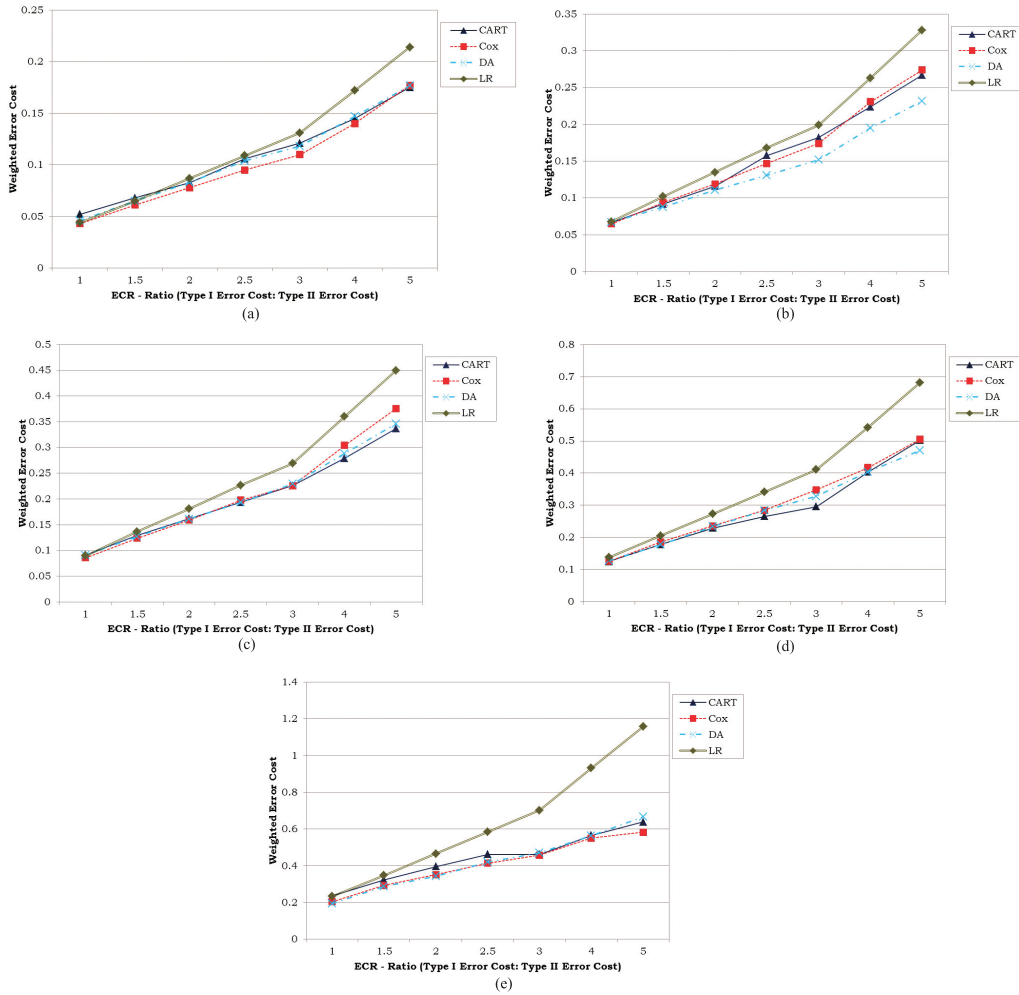
Fig. 3. (a) The cost of 1-year predictions; (b) The cost of 2-year predictions; (c) The cost of 3-year predictions; (d) The cost of 5-year predictions; (e) The cost of 10-year predictions.

the best three-year predictor, but the difference is not substantial enough to place much weight on this finding. It is clear that both the prediction interval and the relative cost of Type I Error influence relative performance, and need to be considered when selecting a model for implementation.

The Cox model has a separate advantage. Unlike survival analysis techniques such as the Cox model, decision trees can only generate a crude probability of failure and cannot easily compare two businesses predicted to be in the same classification group. Further information about the business failure process can also be obtained by analysing (over time) the survival and hazard function output by the Cox survival analysis technique.

The LR model has the highest cost of prediction in all but four cases. Furthermore, the LR model's predictions became relatively worse with higher relative Type I Error costs and with longer prediction intervals. Studying the results in more detail reveals that while the LR model was the best at minimizing Type II Error, it was very poor at avoiding Type I Error.

It is very unusual in financial distress prediction for LR to significantly underperform compared with DA. The explanatory variables found to be statistically significant by LR were compared with those by DA. LR found some relationships that did not occur in DA. In fact, when the *new working capital to total assets* variable was removed
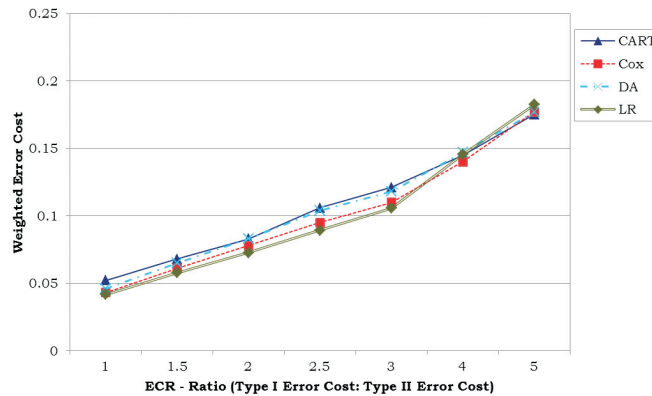
Fig. 4.　The cost of 1-year predictions using a modified LR model.

from the data for prediction intervals of one, two and three years and the *logarithm of deflated sales* removed for five and ten years, the newly estimated LR models had comparable prediction accuracy as demonstrated in Fig. 4. This demonstrates that sample data can have patterns that are not repeated and testing models on hold – out data is critical before real – world implementation.

6. Conclusions

Decision trees, specifically the CART model, had better classification accuracy than the other techniques. More importantly, both the CART decision – tree technique and the Cox survival analysis technique were comparable with each other and discriminant analysis over both a range of misclassification costs and prediction intervals. They were also superior classifiers compared with logistic regression, which performed unusually poorly on the data used in this paper.

The survival analysis and decision – tree techniques investigated here are both useful in financial distress prediction for different purposes. Survival analysis techniques are appropriate for developing a single model to make predictions of varying lengths and to analyse the financial distress process over time. On the other hand, non-parametric decision trees are the best for making accurate predictions without the risk of violating statistical assumptions. These conclusions are based upon a comparison between only one survival analysis and one decision – tree technique and consequently more research should be undertaken to further test these conclusions. For example, this comparison could be extended to include new decision – tree approaches such as Random Forests or time – dependent explanatory variables to exploit the features of survival analysis. Overall, the results presented provide empirical evidence to support the use of survival analysis and decision tree techniques in financial distress warning systems that are useful to most entities in the financial markets.

References

[1] WH. Beaver, Financial Ratios as Predictors of Failure, *Journal of Accounting Research (Supplement)* vol. 4(3), pp. 71–111, (1966).
[2] EI. Altman, Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *Journal of Finance*, vol. 23(4), pp. 589–609, (1968).
[3] E. Deakin, A Discriminant Analysis of Predictors of Business Failure, *Journal of Accounting Research* Spring: pp. 167–179, (1972).
[4] R. Edminster, An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction, *Journal of Financial and Quantitative Analysis 2* vol. 7, pp. 1477–1493, (1972).
[5] V. Agarwal and RJ. Taffler, Twenty-Five Years of the Taffler z-score Model: Does it Really have Predictive Ability? *Accounting and Business Research*, vol. 37(4), pp. 285–300, (2007).
[6] JA. Ohlson, Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, vol. 18(1), pp. 109–131, (1980).
[7] PT. Theodossiou, Predicting Shifts in the Mean of a Multivariate Time Series Process: An Application in Predicting Business Failure, *Journal of the American Statistical Association*, vol. 88(422), pp. 441–449, (1993).

[8] Tan CNW, Artificial Neural Networks: Applications in Financial Distress Prediction & Foreign Exchange Trading, Gold Coast, Australia: *Wilberto Publishing*, (2001).

[9] P. du Jardin and E. Severin, Predicting Corporate Bankruptcy Using a Self-Organizing Map: An Empirical Study to Improve the Forecasting Horizon of a Financial Failure Model, *Decision Support Systems*, vol. 51(3), pp. 701–711, (2011).

[10] H. Frydman, EI. Altman and DL. Kao, Introduction Recursive Partitioning for Financial Classification: The Case of Financial Distress, *Journal of Finance*, vol. 40(1), pp. 269–291, (1985).

[11] P. Joos, K. Vanhoof, H. Ooghe and N. Sierens, Credit Classification: A Comparison of Logit Models and Decision Trees, In *Proceedings Notes of the Workshop on Application of Machine Learning and Data Mining in Finance, 10th European Conference on Machine Learning*, Chemnitz, Germany, pp. 59–72, (1998).

[12] K. Huarng, HK. Yu and CJ. Chen, The Application of Decision Trees to Forecast Financial Distressed Companies, In *2005 International Conference on Intelligent Technologies and Applied Statistics Proceeding*, Taipei, Taiwan, (2005).

[13] A. Gepp, K. Kumar and S. Bhattacharya, Business Failure Prediction Using Decision Trees, *Journal of Forecasting*, vol. 29(6), pp. 536–555, (2010).

[14] MY. Chen, Predicting Corporate Financial Distress Based on Integration of Decision Tree Classification and Logistic Regression, *Expert Systems with Applications*, vol. 38(9), pp. 11261–11272, (2011).

[15] WR. Lane, SW. Looney and JW. Wansley, An Application of the Cox Proportional Hazards Model to Bank Failure, *Journal of Banking and Finance*, vol. 10(4), pp. 511–531, (1986).

[16] HR. Crapp and M. Stevenson, Development of a Method to Assess the Relevant Variables and the Probability of Financial Distress, *Australian Journal of Management*, vol. 12(2), pp. 221–236, (1987).

[17] EK. Laitinen and M. Luoma, Survival Analysis as a Tool for Company Failure Prediction, *Omega*, vol. 19(6), pp. 673–678, (1991).

[18] T. Shumway, Forecasting Bankruptcy More Accurately: A Simple Hazard Model, *The Journal of Business*, vol. 74(1), pp. 101–124, (2001).

[19] A. Gepp and K. Kumar, The Role of Survival Analysis in Business Failure Prediction, *International Research Journal of Finance and Economics*, vol. 16, pp. 13–34, (2008).

[20] JG. Huang and C. Friedman, Modeling Multiperiod Corporate Default Probability When Hazard Ratios Decay, *Journal of Credit Risk*, vol. 5(1), pp. 3–23, (2009).

[21] RJ. Kauffman and B. Wang, Tuning into the Digital Channel: Evaluating Business Model Characteristics for Internet Firm Survival, *Information Technology & Management*, vol. 9(3), pp. 215–232, (2008).

[22] P. Maggiolini and PE. Mistrulli, A Survival Analysis of De Novo Co-Operative Credit Banks, *Empirical Economics*, vol. 30(2), pp. 359–378, (2005).

[23] T. Laitinen and M. Kankaanpää, Comparative Analysis of Failure Prediction Methods: The Finnish Case, *The European Accounting Review*, vol. 8(1), pp. 67–92, (1999).

[24] JN. Morgan and RC. Messenger, THAID: A Sequential Search Program for the Analysis of Nominal Scale Dependent Variables, *Institute for Social Research, University of Michigan, Ann Arbor*, (1973).

[25] L. Breiman, JH. Friedman, R. Olshen and CJ. Stone, Classification and Regression Trees, Wadsworth, (1984).

[26] DR. Cox, Regression Models and Life-Tables, *Journal of the Royal Statistical Society B*, vol. 34, pp. 187–220, (1972).

[27] E. Kahya and PT. Theodossiou, Predicting Corporate Financial Distress: A Time‿series CUSUM Methodology, *Review of Quantitative Finance and Accounting*, vol. 13(4), pp. 323–345, (1999). An earlier version of this paper was presented at the 3rd Annual Multinational Finance Conference (June 1996), and the 156th Annual Meeting of the American Statistical Association, August (1996).

[28] D. Hand, Measuring Classifier Performance: A Coherent Alternative to the Area Under the ROC Curve, *Machine Learning*, vol. 77, pp. 103–123, (2009).