An Evaluation of Decision Tree and Survival Analysis Techniques for Business Failure Prediction

Adrian Charles Benjamin Gepp
BCom, BIT

24 December 2005

Preface

This thesis has been written in fulfilment of the dissertation requirement of the degree of Masters of Information Technology (Honours) at Bond University.

Statement of Originality

The material in this thesis has not been previously submitted for a degree or diploma in any university. To the best of my knowledge this thesis contains no material previously published or written by another person except where due acknowledgement is made in this thesis.

_____

Adrian C B Gepp

Date:

Acknowledgements

A few of the people that have assisted in the production of this thesis are mentioned here.

A very grateful thank-you goes to my supervisor, Associate Professor *Kuldeep Kumar*, for all his guidance and encouragement: completing this thesis has been a fantastic learning journey.

I would like to extend my gratitude to Prof. *Panayiotis Theodossiou* and *Duen-Li Kao* for kindly proving me with an electronic copy of data sets used in their previous research.

Personally, a very special thank-you goes to *Danielle Forrest* for providing an abundance of happiness and being the most supportive friend a person could hope to have.

Finally, I would like to once again thank my parents, *Mum* and *Dad*, for providing the best assistance and friendship possible.

Some Historical Advice

"The next man that comes to you with a good thing [business opportunity], hit him with a club"

Advice from the Victorian Insolvency Court to Dow J Lamont in 1893.

quote from
Murray-Smith, S. (Editor). (1992). *Dictionary of Australian Quotations.* Octopus, Melbourne.

Abstract

Accurate business failure prediction models would be extremely valuable to many industry sectors, particularly in financial investment and lending. The potential value of such models has been recently emphasised by the extremely costly failure of high profile businesses in both Australia and overseas, such as HIH (Australia) and Enron (USA). Consequently, there has been a significant increase in interest in business failure prediction, from both industry and academia.

Statistical models attempt to predict the failure or success of a business based on publicly available information about that business (or its industry and the overall economy), such as accounting ratios from financial statements. Discriminant and logit analyses have been the most popular approaches, but there are also a large number of alternative techniques available. This thesis reviews the various techniques used in previous studies, and presents the results obtained from using two alternative techniques that have shown promise, namely survival analysis and decision trees. These techniques have been compared with the more popular techniques and to each other. Overall, the results suggest that decision trees provide the most accurate predictions of business failure. In addition, while survival analysis techniques are slightly less accurate than other techniques they provide more information that can be used to further the understanding of the business failure process.

Contents

1 Introduction **1**

1.1 Business Failure Prediction . . . 1

1.1.1 BFP Models . . . 2

1.2 Research Aims . . . 4

1.3 Thesis Structure . . . 4

2 The Business Failure Prediction Field **5**

2.1 Different Business Failure Prediction Models . . . 5

2.1.1 Univariate Analysis . . . 5

2.1.2 Standard Multivariate Techniques . . . 6

2.1.3 Discriminant Analysis . . . 8

2.1.4 Logit and Probit Analysis . . . 10

2.1.5 Human Information Processing . . . 13

2.1.6 Artificial Neural Networks . . . 13

2.1.7 Sequential Procedures . . . 16

2.1.8 Other Lesser-Researched Techniques . . . 16

2.2 Decision Tree Techniques . . . 18

2.2.1 Theoretical Analysis . . . 21

2.2.2 Previous Studies in BFP . . . 22

2.3 Survival Analysis Techniques . . . 25

2.3.1 Key SA Concepts . . . 25

2.3.2 Estimation of SA Models . . . 27

2.3.3 Medical Applications . . . 29

2.3.4 Theoretical Analysis . . . 30

2.3.5 Previous Studies in BFP . . . 31

2.4 Potential Research in DTs and SA . . . 34

2.4.1 Decision Trees . . . 34

2.4.2 Survival Analysis . . . 34

3 Decision Tree Research **36**

3.1 Aim . . . 36

3.2 Data . . . 36

3.2.1 Definition of Business Failure . . . 37
3.2.2 Explanatory Variables . . . 37
3.3 Methodology . . . 37
3.3.1 General Methodology . . . 38
3.3.2 DA and RPA Models . . . 40
3.3.3 See5 Models . . . 41
3.3.4 CART Models . . . 41
3.4 Results . . . 42
3.4.1 Effect of Model Complexity . . . 42
3.4.2 Empirical Comparison of Techniques . . . 44
3.4.3 Analysis of the Importance of Variables . . . 50
3.5 Conclusions . . . 54

4 Survival Analysis Research 55
4.1 Aim . . . 55
4.2 Data . . . 55
4.2.1 Definition of Business Failure . . . 56
4.2.2 Explanatory Variables . . . 56
4.2.3 Hold-out Data . . . 57
4.3 Methodology . . . 57
4.3.1 General Methodology . . . 58
4.3.2 DA and LA Models . . . 59
4.3.3 Cox Model . . . 60
4.3.4 Hybrid Models . . . 61
4.4 Results . . . 61
4.4.1 Analysis of The Failure Process . . . 61
4.4.2 Empirical Comparison of Techniques . . . 63
4.4.3 Analysis of the Importance of Variables . . . 75
4.5 Conclusions . . . 77

5 Comparative Research and Overall Conclusions 78
5.1 Comparative Research . . . 78
5.1.1 Data and Methodology . . . 78
5.1.2 Results . . . 79
5.2 Overall Conclusions . . . 83
5.3 Further Research . . . 84

A Decision Tree Research Results 93

B Survival Analysis Research Results 95

C Comparative Research Results 102

Chapter 1

Introduction

1.1 Business Failure Prediction

The field of *business failure prediction* has many aliases, including *bankruptcy prediction*, *firm failure prediction* and *financial (de)stress prediction*. Hereafter it will be referred to as business failure prediction (BFP). As the name suggests, BFP involves developing models that attempt to predict the financial failure of a business before it actually happens. Accurate business failure prediction models would be extremely useful and valuable in the real world. The potential value of such models has been recently emphasised by the extremely costly failure of high profile businesses in both Australia (HIH and OnteTel) and overseas, particularly Enron in the United States. The impact of business failures is highlighted by the fact that over the last eight years more than 4,000 individual Australians, on average, have gone bankrupt due to business failures[1]. Consequently, there has been a significant increase in interest in business failure prediction, from both industry and academia.

Statistical BFP models attempt to predict the failure or success of a business based on publicly available information about that business, such as financial ratios from financial statements. In addition, some studies also include indicators of industry and economy wide performance to aid in the business failure predictions. BFP models aim to produce accurate and informative forecasts about the survival of businesses in order to achieve the following two goals:

- Successful businesses will not be identified as failing businesses; and more importantly,
- Failing businesses will be highlighted as failing businesses

Achieving these goals would be extremely important to many industry sectors for various reasons. With accurate business failure prediction:

[1]The data source used to derive this statement was obtained from the Australian Government, Insolvency and Trustee Service Australia (ITSA) (2005).

- Banks, investment banks, credit unions, and other financial institutions could avoid lending to businesses that will fail, and thus never repay their loans.
- The financial investment sector could avoid losing money by not investing in failing businesses, and consequently improve the investment's risk-return trade-off.
- Businesses could establish long-term relationships with other businesses (such as suppliers) that will be successful, and not fail, in the future. Therefore, the longevity and viability of the business relationship is more secure.
- Regulatory bodies, such as the Australian Securities and Investments Commission (ASIC), can make early identifications of failing businesses. For example, it is interesting to note that bankruptcies in Australia historically increase by almost 10% in the quarter just prior to tax time (June 30)[2]. Early identification of these failures and other failing businesses will assist regulatory bodies to ensure that business failure is 'handled' legally and illegal activities, such as avoiding taxes or diluting debt-holders claim by issuing substantial common stock dividends prior to failure, are avoided.

In addition to all the industry sectors that will profit from accurate BFP models, individuals dealing with businesses could also profit from using accurate BFP models in order to only deal with successful businesses. Moreover, it is possible for businesses to use a BFP model to predict their own failure or survival, and use this information as a measure of financial health for management decision making. Overall, accurate BFP models will increase general confidence in investment, lending and the development of profitable business relationships. Thus, lending and investment (in successful businesses) will increase, which will result in increased stable economic growth for the benefit of everyone involved.

1.1.1 BFP Models

There are three general types of research in the field of BFP (adapted from Laitinen and Kankaanpää (1999)); these are:

1. Theoretical modelling of the failure process of businesses. Only a few studies have focused on this area, particularly Wilcox (1971) and Santomero and Vinso (1977).
2. Searching for the set of explanatory variables (usually accounting and financial ratios) that best explain and predict business failure. Research focusing on this area is common in the accounting literature.
3. Searching for the 'best' empirical method for BFP. Research into this area often studies the predictive power of different explanatory variables as a by-product.

[2]The data source used to derive this statement was obtained from the Australian Government, Insolvency and Trustee Service Australia (ITSA) (2005).

Recently, this 3rd area of research has been the main contributor to the field of BFP. This has been further enhanced by the continuing improvements in computer technology that have enabled extraordinary advances in data analysis. The ramifications of this are that theorists, who develop mathematical and statistical models, no longer have to be concerned about the feasibility of being able to apply their model. Thus, the more computationally intensive models are becoming practically viable options for BFP, and hence the popularity of the 3rd area of research.

The most important characteristics of a BFP model are its two forms of accuracy, classification and prediction. A model's *classification accuracy* is obtained by assessing its accuracy on the data set from which it was developed. Following that, the more important *prediction accuracy* of the model is assessed by its use on a brand new set of data, which reflects how well the model will perform on future predictions. Nevertheless, when measuring either classification or prediction accuracy there is a real-world important consideration that should be noted. It is a more critical error to classify a failing business as successful (Type I Error) than to classify a successful business as failing (Type II Error). The reason for this is that a Type II Error only creates a lost opportunity cost from not dealing with a successful business, for example, missed potential investment gains. In contrast, a Type I Error results in a realised financial loss due to involvement with a business that will fail, for example, losing all money invested in a bankrupt business. Thus, the misclassification costs are not equal in the real world. Therefore, when analysing BFP models a higher weighted penalty should be imposed for a misclassification of a truly failing business (Type I Error). A quantified difference in misclassification costs has not been agreed upon in the literature, as it seems to vary for different circumstances and usually involves subjective decision making.

Additional information to the boolean fail/succeed decision is also a desirable feature of a BFP model. As future predictions can never be made with absolute certainty, the confidence level (probability) of failure predictions is also useful and important. For example,

- The interest premium on loans are calculated based upon the probability that the recipient business will fail.
- Investors often weight a business's expected future cash flows at (1-*probability of failure*) to obtain a probability weighted stream of cash flows, which can then be used to more accurately calculate the fair price of a stock. This fair price is then compared with the market price of the stock to determine whether to buy, sell or hold the stock.
- Regulatory bodies can prioritise their planning and monitoring according to the businesses that are most likely to fail.
- When setting up business relationships, businesses can be compared with each other based on the likelihood (probability) they will fail.

This extra information, such as probabilities of failure, has been criticised as inaccurate due underlying distribution assumptions that are usually violated in the real world. Overall,

extra information is useful, but accuracy still remains the most important characteristic of a BFP model.

1.2 Research Aims

The overall goal of this research work is to advance the field of business failure prediction by significantly contributing to the search for the 'best' business failure prediction technique. *Decision Trees* and *Survival Analysis* are two lesser-studied approaches that have shown promise in predicting business failure. These techniques will be applied to BFP with the following objectives, to determine:

- What information, in addition to the boolean prediction, is provided by these two approaches.
- The classification and prediction accuracy of the two approaches.
- How these two techniques compare with the more popular and traditional approaches of discriminant and logit analyses.

1.3 Thesis Structure

The remainder of this thesis is structured as follows. Chapter 2 contains an introduction into, and detailed review of, the various techniques that have been previously used to predict business failure, with a particular emphasis on decision trees and survival analysis. Chapters 3 and 4 respectively present my research into using decision trees and survival analysis for business failure prediction. Each chapter includes the aims, data used, methodology applied and the results then conclusions obtained. Lastly, Chapter 5 presents a comparison of decision trees and survival analysis, as well as the final conclusions of this research and possible future work.

Chapter 2

The Business Failure Prediction Field

2.1 Different Business Failure Prediction Models

Many different techniques have been applied to business failure prediction (BFP) since its beginning in the 1960s. The field arguably started earlier, but the first statistical and mathematical models for bankruptcy prediction were published in the 1960s. The various techniques used since then are reviewed and referenced in this chapter. The two techniques that are the focus of this thesis, decision trees and survival analysis, will be covered in more detail as separate sections (2.2 and 2.3) following the review of other techniques.

2.1.1 Univariate Analysis

Early attempts to use financial ratios[1] to predict business failure stem from the work of Patrick (1932). This work was later extended by Beaver (1966), who presented the first modern statistical model for BFP. Beaver developed a univariate model, and used a set of 30 financial ratios that were chosen with regard to the following three criteria:

- Prevalence in previous literature,
- Performance in previous studies, and
- Consistence with Beaver's logical (not theoretical) cash flow concept that placed a preference on ratios derived from cash flow statements. The logic behind this is that bankruptcy and insolvency is related to a lack of liquid assets (particularly cash) relative to liabilities due, rather than to accrued accounting assets and liabilities.

Beaver tested his model on 79 failed businesses and 79 similar successful businesses between 1954 and 1964. A cut-off point was identified for each financial ratio, which separated the predicted successful businesses from the predicted failure businesses. From the univariate analysis of each chosen ratio on the data set, a ratio set that consisted of ratios with

[1] The term *financial ratios* used in this thesis refers to both accounting ratios and other cash flow related finance ratios.

the greatest predictive power was established. This set comprised (in declining order of predictive power):

1. Cash flow to total debt,
2. Net income to total assets,
3. Total debt to total assets,
4. Working capital to total assets,
5. Current ratio = *current assets / current liabilities*, and
6. No credit interval = $\frac{immediate\ (or\ quick)\ assets - current\ liabilities}{operating\ costs - depreciation}$.

It is interesting to note that Beaver considered business failure to be much broader than just liquidation: a business was considered to have *failed* financially if **any** of the following were satisfied:

- the business had gone into liquidation,
- there was an overdrawn bank account,
- there had been default on debt (such as a bond), or
- preferred stock dividends had been missed.

Beaver's univariate approach did not contain an overall measure of financial distress, which led to the problem that different ratios made conflicting predictions about a given business. In addition, it was noted that one ratio alone could not encompass the complexity of business failure. This model's error was estimated at 22% Type I Error and 5% Type II Error for one-year ahead predictions. In addition, although the Type II Error remained constant for longer-term prediction, Type I Error increased as the length of the predictions increased. Hence, Beaver's model performs better for shorter predictions. Overall, Beaver's work was pioneering in the BFP field, and sparked the development of using statistical models to predict the failure of a business.

2.1.2 Standard Multivariate Techniques

Researchers in most areas simultaneously collect data on multiple variables, because most phenomena are too complex to be described by one single variable. Therefore, there is a need for multivariate statistical techniques to describe and analyse multivariate data and its internal relationships. It has been shown that multivariate techniques are useful in range of areas, including medicine, sociology, economics, psychology, and biology (Kumar and Ganesalingam, 2001). There are many factors (variables) that affect the financial stress of a business; therefore, multivariate techniques are extremely important in BFP. The main four multivariate techniques applied to BFP are described below.

Component Analysis (also referred to as principal component analysis) involves attempting to express a system with p components in a linear combination of k principal components, where $k < p$ and the k components are representative of the system. That is, the goal is to explain the variance-covariance matrix with a *linear* combination of variables that is less that the number of original variables. It is important to note that every principal component (variable) was previously an original component (variable) of the system. Therefore, a common tool used in component analysis is a correlation matrix to determine the similarity between variables within the system. Component analysis often reveals new and unsuspected relationships between components, and therefore new interpretations can be made. Hence, the objectives of component analysis are:

- data reduction (through less components)
- new interpretation (through discovery of new relationships)

It should also be noted that component analysis is quite often only a means to an end; that is, the resulting k principal component system is often input into a standard multiple regression or cluster analysis. The reason for performing the component analysis first is to identify any different interpretations and also to increase the efficiency of the next statistical process due to the data reduction.

Factor Analysis is similar to component analysis in that the goal is to express a system in a linear combination of a number of variables fewer than in the original system. However, with factor analysis the variance-covariance matrix is estimated by a few (less than the original number of components) underlying, but unobservable, random quantities termed *factors*. That is, factors are not components of the original system (although they may be derived from some of the original components). For example, the intelligence of top management may be an underlying factor in the financial success of a range of businesses in different industries. The way to determine which factors to use is by testing possible random factors and choosing those that have high correlation with the variables being modelled. It is important to note that due to the random quantities involved each application of a factor analysis must be examined by the researcher to determine whether the factors are consistent with the data being modelled.

Discriminant Analysis is a technique used to separate out distinct sets of objects or observations, and then allocate new observations (or objects) to these defined groups (or sets). For example, BFP researchers often separate data into two groups: distressed and non-distressed companies. In general, discriminant analysis is exploratory in nature, and is often used when a causal relationship is not completely understood. The goal is to describe the data (graphically or algebraically) in terms of distinct groupings that are separated from each other by as much as possible. This goal is descriptively

named *separation*; however, it is conventionally named *discrimination* as proposed by Fisher (1938).

Cluster Analysis is similar to discriminant analysis in that it is an exploratory technique that groups data together. The difference comes from the fact that discriminant analysis deals with a known number of groups. However, cluster analysis is more primitive as it makes no assumptions about the number of groups or group structure; that is, groups are created based on natural groupings in the data set being analysed. For example, a cluster analysis may reveal three natural groupings (highly distressed, mid-range and low distress businesses) compared with the two defined groups in discriminant analysis. As cluster analysis is often dependent on the definition of similarity being used it is desirable to develop a quantitative method to define similarity, but often analysis by the 'human eye' is effective. Furthermore, as there is not enough time to choose the best grouping from all the possible groupings (this requires an order factorial algorithm) a 'human eye' cluster analysis is often useful to obtain a good, but not the best, grouping. Overall, although primitive, cluster analysis is useful in providing a simple method to asses dimensionality, outliers, simplify data, and suggest hypotheses about relationships amongst data.

Standard multivariate analysis textbooks such as Johnson and Wichern (1988) and Flury (1997) have more details on all of the above techniques. In addition, Kumar and Ganesalingam (2001) apply and review these four main simple multivariate techniques for application to BFP. Overall, the most prominent simple multivariate technique used in BFP is discriminant analysis, which will be discussed specifically related to BFP in the following section.

2.1.3 Discriminant Analysis

Three forms of Discriminant Analysis (DA) have been applied to BFP: Linear, Multiple and Quadratic DA. Multiple Discriminant Analysis (MDA) involves 2 or more classification groups; however, Linear Discriminant Analysis (LDA), a type of MDA, involves *precisely two* classification groups. For example, LDA can perform boolean classification into successful and failing businesses, but only MDA can classify businesses into low, medium or highly financially stressed businesses. However, both LDA and MDA use a *linear* separating function that separates the businesses into various categories based on a *linear* combination of explanatory variables. On the other hand, Quadratic Discriminant Analysis (QDA) does not impose this restriction and can be used where the separating function is thought to be quadratic; that is, the separating function is a quadratic combination of the explanatory variables. It is important to note that the inclusion of quadratic terms in QDA makes the analysis significantly more computationally intensive, and consequently demands more computing time. However, with the exponential growth in computational power over the last decade, this is unlikely to be a significant issue for researchers or businesses.

Altman (1968) pioneered the use of DA, which was the first multivariate approach applied to BFP. This extended the work of Beaver (1966) by addressing the problem that various ratios made conflicting predictions, and incorporates the concept of a composite measure of business distress. Altman (1968) presented a MDA model to predict business failure, in which the information from several variables (ratios) was combined into a single weighted score[2] for each business. This score is calculated based on the following general formula (the only difference for QDA is that there are quadratic terms):

$$Z = a_1x_1 + a_2x_2 + \ldots + a_nx_n + c$$

where Z is the score, x_i are the independent variables, and a_i and c are the estimated parameters. The separating function for Altman's well-published pioneering MDA model (Altman, 1968) is:

$$Z = 0.012x_1 + 0.0141x_2 + 0.033x_3 + 0.006x_4 + 0.999x_5$$

where
x_1 = working capital over total assets,
x_2 = retained earnings over total assets,
x_3 = earnings before interest and tax over total assets,
x_4 = market value of equity over book value of total debt, and
x_5 = sales over total assets.

It is interesting to note that no cash flow ratios were found to be significant, which contrasts the emphasis placed on cash flow ratios by Beaver (1966).

Cut-off scores are then generated based on sample results and used to classify each observation. Altman's MDA used 2 cut-off scores (1.8 and 2.7) to classify businesses into three categories as shown in the table below.

Z-score lookup	**Prediction**
$Z > 2.7$	Success
$Z < 1.8$	Failure
$1.8 \leq Z \geq 2.7$	Inconclusive

Although the probability of failure or success is not an explicit output of this model, a relative measure of probability can be obtained by calculating the difference between Z-scores and cut-off values. For example, a business with a Z-score of 1.0 is more likely to fail than a business with a Z-score of 1.7. Modern statistical packages, such as SPSS, are able to calculate probabilities of group membership based on the difference between the Z-score and cut-off values.

Altman's DA model outperformed Beaver's univariate model for one-year ahead predictions of business failure; however, Altman's model was not as accurate for longer predictions.

[2]This score is often referred to as a *Z-score*.

It is also interesting to note that Martin (1977) concluded that, using MDA with more than two classification groups was more accurate that using LDA (MDA with two classification groups).

The use of LDA and MDA has been further developed since 1968; principally, Deakin (1972) increased the number of independent variables to 14[3], and Edminster (1972) applied a model similar to Altman's to small businesses. However, linear MDA and LDA have two major statistical requirements which are often violated (Laitinen and Kankaanpää, 1999); they are that

1. The independent variables are multivariate normal, and that
2. The covariance of the two classification groups should be equal.

For example, the first requirement is not met if an independent dummy variable is introduced; however, techniques such as log or square root transformations and elimination of outliers can be used to aid consistency with the first assumption (Laitinen and Kankaanpää, 1999). Richardson and Davidson (1983) found that if the first assumption were violated the model was sensitive to the data used in the estimation of model parameters; however, the practical significance of the difference may be minor. Moreover, to overcome the second requirement quadratic discriminant analysis can be used. However, Altman et al. (1977) and Hamer (1983) found that a quadratic function did not consistently improve practical accuracy even though it improved statistical validity. Furthermore, all discriminant analysis models assume equal misclassification costs of Type I and II Error, which is usually violated (as previously discussed in Section 1.1.1 on page 3). However, once the discriminant analysis model has been generated and the probabilities of failure for each business calculated, the default failure cut-off probability of 50% (for LDA) can be changed. Although this technique does not use the misclassification costs as an input, changing the cut-off probability represents varying the misclassification costs. For example, assigning businesses to the failure group if their probability of failure is greater than 40% represents predicting more failures and reducing the more costly Type I Error. In fact, it represents the following misclassification costs:

- Type II Error Cost $= i$
- Type I Error Cost $= \left(\frac{100\%-60\%}{40\%}\right) \times i = 1.5i$.

This method was used by Frydman et al. (1985), who found that it was not as sensitive to differing misclassification costs as techniques that took the misclassification costs as inputs when developing the model.

2.1.4 Logit and Probit Analysis

Logit Analysis (LA) generates a score for each business similar to DA; however, it is free from the normality and equal covariance assumptions of DA. LA is based upon the cumu-

[3]Deakin (1972) used the same variables as Beaver (1966).

lative logistic function (CLF), and due to its non-linear nature the coefficients are usually estimated using the maximum likelihood method (Kumar and Tan, 2005). Furthermore, unlike the difficult interpretation of the Z-score in the DA model, the score in LA can be directly interpreted as the probability of failure (where the CLF ranges from 0 to 1). The general relationship between the two models is demonstrated below (Z was defined in the previous DA section):

$$\begin{aligned} \ln\left(\frac{P(Z)}{1-P(Z)}\right) &= Z \quad \textit{or rearranged as} \\ P(Z) &= \frac{1}{1+e^{-Z}} \end{aligned}$$

Once each business has an associated probability of failure, n cut-off (or critical) values can be established to separate the businesses into $n+1$ groups similar to DA. However, there is usually one cut-off value that separates businesses into failure and success groups. As in DA, the cut-off values in LA can be changed to cater for different misclassification costs: as Type I Error is more serious, the chosen critical value is often lower than 0.5. Furthermore, inherent in the CLF is that businesses with probabilities closer to 0.5 are more sensitive to changes in the independent variables, as shown by the sleeper slope of the CLF close to values of 0.5 in Figure 2.1. This behaviour is logical as short-term changes will not have as

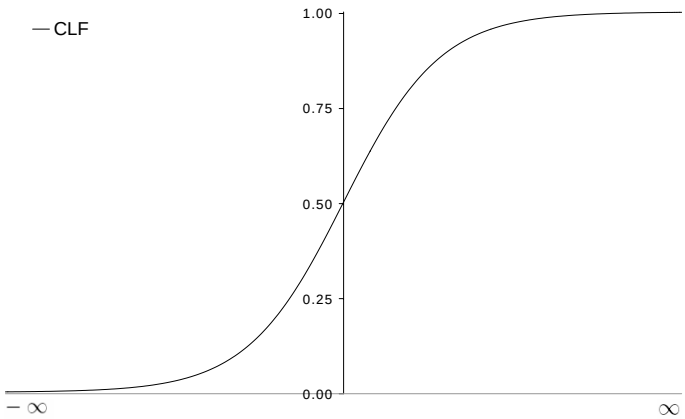

Figure 2.1: The CLF used in Logit Analysis.

great an effect on extremely distressed or successful businesses as it will on those businesses which are borderline.

Ohlson (1980) is the seminal study for applying LA to BFP. He produced three separate LA models to predict one, two and three year ahead failure. Fourteen ratios were used as predictors, consisting of standard accounting ratios[4], dummy variables based comparisons of balance sheet figures, and a variable representing the change in net income over the last year (which comprises time-lagged information). The empirical results for his model were disappointing, but he showed that LA is more statistically valid and easier to interpret than

[4]One of the predictors, *cash provided by operations/total liabilities*, contained cash flow information.

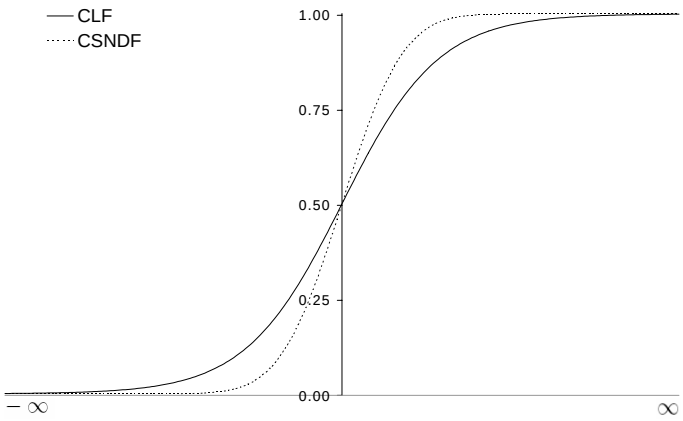

Figure 2.2: A comparison of CLF used in LA and CSNDF used in PA.

DA. In addition, subsequent studies on LA have shown that it is usually slightly empirically superior to DA in both classification and prediction accuracy (Laitinen and Kankaanpää, 1999). However, Martin (1977), Collins and Green (1982) and Hamer (1983) all stated that the overall classification accuracy of DA and LA is not significantly different. Collins and Green went on to further state that Type I Errors appear to be lower in LA, but the additional computational power required for LA is only justified if Type I Errors are very large. This condition may well be justified as, since 1977 when the statement was made, there have been significant increases in computational power and the high cost of Type I Errors has been highlighted by the failure of large businesses (such as HIH and Enron). Overall, it is generally agreed that LA is not less accurate than DA. Therefore, due to its improved statistical validity and straight-forward interpretation, LA has become more popular than DA.

In addition to LA, probit analysis (PA) has also been applied to BFP. The only difference is that PA uses the cumulative standard normal distribution function (CSNDF) instead of the CLF, whereby

$$P(Z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{\frac{-Z^2}{2}} dZ$$

Although more complicated, this function has a very similar shape to that of the CLF, as illustrated in Figure 2.2. This fundamental similarity means that both LA and PA usually produce the same conclusions for the same data. However, PA is more computationally intensive than LA due to the non-linear estimation needed (Gloubos and Grammatikos, 1988); therefore, unlike LA, PA is not prominent within the BFP literature.

For more information on LA refer to a multivariate statistics text such as Flury (1997). There is also an interesting paper by Pacey and Pham (1990), who state the following major problems with DA and LA:

- Arbitrary estimation of cut-off values,
- Assumption of equal misclassification costs in model estimation stage, and
- Use of choice-based samples for model estimation.

Pacey and Pham even suggest that a naive model that assumes all businesses to be bankrupt may outperform both DA and LA. These points are valid, but realistically the cut-off values can be estimated mathematically for given misclassification costs. Therefore, the first problem is not related to the model, but rather caused by a lack of agreement on the actual misclassification costs. In addition, the majority of studies have shown their DA and LA models to be as good as or better (although often not significantly better) than this naive model.

2.1.5 Human Information Processing

Human Information Processing (HIP) involves utilising the existing ability of human decision makers to use information (usually accounting information) to group companies according to their probability of future bankruptcy. Hence, HIP attempts to model the relationship between cues and the decision, rather than the information processing used to form the decisions. Consequently, interviewing and questionnaires are frequently used research techniques.

The pioneering studies in this area are by Libby (1975) and Casey (1980)[5]. Abdel-Khalik and El-Sheshai (1980) presented results of a quantitative comparison that found mathematical and statistical (quantitative) models to be superior decision makers compared with experienced professionals. Moreover, the HIP model produces decisions without a level of certainty. These are the main reasons, together with the recent dramatic increase in computational power, that the emphasis has shifted mainly to developing quantitative models rather than using HIP models. Nevertheless, HIP generates some interesting results that can be used to improve mathematical and statistical models. For example, conclusions such as (1) a greater amount of accounting information does not improve predictive power (Casey, 1980) and (2) the choice (rather than the processing) of information is most important (Abdel-Khalik and El-Sheshai, 1980), are both useful background knowledge when developing mathematical and statistical models.

2.1.6 Artificial Neural Networks

Although somewhat related to HIP, Artificial Neural Networks (ANNs) are far more powerful. ANNs were developed as a computational technique to model the inner-workings of the human brain, as a type of *Artificial Intelligence*. The first major steps towards developing a computer information processing system based on the human brain was made by McCulloch and Pitts in 1943. These models have since been applied to BFP and become somewhat popular in the field of BFP, although not as popular as the major techniques of DA and LA.

Essentially, an ANN is a soft computing technique modelled on the workings of the human brain. The topology of an ANN (in BFP) involves connected layers of neurons:

[5]Laitinen and Kankaanpää (1999) contain a slightly more detailed review of HIP with references to more HIP studies.

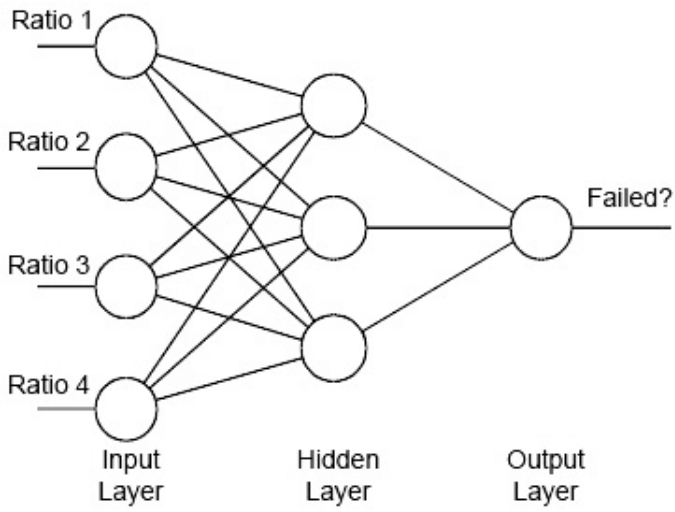

Figure 2.3: A simple example ANN for BFP.

- One layer of input neurons (usually financial ratios),
- One or more hidden layers of interconnected neurons, and
- One output layer of neurons (usually just one boolean fail or success neuron).

A simple example of an ANN for BFP with one hidden layer is shown in Figure 2.3. Each connection between neurons has an associated weight that relates to the likelihood that the connection will be fired and used in the decision. Prior to its use, the ANN is 'trained' like a human brain. This procedure usually involves processing data with known results. An ANN with back-propagation is then 'rewarded' for a correct classification by increasing the weight of the neuron-connections that led to the correct classification. Similarly, the weights of neuron-connections that lead to an incorrect classification are reduced. Eventually, a suitable weighted interconnection of neurons is established that can then be used for making predictions. This is known as supervised learning, but there are also other different training techniques such as an unsupervised and graded learning. Laitinen and Kankaanpää (1999) provide a brief discussion of these training methods, and Tan (2001) covers them in more detail. The training of an ANN plays a large factor in its success. Training an ANN is a difficult and complex process that involves many decisions and refinements: the training method must be chosen along with many other parameter values such as the learning rate, momentum and input noise. Optimising the various parameters of an ANN is an art rather than a science (Kumar and Tan, 2005). Furthermore, the process of optimising the optimal number of hidden layers and neurons in those layers is also an art. The 'trial and error' style of process associated with training an optimal ANN for prediction causes the model quality to vary substantially depending upon the individual undertaking the training process. Thus, an ANN approach is more time consuming than alternative approaches and is usually only used by individuals familiar with them, which has been a barrier to increasing the popularity of ANNs.

The major advantage of ANNs is that they do not have the restrictive assumptions, such as normality and linearity, of the (previously discussed) conventional statistical methods. Elliot and Kennedy (1988) provide a comprehensive review of the assumptions made in other conventional techniques that are not made with ANNs. In addition, due to their flexibility, ANNs can deal with outliers, missing data and multicollinearity better than other techniques. Despite their statistical validity, ANNs have produced mixed results. Laitinen and Kankaanpää (1999) claim that ANNs have not produced significantly superior results comared with other well-known techniques, and show that in their study DA and LA were comparable with ANNs. However, Tan (2001) points out that there have been published applications of ANN in BFP that have outperformed DA and LA. The vast majority of these successful studies have used forward feeding back-propagation neural networks, such as Odom and Sharda (1990), Coleman et al. (1991), Coats and Fant (1992), Tam and Kiang (1992), and Fletcher and Goss (1993). In addition, Salchenberger et al. (1992) demonstrated an ANN model that, at a minimum, performed as well as LA. Furthermore, they showed that if the cut-off value was decreased in an attempt to reduce Type I Errors, then the subsequent increase in Type II Error was less for their ANN model, when compared with LA (Tan, 2001). Overall, the successes that ANNs have had has been in short term prediction (predominantly one year prediction), which is mainly due to the difficulty of adding a time element to ANNs (Kumar and Tan, 2005).

The major criticism of ANNs is that they are a black-box approach. Although they output a continuous score that can be compared with cut-off values to generate a prediction of failure or success (as in DA and LA), the internal logic is hidden from users. That is, it is not possible to gain a full understanding of the significance of each input variable from an ANN and the interpretation of an ANN model is nearly impossible for complex models such as those needed for BFP. However, there has been recent research into methods for converting ANNs into transparent models after they have been trained. For example, RULEX is a rule-extraction ANN program, developed at Queensland University of Technology, that is designed to enable the extraction of rules from an ANN model [RULEX source code available at http://sky.fit.qut.edu.au/~andrewsr/rulexsoftware.html]. The inter-related rules extracted from RULEX can then be efficiently implemented as a decision tree or other logical structure (refer to Section 2.2 for more details on decision trees in BFP).

Genetic algorithms (GAs) have been proposed to overcome the difficulties associated with training ANNs by automating the training process. GAs are an optimisation technique based on Darwinian evolution, which are discussed more in Section 2.1.8. Anandarajan et al. (2001) presented a GA trained ANN model that had greater predictive ability than both a back-propagation ANN and DA model. Fanning and Cogger (1994) also presented a similar model that was comparable to LA and superior to a back-propagation ANN in most instances.

Further information on ANNs, and their application in BFP, is provided in the excellent book by Tan (2001).

2.1.7 Sequential Procedures

Healy (1987) presented cumulative sum (CUSUM) procedures that essentially detect a shift in a series of variables' values from a 'good' distribution to a 'bad' distribution. CUSUM procedures, which date back to 1954, are a set of sequential procedures based on likelihood ratios. These procedures attracted the CUSUM name as they reduce to calculating cumulative sums for many common distributions. CUSUM procedures detect the optimal starting point of the shift and then provide a signal of the shift as soon as possible after the shift occurs. Healy demonstrated the application of CUSUM procedures for detecting shifts in the mean and covariance matrix of a multivariate normal distribution. He noted that this method is very efficient at detecting a shift in the mean, when the mean of the 'good' and 'bad' distributions are known. However, it was also noted that the CUSUM procedures were still applicable in situations when at least the direction of slide towards the 'bad' distribution was known, as is the case in BFP.

This CUSUM concept had obvious applications in BFP, but were not noted by Healy. Nonetheless, Theodossiou (1993) further developed Healy's model and outlined a sequential procedure for detecting when a business shifts from 'good' financial performance to 'bad' financial performance. In terms of BFP, the CUSUM model is a dynamic time-series extension of DA. That is, the model explains a business's tendency towards financial failure over time. A desirable feature of the CUSUM model is that good performance is forgotten quickly in comparison to poor performance, which is remembered longer. This is desirable feature as it inherently treats Type I Error correctly as more serious. Empirically, Theodossiou convincingly showed that this model was superior to DA with his data set.

Kahya and Theodossiou (1999) further extended the work of Theodossiou (1993) by overcoming the problems of non-stationary variables and improving their definition of financial success. In addition, several minor statistical refinements were made to their CUSUM procedures. The best CUSUM model was then chosen by a neural network search procedure from an initial set of 54 explanatory variables[6]. Interestingly, models using the most popular explanatory variables were found to be non-stationary with deteriorating forecasting performance over time. The final chosen CUSUM model had superior predictive power compared to both DA and LA. Despite these positive empirical results, CUSUM procedures have not been widely used in BFP. The exact reason for this is unknown, but the greater complexity of the CUSUM procedures appears to be the most likely reason for its low popularity.

2.1.8 Other Lesser-Researched Techniques

In addition to the techniques already presented in this chapter[7], there are many other lesser-known techniques that have been applied to BFP. Most of these other techniques will

[6]The variables comprised 27 variables popular in the literature and their first differences.

[7]This includes decision trees and survival analysis, which will be discussed in the following two separate sections.

be discussed in this section with references to other literature reviews.

Together with LA and PA, the Linear Probability Model (LPM) was suggested as an alternative to DA, because its score is bounded between 0 and 1 and can be directly interpreted as a probability of failure. Nevertheless, LPM is rarely used in BFP due to its inferior empirical performance. Theodossiou (1991) applied all three techniques to BFP and found that LA and PA had very similar classification and predictive ability that was considerably superior to LPM. More recently, econometric modelling has been applied to BFP (Kumar and Tan, 2005). This technique involves combining a large number of predictor variables, both exogenous and endogenous, into many computationally intensive inter-related regressions. As the number of variables is very high the level of parsimony and predictive ability of econometric models is poor; however, the main goal of econometric modelling is furthering the understanding of the failure process.

In addition to being used to optimise the training of an ANN (see Section 2.1.6, GAs have been independently applied to BFP. GAs are a type of evolutionary algorithm that involve searching for an optimised solution based on Darwinian Evolution, and its concept of survival of the fittest. A comprehensive review of GAs in general has been conducted by Bornholdt (1998). When applied to BFP, GAs search for the optimal set of rules to classify and predict business failure. The final optimised set of rules can easily be extracted from the GA model and then interpreted similar to rules in an expert system. Shin and Lee (2002) applied GAs to BFP and concluded that this GA rule extraction approach has significant promise.

There are also many other lesser-known and usually more complex techniques that have been applied to BFP, such as chaos (or catastrophe) theory (Scapens et al., 1981), multicriteria decision aid (Zopounidis and Dimitras, 1998) and rough sets Dimitras et al. (1999); Zopounidis and Dimitras (1998). These alternative techniques, together with the more popular techniques, have appeared in the recent literature reviews by and Aziz and Dar (2005) and Balcaen and Ooghe (2004). Both of these papers have included excellent summary tables of the literature they have reviewed. Moreover, Balcaen and Ooghe (2004) addressed the question of whether these more complex alternative techniques produce better models. The findings were not definitive, but indicated that there may not be any benefits associated with using more sophisticated alternative techniques.

Altman and Saunders (1988) present an excellent review of the BFP studies up to 1998 from an accounting viewpoint, and Cybinski (2003) has produced a book containing a more recent review of the BFP field. (Zopounidis and Dimitras, 1998) also contains an excellent literature review that discusses most of the techniques discussed in this chapter. Laitinen and Kankaanpää (1999) is another recommended summary paper that presents a good theoretical analysis and empirical comparison of DA, LA, ANNs, HIP, survival analysis and recursive partitioning (a decision tree approach).

2.2 Decision Tree Techniques

Decision Tree (DT) techniques generate a set of tree-based classification rules and construct a decision tree (also known as a classification tree). DTs assign input objects to a group from a predefined set of classification groups: in the case of BFP, a DT usually assigns businesses to either the successful or failing group. In general, DTs are binary trees, which consist of a root node, non-leaf nodes and leaf nodes connected by branches, whereby each non-leaf node has two branches leading to two distinct nodes as shown in Figure 2.4.

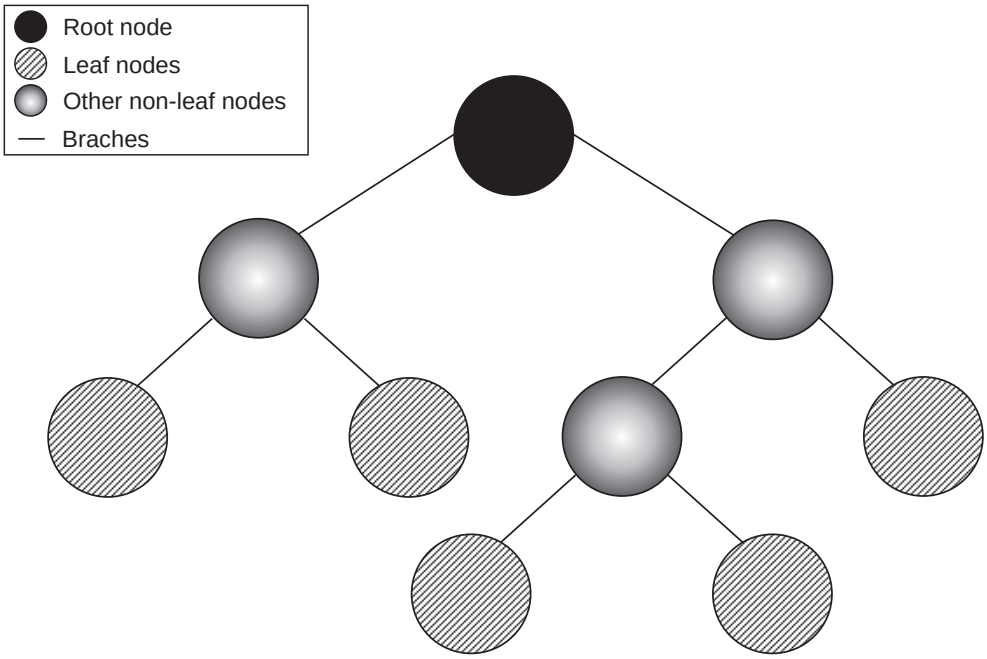

Figure 2.4: The basic structure of a binary tree.

When applied to classification problems such as BFP, leaf nodes that represent classification groups and the non-leaf nodes each contain a splitting (or decision) rule. Thus, the tree is built by a recursive process of splitting the data when moving from a higher level of the tree to a lower level. The splitting rules at each node define the details of the split. The splitting rules are made up of an expression (usually a simple financial ratio) that is evaluated for each case (business) and compared to a cut-off value. For example, a splitting rule might be to classify a business into the

- Left sub-tree if $current\ ratio < 2.1$, or
- Right sub-tree if $current\ ratio \geq 2.1$.

A simple example of a decision tree for BFP is depicted in Figure 2.5. This figure depicts a typical DT for classifying businesses as successful or failing where all the splitting rules are univariate (contain one variable), but the same variable may used in zero, one or many splitting rules. Figure 2.6 then depicts how the initial data set is input into the root node

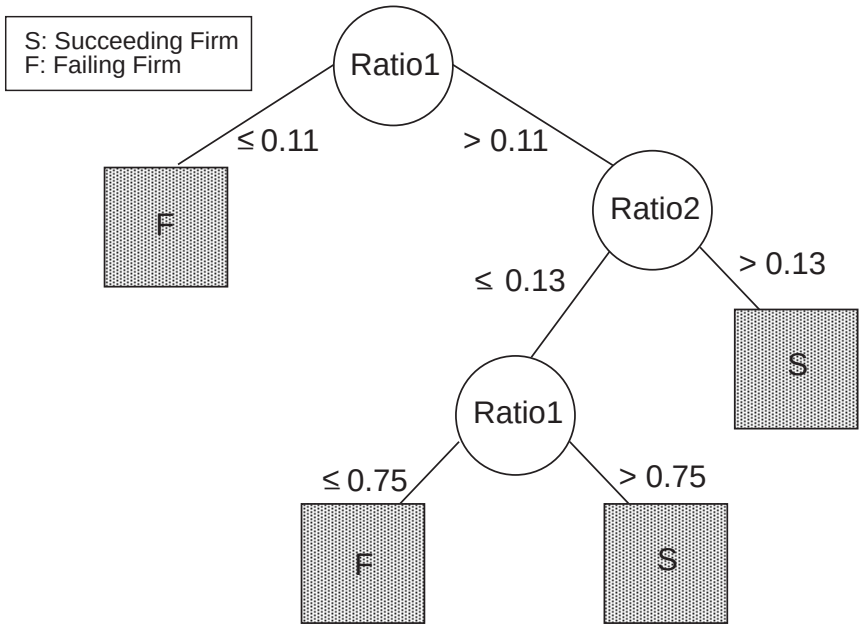

Figure 2.5: An example of a simple decision tree for business failure prediction.

and then split into classification leaf nodes (for a different DT than shown in Figure 2.5). In addition, Type I and II Errors have been identified in Figure 2.6.

When constructing a DT it is possible to continue splitting the data until a very complex tree[8] is created, which perfectly classifies the initial set of data (called the training set) as shown in Figure 2.7. However, this is not the optimal tree as many of the splitting rules have been created based on patterns that are specific to the training set and do not generalise to other data sets. Thus, this complex tree may have 100% classification on the training set, but it will mostly probably be a poor predictor of failure for other businesses. This is known as the problem termed as over-fitting, which is solved by using a less complex tree. Thus, similar to supervised learning with ANNs, *DT building algorithms* are used to manage the creation of DTs. There are two main tasks a DT algorithm performs:

1. Choosing the *best* splitting rule at each non-leaf node that discriminates between successful and failing businesses; and,

2. Managing the complexity of the DT, including deciding when to stop the process and when to use the current DT as the *best* DT.

There are different DT building algorithms that can be used to build the decision tree, each possessing their own method to determine the best splitting rule and best decision tree. Although similar tree structures are created by all building algorithms, the choice of which algorithm to use often has a large influence on the accuracy of the final DT model. The important DT building algorithms from a BFP viewpoint are the recursive partitioning

[8]The complexity of the tree increases as the number of nodes in the tree increases.

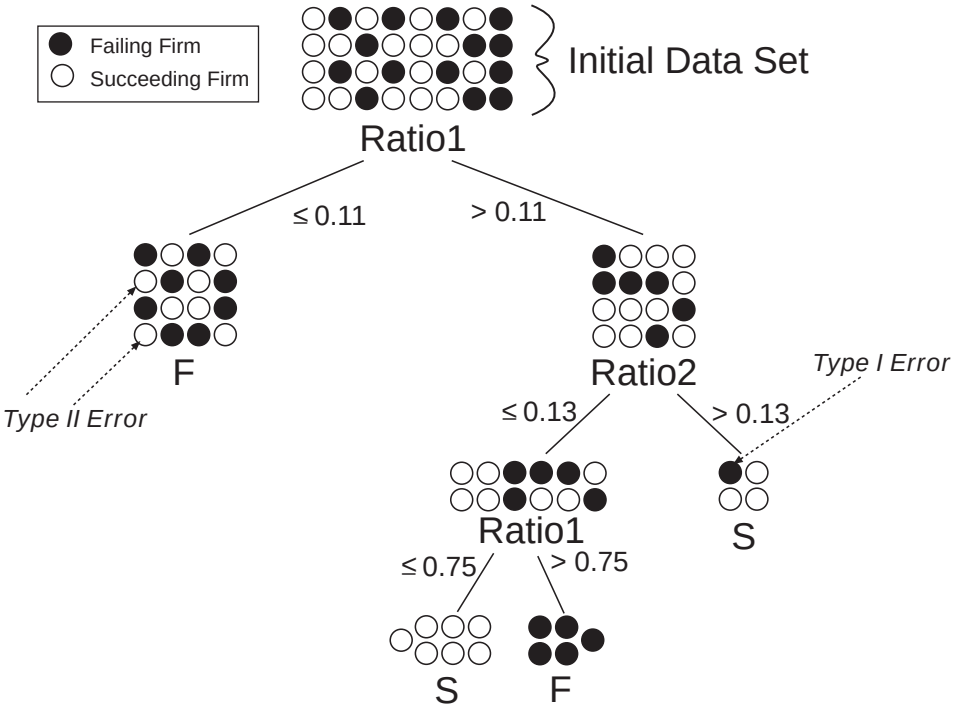

Figure 2.6: Adding data to a decision tree for business failure prediction.

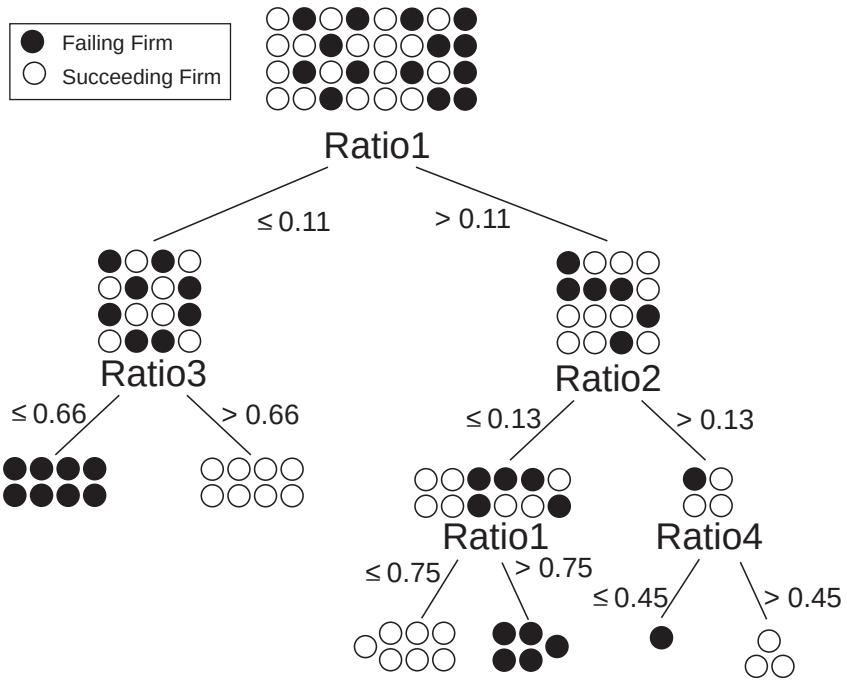

Figure 2.7: A complex decision tree for business failure prediction with perfect classification on the initial data set.

algorithm and entropy algorithms such as Classification and Regression Trees (CART) and See5.

2.2.1 Theoretical Analysis

Unlike the parametric discriminant and logit analysis methods, the major advantage of DTs is that they are non-parametric. This means that DTs make no distribution assumptions about the underlying data, and consequently there is no violation of distribution assumptions and there is no need to consider transforming variables (such as a log transformation). The only assumptions that DTs possess are that the successful and failing groups are discrete, non-overlapping and identifiable, which are common assumptions to all other statistical approaches. DTs can also handle missing values and qualitative data, as well as being be easily represented in a user friendly graphical format (Joos et al., 1998). Another advantage of DT models is that they can take different misclassification costs for Type I and Type II Error as inputs, which can then be incorporated into the DT building process at all stages. This is preferred to adjusting the cut-off values after the model has been generated as is the case for the more popular techniques of DA and LA. However, a disadvantage of many DT building processes is that they require prior probabilities of successful and failed businesses as inputs.

DT models are easy to interpret as the splitting rules are usually univariate. This also allows for easy identification of significant variables, where the root node contains the most significant variable. The relative significance of other variables can be found by comparing their proximity to the root node, with the closer nodes containing more significant variables. This process is even accurate when variables appear in more than one splitting rule, whereby the significant of a variable is measured by the proximity of the root to the closest splitting rule that contains the variable. However, DTs only identify the relative significance of variables, unlike DA and LA that provide quantified statistical figures that represent each variable's significance. Nevertheless, DTs still have the predictive power of a multivariate approach as there are *sequences* of nodes (univariate rules) that lead to each classification leaf-node. Furthermore, these sequences of splitting rules (in nodes) can naturally model interactions between variables without including the interaction term in the model, as has to be done with DA and LA. Although linear combinations of variables could also be used in splitting rules the potentially increased predictive power is not thought to outweigh the reduction in simplicity, ease of interpretation and identification of significant variables.

It can easily deduced from the DT in Figure 2.5 on page 19 that DTs use a type of discrete scoring system, where the leaf nodes are the different discrete scores that are possible. The ramifications of this are that a probability of group membership (failure or success) can not be calculated, and in addition, there is no way to compare and distinguish businesses in the same classification. These are the main disadvantages of DTs compared with the more popular DA and LA models that have a continuous scoring system. However, there is no need for determining cut-off values in a discrete scoring system, such as in DTs. This is an advantage for DT techniques as the arbitrary assignment of cut-off values has

been a major criticism of both DA and LA. It can be argued correctly that DTs have cut-off values at each non-leaf, however, these values are calculated as part of the decision tree building process and are not arbitrarily set.

DTs have been criticised for their forward variable selection process. This criticism is for the case of a single variable being used in multiple splitting rules, when the DT building algorithms do not review the previous rules containing the variable when determining future rules containing that variable (Zopounidis and Dimitras, 1998). This is definitely a theoretical weakness of DTs, but all forward stepwise approaches, such as those commonly used in DA and LA, have this weakness. Furthermore, there is no evidence to suggest that this weakness will significantly reduce the classification and predictive ability of DTs.

2.2.2 Previous Studies in BFP

Despite their advantages, DTs have not yet become as popular and wide-spread in BFP as DA or LA. However, they are considered to be one of the more popular alternatives to DA and LA (along with ANNs and Survival Analysis). Following the successful application of DTs, using the recursive partitioning algorithm, to medical decision making by Goldman et al. (1982, 1988), they were first applied to BFP in the seminal work of Frydman et al. (1985). Since then, other types of DTs have also been applied to BFP with varying success.

Frydman et al.'s RPA method

The application of DTs to BFP by Frydman et al. (1985) was the first application of a non-parametric technique to BFP. As stated above, Frydman et al. used the recursive partitioning algorithm (RPA) to build the DTs, a method that has also become known as recursive partitioning analysis (RPA). The RPA takes as inputs the prior probabilities of businesses failing or being successful, the misclassification costs, and a set of training data (including the actual group classifications). The RPA algorithm then constructs the tree in a way to minimise the expected misclassification cost (Jones, 1987), which is referred to as the *resubstitution risk* by Frydman et al.. The resubstitution risk is calculated using probability theory and Bayesian reasoning. The actual calculation is discussed in more detail in Section 3.3 on page 37.

With the aim of assessing the suitability of DTs (specifically RPA) to BFP, Frydman et al. compared two RPA built DTs with two DA models. The difference between the two DT models and the two DA models was the model size, in relation to the number of explanatory variables incorporated. The larger RPA-DT was created using the approach outlined in the paragraph above, while the smaller tree was chosen as the sub-tree of the larger tree with the best cross validation performance. V-fold cross validation models are calculated by dividing the data into V approximately equal groups and generating a DT for each $V - 1$ group[9]. Each DT was then used to classify the group left out, and the best DT

[9]Frydman et al. used $V = 5$.

is chosen as the tree with lowest average resubstitution risk[10]. Similarly, the smaller and larger DA models were constructed using the 4 and 10 most significant variables respectively, according to a forward stepwise method.

The prior probabilities were set at 2% for failing and 98% for successful businesses, but the techniques were compared over eight different misclassification costs[11]. All the models were formed from a data set of 200 randomly selected manufacturing and retail businesses, of which 58 went bankrupt. The predictors used in this study were 20 standard financial variables (mostly ratios) found to be significant in previous studies. The RPA-DT models were found to be superior classifiers of business failure in the original training data compared with the DA models. In addition, as may be expected, the more parsimonious (smaller) models outperformed their more complex counterparts on the cross-validation analysis. More importantly, the more parsimonious RPA-DT model was found to be a superior predictor to the DA models for almost all misclassification costs on the cross-validation analysis. However, the RPA-DT was found to be the worst technique for prediction in the cross-validation, which highlights the potential risk of over-training a DT model. Overall, after considering their theoretical analysis and empirical results, Frydman et al. concluded that decision trees using RPA were useful tools for predicting business failure.

Other Relevant Studies

Surprisingly, there have only been very few papers that have focused on decision trees since Frydman et al. (1985), despite the existence of many different DT building algorithms, including ID3, See4.5, See5 and CART. See4.5, which is an earlier version of See5, was presented as an improved version of ID3. Thus, See5 is considered to be superior to both See4.5 and ID3; however, an overall 'better than' comparison can not be made with CART as it implements a slightly different approach. All of these DT building algorithms are entropy algorithms based on the concept of *information entropy*. Entropy algorithms select splitting rules that produce the most informative split, which is equivalent to minimising entropy[12], rather than minimising expected misclassification costs as occurs with RPA. The tree size or complexity is then managed by choosing a minimalist DT that retains only splitting rules based on general patterns in the data. There is also another related technique that has been applied to BFP, known as Multivariate Adaptive Regression Splines (MARS). MARS is an extension of the CART technique that incorporates the DT concept of splitting the data into subregions using stepwise regression. Thus, MARS can be thought of as a procedure that searches for the optimal piecewise regression model (which may include higher order terms).

[10] A penalty for higher numbers of leaf-nodes was also incorporated into the selection process of the second DT.

[11] Consistent with the real world, Type I error was always equally or more expensive that Type II Error.

[12] In this context, entropy is analogous to statistical noise or heterogeneity.

Joos et al. (1998) used decision trees to predict credit classifications for one of Belgium's largest banks[13]; specifically, LA was compared with See5. Three models were created for both techniques based on three different data sets comprising a full set of financial variables, a reduced smaller set of financial variables and a set of qualitative variables. Unsurprisingly, Joos et al. concluded that regardless of the technique used, the models based on these three data sets were ranked according to classification ability in the order mentioned in the previous sentence. Furthermore, LA was found to outperform DTs on the full data set. However, as DTs are better able to handle incomplete and qualitative data, See5 outperformed LA on both the reduced and qualitative data set. Finally, Joos et al. analysed the classification accuracy once different misclassification costs were introduced. See5 maintained its superior performance on the reduced and qualitative data sets. In addition, for the eight different misclassification costs trialled[14], See5 was superior for five of these trials with less Type I Error. However, for the trials with the cost of Type I Error set the highest LA was superior. Thus empirically, there was no obvious overall superiority between See5 and LA.

Huarng et al. (2005) also used the See5 package, as well as being the first to apply CART to BFP. They found that CART was empirically superior to See5; however, the data sets comprised less than 12 businesses and 5 variables, which is too small to obtain reliable results. It should also be noted that Shirata (1998) used CART prior to Huarng et al. (2005), but it was simply to select the significant variables to use in their study of Japanese bankruptcies using DA. Shirata used CART for variable selection because it automatically calculates a variable importance score for each variable: this calculation process is discussed more in Section 3.4.3.

Although decision trees have not been a focus of many research papers, various studies have used decision trees as a comparison technique, because Frydman et al. (1985) showed it to be superior to DA. There have been mixed results from these studies. Tam (1991) and Tam and Kiang (1992) found that artificial neural networks (ANNs), DA and LA all outperformed ID3 decision trees. Fernández and Olmeda (1995) also found ANNs to be slightly superior to See4.5 and MARS, which produced similar results. Nevertheless, despite a lower classification accuracy on the initial data set, See4.5 and MARS were found to have superior predictive ability over both LA and DA on the hold-out data set[15]. Unlike most studies, Fernández and Olmeda (1995) also empirically tested combining the techniques: ANNs and LA alone and with other techniques were compared. Overall, models produced by combining techniques usually produced more accurate classifications and predictions. In contrast to these studies, Martinelli et al. (1999) found that See4.5 outperformed ANNs. Furthermore, Laitinen and Kankaanpää (1999) showed that RPA produced similar results to DA and LA (as well as survival analysis). However, all these studies have not considered differing misclassification costs.

[13]Credit classification can be considered identical to BFP from a statistical modelling point of view.

[14]Consistent with the real world, Type I Error was always equally or more expensive than Type II Error.

[15]In addition, MARS was has superior classification accuracy on the initial data set compared to DA.

Overall, it is clear that DTs have not been extensively or exhaustively researched as a BFP technique.

2.3 Survival Analysis Techniques

Survival analysis (SA) techniques is the term applied to the set of dynamic statistical tools used to analyse the time till a certain event. That is, SA models have survival time as the dependent variable. However, some SA models have a derivation of the survival time, such as the probability of survival until time t, as the dependent variable. Thus, the SA approach to BFP is fundamentally different from the DT approach. While DT techniques model BFP as a classification problem, SA models BFP as a time-line, in which the time till the event of business failure is modelled. The following two subsections outline the key concepts in, and ways to develop, SA models with reference to applications in BFP. Then, the prior successful application of SA in medical research is briefly discussed, followed by a theoretical analysis of SA and a review of the studies that have applied SA to BFP.

2.3.1 Key SA Concepts

Censored Data

SA involves estimating the time till an event from a certain starting time. Data is considered censored (or incomplete or latent) if one or both of these two times is unknown (Kaliaperumal, 2005). For example, some businesses do not fail during the study period and consequently do not have a failure time: this is right-censored data, which means that the end-point time is unknown (although known to be greater than the study period). Logically it follows that left-censored data occurs when the starting event (aka origin) is unknown, although it is known to have occurred before a certain point in time. Furthermore, interval censoring occurs when both left and right censored data are present, while non-censored data is termed response data. However, right-censored data is the only common censored data in lifetime distributions (Kaliaperumal, 2005), such as business failure distributions. There are different types of right-censored data, but the concepts mostly apply to the medical field: a good summary of the different approaches is provided by Harrell (2001).

SA models allow the censoring time to vary between businesses in a study. However, a key requirement of SA models with censored data is that the censoring occurs independent of the failure time (Harrell, 2001; Kaliaperumal, 2005). Thus, a pre-planned finish date in a study does not violate this assumption. In addition, delayed entry does not violate the assumption provided that new businesses are not added to the study dependent upon either the failure time or any explanatory variable (which generally is not the case for either situation)[16]. However, this does mean that an SA model will not be valid when used on a

[16]Delayed entry is when a business enters the study after it has begun.

data set containing all, or almost all, right-censored data (successful businesses). Therefore, it is important to select an initial data set (businesses), study origin and study duration that comprises a reasonable number of both failed and successful businesses.

Lifetime Distributions

Lifetime distributions are distributions with a non-negative random variable that represents the lifetimes of individuals (such as businesses, people or animals) in some population. A lifetime distribution can be characterised in a number of ways as shown below, where T is a lifetime distribution variable, t a predefined time period of interest and $Pr(b)$ the probability of b occurring.

- The **Survival Function** represents the probability that an individual (business) will survive past a certain time t. The survival function is mathematically stated as $S(t) = Pr(T > t)$.

- The **Probability Density Function**, or cumulative distribution function, represents the probability that an individual (business) will fail on or prior to a certain time t. The probability density function is mathematically stated as $F(t) = Pr(T \leq t)$, and consequently $F(t) = 1 - S(t)$.

- The **Hazard Function**, also known as the hazard rate and failure rate, represents the *intensity* of death (failure) at a certain time point t. The hazard function is mathematically defined as the limit to the conditional probability (Laitinen and Luoma, 1991):

  $$h(t) =_{\Delta t \to 0}^{lim} \frac{P\left(t < T < t + \Delta t \mid T > t\right)}{\Delta t}$$

  It can then be shown that[17]

  $$\begin{aligned} h(t) = \frac{dFt/dt}{S(t)} &= \frac{F'(t)}{S(t)} = \frac{-S'(t)}{S(t)}, \ and \\ S(t) &= e^{-\int_0^t h(t)\ dt} \end{aligned}$$

  $h(t) \times \Delta t$ represents the likelihood of business failure over the period δt given that the business survives until time t. Thus, the hazard function $h(t)$ approximates the instantaneous rate of business failure at a particular point in time for very small Δt, given that it survives until just prior to that time. The term probability has intentionally not been used here because the hazard function does not represent a probability, as the hazard function may take values greater than one. It is important to note that $h(t)$ can follow many different patterns: it may be increasing, decreasing, constant, U-shaped or any other variant (Kaliaperumal, 2005). Nevertheless, as the

[17] Harrell (2001) presents these derivations in Chapter 16.

survival function can logically never decrease for an increase in time t, the hazard function $h(t)$ is always positive. It is also important to note that the distribution can be described by the cumulative version of the hazard function $H(t)$, which never decreases.

Lifetime distributions can also be described by their mean residual lifetime, but this is not a crucial concept for this review or subsequent analysis. The reader is directed to Kaliaperumal (2005) for more information. The important concept regarding lifetime distributions is, if any one of these functions is known then any of the other functions can be uniquely determined, as shown in the above equations for $h(t)$ and $S(t)$. However, the interpretation of the various functions is very different.

2.3.2 Estimation of SA Models

The are many different SA techniques to estimate the descriptor functions of a lifetime distribution. There are also different SA models that define different relationships between one of the descriptor functions (usually the survival or hazard function) and the set of explanatory variables; hence, there are non-parametric, semi-parametric and parametric models.

Non-parametric Estimation

The simplest, and one of the oldest, ways to describe and analyse failure time data is the life table technique (Kaliaperumal, 2005). This non-parametric technique is similar in concept to a frequency table. The life table is divided into a defined number of time intervals, and the number of failed/successful businesses, as well as the number of cases that were censored, are all calculated for each interval. From these figures the proportions of failed/successful businesses and the cumulative proportion surviving can be calculated. As the cumulative proportion surviving is actually the definition of the survival function, the values of the survival function for each time period have now been calculated. In addition, all the other lifetime distribution descriptors can be derived from the survival function. The calculation formulae for these descriptors are given in Kaliaperumal (2005).

The most popular method for estimating the survival function is the Product-Limit, or Kaplan-Meier, estimator outlined by Kaplan and Meier (1958). This method is actually identical to the life table technique with each time interval only containing one observation. Consequently, the survival function is mathematically defined as

$$\begin{aligned} S(t) &= \prod_{i:t_i<t} \left(\frac{n_i - d_i}{n_i} \right), \text{ then the cumulative hazard function is calculated as} \\ H(t) &= -\ln\left(S(t)\right) \end{aligned}$$

where d_i represents the number of deaths (failures) at time t_i and n_i represents the total number of individuals at risk at time t_i (businesses in study not failed before t_j). Hence,

The main advantage of this method over the life table method is that it does not depend on the grouping of data into subjectively chosen time intervals.

There is another, less-popular, method called the Nelson-Aalen Additive Estimator, which was first proposed by Nelson (1969) and later generalised by Aalen (1976). Using the Nelson-Aalen additive estimator the cumulative hazard function $H(t)$ is calculated as

$$\begin{aligned} H(t) &= \sum \frac{d_i}{n_i}, \text{ then the survival function is calculated as} \\ S(t) &= e^{-H(t)} \end{aligned}$$

where d_i and n_i are the same as defined for the Kaplan-Meier estimator. The Nelson-Aalen estimator has some statistical advantages over the Kaplan-Meier estimator, which are briefly discussed by Harrell (2001) in Chapter 16.

Regression-based Estimation

Essentially, the difference between the various SA models are different assumptions about the relationship between the lifetime distribution descriptors (survival, probability density and hazard function) and the set (vector) of explanatory variables (X). Thus, the general regression formula can be written as

$$h(t) = g(t, X^T\beta)$$

Where X^T is the transpose of X, β is the vector of explanatory variable coefficients (also known as covariates) and g is an arbitrary function. In SA models estimated from regression it is customary to estimate the hazard rate, and then derive the survival rate and other measures as required. Traditionally, SA has been divided into two main types of regression models for SA. These types are the proportional hazards (PH) and accelerated failure time (AFT) models, both of which have fully parametric and semi-parametric versions (refer to Prashanthi (2005) for more details). PH models focus on estimating and understanding the hazard function, while AFT models only focus on the survival time. Due to the increased focus on understanding and interpreting the hazard function at certain time points, PH models are more common in most applications (including BFP). In fact, due to its flexibility the most prominent model applied in the medical and business failure field is the semi parametric PH model defined by Cox (1972). Therefore, Cox's PH model will now be discussed in more detail. An excellent review of other models occurs in chapter 17 of Harrell (2001).

Cox's Proportional Hazards model (Cox, 1972) is defined as:

$$h(t) = h_0(t)e^{X^T\beta+c}$$

where:

- $h_0(t)$ is termed the baseline hazards function and describes how the hazard function changes over time and is the non-parametric part of the model; and,

- $e^{X^T\beta+c}$ describes how the hazard function relates to the business specific explanatory variables and is the parametric part of the model, where c is just an estimated constant. Note that some or all of the explanatory variables can be time dependent.

The regression coefficients β are calculated by an efficient method very similar to the maximum likelihood method (detailed in Kalbfleisch and Prentice (1980)), which is fully parametric wherein the explanatory variables are log-linearly related to the businesses' lifetime distribution. As with traditional regression techniques, the best explanatory variables are chosen from a starting set by forward or backward selection methods. In contrast to traditional techniques, Cox's model does not make any assumptions, such as linearity about the baseline hazard function. As these distribution assumption are often violated, the Cox model is more efficient and robust than traditional models (Balcaen and Ooghe, 2004).

This model can become parametric if assumptions about the distribution of the baseline hazards function are made. Thus, for specific applications of this model, once a large amount of information has been gathered about the lifetime distribution it may be possible to make some educated baseline distribution assumptions and use Cox's model as a parametric model. Hence, the flexibility of the model as its power can be increased for specific problem applications.

2.3.3 Medical Applications

Although survival analysis is relatively new to BFP, it has been used extensively in biomedical sciences (Laitinen and Luoma, 1991). The reason for this is that many biomedical analyses study the time to an event (Kaliaperumal, 2005); for example:

- time from beginning of cancer treatment until death,
- time from birth to first dental problem, and
- time from application of skin treatment cream till healed skin.

This means that questions such as 'how long will a cancer patient live given his current state and treatment?' can now be answered with some degree of accuracy. Survival analysis has also been applied to other medical administration problems such as estimating the hospital stay of psychiatric patients (Kaliaperumal, 2005). The practical applications of these biomedical (and other) analyses are obviously extremely important, which has been (and is) the major reason for the increasing research effort into SA. Moreover, in the latter half of the 20th century, SA's application to the high-profile field of cancer research generated fast growth for the technique in the biomedical area (Kaliaperumal, 2005). Although Cancer studies are the main application for SA, more recently it has had successful application in AIDS research.

Interestingly, researchers have also had success in using ANNs to estimate a SA model (Kumar and Tan, 2005). Ohno_machado et al. (1995) developed an ANN that estimates

survival time more accurately than traditional SA techniques, and Lundin et al. (1999) concluded that this technique was accurate for survival predictions of breast cancer patients.

More information on SA and its applications in the field of (bio-) medicine is contained in Kaliaperumal (2005), which includes a concise summary of the evolution of SA techniques. Initially only the field of research where the event of interest was death was termed survival analysis; however, over time this constraint has relaxed and now survival analysis refers to a collection of techniques to deal with predicting time till an event.

2.3.4 Theoretical Analysis

SA techniques, as are DTs, are more sophisticated than the traditional popular techniques of DA and LA. This added sophistication is the source of advantage for SA models. Except for sequential procedures (discussed previously in Section 2.1.7), SA is the only well-known technique that incorporates the time-series (or longitudinal) nature of BFP data into its model. Thus, SA does not assume that the failure process remains stable over time. All other cross-sectional models are only valid if the underlying failure process remains stable over time (Laitinen and Luoma, 1991; Hans-Peter Blossfeld and Mayer, 1989), which is a problem as the steady failure process assumption is usually violated in the real world (Laitinen and Luoma, 1991). This fundamental difference between the time-series SA models and cross-sectional traditional models also makes empirical comparisons between the techniques difficult. For example, a single SA model can make 1-year, 2-year and 3-year ahead predictions; however, a single DA model can only make one length of prediction based on its input data. Therefore, a single SA model is usually compared with many traditional models such as DA. This is an advantage in itself as one SA model is clearly more powerful making different predictions than one traditional model[18]. Essentially, the cross-sectional models do not utilise the dynamic nature of the failure process.

The built in time factor in SA models allows them to model time-dependent explanatory variables. Zavgren (1985) found that in BFP the signs of the explanatory variable coefficients may change in different years before failure. Laitinen and Luoma (1991) went further and added that the values of the coefficients may also change relative to time before failure: they gave an example that accounts payable may only be important in the final phase of business failure. Thus, an advantage of SA is the capability to model these changes, which can not be done with cross-sectional models. Therefore, SA models appear to be more suited to modelling a dynamic process, such as business failure, than cross-sectional models. This also means that *theoretically* the predictive accuracy of SA models should be greater than that of both DA and LA.

Almost all well-known approaches, including DTs but excluding ANNs, assume that the data (businesses) comes from two distinct populations, which are those that are going to succeed and those going to fail. SA models do not make this assumption, but rather assume

[18]There has also been studies on multilogit analysis (such as Keasey et al. (1990)), which essentially try to add this time-series element to the original logit analysis; however, they are outside the scope of this Thesis.

that all businesses come from the same population distribution, where failing businesses and successful businesses are treated as response and censored data respectively (refer back to Section 2.3.1 on page 25). This assumption is more natural and more accurately models the real world (Laitinen and Luoma, 1991). SA models can also deal with the delayed entry and early exit of businesses from a study, which is likely to happen in business failure studies. Furthermore, SA does not make any of the restrictive distribution assumptions inherent in DA and LA, such as linearity. The semi-parametric and parametric SA models make some distribution assumptions, but they are not as commonly violated.

In addition to the easily interpretable probability of success or failure, SA models also produce the interpretable hazard function that is not available in other techniques. Analysis of the hazard rate can aid understanding of any process of death or failure (Harrell, 2001). Thus, SA is able to provide more information than other techniques, which is a significant component of any good model (Chatfield, 1995). SA also has the relative advantage over DTs of a continuous scoring system similar to that in DA and LA. However, this also means it has the disadvantage relative to DTs of modelling misclassification costs by varying cut-off values that are not incorporated into the model estimation process.

There are also a few other disadvantages associated with the use of SA. There is evidence to suggest that the sample construction, specifically the proportion of failing and successful businesses, may affect the estimation of the SA model. However, this problem seems to be minor as most randomly selected BFP data sets contain a mixture of failing and successful businesses. Some researchers also identified that SA techniques (particularly the Cox model) are subject to multicollinearity problems, but these can be easily avoided by using standard forward and backward variable selection procedures. A more important disadvantage is that SA is designed to focus on determining the effects of explanatory variables on the life of businesses, rather than being designed to predict outcomes such as the failure of businesses. The ramification of this is that obtaining predictions from SA models is more difficult than anticipated.

2.3.5 Previous Studies in BFP

In the context of business failure, SA models, often referred to as hazard models, aim to quantify the relationship between the time till failure and a set of explanatory variables (Laitinen and Kankaanpää, 1999). Similar to DTs and ANN, SA has not yet become as popular in BFP as DA and LA, but it is considered to be a popular alternative to these main techniques. Overall, there have been few studies on the application of survival analysis to BFP, and most the research has used Cox's model, which was previously outlined in Section 2.3.2.

Earlier Studies

The pioneering paper on SA applied to BFP is by Lane et al. (1986), who used the Cox model to predict bank failure. Lane et al. created their model based on a selection of

334 successful and 130 failed banks from the period 1979 to 1983. The model was then tested on a hold-out sample with one and two year predictions, in which the cut-off value was set at the proportion of failed banks in the sample. The prediction accuracy of the Cox model was found to be comparable with DA on the initial and hold-out data, but the Cox model produced lower Type I Errors. Crapp and Stevenson (1987) also followed Cox's model (including a life table analysis). Their study claimed their model overcame the huge problem of selecting what explanatory variables to include in the analysis. They attempted to do this by starting with an "exhaustive" set of variables including financial ratios, macro-economic conditions and business growth; the most relevant variables were then chosen by a stepwise analysis. Crapp and Stevenson then applied their model to some Australian credit unions with similar encouraging results.

Laitinen and Luoma (1991) also applied the Cox model to business failure, due to its superior flexibility. The significance of this paper is that it is the first to critically present the advantages and disadvantages of using survival analysis to predict business failure. Laitinen and Luoma also empirically compared the classification accuracy of the Cox model with DA and LA using 36 failed Finnish limited companies and 36 successful counterparts, along with 12 standard financial ratios as explanatory variables. Their predictions were made by dividing the businesses into two groups based on their hazard ratios, according to the ratio of failed and successful businesses in the original sample (equal groups in this case). Businesses in the group with the higher and lower hazard ratios were then predicted to fail or succeed respectively. Although the techniques were comparable, DA and LA were found to be slightly superior predictors to the Cox model. Nevertheless, Laitinen and Luoma argued that the SA approach was more natural, appropriate and flexible, and used more information. It was also stated that the empirical under-performance could have been due to the small sample or sample bias inadvertently caused by the authors. Therefore, it was the author's belief that further research into SA as a BFP tool would result in SA models becoming superior to traditional models. Earlier support was also given by Keasey et al. (1990) and Ogg (1988) who recommended that SA techniques be used in BFP.

Recent Studies

Rather than create a prediction model, Kauffman and Wang (2001, 2003) used SA techniques to examine the drivers behind the survival of Internet businesses. The data set comprised quarterly data on 100 Internet businesses from the period of 1996 to 2001. Six explanatory variables were used: one industry specific, two business-specific, two e-commerce specific, and one macro-economic variable. Two SA techniques were applied to this data: the non-parametric Kaplan-Meier estimator model was used to perform a descriptive analysis, and Coxs semi-parametric proportional hazards model was used to explore the relative strengths of explanatory variables. The Kaplan-Meier model was stratified by target customers to allow a comparison of survival time of an Internet business with their target customer groups. Different Cox models were also estimated based on different types of failure: bankruptcy, mergers and acquisitions. Useful conclusions, such as businesses

targeting both business and consumer groups are less likely to fail, were drawn from both of the SA techniques applied. Although this work did not develop an SA model for predicting business failure, it demonstrated the usefulness of SA techniques for researching the business failure process.

Shumway (2001) applied the first SA model from a data set of significant size. The model was formed using various financial ratios and market-driven variables for over 2000 companies from the NYSE and AMEX over 31 years. However, less than 10% of these businesses in the data set were failed, which is much lower than the percentage in the real world. This was the pioneering use of a multiperiod logit model[19] to estimate the SA model coefficients. This allowed Shumway to estimate an accelerated failure time (AFT) SA model, which had not been previously applied to BFP. Consistent with previous studies, Shumway noted the theoretical superiority of SA techniques over the more popular techniques (DA and LA). In addition, Shumway's SA model was shown to empirically outperform both DA and LA in out-of-sample predictions. Surprisingly, Shumway also found many previously used financial ratios to be poor predictors, while previously unused market-driven variables were good predictors. However, Shumway only considered Type I Error. This is analogous to a *zero* misclassification cost for Type II Error, which is not the case in reality. In addition, this meant that the accuracy of his model was only tested on the failed businesses that made up less than 10% of this data; thus, the increased validity from using a large data set was somewhat negated.

Laitinen and Kankaanpää (1999) have presented a comparative study, in which SA along with DA, LA, RPA, ANN and HIP were analysed. The six techniques were empirically compared for their 1,2 and 3 year prediction accuracy using a data set containing three explanatory variables from 76 Finnish companies (38 failures). Their analysis showed that SA had superior predictive power for two and three year predictions. However, they concluded that there were no statistically significant differences in the predictive powers of any of the six models, except for LA being superior to SA for one-year predictions. This is somewhat understandable as the major advantage of SA is that it better suits a dynamic application over time, which is emphasised when a larger time-period is used.

Overall, the field of SA in BFP has had limited development since the work done by Laitinen and Luoma (1991): ironically, the purpose of Laitinen and Luoma's paper was to introduce the use of SA techniques in BFP for further application. The reason for this absence of advancement since 1991 has not been attributed to anything in particular in the literature, but is presumably partially due to the 'non-encouraging' empirical results of Laitinen and Luoma (1991). However, as SA techniques have not been explored fully as predictors of business failure, consequently their full power may not yet have been realised.

[19]Each business-year observation is considered as an independent observation.

2.4 Potential Research in DTs and SA

The reviews of SA and DTs above show some promising results for both techniques, especially considering their literature base is very small compared with that of DA and LA. It is also apparent that neither DT or SA techniques have been researched extensively, and that there are still unanswered questions that can be solved by future research. Future research on these techniques for use in BFP should enable comparisons with the more popular techniques to be more reflective of the true predictive power of DTs and SA.

Besides the comparison of RPA built DTs and the SA Cox regression model by Laitinen and Kankaanpää (1999), there has not been a comparison of SA and DT techniques for BFP. In addition, other future research opportunities for DTs and SA are presented in the following two subsections.

2.4.1 Decision Trees

There is opportunity for research to address the following gaps within the DT literature:

- Other than DA and LA, the RPA approach has never been compared with other techniques, such as DT techniques.
- CART has only been applied once to BFP in a study that did not produce reliable results due to a very small sample size (Huarng et al., 2005). In addition, this study did not consider different misclassification costs.
- While See4.5 has been compared with MARS, ANNs, LA and DA (Fernández and Olmeda, 1995; Martinelli et al., 1999), See5 has only been compared with LA (Joos et al., 1998)[20] but not with ANNs, DA or other DT techniques.

In addition, there has only been one study that used MARS to predict business failure (Fernández and Olmeda, 1995). Considering that MARS was not the focus of that particular study, valuable research could involve the use of MARS, including how it can best be used in BFP, how it compares with DA, LA and other DT techniques, and how it handles different misclassification costs. Furthermore, RPA, CART and See5 have never previously been compared with each other. Thus, there is great opportunity for future research on DTs for BFP.

2.4.2 Survival Analysis

Although Lane et al. (1986), Laitinen and Luoma (1991) and many more have indicated that the Cox model was very appropriate for use in BFP, it has not been consistently shown to be superior to traditional techniques. Lane et al. (1986) found the Cox model to slightly empirically outperform DA, but Laitinen and Kankaanpää (1999) found no overall

[20] See5 has also been compared to CART, but the results were not reliable as the sample size was too small (Huarng et al., 2005).

statistical difference between the empirical performance of DA and LA, while Laitinen and Luoma (1991) found both DA and LA empirically superior to SA. Therefore, it would be valuable research to apply the Cox model to another large set of data and compare it to DA and LA again. Furthermore, comparing the techniques across different misclassification costs has never been done for any SA models.

The Kaplan-Meier estimator has only been applied to one set of BFP data (Kauffman and Wang, 2001, 2003). Therefore, it would be valuable to test the technique on a different set of data to further determine how much useful information the technique produces. Moreover, this proposed study should test whether the Kaplan-Meier or the Nelson-Aalen (not previously used in BFP) estimator is more appropriate for BFP.

An accelerated failure time (AFT) SA model has only been applied to BFP in a study that only considered Type II Error (Shumway, 2001). Therefore, Shumway's model could be reapplied to a set of BFP data to analyse both Type I and Type II Errors using this approach. Finally, this approach could then be compared with other SA models to determine the most appropriate for BFP; these could include other AFT models, Cox's model or possibly a more parametric Cox-like model tailored to BFP. These techniques could also be compared amongst themselves and with DT techniques (and the standard DA and LA) across different misclassification costs, which has not yet been investigated. In addition, there is potential for research into ANN estimated SA models for predicting business failure.

Chapter 3

Decision Tree Research

3.1 Aim

The research presented in this chapter has been undertaken to further the assessment of Decision Tree (DT) techniques for use in business failure prediction (BFP). The goal is to assess the empirical classification and prediction accuracy of various DT approaches in BFP. Specifically, the critical comparison of RPA decision trees and DA by Frydman et al. (1985) will be extended to include See5 and CART decision trees, whereby models of both high and low complexity will be compared over various misclassification costs. Thus, the aim of this research is to critically compare RPA, See5 and CART decision trees (and DA) in order to address some of the gaps in the DT literature identified in Section 2.4.1.

3.2 Data

The data set used for this research is an unaltered copy of the data used by Frydman et al. (1985)[1] to ensure the validity of comparisons with Frydman et al.'s results. The main properties of this cross-sectional data set are presented in Table 3.1. It should be noted that the financial year investigated for each business was randomly chosen from within the period 1971–1981.

Property	**Value**
Businesses (Type)	Manufacturing and Retail in the US (in COMPUSTAT database)
Selection Procedure	Random selection
Businesses (Number)	200: 142 successful and 58 failed
Explanatory Variables	20 financial variables: mostly financial ratios

Table 3.1: Properties of the data set used for DT research.

[1]The data set is available from the author upon request.

3.2.1 Definition of Business Failure

In terms of this data set, Frydman et al. considered a business to be failed if it had legally filed for bankruptcy. The Wall Street Journal Index was used to determine whether a business had filed for bankruptcy. The Wall Street Journal Index was used to in preference to the COMPUSTAT database, as failed businesses often remain on the regular COMPUSTAT database. For example, 16 of the failed businesses in this data set that filed for bankruptcy were still listed on the regular COMPUSTAT database.

3.2.2 Explanatory Variables

The 20 explanatory variables in the data set cover a range of financial variables, all obtained by Frydman et al. from the COMPUSTAT database. These variables were chosen as the union of the sets of variables found to be significant in three studies prior to Frydman et al. (1985), namely Altman (1968); Deakin (1972); Altman et al. (1977). Specifically, the data set comprised:

Seven liquidity ratios: cash to total assets (Cash/TA), current assets to current liabilities (CA/CL), current assets to total assets (CA/TA), quick (or immediate) assets to current liabilities (QA/CL), quick assets to total assets (QA/TA), working capital to total assets (WC/TA), and working capital to total sales (WC/SLS).

Four profitability ratios: net income to total assets (NI/TA), earnings before interest and taxes to total assets (EBIT/TA) and the standard deviation (sd) of EBIT/TA, and retained earnings to total assets (RE/TA).

Four activity ratios: cash to total sales (Cash/SLS), current assets to total sales (CA/SLS), quick assets to total sales (QA/SLS) and total sales to total assets (SLS/TA).

Three coverage/leverage ratios: cash flow to total liabilities (CF/TL), total liabilities to total assets (TL/TA), and a log transformation (Ln) of interest coverage+15 (INT+15).

One market structure ratio: market value of equity to total capitalisation[2] (MVE/TC).

One business size variable: log transformation (Ln) of total assets (TA).

An insignificant adjustment was made to all explanatory variables for the capitalisation of leases: Frydman et al. (1985) provides a brief description of this process.

3.3 Methodology

The methodology was based on that applied by Frydman et al. (1985), which was discussed in Section 2.2.2 on page 22. The relevant points of their work will be briefly discussed in

[2]The total capitalisation of a business is the total of its equity and long-term debt.

this section where appropriate. The methodological issues that are relevant to all models will be presented in the first section, while the methodology Frydman et al. (1985) used to generate the DA and RPA decision tree models will then be briefly reviewed. Finally, the methodology used in this research to generate the See5 and CART decision trees will be presented. More detailed information on DTs and their prior application to BFP was presented in Section 2.2 on page 18.

3.3.1 General Methodology

Consistent with Frydman et al. (1985), the parameters used in this research are presented in Table 3.2, where:

- N_i is the sample size of the ith group.
- π_i is the prior probability of a firm belonging to group i.
- c_{ij} represents a misclassification cost measure of assigning a group i firm incorrectly to group j.
- The subscripts s and f correspond to the successful and failing classification groups respectively.

N_f; N_s	58; 142
π_f; π_s	2%; 98%
c_{sf}	1
c_{fs}	1,10,20,30,40,50,60,70

Table 3.2: Parameters for DT Models.

As the models being compared over various misclassification costs, model performance was assessed by analysing a weighted misclassification cost measure, known as resubstitution risk, for which lower values indicate better models. The formula for the resubstitution risk, where n_i and number of group i misclassifications, is

$$\frac{c_{fs}\pi_f n_f}{N_f} + \frac{c_{sf}\pi_s n_s}{N_s}$$

This resubstitution risk formula reveals that the relative misclassification cost measure c_{fs} is only part of the influencing factor on the relative costs of the errors (n_f, n_s), as the number of firms in the sample and the prior probabilities are also factors. Thus, the misclassification cost measure used here is inflated compared to misclassification costs in other models. Comparable misclassification costs can be found by deriving the ratio of

n_f, n_s from the resubstitution cost measure, as shown below.

$$\begin{aligned}
\frac{n_f}{n_s} &= \frac{c_{fs}\pi_f}{N_f} / \frac{c_{sf}\pi_s}{N_s} \\
&= \frac{c_{fs}\pi_f}{N_f} \times \frac{N_s}{c_{sf}\pi_s} \\
&= \frac{c_{fs}\pi_f N_s}{N_f c_{sf}\pi_s} \\
&\quad \text{and if we substitute in the values we get:} \\
&= \frac{c_{fs} \times 0.02 \times 142}{58 \times 1 \times 0.98} \\
&\approx 0.05 c_{fs}
\end{aligned}$$

The uninflated values for the cost of Type I Error relative to Type II Error can be calculated from the equation shown above, and are presented below.

Frydman et al. (1985) measure	1	10	20	30	40	50	60	70
Uninflated equivalent (ECR)	0.05	0.5	1	1.5	2	2.5	3	3.5

These uninflated equivalents will be referred to as ECR values, which stands for the Error Cost Ratio of Type I to Type II Error. The range of ECR values can be interpreted as the Type I Error ranges between 0.05 times and 3.5 times as costly as Type II Error. It can then be seen that the results for the lowest two values for c_{fs} are not as important, as the cost of Type II Error is not greater that the cost of Type I Error in the real world[3].

Measurement of Classification and Prediction Accuracy

The four techniques (DA, RPA, CART and See5) were compared using the resubstitution risk measure for both classification and prediction accuracy. Each technique's performance was first tested on the original training data set to ascertain its classification accuracy. The prediction accuracy of each model was then estimated by using the V-fold cross-validation procedure with $V = 5$. The V-fold cross validation procedure involves dividing the data into V approximately equal groups and generating a DT for each $V - 1$ group. Each DT was then used to classify the group left out to estimate the prediction accuracy. The V-fold cross validation procedure was used in preference to the hold-out sample procedure as it more accurately estimates prediction accuracy when the sample size is small: the sample size used in this research is 200 businesses, which is considered small (Frydman et al., 1985). Simulation studies have shown that the prediction accuracy measured from the V-fold cross validation procedure with V=5 on a small sample is satisfactorily similar to the prediction accuracy measured from a large hold-out sample (Breiman, 1981).

[3]It is interesting to note that Altman (1980) estimated the true value of c_{fs} to lie between 32 and 62, but this is only from the view of commercial bank lending officers.

Generation of Models of Varying Complexity

To be consistent with Frydman et al. (1985), this research has generated two models for each technique that varying in complexity. The complexity of DA is measured by the number of variables included in the model, which is easily controlled using a stepwise procedure that adds variables one at a time. Similarly, the complexity of DTs is measured by the number of terminal nodes. The selection of DTs with different complexities is based upon controlling the amount of pruning done to a DT. Pruning a DT involves removing branches and replacing (small) subtrees with leaf nodes. In brief, DT building algorithms consist of two sequential stages:

1. Create a very complex DT with perfect, or near-perfect, classification. Pre-pruning is possible at this stage. It involves creating a less complex tree at the expense of not not exploring a more complex tree. This increases the efficiency of the tree generation process, but with the risk of a possible loss in classification and prediction accuracy.

2. Choose a sub-tree of the complex tree from Step 1, by systematically pruning (replacing subtrees with leaf nodes) until some criterion is met or no further pruning can be made, that is, a one node tree has been produced. This second step is needed as the tree generated in Step 1 is almost always subject to over-fitting and poor predictive ability.

Hence, DTs of lower or higher complexity are generated by respectively increasing or decreasing the amount of pruning.

3.3.2 DA and RPA Models

The DA and RPA models were constructed by Frydman et al. (1985) as follows. The two DA models were chosen using a forward stepwise procedure, where the

Simple Model (DA_s) included the 4 most statistically significant explanatory variables; and the

Complex Model (DA_c) included the 10 most statistically significant explanatory variables.

The two RPA decision trees were generated with different pruning procedures, whereby the

Simple Model (RPA_s) was chosen as the DT with the smallest resubstitution risk from the 5-fold cross-validation procedure.

Complex Model (RPA_s) underwent less pruning than RPA_s. Frydman et al. (1985) stated that they chose a "relatively complex tree" without providing anymore detail.

3.3.3 See5 Models

Release 2.02, the newest version of the software package available at the time of this research, was the version used[4]. For information on the underlying 'information entropy' algorithm used in See5 refer to Section 2.2.2, and to Quinlan (1993)[5] for further details. The settings used for the See5 research are explained below.

The 'minimum cases per leaf node' was set to 2 to ensure no pre-pruning was done. Furthermore, in-sample and 5-fold cross-validated performance were automatically analysed with the See5 package. Importantly, as See5 does not take prior probabilities as inputs the interpretable misclassification costs had to be used. That is, the following costs of Type I Error relative to Type II Error were used: 0.05, 0.5, 1, 1.5, 2, 2.5, 3 and 3.5.

The difference between the complex and simple See5 decision tree is based on the 'pruning CF' option. The pruning CF option controls the severity of the pruning whereby larger values result in less pruning. This option takes a value ranging from 0 to 100 percent and can be roughly interpreted as a significance level (α) in hypothesis testing, where a sub-tree is not pruned if it is significant at the pruning CF level. The default is 25%, but this value still produces relatively large trees; therefore, the:

Simple Model (See5_s) was generated using a 1% pruning CF; and the,

Complex Model (See5_c) was generated using a 10% pruning CF.

3.3.4 CART Models

Version 5.0 of the CART software package (produced by Salford Systems, San Diego, US) was used. This software claims to be the only DT software based on the original (proprietary) code of the original creators of classification and regression trees (CART) (Breiman et al., 1984). Furthermore, all enhancements made to the original CART system have been made with collaboration with the original creators.

The 'Entropy' method for selecting splitting variables was chosen due to its previous wide popularity and successful application. Similar to RPA, the prior probabilities and misclassification costs are both inputs in a CART model. Thus, the prior probabilities were set to 2% failed and 98% successful businesses and the original misclassification cost measure used by Frydman et al. (1985) ranging from 1 to 70 for the ratio of Type I Error to Type II Error cost. Furthermore, the in-sample and 5-fold cross validation results were both attainable from setting the 'test' method to cross-validation with 5-folds.

Similar to the RPA decision trees, the two CART decision trees were generated under the following conditions; the

[4]The demonstration version of this software was used, but it is important to realise that the algorithm in the demonstration version is *identical* to that in the full version. The only major difference is that the full version can process much more data than the demonstration version.

[5]Quinlan (1993) details See4.5, but the fundamental processes of See4.5 still remain in See5.

Simple Model (CART_s) was chosen as the model with the smallest resubstitution risk from the 5-fold cross-validation procedure.

Complex Model (CART_c) was chosen as the above tree without the last level of pruning as defined by the CART software.

3.4 Results

This section presents the results from implementing the methodology described above. The results have been analysed from three different aspects:

- The less complex models will be compared to the more complex models,
- The main comparison of the four techniques will then be done, and finally,
- An analysis of variable importance will be presented.

These results are presented as an extension to the comparison of RPA decision trees and DA undertaken by Frydman et al. (1985); therefore, for more specific details on the comparison between these two techniques refer to Frydman et al.'s original work. The exact numerical results are presented in a tabular format in Appendix A.

3.4.1 Effect of Model Complexity

The effect of the model complexity has been evaluated by comparing the less and more complex versions of each technique (RPA, DA, CART and See5) with each other. The less and more complex versions of each technique will be hereafter referred to as the simple and complex models respectively. The in-sample classification accuracy of the four complex models is superior to that of their simple model counterparts. The resubstitution risk is lower for the complex models over all misclassification costs, except for a few situations with the See5 package where the models have equal classification accuracy as shown in Table 3.3. These are the exact results that were expected as the simple DTs and DA model were fully contained within their complex counterparts. Thus, as models are generated in a way that only increases the complexity of the model if the classification accuracy increases, more complex models will always outperform less complex models on the in-sample data.

In contrast to the in-sample accuracy comparison, the cross-validation prediction accuracy of the complex models is not theoretically guaranteed to be better than the simple models. Moreover, Table 3.4 shows that in most cases the simple models produce more accurate predictions. Specifically, this table shows that the complex RPA and CART models are never the preferred option. On the other hand, three of the eight different misclassification cost studies resulted in the more complex See5 model outperforming the less complex See5 model. Furthermore, for low and high Type I Error cost, the simple model is the preferred DA model, but for the arguably more realistic mid-range Type I Error Cost the complex DA

ECR	**RPA**	**DA**	**See5**	**CART**
0.05	Complex	Complex	*Equal*	Complex
0.5	Complex	Complex	*Equal*	Complex
1	Complex	Complex	*Equal*	Complex
1.5	Complex	Complex	Complex	Complex
2	Complex	Complex	Complex	Complex
2.5	Complex	Complex	Complex	Complex
3	Complex	Complex	Complex	Complex
3.5	Complex	Complex	Complex	Complex

Table 3.3: A comparison of the in-sample classification accuracy of the simple and complex models for each technique. The table lists the more accurate model (complex, simple or equal) for each technique and misclassification cost. Note that the more accurate model is chosen as that with the lowest resubstitution risk, and ECR refers to the Error Cost Ratio of Type I compared with Type II Error. The actual in-sample resubstitution risk values for each model are shown in Appendix A.

model is preferred. Therefore, this analysis has not conclusively shown less complex models to be superior predictors when the techniques of DA and See5 are being used. However, in terms of predictive power, the less complex models were superior to the more complex models for the majority of the cases, for DA and See5. In addition, generating less complex models has been conclusively shown to increase predictive power when using the RPA and CART decision tree approaches. Thus overall, the less complex models had better predic-

ECR	**RPA**	**DA**	**See5**	**CART**
0.05	Simple	Simple	*Equal*	Simple
0.5	*Equal*	Simple	Simple	Simple
1	Simple	Simple	**Complex**	Simple
1.5	Simple	**Complex**	**Complex**	Simple
2	Simple	**Complex**	Simple	Simple
2.5	Simple	**Complex**	Simple	Simple
3	Simple	Simple	Simple	Simple
3.5	Simple	Simple	**Complex**	Simple

Table 3.4: A comparison of the cross-validated prediction accuracy of the simple and complex models for each technique. The table lists the more accurate model (complex, simple or equal) for each technique and misclassification cost. Note that the more accurate model is chosen as that with the lowest resubstitution risk, and ECR refers to the Error Cost Ratio of Type I compared with Type II Error. The actual in-sample resubstitution risk values for each model are shown in Appendix A.

tive power, even though they are outperformed by more complex models when classifying the original data set. This empirical finding is consistent with the *principle of parsimony*, whereby most statisticians prefer a less complex model given the choice.

3.4.2 Empirical Comparison of Techniques

This section contains a comparison of the in-sample classification and cross-validation prediction ability of RPA, DA, See5 and CART based on resubstitution risks. The resubstitution risk measure was used to compared these techniques to maintain consistency with with the work of Frydman et al. (1985). The resubstitution risk is essentially a weighted error cost measure that takes into account different misclassification costs. Thus, the lower the resubstitution risk the better the model.

The Resubstitution risk measure factors in the different misclassification costs of Type I and Type II errors, but it is not easily interpreted. On the other hand, the correct classification percentage measure, calculated as the number of correct classifications divided by the sample size, is easily interpretable yet does not factor in different misclassification costs. Thus, it is incorrect to compare techniques over different misclassification costs based on their correct classification percentage. Nevertheless, to give the reader a rough idea of the accuracy of the DT techniques being examined the average correct classification percentage is shown Table 3.5. The tabulated percentages in Table 3.5 are based on an average of

	Correct Classification Percentage	
ECR	In-Sample	Cross-Validated
0.05	75.4%	72.5%
0.5	93.8%	83.9%
1	93.4%	83.6%
1.5	93.5%	82.3%
2	92.4%	81.4%
2.5	92.1%	81.1%
3	90.6%	81.5%
3.5	90.4%	80.5%

Table 3.5: The average percentage of correct classifications for the See5 and CART techniques for each misclassification cost. ECR refers to the Error Cost Ratio of Type I compared with Type II Error.

the simple and complex trees generated by See5 and CART[6], and provide an estimate of the overall performance of the See5 and CART techniques. Excluding the results from an unrealistic ECR of 0.05, CART and See5 correctly classified more than 90% of business from

[6]The RPA and DA models were not included as the information needed to calculate the in-sample and cross-validated correct classification percentage was not published by Frydman et al. (1985).

their training sample for all likely misclassification costs. As expected this percentage drops for prediction ability, but the reduction is approximately 10% for all cases (not including ECR of 0.05). Thus, the CART and See5 decision tree approaches are correctly predicting more than 80% of businesses as successful or failing for a wide range of misclassification costs. This is a very good and practically useable level of accuracy for a BFP model.

Table 3.5 also reveals that the accuracy of both classification and prediction decrease for both low and high Type I Error costs (relative to Type II Error cost), that is, for ECR values not close to 1. The reason for this is that when the misclassification costs are not equal the models attempt to reduce the more costly error, even at the expense of a possible overall decrease in the number of correct classifications.

In Sample Classification

There is large variation between the in-sample classification ability of RPA, DA, See5 and CART. The in-sample classification accuracy of the simple models is compared in Figure 3.1. Overall, it is clear that the DA_s model is the worst in terms of classification ability, as all

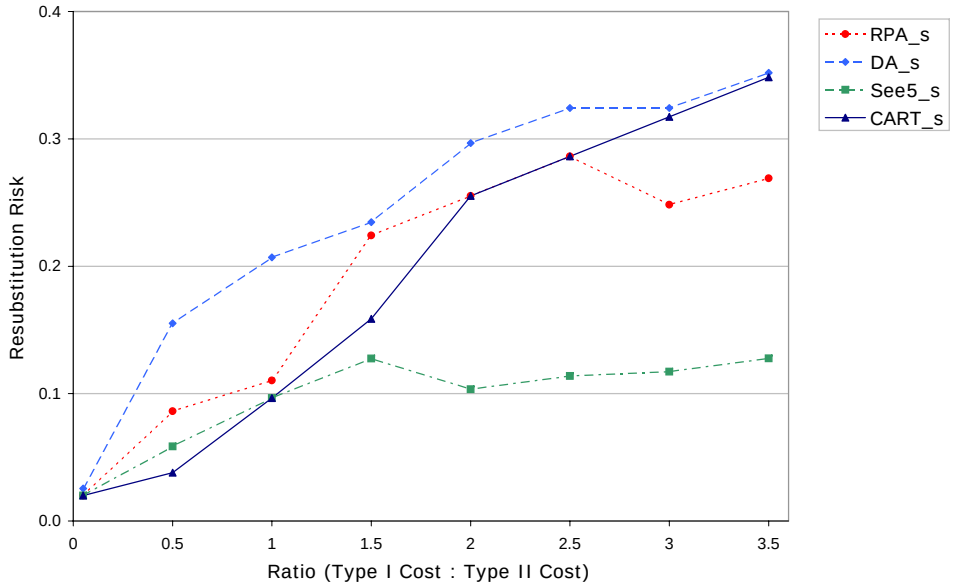

Figure 3.1: Comparison of the Classification Accuracy of the Four Simple Models.

of the three DT models had lower resubstitution risks over all misclassification costs. See5_s had the lowest resubstitution risk except for the following cases: for an ECR of 0.05 all three DT techniques had equal lowest[7], for an ECR of 1 See5_s was equal with CART_s, and for an ECR of 0.5 CART_s's resubstitution risk was slightly lower. As it has already been established that Type I Error is likely to be more costly than Type II Error, the cases where ECR is lower than 1 are not as important. Thus, the See5_s model was significantly better than the other simple models at classifying the original sample. The resubstitution risk

[7]For an ECR of 0.05, all three DT techniques were the same as a naive model that assumes all businesses to be successful.

of the See5_s models does not significantly increase for higher Type I Error costs, which indicates that the technique is adjusting well by reducing Type I Error (at the expense of higher Type II Error). On the other hand, the other models have a significant upward trend in the resubstitution risk for increasing Type I Error cost. Therefore, the classification ability of See5_s is more superior for higher Type I Error costs.

CART_s and RPA_s had similar overall classification accuracy. CART_s had lower resubstitution risk for ECR values up to 1.5, but the RPA_s model had significantly lower resubstitution risks for ECR values of 3 and 3.5. Finally, the two models have equal classification accuracy for ECR values of 2 and 2.5. Hence, a model with superior classification ability can only be chosen from RPA_s and CART_s if the misclassification cost is known.

As discussed in the previous section, all the complex models had lower resubstitution risk for their in-sample classification as they are trained extensions of their simple counterparts. Furthermore, similar comparative results were also obtained amongst the complex models; the in-sample classification accuracy of the complex models is presented in Figure 3.2. The See5 and DA models still remain the best and worst classifiers. The See5_s model

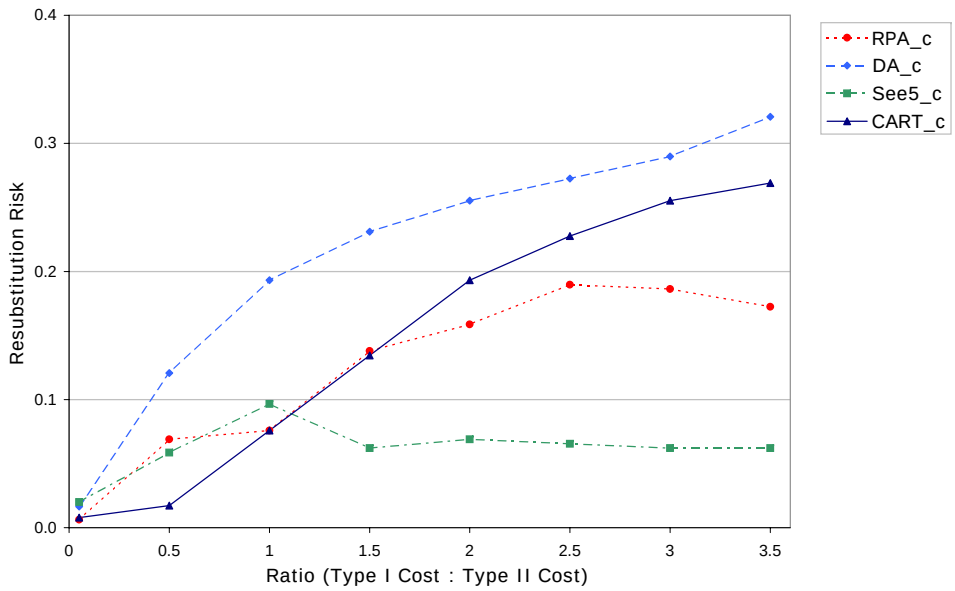

Figure 3.2: Comparison of the Classification Accuracy of the Four Complex Models.

is even superior to the DA_c model over all misclassification costs. Moreover, the RPA_s and CART_s have superior classification ability than the complex DA model for more than 50% of the misclassification cost values. In addition, the CART technique is again the best classifier for ECR values of 1 or less[8]. However, the RPA_c model becomes superior to CART_c for ECR values of 2 and greater, compared with 3 and greater for simple models. Therefore, the RPA_c is considered to have slightly superior overall classification ability than the CART_c model.

[8]The RPA_c model actually has a resubstitution risk 0.002 points lower than CART_c for an ECR of 0.05, but overall CART_c model was superior for ECR of 0.05, 0.5 and 1.

For both the complex and simple models, See5 was remarkably superior in terms of classification ability, especially for higher Type I Error costs. A partial reason for this is that See5 usually generates more complex trees than CART and RPA. This is further discussed at the end of the next section. Figure 3.5 in the next section shows that See5 decision trees increase, rather than decrease like CART, as the Type I Error cost increases. This underlying difference appears to be the main source of See5's superior classification ability, but it should also cause a corresponding reduction in the accuracy of predictions (to be analysed in the next subsection).

Overall, it is clear that all the DT techniques had superior classification ability compared with DA. It was also clear that See5 was the best DT technique in terms of classification ability. RPA and CART were somewhat similar, but RPA appears to be a slightly better classifier when the Type I Error cost is more than double the Type II Error cost.

Cross-Validated Prediction

Compared with the large variability of in-sample classification, the four techniques had more comparable cross-validated prediction ability. The prediction accuracy of the complex models is shown in Figure 3.3 and will be analysed first. It is very noticeable that the lines in

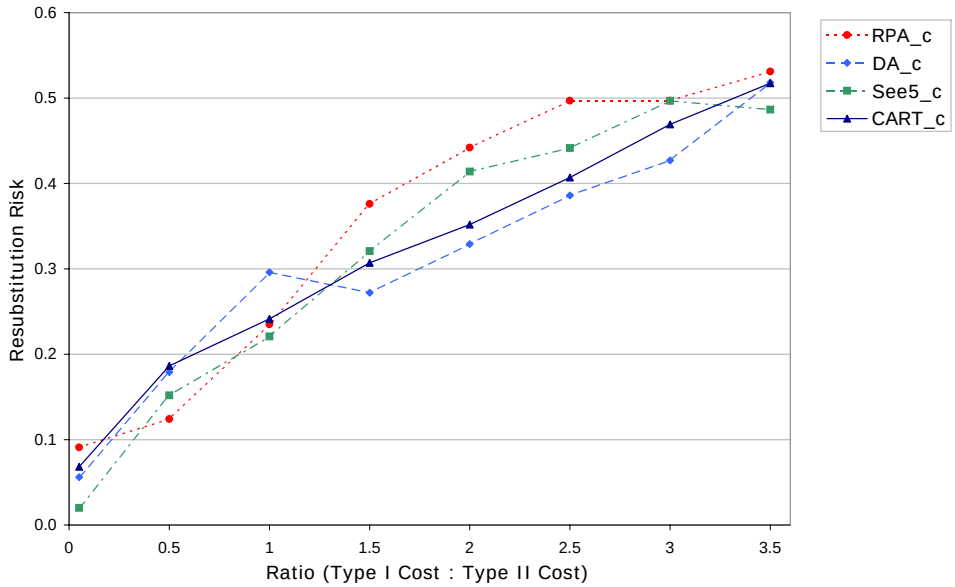

Figure 3.3: Comparison of the Prediction Accuracy of the Four Complex Models.

this graph are much closer together than in the previous graphs that compared the in-sample classification accuracy. Furthermore, there appears to be no obvious relationship between the best classifiers and the best predictors. See5, the best classifier, is only superior to RPA in terms of prediction ability of the larger model. Furthermore, for increasing costs of Type I Error the relative classification accuracy of DA and CART decreased; however, Figure 3.3 reveals that CART_c and DA_c are the best predictors for ECR values greater than 1. Overall, the CART_c model has good prediction accuracy, but the DA_c model appears to

have the best overall prediction accuracy for realistic ECR values of greater than 1. Further analysis of the prediction accuracy of the complex models shown in Figure 3.3 will not be included in this study as the prediction accuracy of the simple models is more relevant. The reason for this is that the simple models were shown to have superior prediction accuracy in Section 3.4.1. Therefore, as searching for the technique with the *best* predictive power is the primary goal of this research, the simple models are more relevant.

The cross-validated prediction accuracy of the simple models is presented in Figure 3.4. This graph reveals that the four simple models have similar prediction ability. Due to this similarity, each model will be analysed separately.

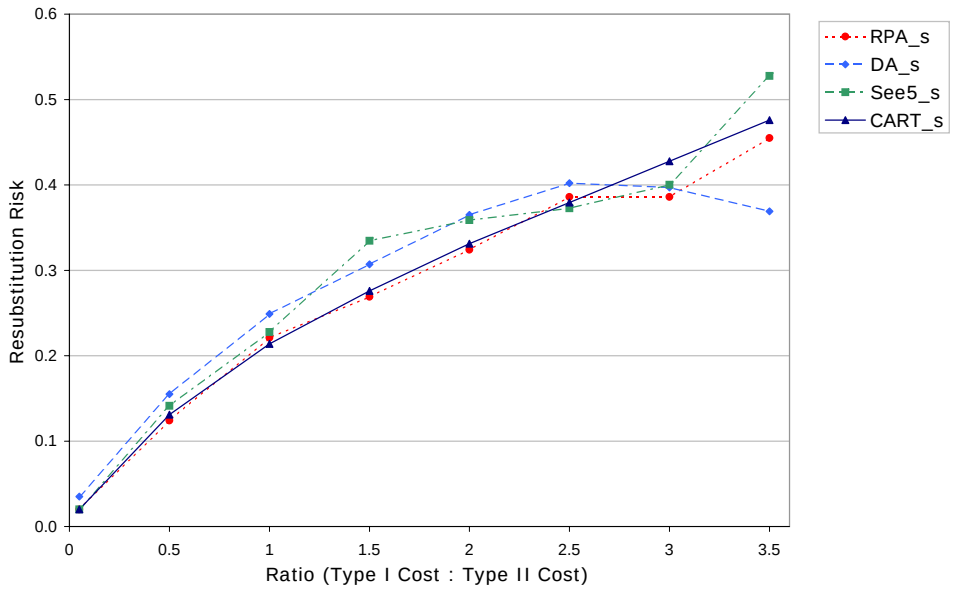

Figure 3.4: Comparison of the Prediction Accuracy of the Four Simple Models.

DA_s The reduction in complexity has not increased the predictive power of the DA model as significantly as the DT models. Consequently, the DA_s model has the biggest resubstitution risk for ECR values less than or equal to 2.5, except for 1.5 where See5_s is slightly larger. Therefore, DA is the worst predictor when Type I Error is not more than 2.5 times costlier than Type II Error. Although the DA_s's resubstitution risk increases as the ECR increases, the resubstitution risk actually decreases when the ECR is greater than 2.5. This causes DA to be the best predictor, with a significantly lower resubstitution risk, when Type I Error is 3.5 times as costly as Type II Error.

See5_s The See5_s model is generally the worst performing DT technique. It has the highest resubstitution risk of all DT models, except for ECR values of 2.5 and 3. This suggests that See5 decision trees might be better for large Type I Error costs, but the See5_s model is the worst predictor by a significant margin for the largest ECR value of 3.5. Thus, See5 is arguably an overall better predictor than DA, but it is the worst predictor amongst the DT techniques.

CART_s The CART_s model has the most consistent predictive ability over different misclassification costs: this is identified as CART_s has the straightest line without any unexpected rises or drops in the resubstitution risk. For ECR values of less than and equal to 2.5, CART had almost identical predictive power to RPA. CART_s and RPA_s are the two best techniques for ECR values up to 2, with the individual best technique changing between the two techniques. See5 actually has a slightly smaller resubstitution risk than CART when the ECR is 2.5, and CART is the worst technique when the ECR is 3. Nevertheless, CART's consistent performance causes it to be better than See5 and DA, but slightly worse than RPA for an ECR of 3.5. Overall, CART has superior predictive power compared with DA and See5. In addition, CART is almost identical to RPA, but RPA adjusted slightly better the highest two Type I Error costs.

RPA_s The simple RPA_s model handles the changing misclassification costs much better than the complex RPA_c model. The reduction in complexity from the complex to the simple model had the largest corresponding increase in predictive power for the RPA technique. Consequently, the RPA_s model had superior predictive power compared with the other three techniques, although CART was almost identical for ECR values of 2.5 and lower.

It is important to discuss the remarkable difference between the classification and prediction ability of the See5 technique. See5 was a significantly superior classification technique, but the predictive ability of See5 was not as good (as discussed above). In addition, the large increase in resubstitution risk for the highest ECR value of 3.5 is concerning. The reason for this is partially due to the fact that See5 generates more complex trees compared with other DT techniques, as is shown in a comparison with CART in Figure 3.5[9]. This graph reveals that the DTs generated by See5 are significantly more complex than the DTs generated by CART. Furthermore, while the CART decision trees on average reduce complexity the See5 decision trees increase complexity, for increases in ECR. The resulting more complex See5 models then have better classification ability yet worse prediction ability. However, the See5 technique still remains somewhat comparable with the other DT techniques in terms of prediction ability. Hence, this increased complexity is only a partial reason for the relative change between the classification and prediction ability of See5, and there are other elements of the See5 algorithm that make its classification ability significantly superior to its prediction ability.

Overall, this research has found DTs to have slightly better predictive power than DA. CART and DA were also found to be better predictors than See5. Finally, RPA and CART were almost identical except for the highest levels of Type I Error cost where RPA adjusted better. The selection of RPA as the technique with the best predictive power is cautioned

[9]The information needed to create a similar graph including RPA was not published by Frydman et al. (1985). However, from partial information provided it is concluded that the RPA decision trees were more comparable to the complexity of the CART decision trees.

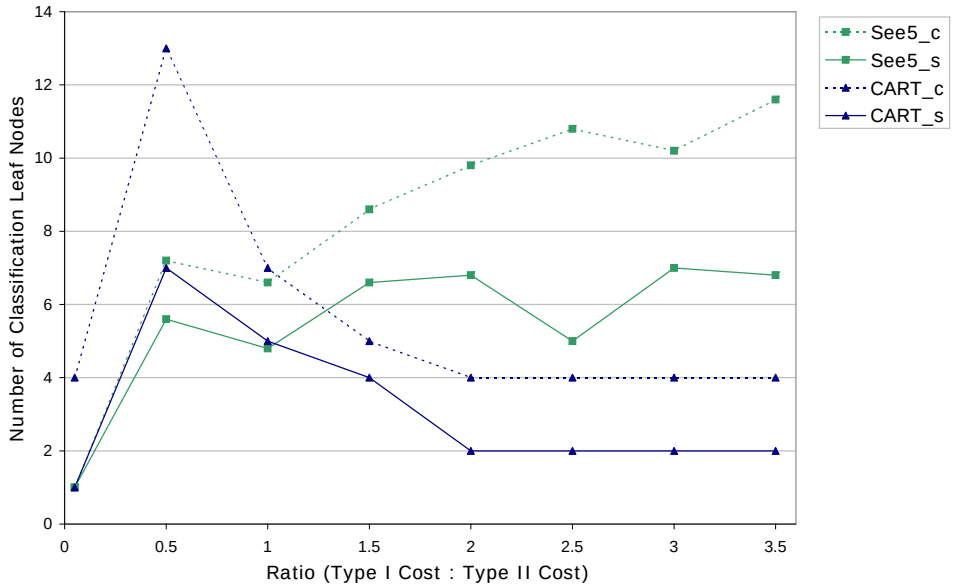

Figure 3.5: Comparison of the Complexity of CART and See5 decision trees.

by ensuring that the complexity of the RPA built DTs is limited as was done for the RPA_s decision trees.

3.4.3 Analysis of the Importance of Variables

The variables found to be significant and important depend upon the technique used. Therefore, variable importance will first be analysed separately for each technique before overall conclusions are made.

DA The four most significant and important variables according to DA are shown below.

Variable	Variable Type	Coefficient	Standardised Coefficient
CA/CL	Liquidity	0.622	0.528
NI/TA	Profitability	5.322	0.514
Ln(TA)	Business Size	0.712	0.491
MVE/TC	Market Structure	1.149	0.328

The standardised coefficients indicate the importance of each variable in predicting business failure. Thus, the CA/CL ratio is found to be the most important variable by DA.

RPA All of the DTs generated by RPA had the leverage ratio CF/TL as the root, except for the case of equal misclassification costs when Cash/SLS was the most important variable used in the root of the DT[10]. However further analysis of the structure of the RPA decision trees reveals that after the root, the DTs vary considerably regarding

[10]The structure of a couple of RPA built DTs is shown in Frydman et al. (1985).

the number of variables, the variables used, the sequence of variables and the splitting points (Frydman et al., 1985). It was a general trend that Cash/SLS and TL/TA were important for most misclassification costs, followed by Log(INT+15) and MVE/TC that were often also important. Nevertheless, the CF/TL ratio was by far the most important variable according to RPA. Overall, the variables found to be important by DA and RPA are completely different.

CART The variables important in CART decision trees are different from both DA and RPA. The CA/SLS variable was found at the root of all trees generated by CART. Furthermore, except for the lowest Type I Error cost that found QA/SLS to be the second most important variable, all simple DTs generated by CART were sub-trees of the DT developed for an ECR of 0.5 as shown in Figure 3.6. Unlike RPA decision trees, CART decision trees generally used the same variables in the same sequence with the same splitting points. Therefore it is easy to identify the most important

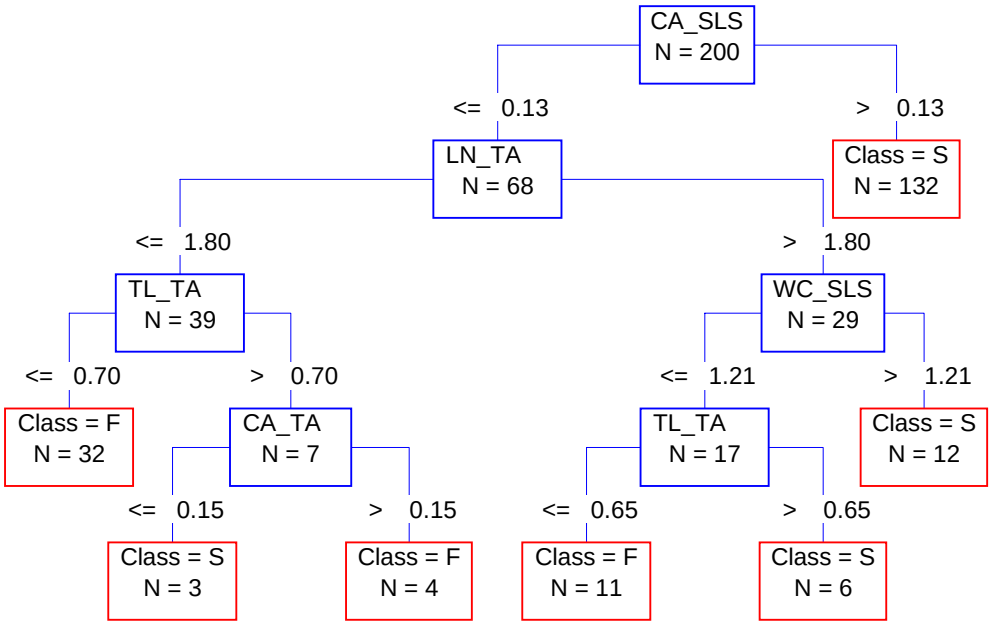

Figure 3.6: DT generated by CART with simple model setting and ECR of 0.5. Note that this graphical tree representation was automatically created by the CART software package, a feature that is not present with See5.

variables in CART models, which are in order of importance,

- The activity ratio CA/SLS,
- The business size variable Ln(TA) also found to be important by DA,
- The coverage ratio TL/TA, and
- The liquidity ratio WC/SLS.

See5 Except for an ECR value of 0.5 where the WC/TS was used at the root, the activity ratio CA/SLS was used at the root of every See5 decision tree. This similarity with CART is probably explained as both of these techniques are types of entropy DT algorithms (refer back to Section 2.2.2). However, the remainder of See5 decision trees vary considerably over different misclassification costs. It is also very common for See5 decision trees to have variables repeated in more than one splitting rule. All variables were found to be at least somewhat important in at least one See5 decision tree, but the majority of the important variables were profitability and liquidity ratios. Specifically, the liquidity ratio WC/TA and the profitability ratio EBIT/TA, then the profitability ratio NI/TA (also found as important by DA), were found to be important in See5 decision trees after CA/SLS.

See5 and CART are the only techniques with the most important variable (CA/SLS) in common. The CA/CL variable that was found to be the most important by DA was not found to be important in other models. Similarly, the most important variable (CF/TL) in RPA models was not found to be very important in any other models. This is an important finding as CF/TL was the only cash flow variable in the data set, and it was also the found to be the best predictor of business failure by Beaver (1966). Therefore, the question of whether cash flow variables are the best predictors of business failure remains without a conclusive answer. Hence, it is advisable to continue to include cash flow variables in business failure data sets so that the model estimation procedures can decide whether or not to include them in the final model.

Overall, there is a wide range of variables found to be important in the DA, RPA, CART and See5 models. The only common element appears to be that DA, RPA, CART and See5 models tend to include a range of variables types rather than just variables from one classification group.

An Alternative Approach

The large variation in variables found to be important in different models reduces confidence in creating a ranking based on their ability to predict business failure. Part of the reason for this large variation is that the variables included in statistical models are dependent on the variables already included in the model. For example, the correlation between variables plays a large role, where usually only one variable out of a highly correlated pair of variables will be included in the final model. However, the variable that has been left out is not a bad predictor, but rather there is just a very similar alternative that is slightly better. Thus, many variables that have good predictive power in terms of business failure are often not included in the final models.

A different way of identifying important variables, rather than interpreting the usefulness of variables in one or more DT models, is required to produce a general ranking by predictive power. Contrary to the general statement made in Section 2.2.1 that 'DT techniques do not provide a numeric measure of variable important', CART provides a

unique feature, whereby it produces a statistical ranking of variable importance based on how accurate the model would have been produced if it did include each variable[11]. CART generates a measure between 0 and 100 with higher numbers representing increased performance. These importance levels are shown in Table 3.6 for two different misclassification cost levels. It is important to note that variables not included in CART decision trees can

1 ECR		**3.5 ECR**	
CA/SLS	100	CA/SLS	100
WC/SLS	94	WC/SLS	84.34
Cash/TA	75.75	Cash/TA	71.89
WC/TA	65.65	WC/TA	63.98
TL/TA	21.77	TL/TA	6.02
Ln(TA)	19.47	Ln(INT+15)	3.84
QA/SLS	12.16	Cash/SLS	-
MVE/TC	11.16	CF/TL	-
CA/TA	9.31	CA/CL	-
SLS/TA	8.36	CA/TA	-
Ln(INT+15)	8.31	EBIT/TA	-
QA/TA	4.28	Ln(TA)	-
RE/TA	4.28	MVE/TC	-
EBIT/TA	0.36	NI/TA	-
Cash/SLS	-	QA/CL	-
CF/TL	-	QA/TA	-
CA/CL	-	QA/SLS	-
NI/TA	-	RE/TA	-
QA/CL	-	sd(EBIT/TA)	-
sd(EBIT/TA)	-	SLS/TA	-

Table 3.6: Variable Importance for CART for 1 and 3.5 misclassification costs.

be found as important by this procedure; for example, while Cash/TA did not appear in any CART decision trees it was found by this procedure to be the third most important variable. The figures in Table 3.6 are representative of all misclassification costs: CA/TS, WC/TS, Cash/TA and WC/TA are the four most important variables for all misclassification costs, except for an ECR of 0.05 where QA/TS is the third most important variable instead of Cash/TA. The table also reveals the general trend that the relative importance of these four main variables increases for higher Type I Error costs. In addition, TL/TA and Log(TA) are also shown to be somewhat important, with Ln(Int+15) also becoming somewhat important for high Type I Error cost levels.

[11]The variable importance ranking process is further explained in the manual for the CART software package.

These results are still contrary to the variables found important by DA and RPA. In addition, this ranking scheme does not have an interpretation associated with it like the variable importance rankings based on specific models. For example, based on Table 3.6, it can not be concluded that if the CA/SLS ratio is greater than 0.13 then the business will be successful, as can be concluded from the common CART decision tree shown in Figure 3.6 (on page 51). Nevertheless, it is considered by some researchers that the unique variable importance ranking developed automatically by CART is the best way for determining variable importance. For example, Shirata (1998) used CART for the sole purpose of deciding what variables were important enough to include them in their DA model to predict business failure.

3.5 Conclusions

Extending the work of Frydman et al. (1985) to include See5 and CART provided further empirical evidence to support the claim that less complex, more parsimonious, models are better predictors than more complex models. More importantly, this research has provided further evidence to suggest that DT techniques are superior predictors of business failure. Amongst the DT techniques, See5 had the worst overall predictive power, even though it had the best in-sample classification ability. The CART and RPA decision tree techniques produced very similar results and had the best overall predictive ability, but the RPA technique had slightly better overall predictive ability as it performed better with high Type I Error costs. Nevertheless, the best technique still remains dependent on the specification of the misclassification cost, which depends upon the context of the prediction and will most probably involve subjective judgement.

Chapter 4

Survival Analysis Research

4.1 Aim

The research presented in this chapter has been undertaken to further the assessment of Survival Analysis (SA) techniques for use in business failure prediction (BFP). The main goal is to assess the empirical classification and prediction accuracy of the Cox model applied to BFP. This analysis has been undertaken on a sufficiently large data set and over different years of predictions with different misclassification costs, in order to address some of the gaps in the SA literature identified in Section 2.4.2.

4.2 Data

A large time-series data set has been used for this research[1]. This data set was acquired from researchers that have previously undertaken BFP studies (Kahya and Theodossiou, 1999): their work was summarised in Section 2.1.7. The main properties of this time-series data set are presented in Table 4.1. The time-series nature of the data set means that more than a boolean variable describing whether a business is successful or failing is required. Therefore, in addition to the boolean variable representing success and failure, there is also a variable that stores the number of years till failure for the failing group of businesses. Businesses were chosen from large stock exchanges to impose a minimum business size restriction as Kahya and Theodossiou believe that the explanatory variables will relate differently to significantly smaller businesses, as has been shown in previous studies such as that by Edminster (1972). In addition, petroleum and mining businesses were not considered as they have statistically different financial attributes when compared with other manufacturing and retail firms. Within these boundaries, Kahya and Theodossiou state that this data set is large enough to be representative of the population. Furthermore, there is no sampling bias in the data, which is a weakness of a major prior BFP study (Laitinen and Luoma, 1991) of the Cox model.

[1]The data set is available from the author upon request.

Property	**Value**
Businesses (Type)	Manufacturing and Retail (listed on the NYSE and AMEX)
Selection Procedure	Random selection
Businesses (Number)	189: 117 successful and 72 failed
Explanatory Variables	27 financial variables: mostly financial ratios
Time–line	1974–1991 (18 years)
Number of Business–Years (Instances)	2,954

Table 4.1: Properties of the data set used for SA research. The number of business–years does not equal the (number of businesses × number of years) as some businesses fail before the end of the study period and some businesses enter the study after its beginning. NYSE and AMEX are acronyms of two US financial exchanges, the New York Stock Exchange and the American Stock and Options Exchange.

4.2.1 Definition of Business Failure

The time of failure for a business in this data set is determined as the time when a business first experiences one of the signs of failure, which in this study are based on debt default as stated by Kahya and Theodossiou (1999). That is, a business is considered a failure if any debt default or debt renegotiation attempts occur with any of its creditors. Information about the debt default or debt renegotiation was gathered from the pertinent annual issues of the Wall Street Journal Index. This definition of business failure was used in preference to the legal definition of business failure to avoid the data contamination that occurs with the latter definition. Basically, it is possible and sometimes common for financially failing businesses to be taken over by other businesses and never be legally considered failed, but for our modelling purposes they should be considered as failed. Conversely, businesses also file for liquidation or reorganisation to satisfy the legal definition of bankruptcy, when they are not financially in trouble but rather making a strategic choice. Further reasoning behind the decision to use this definition of business failure is explained by Kahya and Theodossiou (1999).

4.2.2 Explanatory Variables

The 27 explanatory variables in this data set consist of a broad range of financial ratios, including some variables that measure business size. All the variables were obtained by Kahya and Theodossiou from the 1993 annual industrial and research COMPUSTAT databases. Most (22) of the ratios were chosen as they had appeared as significant explanatory variables in BFP models from previous studies. The last 5 variables were chosen by Kahya and Theodossiou as they considered them to be potentially important predictors of business failure. Specifically, the data set comprised:

Eight profitability ratios: gross profit to sales (GP/SLS), net income to book value of equity (NI/BVE), net income to fixed assets (NI/FA), net income to total assets (NI/TA), operating income to fixed assets, (OPI/FA), operating income to sales (OPI/SLS), operating income to total assets (OPI/TA), and retained earnings to total assets (RE/TA).

Seven liquidity ratios: cash to current liabilities (Cash/CL), cash to total assets (Cash/TA), current assets to current liabilities (CA/CL), current assets to total assets (CA/TA), net working capital to sales (WC/SLS), net working capital to total assets (WC/TA), and quick (or immediate) assets to current liabilities (QA/CL).

Four business size variables: log transformations (Ln) of: deflated fixed assets (FA), deflated sales (SLS), deflated total assets (TA) and the number of employees (EMPL).

Three management efficiency ratios: accounts receivable to current assets (RCV/CA), accounts receivable to sales (RCV/SLS) and inventory to sales (INV/SLS).

Two financial leverage ratios: long-term liabilities to total assets (LTL/TA), and total liabilities to total assets (TL/TA).

One activity ratio: sales to total assets (SLS/TA).

One operating leverage ratio: fixed assets to total assets (FA/TA).

One market structure ratio: market value of equity to total liabilities (MVE/TL).

Table 1 in Kahya and Theodossiou (1999) also details the specific studies that found each variable to be significant, as well as the computation used to calculate each variable from the fields in the COMPUSTAT database.

4.2.3 Hold-out Data

For the purpose of estimating a model's ability to predict the failure or success of businesses not used in the model estimation process, a separate hold-out data set was created from the overall data set. The hold-out data set was created as approximately 10% of the size of the initial data set with similar proportions of successful and failed businesses. The 11 successful and 7 failed businesses were randomly chosen from the initial data set of 117 successful and 72 failed businesses; therefore, the training data set comprised 106 successful businesses and 65 failed businesses. Overall, this meant that the training data set comprised 2,669 instances, and the hold-out data set comprised 285 instances.

4.3 Methodology

The methodology used in this research applied the Cox model to the data set described in the previous section. The Cox model was then compared with the most popular techniques

for BFP, that is, discriminant analysis (DA) and logit analysis (LA). In addition, two hybrid models integrating Cox's model into each of DA and LA were created to test the success of combining the more popular techniques with SA techniques. The relevant aspects of these techniques will be reviewed in this section, but more information about DA, LA and the Cox model is available in the reviews in Chapter 2.

The methodological issues that are relevant to all models will be presented in the first section. Then, the methodology used in this research to generate the DA, LA and Cox models will be presented.

4.3.1 General Methodology

The software package used to develop all of the statistical models was SPSS version 13.0 for Windows. In addition, Microsoft Office Excel 2003 (SP2) was used to efficiently analyse the accuracy all of the models. Both the classification and prediction ability of the DA, LA and Cox models were analysed. The classification ability was determined by how accurately a model classified businesses in the original training data set, known as 'in-sample' classification. The prediction ability was determined by how accurately a model classified new businesses from the hold-out data set.

The different models were compared based on prediction intervals of 1-year, 2-year, and every subsequent yearly period for up to 10-years. The purpose of this extensive comparison was to reveal whether certain techniques perform better at shorter or longer term predictions. This analysis was made possible due to the time series nature of the data. However, the comparison of the time-series Cox model with the cross-sectional DA and LA techniques was not straight forward[2]. This difficulty stems from the fact that one Cox model can make all the different (time) length predictions; that is, the same Cox model can make 1-year, 2-year and up to 10-year predictions. However, DA and LA models are not this versatile and can only make the specific length of prediction that was presented in the training data. Therefore, a single Cox model was compared with a separate DA and LA model for each different prediction length. It is arguable that this could bias the results towards the traditional models, but this is the process that would be undertaken for real-world predictions of different length with both DA and LA. Thus, the ability of one Cox model to do the same as many DA and LA models is an efficiency advantage of the Cox model, which also has advantages in simplifying model interpretation and identification of significant variables.

The different models have also been compared for various misclassification costs over each prediction length to assess how they adapt to higher Type I Error costs. As the misclassification costs are not inputs integrated into the model generating process for DA, LA or the Cox model, the different misclassification costs were achieved by varying the cut-off value of the probability of success. All three techniques output a probability of success or survival[3], which is then compared with a cut-off value ranging between 0 and 1

[2]Cross-sectional models are those that only model a snapshot in time.

[3]The probability of failure is simply calculated as 1 − probability of success/survival.

to determine whether the prediction is for failure or success. Usually this cut-off value is set to 0.5 representing equal misclassification costs, whereby a business with a probability of success/survival greater than 0.5 results in a successful prediction (else a fail prediction). The DA, LA and Cox models have been compared over 8 different cut-off values, ranging from 0.5 up to 0.85, in 0.05 intervals. Larger cut-off values increase the chance that a given business will be classified as a failure, and consequently reduce the amount of Type I Error. Therefore, higher cut-off values represent higher Type I Error costs. Thus, cut-off values lower than 0.5 were not studied as they represent Type II Error being more costly than Type I Error, which is not a realistic scenario (as discussed in Section 1.1.1 on page 3). Specifically, the relative cost of Type I Error compared with Type II Error, termed the ECR in Chapter 3, is estimated as the likelihood of a business being classified as failed by calculating cut-off/(1 − cut-off. The ECR values corresponding to each cut-off level are presented in the table below.

Cut-off	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85
ECR	1	1.22	1.5	1.86	2.33	3	4	5.67

For the equal misclassification cost case the techniques can be compared based on their percentage of correct classifications. However, as previously discussed (in Section 3.4.2 on page 44) a measure simply based on the number of correct classifications can not be used to compare techniques in a situation of unequal misclassification costs as it assumes equal costs. Therefore, for the cases of unequal misclassification costs the techniques have been compared by a weighted error cost measure, for which lower values indicated better performance. The measure is calculated as

$$(ECR \times n_f + n_s)/N$$

where,

- The subscripts s and f correspond to the successful and failing classification groups respectively,
- n_i is the number of group i misclassifications, and
- N is the number of instances in the relevant data set (in-sample or hold-out data set), and the division by N is performed simply to reduce the range of the weighted error cost measure.

4.3.2 DA and LA Models

Each of the 10 DA and LA models produced for each prediction length were generated by standard SPSS procedures. As the DA and LA models are cross-sectional models they require cross-sectional data sets. Therefore, 10 different data sets were produced, one for each prediction length. The only difference between the data sets was the boolean response variable, which was set to represent 'successful' if the years to failure variable was greater

than the prediction length. For example, if the years to failure for a particular instance was 5, then the response variable would be successful for the 1-year prediction data set, but failed for the 10-year prediction data set.

The DA models were estimated using a forward stepwise procedure so that only significant variables were included. The prior probabilities were set to be computed from the proportion in the initial data set, as this has previously been shown to increase the predictive ability of DA models. The LA models were estimated using a similar forward stepwise procedure based on likelihood ratios (termed the 'Forward: LR' method in SPSS). For the stepwise procedures in both the LA and DA models, the maximum level of significance for entry was set at 5% and the minimum level of significance for removal was set at 10%. In addition, constant terms were included for both the LA and DA models. The output variable measuring the probability of belonging to the successful group from the final model was then exported to Microsoft Excel to analyse the results for the 8 different cut-off values.

4.3.3 Cox Model

As stated previously, only one Cox model needed to be developed due to its flexibility. The model's parameters were estimated using the in-built Cox regression procedure in SPSS[4]. As the Cox model is a time-series model the procedure took as inputs, both the boolean variable representing failure or success and the variable containing the number of years till failure, as well as the explanatory variables. The model building process was conducted using a forward stepwise procedure based on likelihood ratios (termed 'Forward: LR'), where the significance level boundaries for entry and removal were set to 5% and 10% as with the DA and LA models.

The Cox model was designed primarily for analysing given data sets rather than generating a model to make future predictions. Therefore, the model's parameters were exported to Microsoft Excel to assess the model's accuracy over various misclassification costs and prediction intervals. With the parameters estimated from SPSS, the Cox model was created based on the Cox equation (previously defined in Section 2.3.1 on page 26) for the cumulative hazard function $H(t)$:

$$H(t) = h_0(t)e^{X^T\beta}$$

As previously defined (in Section 2.3.2 on page 28) $h_0(t)$ is the baseline hazard function and $X^T\beta$ is the multiplication of the coefficient and variable matrices. The survival function $S(t)$, which represents the probability of survival until time t, was then created using its relationship with $H(t)$ (also defined in Section 2.3.1), where:

$$\begin{aligned} S(t) &= e^{-H(t)} \text{ , that is,} \\ S(t) &= e^{-h_0(t)e^{X^T\beta}} \end{aligned}$$

[4]The Cox regression procedure in SPSS does not include constant terms

Rather than applying the cut-off comparison to the hazard function as done previously by Laitinen and Luoma (1991), this study has compared the cut-off values with the survival function $S(t)$. The reason for this is that the survival function is conceptually related to the probability of belonging to the successful group used in both DA and LA to compare with the cut-off values.

4.3.4 Hybrid Models

To determine whether a combination of the Cox model with one of the most popular techniques (DA or LA) would better predict business failure, two hybrid models were developed. These hybrid models were generated by including the values of the survival function as an explanatory variable in the DA or LA model estimation process. The survival function values included in the data set matched the prediction length of the DA or LA model; for example, $S(2)$ values were used for the DA and LA model used for 2-year ahead predictions. The DA and LA models were then developed (and analysed) for each prediction length in the same way as described above, but with an extra explanatory variable.

4.4 Results

This section presents the results from implementing the methodology described above. Firstly, the change in the business failure process over time will be analysed by using information obtained from the Cox model. The empirical performance of the techniques will then be investigated: the DA, LA and Cox models will be compared and then the hybrid models will be analysed. Finally, the last subsection discusses the relative significance and importance of the explanatory variables. The numerical results are presented in a tabular format in Appendix B.

4.4.1 Analysis of The Failure Process

As a Survival Analysis (SA) technique, the Cox model automatically outputs the survival function and the hazard function, and information on how they change over time. This information can be interpreted easily to gain further understanding of how the business failure process changes over time.

The survival function for each of the 18 years of the time-series data set used in this research is shown in Figure 4.1. The survival probabilities that were used to create this figure were calculated for an average business, that is, based on the mean of the explanatory variables. This graph reveals a linear decline in the survival rate of a business over the first 15 years of it being included in the study. The survival probability of a business reduces by approximately 8% each year. However, after 15 years the survival probability does not drop significantly and remains steady at about 77%. Nevertheless, this survival probability is still reasonsably high. The opposing, but consistent, trend was found from analysing

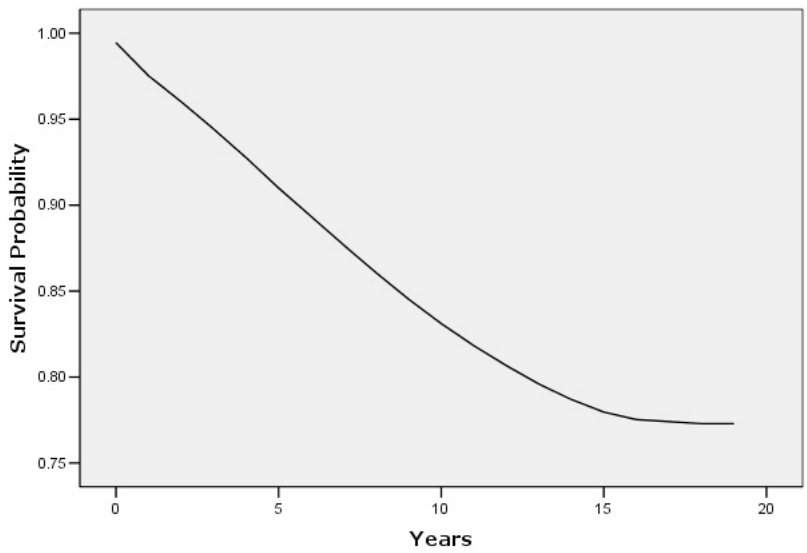

Figure 4.1: Survival function at the mean of explanatory variables.

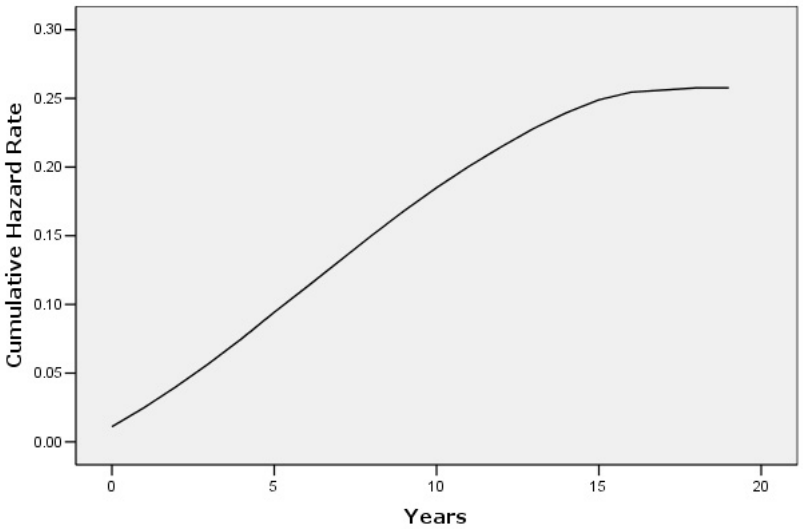

Figure 4.2: Cumulative hazard function at the mean of explanatory variables.

the cumulative hazard function shown in Figure 4.2. The cumulative hazard function is shown to increase in a linear trend for the first 15 years, but then remains somewhat constant after 15 years. This means that assuming a business has survived for 15 years or more, the likelihood it will fail in the near future is extremely low, and the probability of survival for *more than* 15 years is not significantly lower than the probability of survival *for* 15 years. This seems intuitively logical as established businesses have historically been considered much less likely to fail than newly established businesses. In addition, businesses that have survived longer than just 5 years are usually considered to be at least somewhat established. Therefore, it is surprising that it takes 15 years before there is a reduction

in the instantaneous hazard rate (indicated by a reduction in the slope of the cumulative hazard function). That is, the linear increase in the cumulative hazard function over the first 15 years indicates that a business is as likely to fail in the first year as it is in the fourteenth year.

This analysis is an example of how the extra information automatically generated by the Cox model (as a SA technique) aids the understanding of the business failure process.

4.4.2 Empirical Comparison of Techniques

For simplicity, the empirical analysis of the techniques has been subdivided into the following sequence of subsections:

- An analysis of the Cox, DA and LA models where the misclassification costs are equal. This allows the analysis to be undertaken based on the correct classification percentage measure that is easily interpretable.
- A comprehensive analysis of the in-sample classification accuracy of the three techniques over all the different misclassification costs and different length classifications. Due to the varying misclassification costs, the weighted error cost measure was the comparison measure used.
- A comprehensive analysis of the hold-out prediction accuracy, as undertaken for the classification accuracy.
- An analysis of the usefulness of the hybrid models.

All the graphs presented in this section were created from the numerical results presented in Appendix B.

Equal Misclassification Costs

The in-sample classification accuracy of the three models over the 10 different prediction lengths is presented in Figure 4.3. Overall, the performances of the LA, DA and Cox models are very similar with equal misclassification costs. It is also apparent from this graph that the classification ability of all techniques reduces as the prediction interval increases. All three models correctly classify about 96% of businesses according to their possible failure within 1 year; however, this accuracy decreases constantly to approximately 82% accuracy when classifying businesses according to their possible failure within 10 years. Although the decline in classification accuracy occurs mostly in a linear trend, the graph suggests that the accuracy level may flatten out for longer prediction intervals, especially for the DA and LA techniques. Nevertheless, this decline in accuracy is not a large problem, as an 82% correct classification percentage indicates a good model that will be very useful in the real world.

Figure 4.3 also shows that the DA and LA models always have a slightly higher number of correct classifications compared with the Cox model. However, the difference between

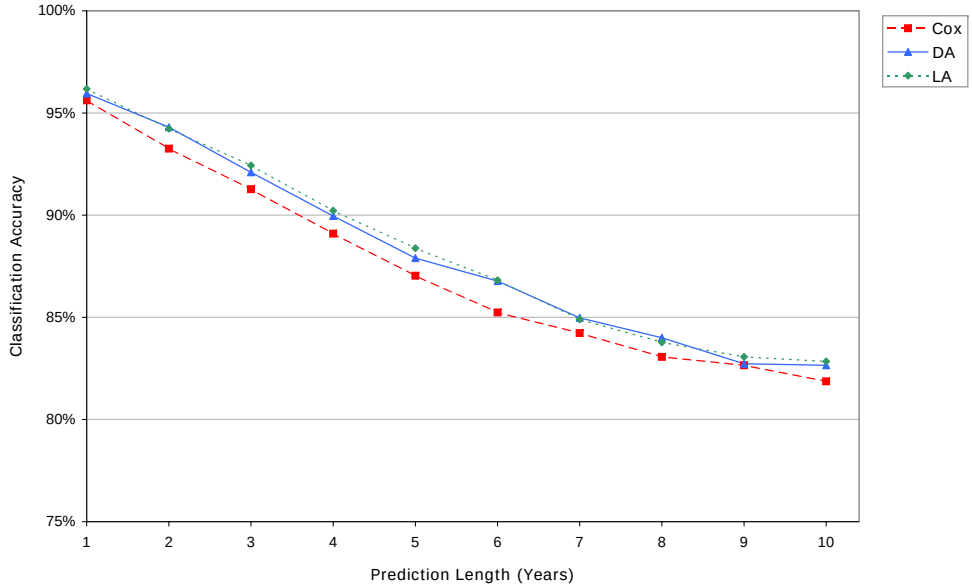

Figure 4.3: The percentage of correct in-sample classifications for the DA, LA and Cox models, compared over prediction lengths of 1 to 10 years.

the correct classification percentages is less than 1% on average, with the largest difference when classifying businesses based on their possible failure within 6 years: the DA model and LA model correctly classified 1.54% and 1.57% more businesses respectively. It is also interesting to note that the relative performance of the Cox model did not improve for longer prediction intervals (up to 10 years). Although the DA and LA lines in Figure 4.3 are very similar, the LA models had a slightly higher correct classification percentage than the DA models for 6 out of the 10 different prediction intervals. The table below shows the number of correct classifications made out of the 2,669 in-sample classifications performed by each technique, averaged over the different prediction intervals.

Cox	DA	LA
2331	2352	2356

These results indicate there is consistency with previous DA and LA research that has found the two techniques to be very similar in terms of classification ability, and the LA models are more often slightly better. Overall, the Cox model is very similar to both DA and LA in terms of classification ability, but the relative order of the techniques according to classification ability with equal misclassification costs is LA, DA and then the Cox model.

The prediction accuracy is even more similar than the classification accuracy when the cost of misclassifications are equal, as shown in Figure 4.4. It is again evident that the accuracy of the models decreases as the prediction interval increases. This result is consistent with the intuitive thought that it is harder to classify and predict events that occur further into the future. From comparing this graph with the previous graph of classification accuracy (Figure 4.3) it can be seen that there has only been a small decline in accuracy when moving from the in-sample to hold-out data. This suggests that all the

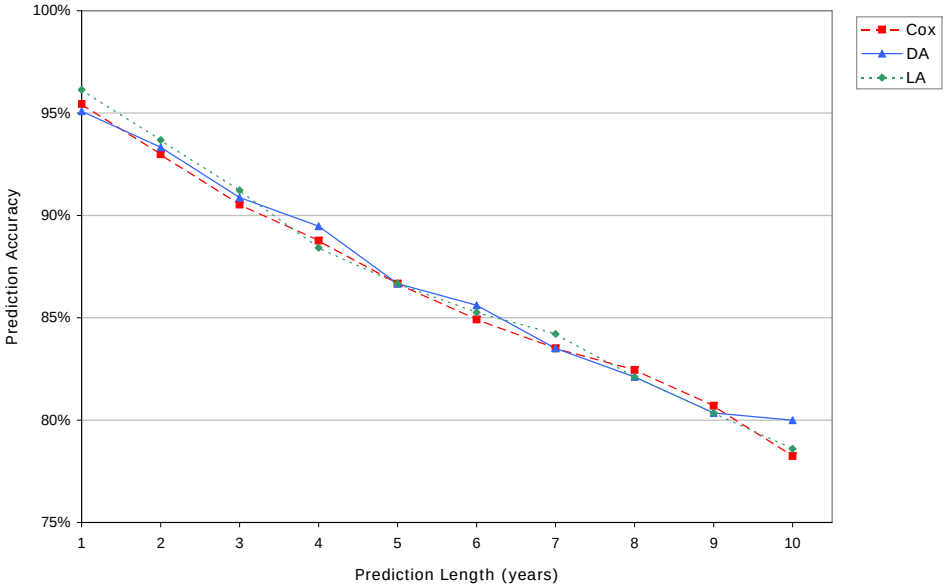

Figure 4.4: The percentage of correct hold-out predictions for the DA, LA and Cox models, compared over prediction lengths of 1 to 10 years.

techniques have successfully generated models predominantly based on general trends in the training data set, and overly complex models have not been created. However, it is observable that the loss in accuracy between classification and prediction is more evident with longer prediction intervals.

When moving from in-sample classification to hold-out prediction, the LA, DA and Cox models respectively lost 1.62%, 1.43% and 0.91% accuracy on average. That is, the techniques declined in accuracy proportional to their in-sample classification accuracy. Hence, the predictive ability of all the three techniques is almost the same, and the best predictive technique depends upon the prediction inverval. The Cox model is the best for 8 and 9 year ahead predictions, but DA is more than 1% more accurate than both the Cox and LA model. Therefore, again there is no evidence to suggest that the Cox model performed better with longer prediction intervals. From 285 predictions made, the LA and DA techniques averaged 247 correct, while the Cox model averaged 246. Thus for equal misclassification costs, the Cox model has produced almost identical prediction accuracy compared with LA and DA models.

In-sample Classification

The in-sample classification accuracy over different misclassification costs was analysed over each prediction length. The comparison of the Cox, DA and LA models for classification based on failure or success in one year is presented in Figure 4.5. The Cox model is clearly the worst performing technique with greater weighted error cost over all misclassification costs. In addition, the Cox model does not adapt as well as the other techniques to increasing Type I Error costs. Therefore, the relative superiority of DA and LA as classifiers

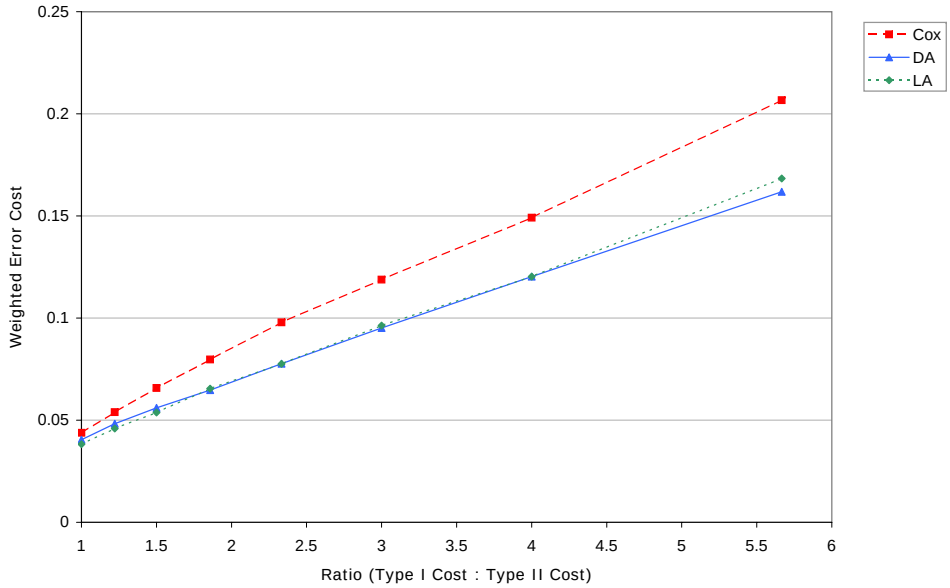

Figure 4.5: The cost of one year ahead in-sample classifications.

increases as the cost of Type I Error increases. The DA and LA models have very similar performance, where LA was marginally better for lower Type I Error costs, but DA was marginally better for the highest Type I Error costs.

The observation that the Cox model is more costly than DA and LA, especially for high Type I Error costs, continues for up to prediction intervals of six years. This is shown in Figure 4.6 and the first two graphs of Figure 4.7. However, the performance gap between the Cox model and DA and LA models decreases as the prediction interval increases from one year to six years. Thus, the relative superiority of DA and LA over the Cox model decreases as the prediction interval increases. In addition, the increase in the performance gap for higher Type I Error costs becomes less significant as the prediction interval increases from one to six years, which means that the Cox models adapts better to higher Type I Error costs as the prediction interval increases.

The classification accuracy of the Cox model still remains slightly worse than the DA and LA models for 7 to 10 years ahead, as shown in the last graph in Figure 4.7 and all of Figure 4.8. However, the error cost of the Cox model is much closer to that of DA and LA models for these longer prediction intervals. This is especially evident for high Type I Error costs, for which the Cox model was previously a significantly worse classifier than DA and LA.

Figures 4.6, 4.7 and 4.8 also reveal that the overall classification accuracy of LA is slightly superior to DA, although they are both very similar. For two year prediction intervals, LA is slightly less costly than DA for higher Type I Error costs. LA and DA are very similar for 3 to 6 year ahead classifications: LA performs slightly better than DA more than 50% of the time for 3 to 5 year ahead classifications, but DA performs better than LA over most misclassification costs for 6 year ahead classifications. DA and LA are again very similar for 7 and 8 year prediction intervals, but the LA model becomes slightly superior with a

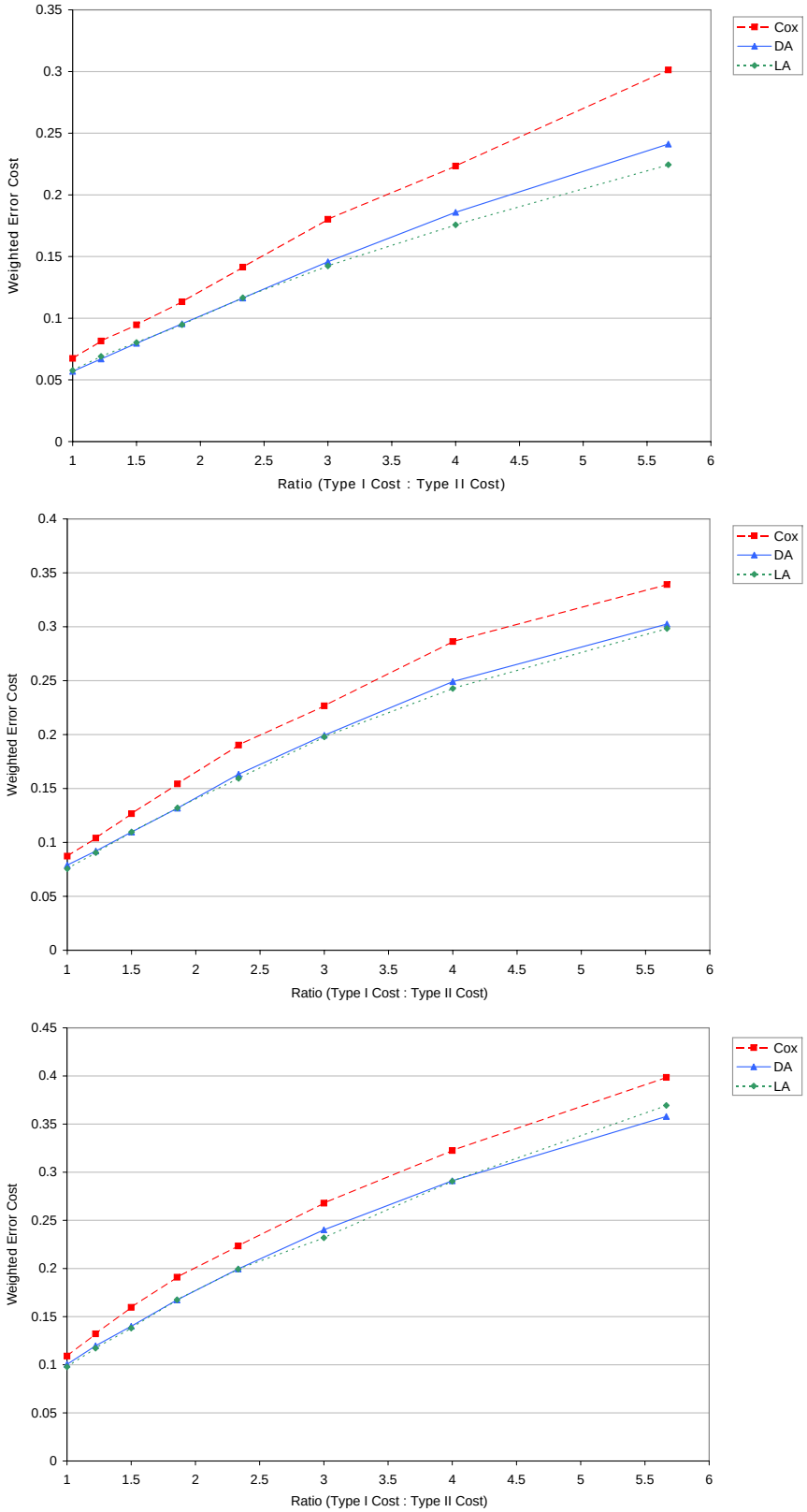

Figure 4.6: The cost of in-sample classifications for prediction intervals of two (top), three (middle), and four (bottom) years.

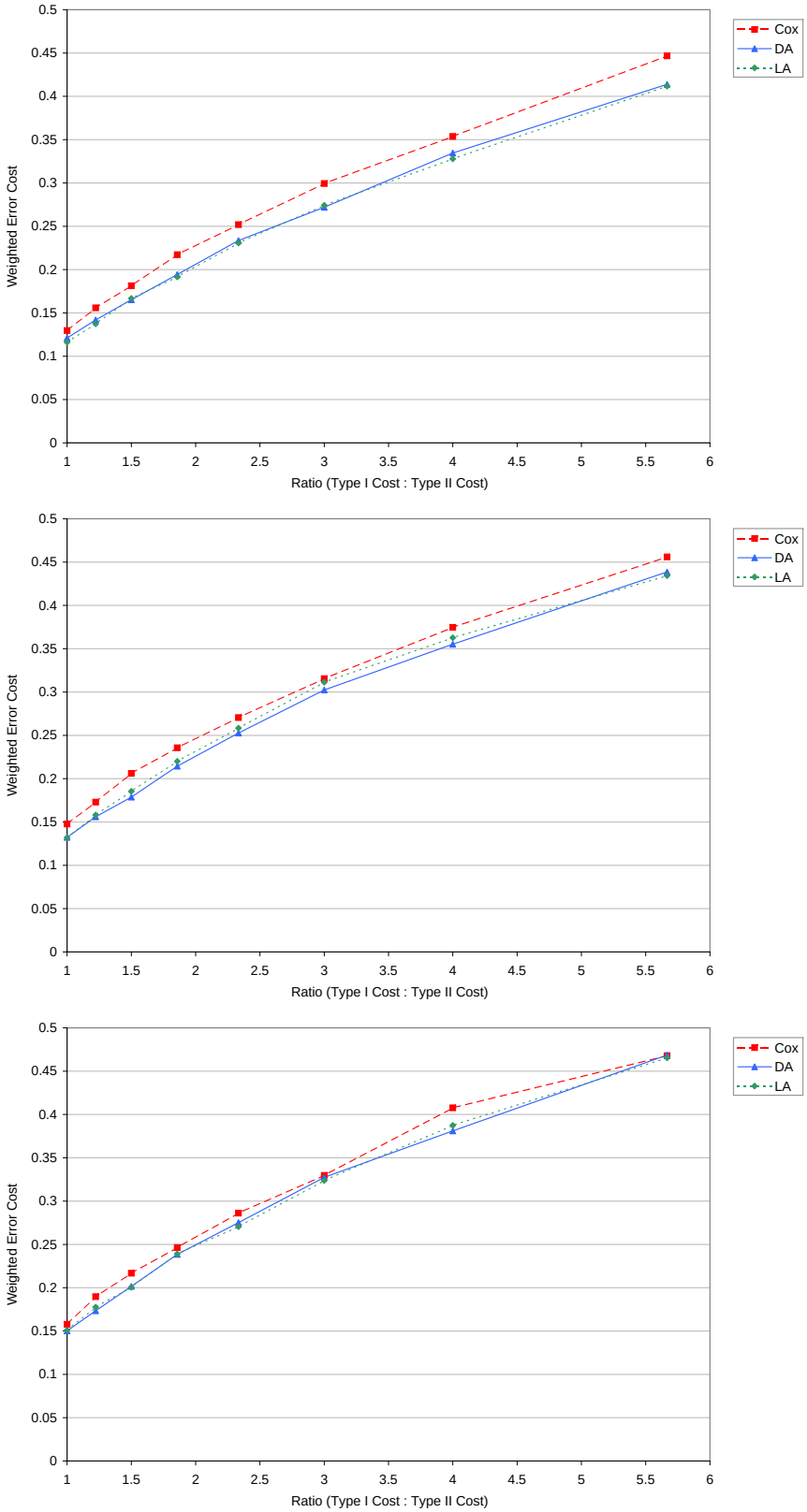

Figure 4.7: The cost of in-sample classifications for prediction intervals of five (top), six (middle) and seven (bottom) years.

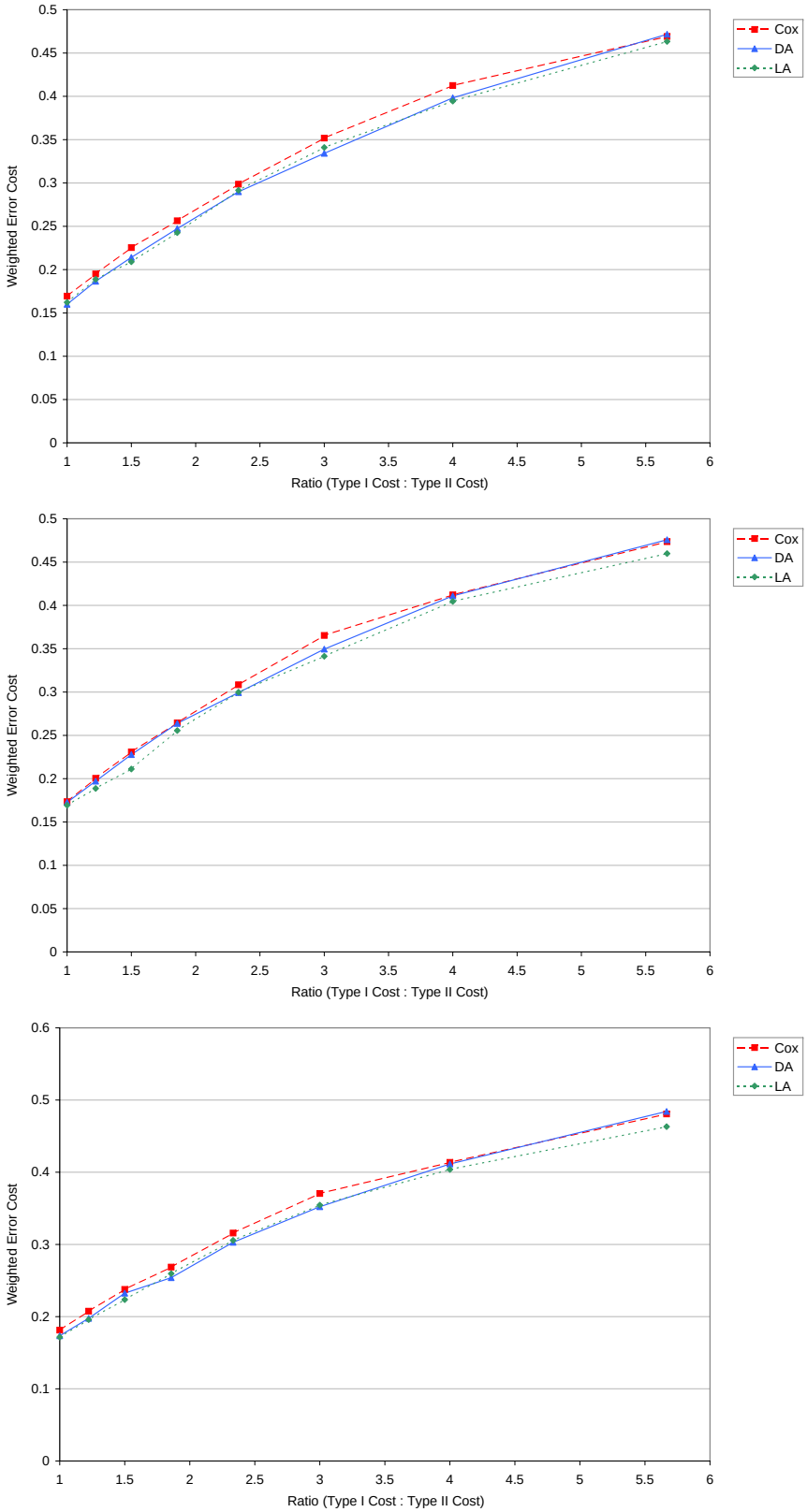

Figure 4.8: The cost of in-sample classifications for prediction intervals of eight (top), nine (middle) and ten (bottom) years.

lower cost as the Type I Error increases for 9 and 10 years ahead.

Overall, while the Cox model has comparable classification power, the DA and LA have better classification ability. In general, it seems that DA and LA adapt better to higher Type I Error costs, especially when classifying business over shorter terms. However, the Cox model does become much more comparable with the more popular techniques of DA and LA when classifying business based on their possible failure within 6 years time or for longer periods (up to 10 years).

Hold-Out Prediction

The prediction accuracy was analysed over each prediction length as performed for the classification accuracy. However, the prediction accuracy is far more important than the classification accuracy as it better approximates how useful the models would be on new data in the real world.

The prediction accuracy is generally more varied than the classification accuracy presented in the previous section. The one-year prediction accuracy, estimated on the hold-out data, for the Cox, DA and LA models is presented in Figure 4.9. Although the Cox model

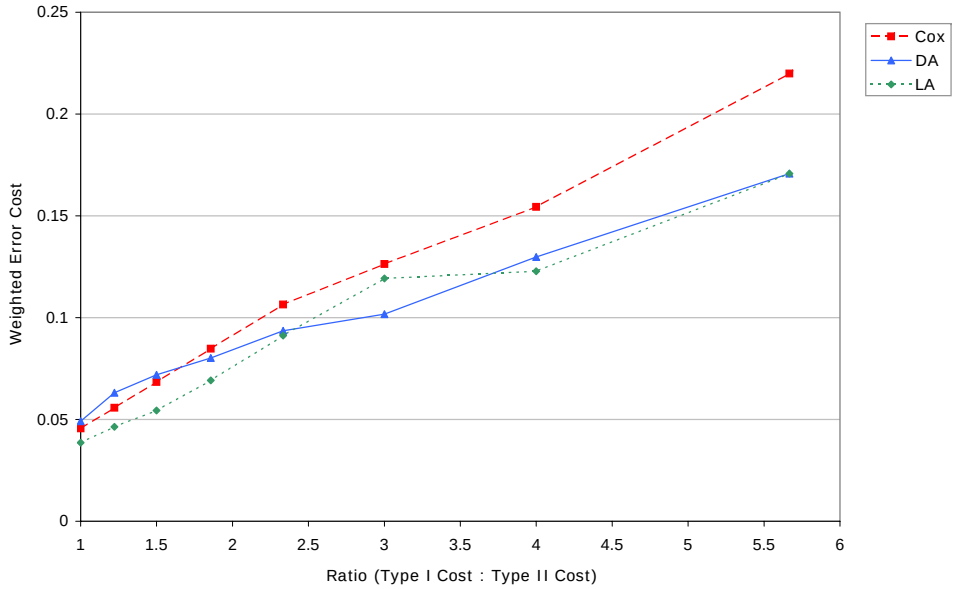

Figure 4.9: The cost of one year ahead hold-out predictions.

had a lower risk compared with the DA model for lower Type I Error costs, the Cox model did not adapt well to larger costs of Type I Error. Therefore, the Cox model became significantly more costly than its DA and LA model alternatives for higher Type I Error costs. The best predictor between LA and DA was LA for all but one misclassification cost where the cost of Type I Error was 3 times as high as Type II Error.

Figures 4.10, 4.11 and 4.12 show the general trend of higher costs for Cox models that became worse with higher Type I Error cost, and continued for the longer prediction intervals. In addition, the predictive power of LA and DA models was similar for lower

Type I Error costs, but the LA model was usually less costly for higher Type I Error costs. However, there are some slight variations in these general trends that are explained below for the specified prediction intervals.

Two Years: (Figure 4.10) Although the cost of the Cox model increases with the cost of Type I Error compared with the LA model, the Cox model actually became slightly better than the DA model for the highest Type I Error cost.

Three and Four Years: (Figure 4.10) The Cox model had the same predictive power as the DA and LA models when the cost of Type I Error was less than or equal to three times the Type II Error. However, for higher Type I Error costs the Cox model was outperformed by the other two models. As usual, the LA model was significantly superior for the highest Type I Error cost, but the DA model had greater prediction accuracy for error cost ratios of 3 and 4, for 3 year prediction intervals.

Five, Six and Seven Years: (Figure 4.11) The Cox model became, relatively, very costly for higher Type I Error Costs, but was comparable to both the LA and DA models for lower Type I Error costs. However, as the prediction interval increases the Cox model became costlier than the LA and DA models for lower Type I Error costs. In addition, while the DA and LA models were again comparable, the LA model is only convincingly superior to the DA model for the six year prediction interval. For five and seven year prediction intervals, the technique with the better prediction accuracy varies for differing misclassification costs, with the DA model being the best technique for the highest Type I Error cost.

Eight Years: (Figure 4.12) The Cox model is only comparable with the DA and LA models for error cost ratios of 1.5 and less, before it becomes more costly. For the eight year prediction interval, the DA model is superior to the LA model, especially for higher Type I Error costs.

Nine and Ten Years: (Figure 4.12) The DA and LA techniques have very similar predictive power, but the LA technique is the best for the highest Type I Error cost. The Cox model followed a similar pattern of becoming the most costly alternative for error cost ratios of above 1.5, but it adapted better to the highest Type I Error costs. The Cox model had better prediction accuracy than LA for the second highest cost of Type I Error for nine year predictions, and DA for the highest cost of Type I Error for ten year predictions.

The technique with the best predictive ability varies between DA and LA depending upon the prediction intervals and misclassification costs. It is not possible to choose an overall 'best technique' as the prediction interval and misclassification costs are factors specific to each particular situation. Nevertheless, if a ranking must be made the LA would be slightly superior to DA in terms of prediction accuracy as it was superior in more situations and, on average, adapted better to rising Type I Error costs.

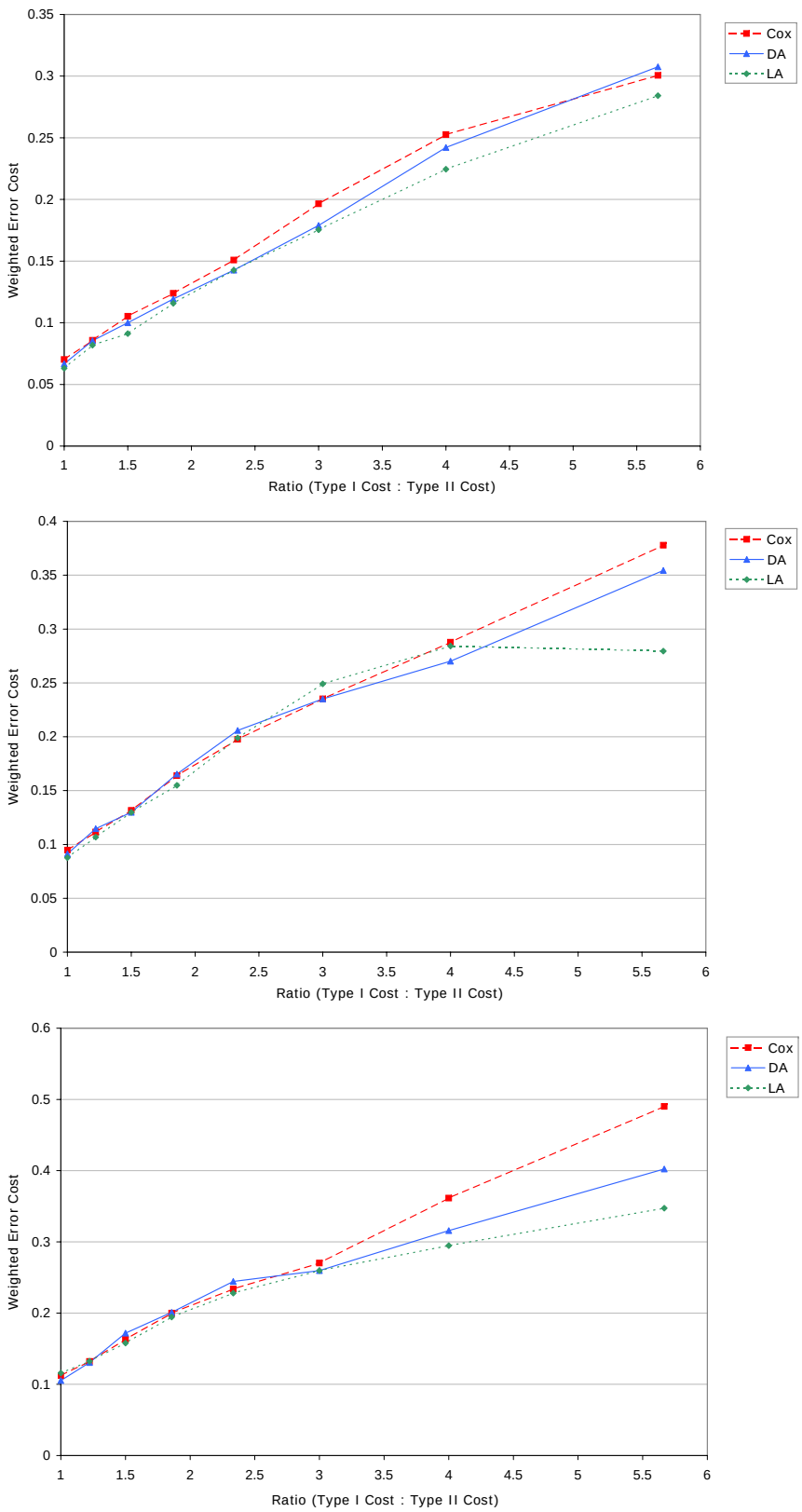

Figure 4.10: The cost of hold-out predictions for prediction intervals of two (top), three (middle), and four (bottom) years.

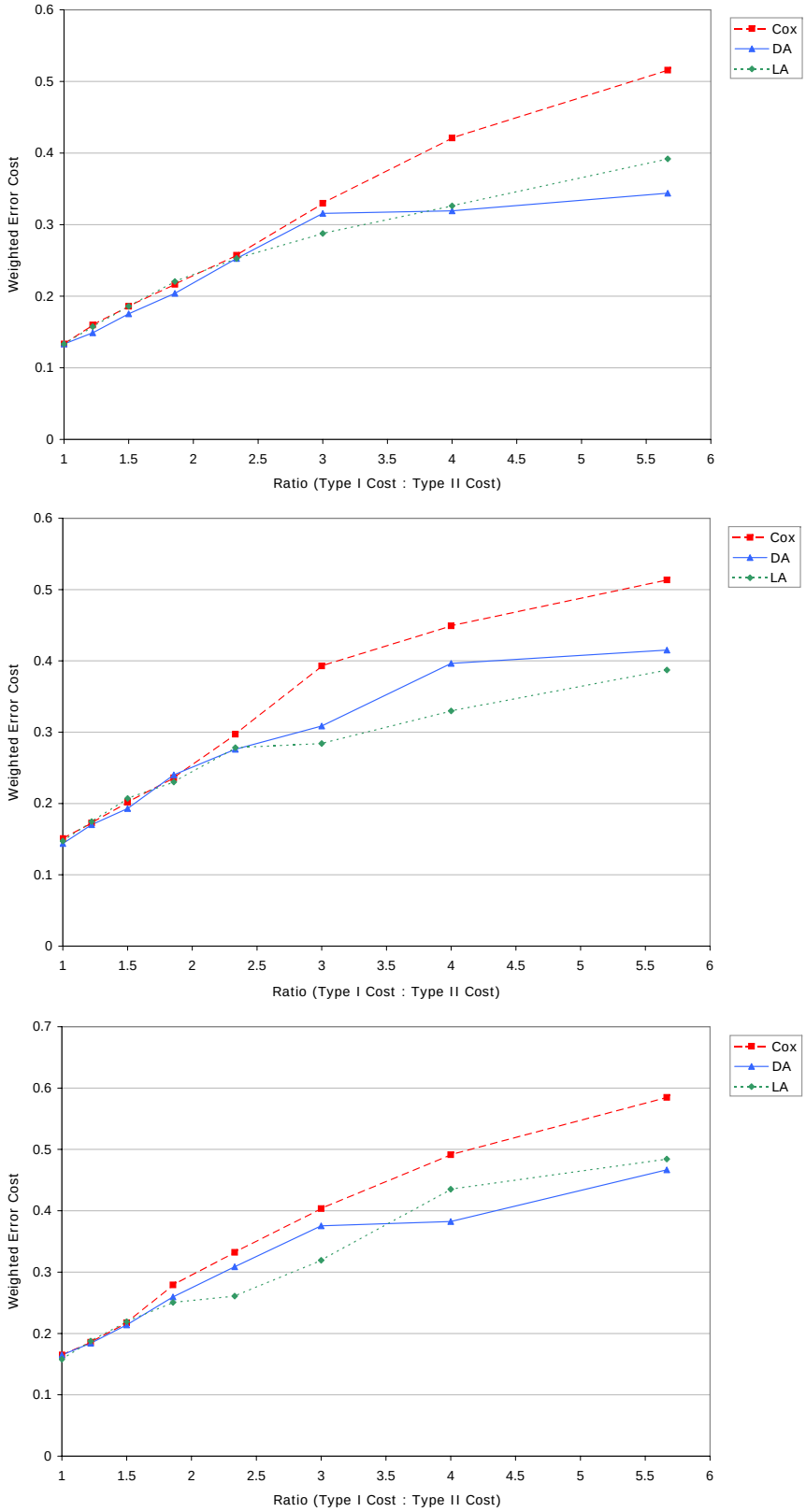

Figure 4.11: The cost of hold-out predictions for prediction intervals of five (top), six (middle), and seven (bottom) years.

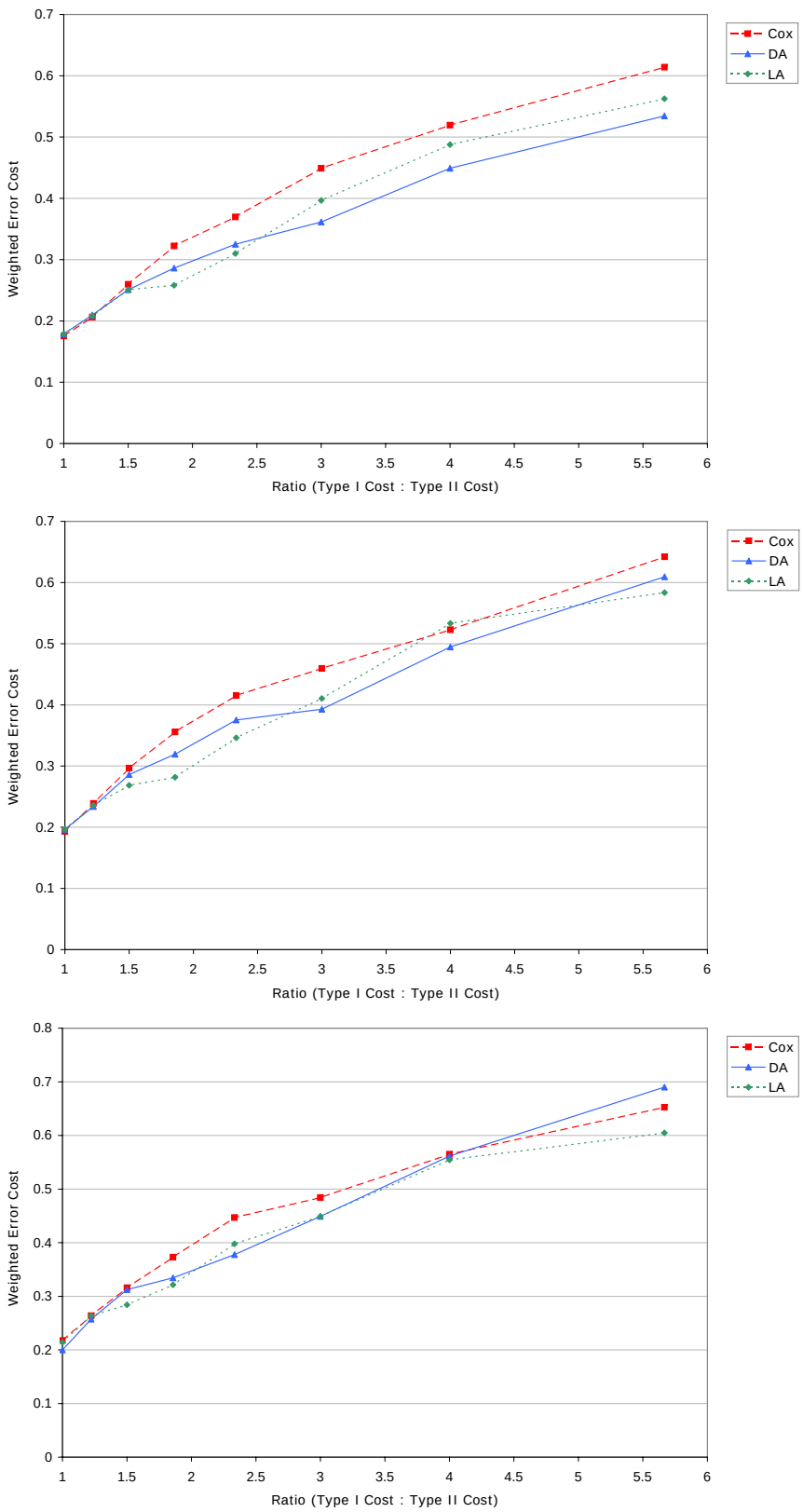

Figure 4.12: The cost of hold-out predictions for prediction intervals of eight (top), nine (middle), and ten (bottom) years.

Although the Cox model was comparable in many situations, the predictive power of the Cox model was slightly worse than DA and LA, especially for high costs of Type I Error. Furthermore, the relative performance of the Cox model as the prediction level increased was dependent upon the misclassification costs. For lower Type I Error costs there was no improvement, or even a decline, in the prediction accuracy of the Cox model as the prediction interval increased. However with higher Type I Error costs, the prediction accuracy of the Cox model improved for the longer prediction intervals.

Hybrid Models

In most cases, the LA stepwise procedures did not even include the survival probabilities as an explanatory variable in the LA model. The survival probability was only included by the LA stepwise procedure for the 10 year ahead predictions. Furthermore, the hybrid model that was developed for 10-year predictions did not adapt as well to higher costs of Type I Error, as it produced more Type I Error than the simple LA model without a corresponding reduction in Type II Error for both classification and prediction.

The survival probabilities were included in all the DA models, but the results were similar to the 10 year LA hybrid model. For nearly all cases, the weighted error cost measure was higher for the hybrid model than the simple DA model. This reason for this was that the Type I Error had increased in the hybrid model, particularly for higher Type I Error costs and longer prediction intervals.

Hence, this method of producing hybrid models did not improve the ability to classify or predict business failure, and consequently is not appropriate for classifying or predicting business failure. This does not mean however that there are not other ways of developing hybrid models that could be appropriate for business failure.

4.4.3 Analysis of the Importance of Variables

The simplest analysis of variable importance is obtainable from the Cox model, as it was the only one model capable of being used for all the different prediction intervals. The importance of the variables is even easier to analyse as the variable coefficients do not change, only the baseline hazards function changed over time. The details of the variables included in the Cox model are presented in Table 4.2, where the Exp(Δ) column represents the estimated percentage change in the hazard of failure for a 0.1 increase in a given explanatory variable. Therefore, this value can be used to determine the importance of each variable, where a higher (absolute) value indicates that the variable has a bigger impact on the failure of a business. Hence, it is clear that the financial leverage ratio of long-term liabilities to total assets (LTL/TA) is the most important variable, whereby for each 0.1 increase in the ratio the hazard of failure will increase by 42.95%. It is also observable that a range of variable types are shown to be important in the Cox model, as is the case for most BFP models.

The DA models generated for the different prediction intervals were very different in

Variable	Variable Type	Coefficient	Significance	Exp(Δ)
LTL/TA	Financial Leverage	1.667	0.00	42.95%
OPI/TA	Profitability	-6.855	0.00	-9.99%
RCV/CA	Management Efficiency	-1.482	0.00	-7.73%
OPI/FA	Profitability	0.466	0.00	5.93%
CA/TA	Liquidity	-0.718	0.00	-5.12%
Ln(SLS)	Business Size	-0.586	0.00	-4.43%
RE/TA	Profitability	-0.492	0.00	-3.88%
QA/CL	Liquidity	-0.349	0.00	-2.95%
Ln(EMPL)	Business Size	0.202	0.00	2.23%
MVE/TL	Market Structure	-0.100	0.00	-0.95%

Table 4.2: Explanatory variables included in the Cox model, ordered by importance from top (most) to bottom (least). Note that all the variables are significant at the 1% significance level.

terms of the specific variables included, the number of variables included and the importance of specific variables. The NI/TA (profitability) variable was the most important variable for one, two and three year prediction intervals, but it was unimportant or not even included in the model, for prediction intervals of greater than four years. The variable with the most consistent importance was the business size variable Ln(SLS) that was relatively important in all DA models and the most important variable for prediction intervals of 4,5 and 7-10. The only other common trend was that the variables found to be important in the Cox model were not important in the DA models. For example, the LTL/TA variable was not included in any DA model.

Similar to DA the number of variables included in LA models varies, but there is a trend for more variables to be included for the longer prediction intervals. However, in contrast to the DA models, the LA models found similar results to the Cox model in terms of variable importance. In fact, the LA model for eight, nine and ten year prediction intervals includes all the variables included in the Cox model with the addition of the profitability variable NI/FA. The LTL/TA variable was found to be the most important in these three LA models, but the relative importance of the remaining variables varies with the prediction interval. Although the LTL/TA variable is not included in LA models for prediction intervals from one to seven years, the most important variable in these models is TL/TA. This makes logical sense as LTL/TA is simply a long-term focused version of TL/TA. In addition, the only variable to be included in all of the LA models with a high degree of importance was the OPI/TA variable, which was found to be the second most important variable in the Cox model.

Overall, it can be seen that the Cox and LA models found similar variables to be important in their models. On the other hand, the variables found to be important by DA models varied with the prediction interval; but nonetheless, they were mostly different from

those found as important by the Cox and LA models. Overall, it seems that all techniques use a variety of variable types to predict business failure. Finally, this analysis has shown that in the majority of models a financial leverage ratio of liabilities to assets is a very good indicator of business failure, where the higher the financial leverage of a business the higher the likelihood of the business failing in the future.

4.5 Conclusions

The Cox model was able to classify and predict business failure as well as both DA and LA models for equal misclassification costs. Moreover, the prediction accuracy of the Cox model was very comparable with DA and LA for unequal misclassification costs, when Type I Error was no more than 1.5 times greater than Type II Error. Although the Cox model was comparable with DA and LA in many other situations, it can be stated that DA and LA are better predictors of business failure than the Cox model, especially for high costs of Type I Error. However, the ability to adapt to high costs of Type I Error improved for the Cox model, relative to DA and LA, as the prediction interval increased. It should also be noted that the DA and LA models had very similar predictive ability as has been found in many previous studies.

Although the prediction accuracy of the Cox model is not as good as the alternative techniques, it provides more information about the business failure process through the interpretation of the hazard and survival function over time. In addition, for the situation of different prediction intervals, the Cox model is much easier to interpret and assess the importance of variables as only one model is needed to cater for different prediction intervals.

Chapter 5

Comparative Research and Overall Conclusions

5.1 Comparative Research

This comparative study aims to compare a Decision Tree (DT) technique with a Survival Analysis (SA) technique to help address the lack of comparisons between the DT and SA techniques identified in Section 2.4. Specifically, the CART and Cox models were compared based on their ability to predict business failure, which has not previously been done. The See5 technique was not included in this analysis, because only the demonstration version of the See5 package was available for this research, which was unable to process the large data set used in this comparative study.

5.1.1 Data and Methodology

The same Cox model, developed and tested on the data sets in the previous chapter, was used for this comparison. The CART model was then developed and tested on the same in-sample and hold-out data sets as the Cox model (outlined in Chapter 4). This time-series data set (from the SA research in Chapter 4) was used as it allows for comparisons over different prediction lengths as well as various misclassification costs. To gain an understanding of the comparative adaptability of the CART and Cox models to different prediction intervals, 1, 2, 3, 5 and 10 year intervals were analysed. In addition, the following set of ratios of Type I to Type II Error cost were used: 1, 1.5, 2, 2.5, 3, 4, 5. Thus, as the study included analysis over different misclassification costs the two techniques were compared based on the weighted error cost measure used previously in Chapter 4 (outlined in Section 4.3.1). These misclassification costs were given as inputs to the CART model, and were implemented for the Cox model by using the appropriate cut-off value. The corresponding cut-off values shown in the table below, were calculated as described in Section 4.3.1.

Error Cost Ratio	1	1.5	2	2.5	3	4	5
Cut-off Value	0.5	0.6	0.67	0.71	0.75	0.8	0.83

Separate CART models were developed for each prediction interval as was undertaken for the DA and LA models when comparing them to the Cox model. Each CART model was developed using the same Entropy method as used in Chapter 3, and prior probabilities were set to the proportions of successful and failed businesses in the data set. In addition, as simpler DTs were found to be better predictors the smallest error cost tree was chosen as the final CART model (analogous to the CART_s model in Chapter 3).

5.1.2 Results

The numerical data used to generate the graphs presented in this section is presented in a tabular format in Appendix C. As shown in Figures 5.1 and 5.2 the in-sample classification of the CART model is almost always superior to the Cox model. For all the prediction

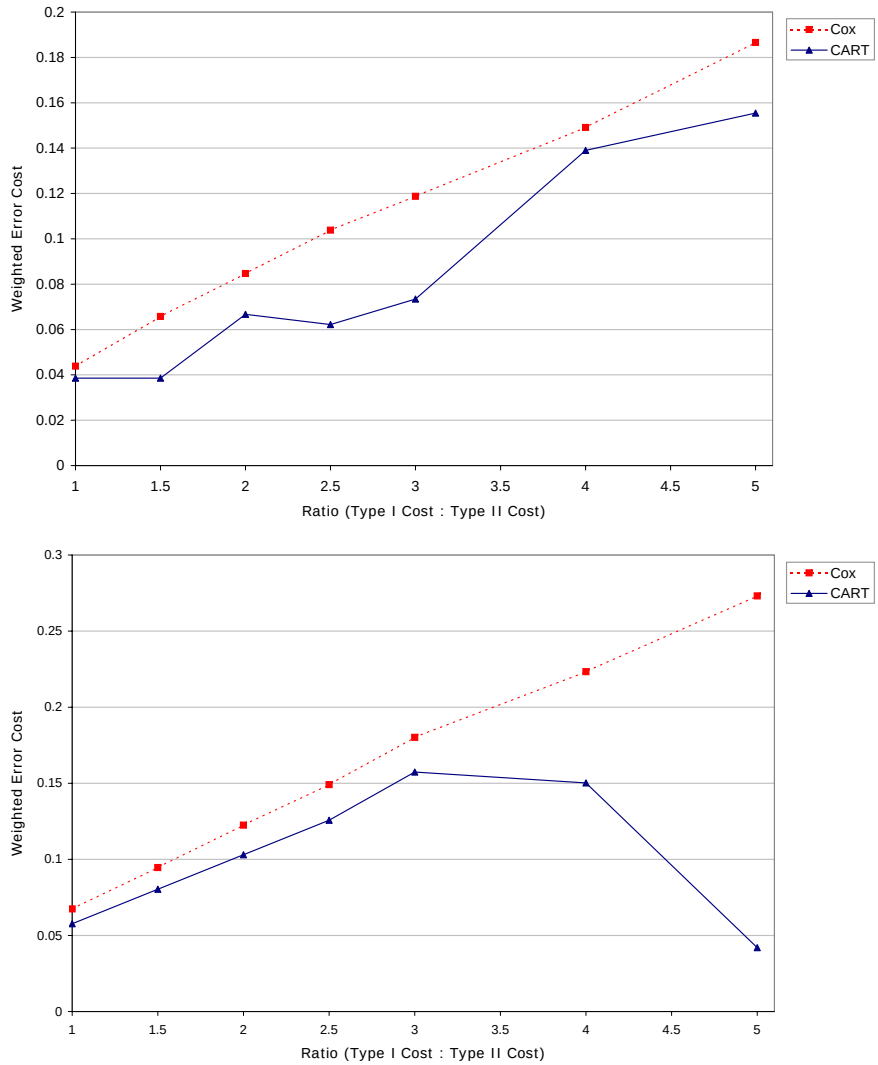

Figure 5.1: The cost of in-sample classifications for prediction intervals of one (top) and two (bottom) years.

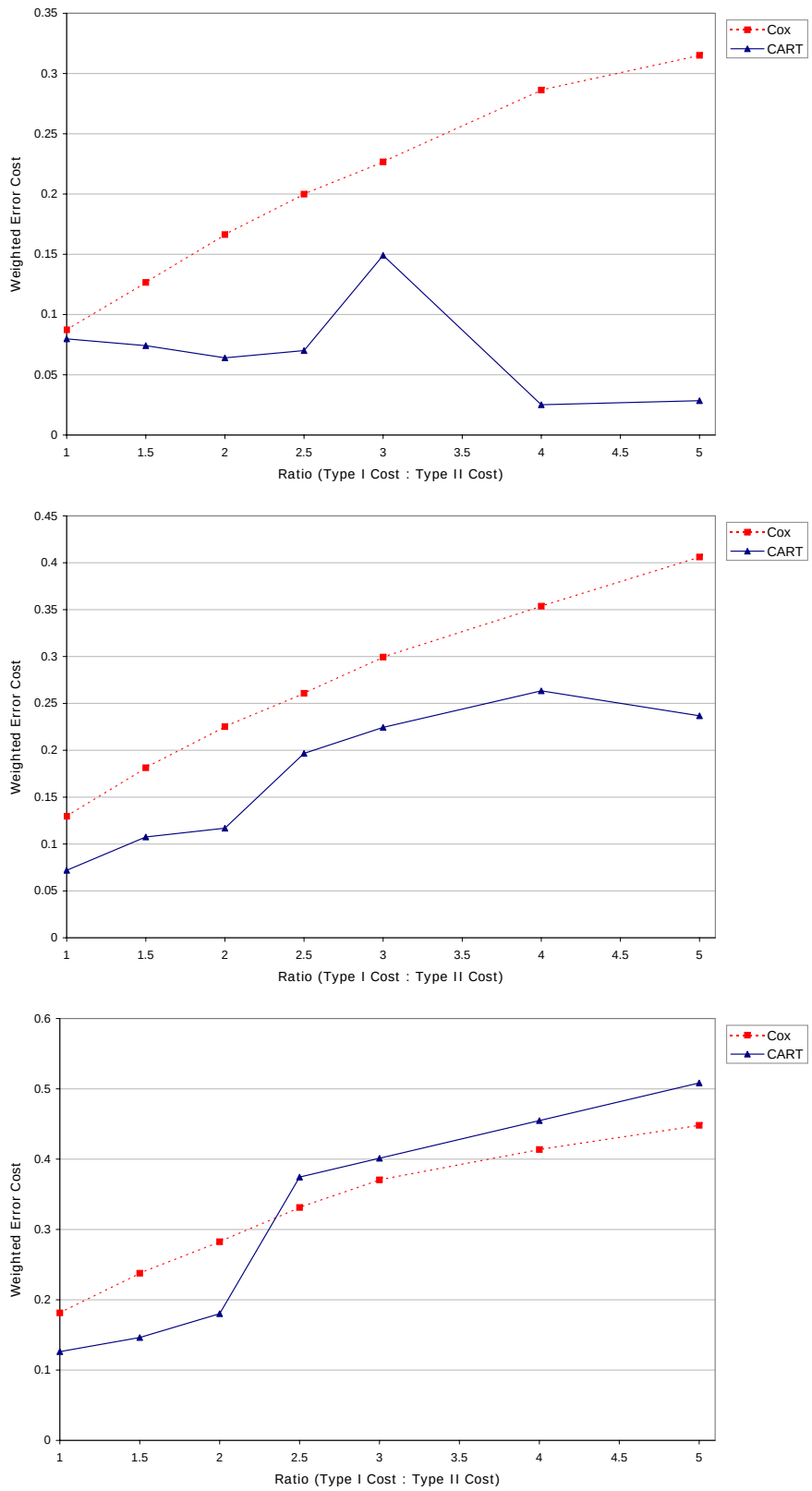

Figure 5.2: The cost of in-sample classifications for prediction intervals of three (top), five (middle) and ten (bottom) years.

intervals other than ten years, the Cox model is costlier than the CART model for all misclassification costs. In addition, the cost of the CART model becomes significantly less than the Cox model for higher Type I Error costs and 2, 3 and 5 year prediction intervals. The exception to the superiority of the CART model occurs for the longest prediction interval of 10 years (shown in Figure 5.2), where the classification accuracy of the Cox model becomes superior when the ratio of the error costs is greater than or equal to 2.5. Therefore, the Cox model is shown to have improved classification for longer prediction intervals relative to the CART model, as was also found relative to both the DA and LA models in Chapter 4.

The prediction accuracy of the Cox and CART models are compared in Figures 5.3 and 5.4. The CART model has a lower cost of predictions for all misclassification costs over all prediction intervals. In addition, the superiority of the CART model over the Cox

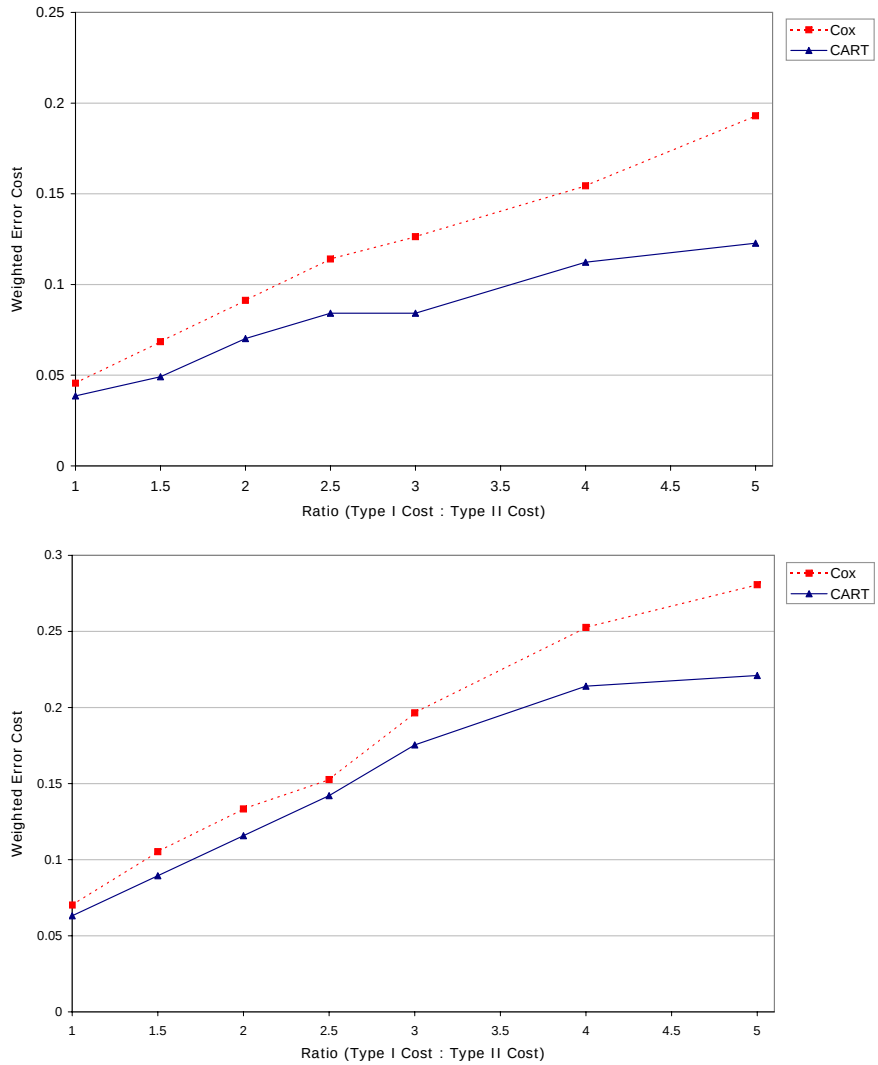

Figure 5.3: The cost of hold-out predictions for prediction intervals of one (top) and two (bottom) years.

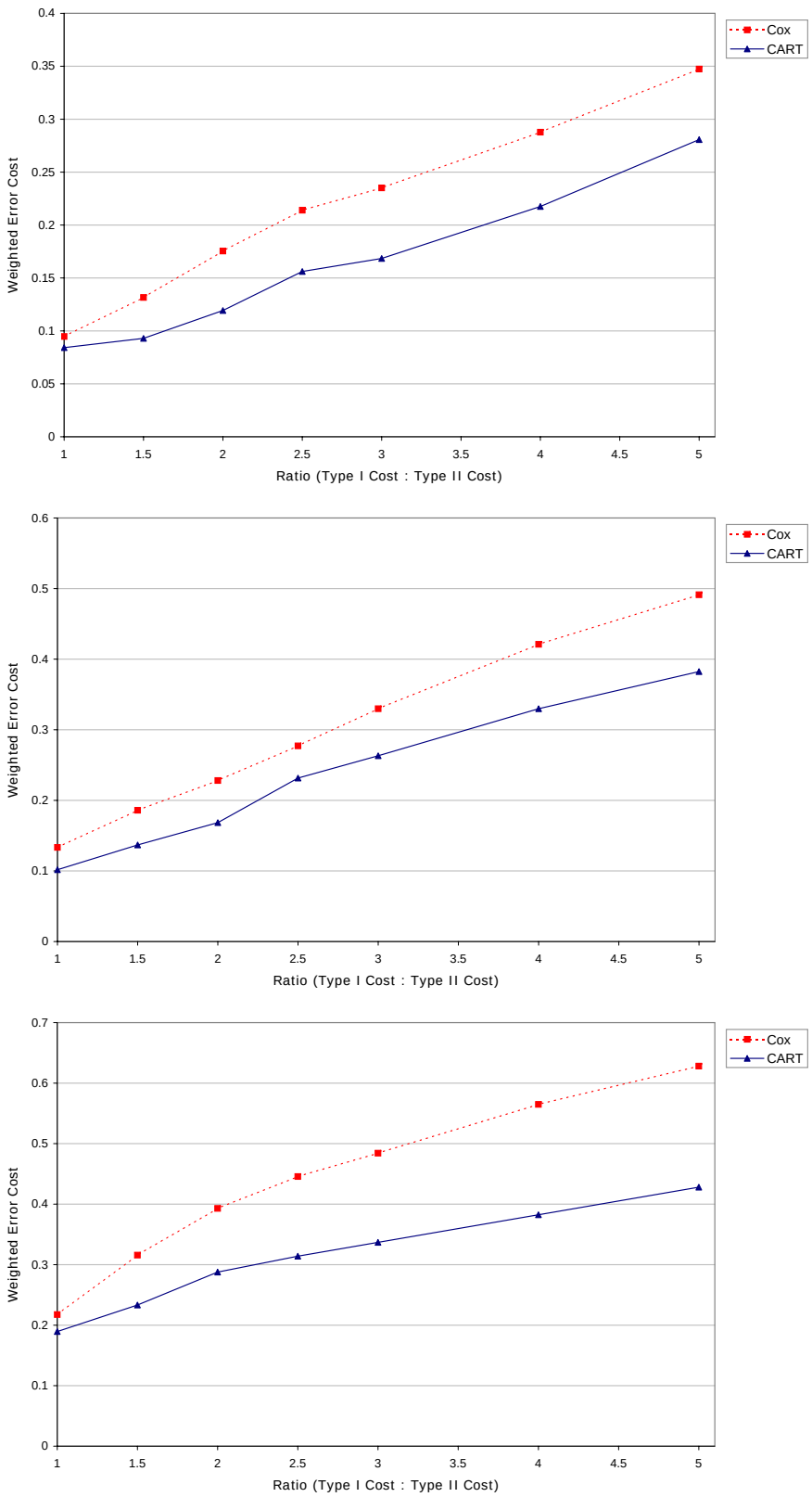

Figure 5.4: The cost of hold-out predictions for prediction intervals of three (top), five (middle) and ten (bottom) years.

model increases for higher costs of Type I Error and longer prediction intervals.

In general, the CART decision tree technique is a better classifier, and more importantly a better predictor, of business failure than the Cox survival analysis technique.

5.2 Overall Conclusions

This research evaluated the decision tree and survival analysis techniques for their application in business failure prediction. It is also believed to be the most extensive research that has been undertaken on these two approaches in their application to predicting business failure. No conclusive result has been found in terms of one approach being superior to all other models in all aspects. Rather, decision trees and survival analysis both possess distinct advantages and disadvantages.

Decision Trees

Decision trees have been shown to have superior classification and prediction accuracy compared with the more popular discriminant analysis technique, over most misclassification costs. Decision trees, specifically the CART model, also had better classification accuracy than the Cox survival analysis technique, over most misclassification costs and prediction intervals. More importantly, the CART decision tree technique was always superior to the Cox survival analysis technique in terms of prediction accuracy, over various misclassification costs and prediction intervals.

Within the general decision tree approach, no single tree-building technique was found to be superior for all misclassification costs. The See5 technique had the best ability to classify businesses, yet in most cases it also had the worst prediction ability. The CART and RPA techniques both had very similar predictive power over most misclassification costs, and together were the best predictors of business failure. Specifically, the RPA technique was slightly better at predicting when the cost of Type I Error was relatively high compared with Type II Error.

Overall, the decision tree techniques had the best empirical accuracy of any technique tested, but they have the disadvantage of not providing as much information. Unlike the more popular techniques of discriminant and logit analyses, decision trees can not generate the probability of failure nor compare two businesses predicted to be in the same classification group. However, this additional information is often considered of secondary importance compared with the accuracy of the technique. Furthermore, the fact that decision tree techniques are non-parametric is also a major advantage. This means that, unlike the more popular techniques of discriminant and logit analyses, decision tree techniques are not based upon distribution assumptions which are usually violated with real-world data. Thus, decision trees can provide accurate, and statistically valid, predictions of business failure.

Survival Analysis

The Cox model, the most popular survival analysis technique, was outperformed by the CART decision tree technique in terms of classification and prediction ability, for almost all misclassification costs and prediction intervals. Furthermore, compared with the Cox survival analysis technique, both discriminant analysis and logit analysis were slightly superior classifiers and predictors of business failure, for the majority of misclassification costs and prediction intervals. However, in the case of equal (or near equal) misclassification costs, the Cox survival analysis technique had the same classification and prediction accuracy as discriminant and logit analyses, even for different time-length predictions.

Although the Cox survival analysis technique has accuracy comparable with the more popular discriminant and logit analyses, the empirical accuracy of the technique is a relative weakness. However, survival analysis techniques are still useful as they provide more information and are more versatile. In addition to the probabilities of failure also obtained with the more popular techniques, further information about the business failure process can be obtained by analysing (over time) the survival and hazard function output by the Cox survival analysis technique. The Cox model, as a time-series survival analysis technique, is also more versatile in the sense that one model is capable of various time-length predictions unlike the other cross-sectional models that are limited to one prediction length per model. Hence, the Cox survival analysis technique can provide a reasonably accurate model, which is very useful for studying the failure process and efficiently making different time-length predictions.

Summary

The survival analysis and decision tree techniques investigated here are both useful in business failure prediction for different purposes. Survival analysis techniques are appropriate for making predictions of varying lengths and analysing the business failure process over time. On the other hand, decision trees are the best techniques for making very accurate, and statistically valid, predictions about the failure or success of businesses.

5.3 Further Research

There is much research that could be undertaken in the future within the field of business failure prediction. The recommendations for further research presented in this section will be limited to survival analysis and decision trees, and will primarily be extensions to the work presented in this thesis.

The research presented in this thesis on decision trees (Chapter 3), survival analysis (Chapter 4) and a comparison of these techniques (Section 5.1 earlier in this chapter) could all be replicated on new and different sets. For example, data from Australian businesses could be used to determine whether the same results are obtained for Australian businesses as US businesses. Addition explanatory variables could also be included that indicate the

state of the industry and economy in which each business resides, and again the results could be compared with those presented in this thesis.

Additional techniques from within the decision tree and survival analysis approaches could also be studied. The decision tree research presented in Chapter 3 could be extended to include the Multivariate Adaptive Regression Splines (MARS) technique, which has not yet been extensively evaluated for its usefulness in business failure prediction. The survival analysis research presented in Chapter 4 could be extended to include survival analysis techniques other than the Cox model. For example, an accelerated failure time model or survival analysis model estimated by an artificial neural network could be added, both of which have had little application to business failure prediction. In addition, time-dependent explanatory variables could be added to the survival analysis model to observe whether they improve the empirical performance of the models.

The comparative study presented in Section 5.1 could also be extended to include the additional techniques mentioned in the previous paragraph. In addition, the See5 technique could be added by obtaining the full version of the See5 package and the RPA technique could also be included. Although the Cox model and discriminant and logit analyses hybrid models were not useful for predicting business failure, hybrid decision tree and survival analysis models may be more successful. This could be done, for example, by including the probability of survival as an explanatory variable in a decision tree's data set. Other such means of creating hybrid models could also be explored.

A potentially interesting extension to the idea of predicting business failure has been previously suggested by Frydman et al. (1985). That is, future research could be undertaken on predicting whether already failed businesses will be successfully reorganised into successful new businesses. It is thought that research into this area may further the understanding of the business failure process.

Bibliography

Aalen, O. O. (1976). Nonparametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics*, 3:15–27.

Abdel-Khalik, A. R. and El-Sheshai, K. M. (1980). Information choice and utilization in an experiment on default prediction. *Journal of Accounting Research*, 18(2).

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4):589–609.

Altman, E. I. (1980). Commercial bank lending: Process, credit scoring, and costs of errors in lending. *Journal of Financial and Quantitative Analysis*, 15:813–832.

Altman, E. I., Heldeman, R. G., and Marayanan, P. (1977). ZETA analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, 1:29–54.

Altman, E. I. and Saunders, A. (1988). Credit risk measurement: Developments over the last 20 years. *Journal of Banking and Finance*, 21:1721–1742.

Anandarajan, M., Lee, P., and Anandarajan, A. (2001). Bankruptcy prediction of financially stressed firms: an examination of the predictive accuracy of artificial neural networks. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 10(2):69–81.

Australian Government, Insolvency and Trustee Service Australia (ITSA) (Last Updated July 2005). Quarterly statistics. Available Online. URL of ITSA Homepage: http://www.itsa.gov.au/.

Aziz, M. A. and Dar, H. A. (2005). Predicting corporate bankruptcy: Where we stand? *Corporate Governance*, 5 forthcoming.

Balcaen, S. and Ooghe, H. (2004). Alternative methodologies in studies on business failure: do they produce better results than the classic statistical methods? Technical Report 04/249, Faculty of Economics and Business Administration, Ghent University.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research (Supplement)*, 4(3):71–111.

Bornholdt, S. (1998). Genetic algorithms. In Gramß, T., Bornholdt, S., Groß, M., Mitchell, M., and Pellizzari, T., editors, *Non-Standard Computation*, pages 59–94. WILEY-VCH, Weinheim, Germany.

Breiman, L. (1981). Automatic identification of chemical spectra. Technical report, Technology Service Corporation, Santa Monica, CA.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks.

Casey, C. J. (1980). Variation in accounting information load: The effect on loan officers' predictions of bankruptcy. *The Accounting Review*, LV(1).

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society A*, 158:419–466.

Coats, P. K. and Fant, L. F. (1992). A neural network approach to forecasting financial distress. *Journal of Business Forecasting*, 10(4):9–12.

Coleman, K. G., Graettinger, T. J., and Lawrence, W. F. (1991). Neural networks for bankruptcy prediction: the power to solve financial problems. *Artificial Intelligence Review*, pages 48–50.

Collins, R. A. and Green, R. D. (1982). Statistical methods for bankruptcy forecasting. *Journal of Economics and Business*, 32:349–354.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34:187–220.

Crapp, H. R. and Stevenson, M. (1987). Development of a method to assess the relevant variables and the probability of financial distress. *Australian Journal of Management*, 12(2):221–236.

Cybinski, P. J. (2003). *Doomed Firms: An econometric analysis of the path to failure*. Prentice Hall, New Jersey, second edition.

Deakin, E. (1972). A discriminant analysis of predictors of business failure. *A Journal Accounting Research*, pages 167–179.

Dimitras, A. I., Slowinski, R., Susmaga, R., and Zopounidis, C. (1999). Business failure prediction using rough sets. *European Journal of Operational Research*, 114:263–280.

Edminster, R. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis (2)*, pages 1477–1493.

Elliot, J. A. and Kennedy, D. B. (1988). Estimation and prediction of categorical models in accounting research. *Journal of Accounting Literature*, 7:202–242.

Fanning, K. M. and Cogger, K. O. (1994). A comparative analysis of artificial neural networks using financial distress prediction. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 3:241–252.

Fernández, E. and Olmeda, I. (1995). Bankruptcy prediction with artificial neural networks. In *International Workshop on Artificial Neural Networks*, pages 1142–1146.

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386.

Fletcher, D. and Goss, E. (1993). Forecasting with neural networks: an application using bankruptcy data. *Journal of Information and Management*, 24(3):159–167.

Flury, B. (1997). *A first course in multivariate statistics*. Stringer Texts in Statistics. Springer.

Frydman, H., Altman, E. I., and Kao, D.-L. (1985). Introduction recursive partitioning for financial classification: the case of financial distress. *Journal of Finance*, 40(1):269–291.

Gloubos, G. and Grammatikos, G. (1988). Predicting bankruptcy of industrial firms in Greece. *Studies in Banking and Finance*, 7:37–46.

Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rouan, G. W., Weisberg, M. C., Acampora, D., Stasiulewicz, C., Walshon, J., Terranova, G., Gottlieb, L., Kobernick, M., Goldstein-Wayne, B., Copen, D., Daley, K., Brandt, A. A., Jones, D., Mellors, J., and Jakubowski, R. (1982). A computer protocol to predict myocardial information in emergency department patients with chest pain. *The New England Journal of Medicine*, 307:588–597.

Goldman, L., Weinberg, M., Olshen, R., Cook, E. F., Sargent, R. K., Lamas, G. A., C, C. D., Wilson, C., Deckelbaum, L., Fineberg, H., and Stiratelli, R. (1988). A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *The New England Journal of Medicine*, 318:797–803.

Hamer, M. (1983). Failure prediction: sensitivity of classification accuracy to alternative statistical methods and variable sets. *Journal of Accounting and Policy*, 2:289–307.

Hans-Peter Blossfeld, A. H. and Mayer, K. U. (1989). *Event History Analysis: Statistical Theory and Application in the Social Sciences*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Harrell, F. E. (2001). *Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis*, chapter 16. Springer Series in Statistics. Springer, New York.

Healy, J. D. (1987). A note of multivariate CUSUM procedures. *Technometrics*, 29(4):409–412.

Huarng, K., Yu, H.-K., and Chen, C.-J. (2005). The application of decision trees to forecast financial distressed companies. In *2005 International Conference on Intelligent Technologies and Applied Statistics Proceeding*, Taipei, Taiwan.

Johnson, R. A. and Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis*, chapter 8–9,11–12. Prentice Hall, New Jersey, second edition. There is a newer edition that covers the same material with fewer errors.

Jones, F. L. (1987). Current techniques in bankruptcy prediction. *Journal of Accounting Literature*, 6:131–164.

Joos, P., Vanhoof, K., Ooghe, H., and Sierens, N. (1998). 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*, pages 59–72, Chemnitz, Germany.

Kahya, E. and Theodossiou, P. T. (1999). Predicting corporate financial distress: A time-series CUSUM methodology. *Review of Quantitative Finance and Accounting*, 13(4):323–345. 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).

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kaliaperumal, V. G. (2005). Survival analysis -an overview. In Sahai, A., Kaliaperumal, V. G., and Venkatesan, P., editors, *22nd Annual National Conference of Indian Society for medical statistics: Pre Conference Workshop/Seminar on Survival Analysis*. Indian Society for Medical Statistics (ISMS), Jawaharlal Institute of Postgraduate Medical Education and Research.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Kauffman, R. and Wang, B. (2001). The success and failure of dotcoms: A multi-method survival analysis. In *Proceedings of the 6th INFORMS Conference on Information Systems and Technology (CIST)*, Miami, FL, USA.

Kauffman, R. and Wang, B. (2003). Duration in the digital economy: Empirical bases for the survival of internet firms. In *36th Hawaii International Conference on System Sciences (HICSS)*, Hawaii.

Keasey, K., McGuinness, P., and Short, H. (1990). Multilogit approach to predicting corporate failure: Further analysis and the issue of signal consistency. *Omega*, 18(1):85–94.

Kumar, K. and Ganesalingam, S. (2001). Detection of financial distress via multivariate statistical analysis. *Detection and Prediction of Financial Distress*, 27(4):45–55.

Kumar, K. and Tan, C. (2005). Some recent developments in financial distress prediction. *Bulletin of the International Statistical Institute*.

Laitinen, E. K. and Luoma, M. (1991). Survival analysis as a tool for company failure prediction. *Omega*, 19(6):673–678.

Laitinen, T. and Kankaanpää, M. (1999). Comparative analysis of failure prediction methods: the Finnish case. *The European Accounting Review*, 8(1):67–92.

Lane, W. R., Looney, S. W., and Wansley, J. W. (1986). An application of the Cox proportional hazards model to bank failure. *Journal of Banking and Finance*, 10(4):511–531.

Libby, R. (1975). Accounting ratios and the prediction of failure: some behavioural evidences. *Journal of Accounting Research*, Spring:150–161.

Lundin, M., Lundin, J., Burke, H. B., and Joensuu, H. (1999). Artificial Neural Network applied to survival prediction in breast cancer. *Oncology*, 57:281–286.

Martin, D. (1977). Early warning of bank failure-a logit regression approach. *Journal of Banking and Finance*, 1:249–276.

Martinelli, E., de Carvalho, A., Rezende, S., and Matias, A. (1999). Rules extractions from banks bankrupt data using connectionist and symbolic learning algorithms. In *Proc. Computational Finance Conf.*, New York.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas imminent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1:27–52.

Odom, M. D. and Sharda, R. (1990). A neural network model for bankruptcy prediction. In *Proceedings of the IEEE international conference on neural networks*, pages 1163–1168, San Diego.

Ogg, P. J. (1988). Quantitative aspects of modelling financial distress. Working Paper. Submitted to The Inaugural Australasian Finance and Banking Conference without acceptance. This is also the authors maiden name, her name in publications is now Cybinski.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131.

Ohno_machado, L., Qalker, M. G., and Musen, M. A. (1995). Hierarchical neural network for survival analysis. In *MEDINFO 95*, pages 828,832, Vancouver, Canada.

Pacey, J. and Pham, T. (1990). The predictiveness of bankruptcy models: Methodological problems and evidence. *Journal of Management*, 15:315–337.

Patrick, P. (1932). A comparison of ratios of successful industrial enterprises with those of failed firms. *Certified Public Accountant*, October, November, and December:598–605, 656–662, and 727–731 respectively.

Prashanthi (2005). Survival regression. In Sahai, A., Kaliaperumal, V. G., and Venkatesan, P., editors, *22nd Annual National Conference of Indian Society for medical statistics: Pre Conference Workshop/Seminar on Survival Analysis*. Indian Society for Medical Statistics (ISMS), Jawaharlal Institute of Postgraduate Medical Education and Research.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, California.

Richardson, F. M. and Davidson, L. F. (1983). An exploration into bankruptcy discriminant model sensitivity. *Journal of Business Finance and Accounting*, 10(2):195–207. Summer Edition.

Salchenberger, L. M., Cinar, E. M., and Lash, N. A. (1992). Neural networks: A new tool for predicting thrift failures. *Decision Sciences*, 23(4):899–916.

Santomero, A. and Vinso, J. (1977). Estimating the probability of failure for firms in the banking system. *Journal of Banking and Finance*, pages 185–205.

Scapens, R. W., Ryan, R. J., and Flecher, L. (1981). Explaining corporate failure: A catastrophe theory approach. *Journal of Business Finance and Accounting*, 8:1–26.

Shin, K. and Lee, Y. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems Applications*, 23(3):321–328.

Shirata, C. Y. (1998). Financial ratios as predictors of bankruptcy in Japan: An empirical research. In *Proceedings of The Second Asian Pacific Interdisciplinary Research in Accounting Conference*, pages 437–445, Osaka, Japan.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1):101–124.

Tam, K. Y. (1991). Neural networks and the prediction of bank bankruptcy. *Omega*, 19:429–445.

Tam, K. Y. and Kiang, M. (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management Science*, 38(7):926–947.

Tan, C. N. W. (2001). *Artificial neural networks: applications in financial distress prediction & foreign exchange trading*. Wilberto Publishing, Gold Coast, Australia.

Theodossiou, P. T. (1991). Alternative models for assessing the financial condition of business in Greece. *Journal of Business Finance and Accounting*, 18(5):697–720.

Theodossiou, P. T. (1993). Predicting shifts in the mean of a multivariate time series process: an application in predicting business failure. *Journal of the American Statistical Association*, 88(422):441–449.

Wilcox, J. (1971). A gambler's ruin prediction of business failure using accounting data. *Sloan Management Review*, pages 1–10. Spring copy of the journal.

Zavgren, C. V. (1985). Assessing the vulnerability to failure of American industrial firms: A logistic analysis. *Journal of Business, Finance and Accounting*, pages 19–45. Spring Edition.

Zopounidis, C. and Dimitras, A. I. (1998). *Multicriteria decision aid methods for the prediction of business failure*. Kluwer Academic Publishers, Boston.

Appendix A

Decision Tree Research Results

Appendix A contains the complete set of results obtained from the Decision Tree (DT) Research presented in Chapter 3. These results are presented in a tabular format based on the tables in Frydman et al. (1985) that this research extends.

	In-Sample							
ECR	RPA_c	RPA_s	DA_c	DA_s	See5_c	See5_s	CART_c	CART_s
0.05	0.006	0.020	0.017	0.026	0.020	0.020	0.008	0.020
0.5	0.069	0.086	0.121	0.155	0.059	0.059	0.017	0.038
1	0.076	0.110	0.193	0.207	0.097	0.097	0.076	0.097
1.5	0.138	0.224	0.231	0.235	0.062	0.128	0.134	0.159
2	0.159	0.255	0.255	0.297	0.069	0.103	0.193	0.255
2.5	0.190	0.286	0.272	0.324	0.066	0.114	0.228	0.286
3	0.186	0.248	0.290	0.324	0.062	0.117	0.255	0.317
3.5	0.173	0.269	0.321	0.352	0.062	0.128	0.269	0.348
	Cross-Validated							
ECR	RPA_c	RPA_s	DA_c	DA_s	See5_c	See5_s	CART_c	CART_s
0.05	0.091	0.020	0.056	0.035	0.020	0.020	0.068	0.020
0.5	0.124	0.124	0.179	0.155	0.152	0.141	0.186	0.131
1	0.235	0.221	0.296	0.249	0.221	0.228	0.241	0.214
1.5	0.376	0.269	0.272	0.307	0.321	0.335	0.307	0.276
2	0.442	0.324	0.329	0.365	0.414	0.359	0.352	0.331
2.5	0.497	0.386	0.386	0.402	0.441	0.373	0.407	0.379
3	0.497	0.386	0.427	0.397	0.497	0.400	0.469	0.428
3.5	0.531	0.455	0.518	0.369	0.486	0.528	0.517	0.476

Table A.1: Resubstitution risk of each model for all the Error Cost Ratios (ECR) of Type I to Type II Error. These numbers were calculated using the number of Type I and Type II Errors produced by each model, which is shown in Table A.2 on the next page.

	In-Sample																							
ECR:	0.05			0.5			1			1.5			2			2.5			3			3.5		
Error Type:	I	II	T	I	II	T	I	II	T	I	II	T	I	II	T	I	II	T	I	II	T	I	II	T
RPA_c	18	0	18	14	3	17	9	2	11	6	11	17	6	11	17	5	15	20	5	12	17	2	18	20
RPA_s	58	0	58	19	3	22	14	2	16	9	19	28	9	19	28	9	19	28	9	9	18	4	25	29
DA_c	48	0	48	25	5	30	17	11	28	13	14	27	11	15	26	9	17	26	8	18	26	7	22	29
DA_s	54	1	55	33	6	39	18	12	30	12	16	28	11	21	32	8	27	35	6	29	35	6	30	36
See5_c	58	0	58	17	0	17	11	3	14	6	0	6	5	0	5	3	2	5	0	9	9	0	9	9
See5_s	58	0	58	17	0	17	11	3	14	9	5	14	5	5	10	3	9	12	3	8	11	3	8	11
CART_c	23	0	23	5	0	5	11	0	11	11	3	14	10	8	18	10	8	18	5	22	27	4	25	29
CART_s	58	0	58	11	0	11	11	3	14	10	8	18	9	19	28	9	19	28	9	19	28	9	19	28
	Cross-Validation																							
ECR:	0.05			0.5			1			1.5			2			2.5			3			3.5		
Error Type:	I	II	T	I	II	T	I	II	T	I	II	T	I	II	T	I	II	T	I	II	T	I	II	T
See5_c	58	0	58	22	11	33	18	14	32	19	18	37	18	24	42	16	24	40	16	24	40	13	25	38
See5_s	58	0	58	27	7	34	17	16	33	23	14	37	15	22	37	10	29	39	12	22	34	13	31	44
CART_c	38	8	46	16	19	35	22	13	35	17	19	36	15	21	36	14	24	38	14	26	40	14	26	40
CART_s	58	0	58	16	11	27	21	10	31	16	16	32	14	20	34	14	20	34	14	20	34	14	20	34

Table A.2: Errors produced by each model for all the Error Cost Ratios (ECR) of Type I to Type II Error. The abbreviations I, II and T indicate Type I, Type II, and Total Error respectively. Note that Frydman et al. (1985) did not publish the number of errors produced by the RPA and DA models in their cross-validation trials.

Appendix B

Survival Analysis Research Results

Appendix B contains the complete set of results obtained from the Survival Analysis (SA) Research presented in Chapter 4. These results are presented in the following tables:

Tables B.1 and B.2 contain the weighted error cost measures of each technique for prediction intervals of 1 to 5 and 6 to 10 years respectively. This includes both in-sample and hold-out results over the complete set of 8 different misclassification costs.

Tables B.3 and B.4 contain the breakdown of in-sample classifications of each technique into correct classifications, Type I Error and Type II Error for prediction intervals of 1 to 5 and 6 to 10 years respectively.

Tables B.5 and B.6 contain the breakdown of hold-out predictions of each technique into correct classifications, Type I Error and Type II Error for prediction intervals of 1 to 5 and 6 to 10 years respectively.

	In-Sample														
Cut-off / ECR	Cox1	DA1	LA1	Cox2	DA2	LA2	Cox3	DA3	LA3	Cox4	DA4	LA4	Cox5	DA5	LA5
0.5 / 1	0.044	0.040	0.038	0.067	0.057	0.058	0.087	0.079	0.076	0.109	0.100	0.098	0.130	0.121	0.116
0.55 / 1.2	0.054	0.048	0.046	0.082	0.067	0.069	0.104	0.092	0.091	0.132	0.120	0.117	0.156	0.142	0.137
0.6 / 1.5	0.066	0.056	0.054	0.095	0.080	0.080	0.127	0.110	0.110	0.160	0.140	0.138	0.181	0.165	0.167
0.65 / 1.9	0.080	0.065	0.065	0.113	0.095	0.095	0.154	0.132	0.132	0.191	0.167	0.167	0.217	0.194	0.192
0.7 / 2.3	0.098	0.078	0.078	0.141	0.116	0.117	0.190	0.163	0.159	0.223	0.200	0.199	0.252	0.234	0.231
0.75 / 3	0.119	0.095	0.096	0.180	0.146	0.142	0.227	0.199	0.198	0.268	0.240	0.232	0.299	0.272	0.274
0.8 / 4	0.149	0.120	0.120	0.223	0.186	0.176	0.286	0.249	0.243	0.323	0.291	0.291	0.354	0.335	0.328
0.85 / 5.7	0.207	0.162	0.168	0.301	0.241	0.224	0.339	0.302	0.298	0.399	0.358	0.369	0.447	0.414	0.412
	Hold-Out														
Cut-off / ECR	Cox1	DA1	LA1	Cox2	DA2	LA2	Cox3	DA3	LA3	Cox4	DA4	LA4	Cox5	DA5	LA5
0.5 / 1	0.046	0.049	0.039	0.070	0.067	0.063	0.095	0.091	0.088	0.112	0.105	0.116	0.133	0.133	0.133
0.55 / 1.2	0.056	0.063	0.046	0.086	0.085	0.082	0.112	0.115	0.107	0.132	0.131	0.133	0.160	0.149	0.158
0.6 / 1.5	0.068	0.072	0.054	0.105	0.100	0.091	0.132	0.130	0.130	0.163	0.172	0.158	0.186	0.175	0.186
0.65 / 1.9	0.085	0.080	0.069	0.124	0.119	0.116	0.164	0.165	0.155	0.200	0.201	0.194	0.217	0.204	0.221
0.7 / 2.3	0.106	0.094	0.091	0.151	0.143	0.143	0.198	0.206	0.199	0.234	0.244	0.228	0.257	0.253	0.253
0.75 / 3	0.126	0.102	0.119	0.196	0.179	0.175	0.235	0.235	0.249	0.270	0.260	0.260	0.330	0.316	0.288
0.8 / 4	0.154	0.130	0.123	0.253	0.242	0.225	0.288	0.270	0.284	0.361	0.316	0.295	0.421	0.319	0.326
0.85 / 5.7	0.220	0.171	0.171	0.301	0.308	0.284	0.378	0.354	0.280	0.490	0.402	0.347	0.516	0.344	0.392

Table B.1: Weighted error cost measures of all techniques with prediction intervals of 1 to 5 years, for all the Error Cost Ratios (ECR) of Type I to Type II Error. The number written following each technique's name represents the prediction interval, for example, the heading Cox2 represents a prediction interval of 2 years with the Cox model. The values in this table are derived from the number of Type I and Type II Errors for each technique, which is presented in Table B.3 and Table B.5 for prediction intervals of 1 to 5 years.

	In-Sample														
Cut-off / ECR	Cox6	DA6	LA6	Cox7	DA7	LA7	Cox8	DA8	LA8	Cox9	DA9	LA9	Cox10	DA10	LA10
0.5 / 1	0.148	0.132	0.132	0.158	0.150	0.151	0.169	0.160	0.162	0.173	0.173	0.169	0.181	0.173	0.172
0.55 / 1.2	0.173	0.156	0.158	0.190	0.173	0.178	0.195	0.187	0.189	0.200	0.197	0.189	0.208	0.198	0.196
0.6 / 1.5	0.206	0.179	0.185	0.217	0.202	0.201	0.225	0.214	0.209	0.231	0.228	0.211	0.238	0.232	0.223
0.65 / 1.9	0.236	0.214	0.220	0.246	0.239	0.239	0.256	0.247	0.243	0.264	0.264	0.256	0.269	0.254	0.260
0.7 / 2.3	0.271	0.253	0.258	0.286	0.275	0.271	0.299	0.290	0.292	0.308	0.299	0.300	0.316	0.303	0.306
0.75 / 3	0.315	0.302	0.311	0.330	0.327	0.324	0.352	0.334	0.341	0.365	0.350	0.341	0.371	0.352	0.354
0.8 / 4	0.375	0.355	0.363	0.408	0.381	0.387	0.413	0.398	0.395	0.412	0.411	0.405	0.414	0.411	0.404
0.85 / 5.7	0.456	0.439	0.434	0.468	0.469	0.465	0.469	0.472	0.463	0.474	0.476	0.460	0.481	0.484	0.463
	Hold-Out														
Cut-off / ECR	Cox6	DA6	LA6	Cox7	DA7	LA7	Cox8	DA8	LA8	Cox9	DA9	LA9	Cox10	DA10	LA10
0.5 / 1	0.151	0.144	0.147	0.165	0.165	0.158	0.175	0.179	0.179	0.193	0.196	0.196	0.218	0.200	0.214
0.55 / 1.2	0.173	0.170	0.175	0.186	0.184	0.188	0.206	0.210	0.208	0.239	0.234	0.236	0.264	0.257	0.264
0.6 / 1.5	0.202	0.193	0.207	0.218	0.214	0.219	0.260	0.251	0.251	0.296	0.286	0.268	0.316	0.312	0.284
0.65 / 1.9	0.236	0.240	0.230	0.279	0.260	0.251	0.322	0.286	0.258	0.356	0.319	0.282	0.373	0.334	0.321
0.7 / 2.3	0.297	0.276	0.278	0.332	0.309	0.261	0.370	0.325	0.310	0.415	0.375	0.346	0.447	0.378	0.398
0.75 / 3	0.393	0.309	0.284	0.404	0.375	0.319	0.449	0.361	0.396	0.460	0.393	0.411	0.484	0.449	0.449
0.8 / 4	0.449	0.396	0.330	0.491	0.382	0.435	0.519	0.449	0.488	0.523	0.495	0.533	0.565	0.561	0.554
0.85 / 5.7	0.513	0.415	0.387	0.585	0.467	0.484	0.614	0.535	0.563	0.642	0.609	0.584	0.653	0.690	0.605

Table B.2: This table extends Table B.1 to include the weighted error cost measures of all techniques for prediction intervals of 6 to 10 years. As in Table B.1, the number written following each technique's name represents the prediction interval. The values in this table are derived from the number of Type I and Type II Errors for each technique, which is presented in Table B.4 and Table B.6 for prediction intervals of 6 to 10 years.

	Cox1			Cox2			Cox3			Cox4			Cox5		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	2552	113	4	2489	169	11	2436	219	14	2378	266	25	2323	312	34
0.55	2550	112	7	2488	165	16	2438	210	21	2374	259	36	2320	301	48
0.6	2549	111	9	2495	157	17	2433	204	32	2368	250	51	2323	276	70
0.65	2548	107	14	2495	150	24	2426	197	46	2360	234	75	2307	254	108
0.7	2545	103	21	2485	145	39	2412	188	69	2350	208	111	2294	223	152
0.75	2546	97	26	2468	140	61	2390	163	116	2316	181	172	2242	186	241
0.8	2535	88	46	2439	122	108	2334	143	192	2246	146	277	2163	146	360
0.85	2500	82	87	2355	105	209	2240	102	327	2100	106	463	1990	110	569
	DA1			DA2			DA3			DA4			DA5		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	2561	77	31	2517	125	27	2458	175	36	2401	229	39	2346	278	45
0.55	2557	76	36	2517	120	32	2461	169	39	2399	222	48	2348	260	61
0.6	2557	75	37	2515	117	37	2458	163	48	2400	209	60	2349	242	78
0.65	2557	71	41	2511	113	45	2451	156	62	2392	198	79	2344	226	99
0.7	2555	70	44	2505	110	54	2432	149	88	2379	182	108	2319	205	145
0.75	2549	67	53	2494	107	68	2417	140	112	2364	168	137	2293	175	201
0.8	2543	65	61	2479	102	88	2394	130	145	2333	147	189	2232	152	285
0.85	2531	63	75	2450	91	128	2361	107	201	2269	119	281	2129	121	419
	LA1			LA2			LA3			LA4			LA5		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	2567	89	13	2515	136	18	2467	184	18	2408	234	27	2359	271	39
0.55	2566	88	15	2514	132	23	2467	178	24	2406	224	39	2359	255	55
0.6	2566	81	22	2518	127	24	2462	171	36	2405	209	55	2345	242	82
0.65	2562	79	28	2519	120	30	2454	160	55	2390	196	83	2343	216	110
0.7	2562	75	32	2511	115	43	2441	148	80	2377	180	112	2308	191	170
0.75	2558	73	38	2495	103	71	2413	136	120	2352	151	166	2267	165	237
0.8	2549	67	53	2470	90	109	2375	118	176	2277	128	264	2181	129	359
0.85	2509	62	98	2420	75	174	2288	89	292	2145	99	425	2032	99	538

Table B.3: Breakdown of the in-sample classifications of each technique into correct classifications (OK), Type I Error (I) and Type II Error (II). This table includes prediction intervals of 1 to 5 years, and Table B.4 on the next page includes prediction intervals of 6 to 10 years. The number written following each technique's name represents the prediction interval, for example, the heading Cox3 represents a prediction interval of 3 years with the Cox model.

	Cox6			Cox7			Cox8			Cox9			Cox10		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	2275	349	45	2248	364	57	2217	382	70	2206	391	72	2185	401	83
0.55	2279	322	68	2239	343	87	2226	348	95	2213	355	101	2194	355	120
0.6	2269	300	100	2245	309	115	2224	313	132	2207	308	154	2189	309	171
0.65	2268	266	135	2242	269	158	2212	265	192	2186	260	223	2175	260	234
0.7	2247	225	197	2208	227	234	2172	225	272	2138	219	312	2113	215	341
0.75	2189	181	299	2147	179	343	2092	181	396	2056	181	432	2030	175	464
0.8	2092	141	436	2022	147	500	1976	136	557	1938	123	608	1907	114	648
0.85	1914	99	656	1845	91	733	1795	81	793	1755	75	839	1722	72	875
	DA6			DA7			DA8			DA9			DA10		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	2316	294	59	2268	339	62	2242	351	76	2208	368	93	2206	355	108
0.55	2314	276	79	2275	310	84	2243	324	102	2217	334	118	2212	317	140
0.6	2319	254	96	2271	280	118	2241	287	141	2209	296	164	2192	287	190
0.65	2299	236	134	2250	254	165	2225	252	192	2186	258	225	2197	240	232
0.7	2280	214	175	2224	217	228	2186	218	265	2166	222	281	2147	215	307
0.75	2232	185	252	2169	187	313	2145	184	340	2106	185	378	2079	175	415
0.8	2174	151	344	2096	148	425	2053	149	467	2013	147	509	1982	137	550
0.85	2063	121	485	1969	118	582	1914	108	647	1875	102	692	1834	98	737
	LA6			LA7			LA8			LA9			LA10		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	2317	306	46	2266	335	68	2236	350	83	2217	346	106	2211	341	117
0.55	2310	283	76	2263	306	100	2234	310	125	2232	300	137	2214	305	150
0.6	2305	262	102	2268	269	132	2245	268	156	2234	257	178	2205	265	199
0.65	2285	237	147	2240	244	185	2221	233	215	2191	238	240	2176	233	260
0.7	2250	203	216	2215	201	253	2164	205	300	2141	204	324	2120	200	349
0.75	2184	173	312	2145	170	354	2095	168	406	2078	160	431	2051	164	454
0.8	2100	133	436	2037	134	498	2003	129	537	1967	126	576	1951	120	598
0.85	1953	95	621	1889	99	681	1857	91	721	1843	86	740	1811	81	777

Table B.4: This table extends Table B.3 to include the breakdown of in-sample classifications for prediction intervals of 6 to 10 years. As in Table B.3, the number written following each technique's name represents the prediction interval.

	Cox1			Cox2			Cox3			Cox4			Cox5		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	272	13	0	265	20	0	258	27	0	253	32	0	247	35	3
0.55	272	13	0	265	20	0	259	26	0	254	30	1	247	34	4
0.6	272	13	0	265	20	0	260	25	0	253	29	3	248	32	5
0.65	272	13	0	266	19	0	258	23	4	252	28	5	249	30	6
0.7	272	13	0	266	18	1	258	22	5	253	26	6	249	28	8
0.75	273	12	0	263	17	5	258	20	7	254	23	8	245	27	13
0.8	274	11	0	261	16	8	257	18	10	248	22	15	237	24	24
0.85	269	10	6	260	13	12	252	16	17	234	19	32	222	18	45
	DA1			DA2			DA3			DA4			DA5		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	271	9	5	266	15	4	259	21	5	255	28	2	247	32	6
0.55	269	9	7	264	15	6	257	21	7	254	28	3	249	29	7
0.6	269	9	7	264	15	6	258	20	7	250	28	7	249	28	8
0.65	269	8	8	263	14	8	255	20	10	250	26	9	250	27	8
0.7	269	8	8	263	14	8	253	20	12	250	26	9	249	27	9
0.75	270	7	8	262	14	9	254	18	13	253	21	11	245	25	15
0.8	269	7	9	258	14	13	256	16	13	252	19	14	245	17	23
0.85	269	7	9	258	13	14	254	15	16	245	16	24	243	12	30
	LA1			LA2			LA3			LA4			LA5		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	274	10	1	267	16	2	260	21	4	252	27	6	247	32	6
0.55	274	10	1	265	15	5	259	20	6	253	26	6	247	31	7
0.6	274	9	2	266	14	5	258	20	7	253	26	6	247	30	8
0.65	273	9	3	264	14	7	258	20	7	251	25	9	247	29	9
0.7	271	9	5	263	14	8	255	20	10	252	24	9	249	27	9
0.75	269	9	7	263	14	8	254	20	11	253	21	11	247	22	16
0.8	271	7	7	260	13	12	255	17	13	249	16	20	243	17	25
0.85	269	7	9	260	12	13	252	10	23	242	12	31	234	13	38

Table B.5: Breakdown of the hold-out predictions of each technique into correct predictions (OK), Type I Error (I) and Type II Error (II). This table includes prediction intervals of 1 to 5 years, and Table B.6 on the next page includes prediction intervals of 6 to 10 years. The number written following each technique's name represents the prediction interval, for example, the heading Cox4 represents a prediction interval of 4 years with the Cox model.

	Cox6			Cox7			Cox8			Cox9			Cox10		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	242	39	4	238	43	4	235	46	4	230	50	5	223	56	6
0.55	244	37	4	241	40	4	236	44	5	228	50	7	222	55	8
0.6	245	35	5	242	38	5	233	44	8	225	49	11	221	52	12
0.65	246	33	6	238	38	9	230	43	12	223	46	16	219	47	19
0.7	243	32	10	237	35	13	229	37	19	220	40	25	215	43	27
0.75	235	31	19	230	30	25	223	33	29	218	32	35	215	34	36
0.8	229	24	32	220	25	40	215	26	44	214	26	45	205	27	53
0.85	218	17	50	207	19	59	194	18	73	186	18	81	183	18	84
	DA6			DA7			DA8			DA9			DA10		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	244	35	6	238	42	5	234	45	6	229	50	6	228	52	5
0.55	244	34	7	241	38	6	235	44	6	229	48	8	223	51	11
0.6	246	32	7	242	36	7	235	43	7	227	47	11	221	50	14
0.65	244	32	9	241	35	9	236	38	11	230	42	13	224	40	21
0.7	245	29	11	241	33	11	235	32	18	226	36	23	224	35	26
0.75	247	25	13	236	29	20	232	25	28	229	28	28	221	32	32
0.8	238	22	25	236	20	29	226	23	36	216	24	45	206	27	52
0.85	232	14	39	222	15	48	212	17	56	200	19	66	191	22	72
	LA6			LA7			LA8			LA9			LA10		
Cut-off	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
0.5	243	36	6	240	39	6	234	44	7	229	49	7	224	54	7
0.55	243	35	7	240	38	7	235	42	8	228	46	11	221	50	14
0.6	243	34	8	241	37	7	233	39	13	229	41	15	226	44	15
0.65	246	31	8	241	32	12	238	31	16	233	33	19	226	38	21
0.7	243	28	14	244	25	16	234	28	23	229	32	24	221	37	27
0.75	248	22	15	240	23	22	226	27	32	224	28	33	219	31	35
0.8	242	17	26	227	22	36	215	23	47	208	25	52	205	26	54
0.85	226	11	48	217	15	53	204	17	64	198	17	70	192	17	76

Table B.6: This table extends Table B.5 to include the breakdown of hold-out predictions for prediction intervals of 6 to 10 years. As in Table B.5, the number written following each technique's name represents the prediction interval.

Appendix C

Comparative Research Results

Appendix C contains the complete set of results obtained from the comparative study of the CART and Cox model presented in Section 5.1 of Chapter 5. Table C.1 presents the weight error cost measures for each technique, and Table C.2 presents the breakdown of the technique's classification and prediction.

	In-Sample									
Cut-off / ECR	CART1	Cox1	CART2	Cox2	CART3	Cox3	CART5	Cox5	CART10	Cox10
0.5 / 1	0.039	0.044	0.058	0.067	0.080	0.087	0.072	0.130	0.126	0.181
0.6 / 1.5	0.039	0.066	0.080	0.095	0.074	0.127	0.108	0.181	0.146	0.238
0.67 / 2	0.067	0.085	0.103	0.123	0.064	0.166	0.117	0.225	0.180	0.283
0.71 / 2.5	0.062	0.104	0.126	0.149	0.070	0.200	0.197	0.261	0.374	0.331
0.75 / 3	0.073	0.119	0.157	0.180	0.149	0.227	0.224	0.299	0.401	0.371
0.8 / 4	0.139	0.149	0.150	0.223	0.025	0.286	0.263	0.354	0.455	0.414
0.83 / 5	0.155	0.187	0.042	0.273	0.028	0.315	0.237	0.406	0.508	0.448
	Hold-Out									
Cut-off / ECR	CART1	Cox1	CART2	Cox2	CART3	Cox3	CART5	Cox5	CART10	Cox10
0.5 / 1	0.039	0.046	0.063	0.070	0.084	0.095	0.102	0.133	0.189	0.218
0.6 / 1.5	0.049	0.068	0.089	0.105	0.093	0.132	0.137	0.186	0.233	0.316
0.67 / 2	0.070	0.091	0.116	0.133	0.119	0.175	0.168	0.228	0.288	0.393
0.71 / 2.5	0.084	0.114	0.142	0.153	0.156	0.214	0.232	0.277	0.314	0.446
0.75 / 3	0.084	0.126	0.175	0.196	0.168	0.235	0.263	0.330	0.337	0.484
0.8 / 4	0.112	0.154	0.214	0.253	0.218	0.288	0.330	0.421	0.382	0.565
0.83 / 5	0.123	0.193	0.221	0.281	0.281	0.347	0.382	0.491	0.428	0.628

Table C.1: Weighted error cost measures of the CART and Cox models for all prediction intervals and Error Cost Ratios (ECR) of Type I to Type II Error. The number written following each technique's name represents the prediction interval, for example, the heading Cox5 represents a prediction interval of 5 years with the Cox model. The values in this table are derived from the number of Type I and Type II Errors for each technique, which is presented in Table C.2 on the next page.

In-Sample															
	CART1			CART2			CART3			CART5			CART10		
ECR	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
1	2566	75	28	2515	121	33	2456	163	50	2477	156	36	2332	228	109
1.5	2597	62	10	2515	121	33	2530	118	21	2459	154	56	2359	161	149
2	2566	75	28	2515	121	33	2558	60	51	2477	120	72	2342	154	173
2.5	2584	54	31	2515	121	33	2536	36	97	2363	146	160	1884	143	642
3	2583	55	31	2421	86	162	2477	103	89	2358	144	167	1884	143	642
4	2430	44	195	2490	74	105	2617	5	47	2356	130	183	1884	143	642
5	2430	44	195	2601	11	57	2605	3	61	2321	71	277	1884	143	642
	Cox1			Cox2			Cox3			Cox5			Cox10		
ECR	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
1	2552	113	4	2489	169	11	2436	219	14	2323	312	34	2185	401	83
1.5	2549	111	9	2495	157	17	2433	204	32	2323	276	70	2189	309	171
2	2549	106	14	2491	149	29	2421	196	52	2310	242	117	2158	243	268
2.5	2545	102	22	2484	142	43	2407	181	81	2291	212	166	2094	206	369
3	2546	97	26	2468	140	61	2390	163	116	2242	186	241	2030	175	464
4	2535	88	46	2439	122	108	2334	143	192	2163	146	360	1907	114	648
5	2515	86	68	2388	112	169	2284	114	271	2057	118	494	1801	82	786

Hold-Out															
	CART1			CART2			CART3			CART5			CART10		
ECR	OK	I	II	OK	I	II	OK	I	II	OK	I	II	OK	I	II
1	274	9	2	267	15	3	261	19	5	256	21	8	231	40	14
1.5	275	8	2	267	15	3	265	13	7	256	20	9	235	33	17
2	274	9	2	267	15	3	262	11	12	256	19	10	234	31	20
2.5	273	8	4	267	15	3	257	11	17	246	18	21	215	13	57
3	275	7	3	255	10	20	259	11	15	246	18	21	215	13	57
4	262	3	20	254	10	21	256	11	18	245	18	22	215	13	57
5	262	3	20	258	9	18	253	12	20	240	16	29	215	13	57
	Cox1			Cox2			Cox3			Cox5			Cox10		
1	272	13	0	265	20	0	258	27	0	247	35	3	223	56	6
1.5	272	13	0	265	20	0	260	25	0	248	32	5	221	52	12
2	272	13	0	266	19	0	258	23	4	249	29	7	218	45	22
2.5	272	13	0	267	17	1	257	22	6	248	28	9	215	38	32
3	273	12	0	263	17	5	258	20	7	245	27	13	215	34	36
4	274	11	0	261	16	8	257	18	10	237	24	24	205	27	53
5	270	10	5	261	14	10	254	17	14	229	21	35	190	21	74

Table C.2: Breakdown of the in-sample classifications and hold-out predictions of the CART and Cox models into correct classifications (OK), Type I Error (I) and Type II Error (II). The number written following each technique's name represents the prediction interval, for example, the heading Cox10 represents a prediction interval of 10 years with the Cox model.