Bond University

DOCTORAL THESIS

Topics on Financial Distress Prediction Modelling

Halteh, Khaled

Award date:
2019

Topics on Financial Distress Prediction Modelling

Khaled Jamal Fadel Halteh

BOND BUSINESS SCHOOL

Gold Coast, Queensland

Australia

Submitted in total fulfillment of the requirements of the Degree of Doctor of Philosophy on the 13th day of May, 2019.

Under the Supervision of Professor Kuldeep Kumar & Doctor Adrian Gepp

This research was funded by an Australian Government Research Training Program Scholarship

To my father, Jamal; mother, Jumana; & late uncle, Hanna

Abstract

Financial distress is a critical social and economic problem that affects innumerable businesses the world over. Consequences of such an occurrence can go beyond the business owners and stakeholders – as was evident in the 2008 Global Financial Crisis (GFC), it can lead to a much larger macroeconomic calamity. Therefore, having the power to predict – and hence aid businesses from failing, has the potential to save not only the business, but whole economies from collapsing. This research's academic contribution is to advance the field of Financial Distress Prediction (FDP) by tackling this issue from multiple angles – each being explored in a separate chapter – including: industry-specificity, index development, Islamic banking, variables affecting bankruptcy, class imbalance in data-sets, and Large Companies (LCs) vis-à-vis Small and Medium Enterprises (SMEs). This was achieved through utilising cutting-edge machine learning techniques, such as: Artificial Neural Networks (ANNs), Decision Trees (DTs), Random Forests (RFs), and Stochastic Gradient Boosting (SGB); and comparing their outcomes with results achieved from using well-established benchmark statistical techniques, such as: Multivariate Discriminant Analysis (MDA) and Logistic Regression (LR).

Two major databases were used in this thesis to extract more than 60 explanatory variables derived from financial statement data pertaining to thousands of existing and failed Australian and international companies across various industries in the marketplace. The extracted data were used to test for the validity and predictive power of the developed statistical models. The results in Chapter 3 empirically showcase that industry-specific models are superior to a *one-size-fits-all* model. The chapter also presents the most important variables in predicting financial distress pertaining to each industry. The results in Chapter 4 show that all FDP models built using machine learning techniques outperform a model built using the traditional LR statistical technique. Chapter 5 reveals that FDP models built using a data-set via the Synthetic Minority Oversampling Technique (SMOTE) outperform those using a standard data-set that is imbalanced. Chapter 6 presents a series of novel and user-friendly FDP indices that provide a standardised score for companies according to their success or

distress potential. Chapter 7 explores the differences between conventional and Islamic banking, then proceeds to build FDP models using machine learning techniques, each with a different measure of Islamic banks' financial distress. The aim was to present the most important variables in forecasting financial distress relating to Islamic banks. Chapter 8 creates FDP models using machine learning techniques on data-sets comprised of LCs and SMEs that are listed on the Australian Stock Exchange (ASX). These models are then compared with models that were built using data that have been *SMOTEd*, in order to establish the empirically superior FDP model, as well as outlining the most important variables in determining the successes or failures of SMEs and LCs.

The multifaceted approach used in this dissertation contains many important practical contributions, including: aiding lenders in accurately determining the economic viability of providing loans to prospective borrowers, offering investors with invaluable insight on their existing and/or potential investment, enabling governmental agencies to monitor businesses with high chances of bankruptcy, and providing managers and decision makers with invaluable insight to be used in conjunction with their expertise, in order to install proactive measures to mitigate the chances of falling into financial distress. These benefits have the potential to assist whole economies from falling into a recession as a result of increased business failure.

Keywords

Financial distress prediction; industry-specificity; statistical modelling; machine learning techniques; decision trees; random forests; stochastic gradient boosting; Islamic banking; Synthetic Minority Oversampling Technique (SMOTE); Australia; index construction; Small and Medium Enterprises (SMEs).

Declaration by Author

This thesis is submitted to Bond University in fulfilment of the requirements of the degree of Doctor of Philosophy.

This thesis represents my own original work towards this research degree and contains no material which has been previously submitted for a degree or diploma at this University or any other institution, except where due acknowledgement is made.

Khaled Jamal Fadel Halteh *13th of May, 2019*

Declaration of Author Contributions

Publication Co-authored	Statement of Contribution
Halteh, K., Kumar, K., & Gepp, A. (2018). Using Cutting-Edge Tree-Based Stochastic Models to Predict Credit Risk. *Risks, 6*(2), 55. Available at: http://www.mdpi.com/2227-9091/6/2/55/pdf	KH 80%, KK10%, AG10%
Halteh, K., Kumar, K., & Gepp, A. (2018). Financial distress prediction of Islamic banks using tree-based stochastic techniques. *Managerial Finance*, 44(6), 759-773 https://doi.org/10.1108/MF-12-2016-0372	KH 75%, KK 10%, AG15%
Halteh, K. (2015). Bankruptcy Prediction of Industry-Specific Businesses Using Logistic Regression. *Journal of Global Academic Institute Business & Economics, 1*(2), 151-163	KH 100%

Research Outputs and Publications during Candidature

Peer-Reviewed Publications:

- Halteh, K., Kumar, K., & Gepp, A. (2018). Using Cutting-Edge Tree-Based Stochastic Models to Predict Credit Risk. *Risks, 6*(2), 55. Available at: http://www.mdpi.com/2227-9091/6/2/55/pdf
- Halteh, K., Kumar, K., & Gepp, A. (2018). Financial distress prediction of Islamic banks using tree-based stochastic techniques. *Managerial Finance*, 44(6), 759-773 https://doi.org/10.1108/MF-12-2016-0372
- Halteh, K. (2015). Bankruptcy Prediction of Industry-Specific Businesses Using Logistic Regression. *Journal of Global Academic Institute Business & Economics, 1*(2), 151-163.

Published and Presented Conference Paper:

- Halteh, K. (2015). Bankruptcy Prediction of Industry-Specific Businesses Using Logistic Regression. *Proceedings of Prague International Academic Conference.* Prague, 119-129.

Published and Presented Conference Abstracts:

- Halteh, K., Gepp, A., & Kumar, K. (2017). Financial Distress Prediction in the Australian Mining Industry using Tree-based Stochastic Techniques. *Proceedings of 29th Asian-Pacific Conference on International Accounting Issues*. Kuala Lumpur, 9.
- Halteh, K., Gepp, A., & Kumar, K. (2016). Financial Distress Prediction using Cutting-Edge Statistical Techniques. *Proceedings of 28th Asian-Pacific Conference on International Accounting Issues*. Maui, 71.
- Halteh, K., & Kumar, K. (2015). Bankruptcy Prediction using Industry-Specific Variables. *Proceedings of 27th Asian-Pacific Conference on International Accounting Issues*. Gold Coast, 36.

Acknowledgements

First and foremost, I would like to thank God for giving me strength, resilience, endurance, and for His guidance every step of the way throughout my PhD journey. Dear Lord, I am eternally grateful for the copious amounts of blessings that You have bestowed upon me.

The past four years have been filled with a constellation of emotions – joyful and tearful alike; starting with the acceptance letter for the PhD program at Bond University, and ending with the last period in this dissertation. Let me not forget all the memorable moments in-between – from achieving two first-in-class awards, to attending and presenting in domestic and international conferences, to passing the Confirmation of Candidature with the panel's recommendation to upgrade me from MPhil to PhD, to my first journal publication, and to not-so-joyful moments like sustaining a serious neck injury and missing out on family and friends' events for being inundated with research. Now that I have finally completed this arduous task, I feel like I have reached the crowning point of my career. I can now look back and confidently say: *it was all worth it!*

To my father, Jamal, and my mother, Jumana – if it were not for your daily phone calls, persistence, and perseverance (*euphemisms for nagging*), I would not be where I am today. I would like to thank you from the depths of my heart for your continued support – I am forever indebted to you.

Special thanks to my supervisors and mentors, Professor Kuldeep Kumar and Doctor Adrian Gepp, for their continual counsel, encouragement, and unrivalled support during my time at Bond University. Professor Kumar has been helping me achieve my potential ever since I was completing a Master of Business Administration (MBA) degree. During the PhD degree, he continually pushed me to publish papers in reputable journals whenever any opportunity presented itself, since he knew the peerreview process will greatly benefit me not only to complete my degree, but also prepare me for future academic publications. Doctor Adrian was always available to aid me every step of the way, especially pertaining to software and academic-writing issues. My supervisors were always there to guide me through any difficulties and offer their prompt leadership and wisdom. So, thank you both dearly; without your support, I would not have been able to complete my PhD. I am endlessly appreciative for everything you have done for me.

I would like to extend my thanks and gratitude to my sisters, Dina and Siwar; my cousins Dr. Bashaar, Dr. Firas, Ibrahim; my grandparents Michael and Laila; my great-uncle Dr. Odeh and his wife Linda; my uncle Daniel and his wife Dina; my aunties Siham and Su'ad; and my dear friends Rami, Yasmine, Louie, Dr. Abdallah, and Dr. Burhan. Their unrelenting encouragement, hospitality, and prayers helped me complete my doctorate degree.

Last, but not least, I would like to sincerely thank the Australian Government for funding this research ever since January, 2015, through the Research Training Program Scholarship.

Copyright Information

Parts of Chapter 3, Chapter 4, and Chapter 7 are published material in peer-reviewed journals. Permission has been granted by the publishers to be used in this thesis.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

Acronym/Abbreviation	Meaning
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASIC	Australian Securities and Investments Commission
ASX	Australian Stock Exchange
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic
BFP	Business Failure Prediction
BNN	Back-propagation Neural Networks
CART	Classification and Regression Trees
CDP	Credit Default Prediction
CRM	Credit Risk Modelling
DT	Decision Trees
FDI	Foreign Direct Investment
FDP	Financial Distress Prediction
FDPI	Financial Distress Prediction Index
FRP	Financial Risk Prediction
FWI	Factor Weighted Index
GFC	Global Financial Crisis
H	Hypothesis
IT	Information Technology
KMO	Kaiser-Meyer-Olkin
K-NN	K-Nearest Neighbours
LC	Large Companies
LR	Logistic Regression
MDA	Multivariate Discriminant Analysis
NSI	Non-Standardised Index
NZSE	New Zealand Stock Exchange
PCA	Principal Component Analysis
RF	Random Forests
ROC	Receiver Operating Characteristic
ROI	Return on Investment
RPA	Recursive Partitioning Analysis
RQ	Research Question
RSF	Random Survival Forests
SGB	Stochastic Gradient Boosting
SI	Standardised Index
SME	Small and Medium Enterprises

SMOTE	Synthetic Minority Oversampling Technique
SPM	Salford Predictive Modeller
SVM	Support Vector Machines
WFLI	Weighted Factor Loading Index

Chapter 1: Introduction

Financial distress is a critical indicator of a company's financial health because it can prove to be detrimental if it is not addressed promptly. Consequences of such an occurrence can go beyond the business owners and stakeholders – as was evident in the 2008 Global Financial Crisis (GFC), it can lead to a much larger macroeconomic calamity. Therefore, having the power to predict business failure has the potential to save not only the business, but whole economies from collapsing. There are many causes of financial distress; some of these causes include reasons that are within the company's control, such as: fraud, managerial ineptness, neglect, and financial (Anderson, 2006); and others that are extraneous to the company, including: government laws and regulations, economic stability, natural disasters, and political turmoils. To allay the chances of falling into financial distress, Financial Distress Prediction (FDP) models can be an invaluable asset.

FDP models attempt to predict the financial failure or success of a business based on data, usually from publicly available information, such as financial ratios from financial statements (Gepp & Kumar, 2012). Such models can provide an early warning signal of probable financial distress, as well as showcasing the variables that have the strongest effect on determining a company's financial standing. This can help managers, investors, and other stakeholders to make educated decisions and install proactive measures to prevent possible insolvency, thus reducing realised incurred losses (Jaikengkit, 2004). Due to the models' wide applicability and important implications, the literature is quickly becoming inundated with studies across various disciplines, including but not limited to: finance, accounting, statistics, and actuarial studies (Cybinski, 2001; Yu, Miche, Séverin, & Lendasse, 2014).

Researchers on this topic have utilised a variety of statistical and machine learning techniques – Multivariate Discriminant Analysis (MDA), Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), and Stochastic Gradient Boosting (SGB), to name a few, in order to find the most accurate model. This thesis explores the literature and mechanics pertaining to FDP models, describes the pros and cons of each, employs numerous techniques on a variety of data-sets, and compares the generated FDP models' accuracies. The findings in this thesis will empirically showcase which technique(s) have superior predictive power, which variables are the most important in the models, and present various methodologies that aim at further enhancing the predictive accuracy of FDP models. As per West, Dellana, and Qian (2005), when comparing models, even an infinitesimal improvement in percentage accuracy can lead to huge savings. Therefore, when an almost negligible improvement in prediction accuracy across different models is presented in this thesis, a valid conclusion towards the superiority of the technique used can be inferred.

Financial Distress Prediction is known by many names, including: Business Failure Prediction (BFP), bankruptcy prediction, Financial Risk Prediction (FRP), Credit Risk Modelling (CRM), insolvency prediction, and Credit Default Prediction (CDP). For consistency purposes, Financial Distress Prediction, and its acronym FDP, will be regularly used in this thesis to refer to the aforementioned synonyms.

According to Gepp and Kumar (2012), some of the gains of utilising FDP models include:

- Allowing banks and lenders to assess a business's financial distress probability before determining whether a loan is suitable, and if so, how much excess and premium to charge;

- Governments and watchdog institutions can utilise the models to focus on businesses with high financial distress probabilities;

- Existing and potential stockholders can use the FDP models to make informed decisions about their investments for best Return on Investment (ROI) opportunities;

- Enabling potential merger companies and other stakeholders to assess the likelihood of a business's failure or success as an indicator of whether there will be sustainable benefits gained from operating or continuing to operate with the company at hand.

1.1 Research Questions and Hypotheses

This study will answer each Research Question (RQ) and Hypothesis (H) outlined below. These questions were based on an extensive review of the literature relating to FDP, which included reviewing 220 journal publications, books, theses, news articles, web pages, and conference proceedings. These will be explored in Chapter 2, as well as in the Literature Review sections of each proceeding chapter. After reviewing the literature, it was evident that there was a shortage of FDP studies focusing on certain aspects. Therefore, this provided the impetus and motivation to dedicate this research towards expanding on the available literature, especially due to the fact that there are vast potential contributions to be gained, not only on a local scale, but globally. These gaps in the literature helped formulate the research questions and hypotheses presented below. The research questions will be addressed throughout the thesis; each chapter's introduction and conclusion section will indicate which hypothesis/hypotheses were addressed in that chapter.

RQ1: Do industry-specific models have a greater ability to predict financial distress vis-à-vis a *one-size-fits-all* model?

- **Justification:** After reviewing the FDP literature, less than 5% mentioned industry-specific FDP models, and of those, none were scoped around Australian businesses. Hence, the first hypothesis is as follows:

  H_1: Industry-specific models have a greater ability to predict financial distress when compared to a *one-size-fits-all* industry-wide model.

RQ$_2$*:** Do independent variables differ in predictive importance across the models mentioned in ***RQ$_1$/H$_1$?

- **Justification:** Through reviewing the literature, it was found that less than 5% of studies pertaining to FDP mentioned variable predictive importance by industry, and of those, none were scoped around Australia. Hence, the second hypothesis is as follows:

  > ***H$_2$*:** Independent variables differ in predictive importance across the models mentioned in ***RQ$_1$/H$_1$***.

***RQ$_3$*:** Will using cutting-edge recursive partitioning techniques yield more accurate results vis-à-vis traditional statistical techniques?

- **Justification:** Through reviewing the FDP literature, around 30% of studies compared the accuracy of statistical models with recursive partitioning techniques, and of those, around 1% were centred around Australia. Hence, the third hypothesis is as follows:

  > ***H$_3$*:** Using cutting-edge recursive partitioning techniques will yield empirically superior results compared to traditional statistical techniques.

***RQ$_4$*:** Does class imbalance affect detection accuracy of the statistical models, and if so, how can it be enhanced?

- **Justification:** Class imbalance occurs when there is a substantial difference in the ratio between the classes in a data-set; therefore, it may have an effect on the predictive accuracy of FDP models. Through reviewing the FDP literature, less than 20% of studies were centred around class imbalance. Hence, the fourth hypothesis is as follows:

***H_4*:** Class imbalance does affect the detection accuracy of FDP models, and it can be enhanced by optimising the cut-off points or using SMOTE vis-à-vis a model that is built on a standard imbalanced data-set.

***RQ_5*:** Does the importance of independent variables vary between FDP models for Small and Medium Enterprises (SMEs) vis-à-vis Large Companies (LCs)?

- **Justification:** Through reviewing the FDP literature, less than 5% of FDP studies concentrated on SMEs, and less than 1% concentrated on independent variables differences between FDP models for SMEs and large companies. Hence, the fifth hypothesis is as follows:

***H_5*:** Independent variables' importance vary between FDP models for SMEs vis-à-vis LCs.

***RQ_6*:** Are there any benefits for creating an FDP index?

- **Justification:** Through reviewing the FDP literature, there were no studies that presented an FDP index, despite the existence of studies within the literature regarding the creation of indices. Therefore, this presents the potential for a pioneering study in this area. Hence, the sixth hypothesis is as follows:

***H_6*:** Creating an FDP index is more accurate, informative, and user-friendly than solely relying on standard FDP models.

RQ*$_7$*: Does varying the measure of banks' financial distress yield different important variables pertaining to Islamic banks?

- **Justification:** Through reviewing the FDP literature, around 20% of studies pertaining to FDP were centred around banks, and of those, around 1% focused on Islamic banks. Hence, the seventh hypothesis is as follows:

  H*$_7$*: The most important variables in FDP models for Islamic banks vary according to the measure of financial distress used.

1.2 Data

The data for the companies used in this research were extracted from several sources, including MorningStar and Capital IQ, which provide readily available archival data. According to Shultz, Hoffman, and Reiter-Palmon (2005), using archival data in the research has many benefits, including: ease of extraction, global accessibility, generally containing large amounts of data over many years, and most importantly, its ease of reproducibility and verifiability/falsifiability – key components of empirical tests. This enhances data quality by enabling more efficient and effective data extraction, cleaning, and analysis before commencing FDP modelling.

MorningStar offers archival data on publicly listed companies in the Australian Stock Exchange (ASX) and New Zealand Stock Exchange (NZSE), as well as data on approximately half-a-million investment offerings, in addition to real-time international market data on millions of commodities, foreign exchange, indices, and numerous others (MorningStar, 2015). Data from MorningStar has been extensively used in prior research across various fields, some of which are by: Halteh (2015); Halteh, Kumar, and Gepp (2018b); Shah (2014); Smith, Ren, and Dong (2011).

Capital IQ provides web-based information services that combine information on companies worldwide along with a variety of software applications that allow financial professionals to analyse company fundamentals, build financial models, screen for investment ideas, and execute other financial research tasks (Phillips, 2012). Capital IQ has been used in previous studies across various disciplines in the literature, some of these include: Feldman and Zoller (2012); Halteh, Kumar, and Gepp (2018a); Kahle and Stulz (2013).

1.3 Study Scope and Research Objectives

The sole data analysis methodology for this study is quantitative based. According to Kruger (2003), there are many advantages to using quantitative data analysis including: ease of replication; more accurate analysis and comparison to existing literature; efficient summarisation of huge sources of information; allowance of a wider scope of study, involving many subjects; mitigation of personal biases by researchers due to objective data, resulting in greater validity, reliability, and accuracy of results.

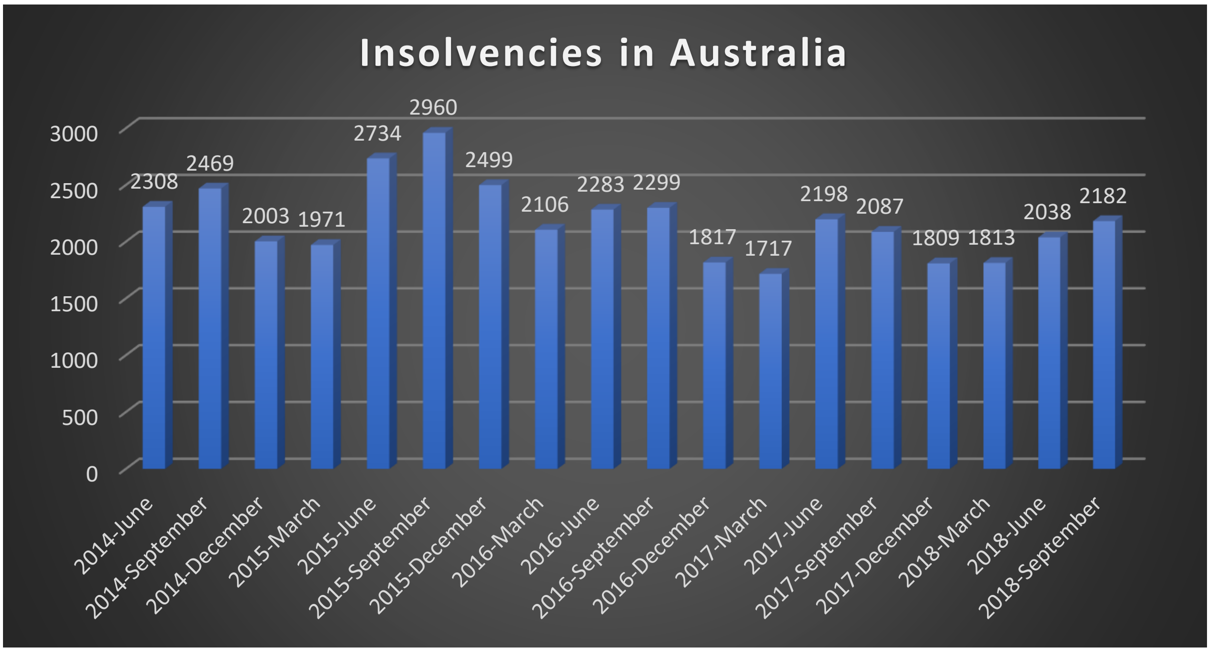
This research focuses primarily on the Australian marketplace, with the exception of Chapter 7 which covers Islamic banking on a global scale. The applicability of this research, however, is not at all limited to Australia; on the contrary, the research methodologies can be applied to any international setting that has data available. The reasons why Australia was chosen are because:

- Australia is the country of residence of the researcher – this entails having a direct and vested interest in investigating FDP in the context of Australia, in order to benefit the Australian economy;

- Paucity of FDP literature focusing on Australia – this research makes a significant contribution to the limited literature available, and aims to encourage future studies to have an Australian-centric approach;

- High insolvency rates – Australia is considered one of the largest mixed economies in the world, with a GDP of A$1.6 trillion in 2015 and US$1.5 trillion in 2018; it has a AAA credit rating and an unemployment rate below 6% (ABC, 2011; ABS, 2017). At the same time, paradoxically, according to the Australian Securities and Investments Commission (ASIC, 2015), around 3,000 businesses went insolvent in the September quarter of 2015 – that equates to almost 1,000 bankrupt businesses per month – that is an increase of 8.3% from the June quarter, and an increase of 20% from the September quarter in 2014. Four years later, the statistics are slightly more promising, but still far from significantly alleviated. In the September quarter of 2018, more than 2,180 companies went insolvent, an increase of 7.1% from the previous quarter, and an increase of 4.6% from the September quarter in 2017 (ASIC, 2018). Refer to Figure 1.1 for a visual representation of insolvency figures in Australia for the time-period 2014-2018 according to ASIC.

Figure 1. 1 Insolvencies in Australia from June 2014 till September 2018



There is sufficient empirical evidence to suggest a sustained large number of business bankruptcies in Australia. If this perpetuates, it may lead to a number of negative outcomes, including: higher unemployment rates and a potential lowering of the AAA credit rating status of the country – which can have a deleterious impact on foreign investment, or as often referred to in the literature – Foreign Direct Investment (FDI). According to the Australian Department of Foreign Affairs and Trade and the Australian Trade and Investment Commission, foreign investment is an integral component of the Australian economy, that helps boost employment, fund hospitals, schools, and other government services. Between 2014-2015, FDI contributed to 41% of Australia's goods and services exports, accounted for $2.7 trillion in assets, and contributed $286 billion to Australia's Industry Value Added. In 2017, foreign investment contributed $43 billion to the total investment flows of $433 billion, that is, approximately 10% (Austrade, 2015; DFAT, 2018). Thus, FDI is critical to Australia's economy, hence, and any kind of instability that may lead to a drop in FDI will have an unfavourable effect on the Australian economy.

A real-world example of these dire consequences occurred in the United States of America following the 2008 Global Financial Crisis (GFC), which brought about the collapse of titans like Lehmann Brothers, AIG, and Enron. In the years that followed, the United States' economy continued to suffer, which eventually led to S&P downgrading the USA's 70-year-long AAA credit rating to AA+, following unsuccessful plans to fix the debt crisis (Elliott, Treanor, & Rushe, 2011). After the announcement, all three major U.S. indexes – Dow Jones, NASDAQ, and S&P500 – declined between five and seven percent in one day, erasing around $2.5 trillion from global equity (Bloomberg, 2011).

This study addresses the following research objectives:

- To discover whether financial distress prediction of businesses can be more accurately achieved using industry-specific models vis-à-vis a *one-size fits all* approach (Chapter 3);

- To compare the predictive accuracy of various statistical and machine learning models in order to determine which model, or set of models, is/are optimal, and identify the inferior models (Chapter 3, Chapter 4, Chapter 5, Chapter 7, Chapter 8);

- To determine the variables which are most important for each industry-group in predicting financial distress and check for variable differences across industries (Chapter 3, Chapter 4, Chapter 7, Chapter 8);

- Analyse and compare the differences between conventional and Islamic banks, if any, in terms of FDP models and variable differences (Chapter 7);

- To check for differences between large companies vis-à-vis Small and Medium Enterprises (SMEs), in terms of FDP models and variable differences (if any), and develop an FDP model for SMEs (Chapter 8);

- To check for issues associated with class imbalance and how to remedy them (Chapter 4, Chapter 5, Chapter 8);

- To develop an index which can rank companies based on their financial health (Chapter 6).

1.4 Thesis Structure

This treatise is structured in the following manner: Chapter 1 introduces the topic of FDP, outlines the research questions and hypotheses that will be explored throughout the thesis, shows the sources from which the data used in this thesis were extracted, and presents the scope of the study and the research objectives; Chapter 2 presents an overarching literature survey regarding seminal and contemporary studies centred around FDP and the various techniques used by the researchers; Chapter 3 investigates whether companies' financial health is best explained by using a *one-size-fits-all,* or an industry-specific approach, and whether independent variable importance differ amongst industries; Chapter 4 presents an FDP case study on the Australian mining industry, through examining whether machine learning techniques outperform traditional statistical techniques, as well as presenting a method for dealing with class imbalance; Chapter 5 inspects how to deal with a class imbalanced data-set through applying Synthetic Minority Oversampling Technique (SMOTE) to create a balanced data-set, then testing whether the *SMOTEd* data-set outperformed the original data-set using a variety of techniques; Chapter 6 focuses on constructing a novel and user-friendly Financial Distress Prediction Index (FDPI) which ranks companies as per their financial health; Chapter 7 briefly examines the differences between Islamic and conventional banking, and creates three FDP models to outline the most important variables in predicting Islamic banks' financial distress; and finally, Chapter 8 applies SMOTE to imbalanced data-sets comprised of SMEs and LCs, and creates FDP models to test for variable differences amongst LCs and SMEs, as well as presenting the empirically superior model. Chapter 9 presents overarching conclusions of the studies carried out in this thesis, the limitations of the research conducted, and prospects for future works.

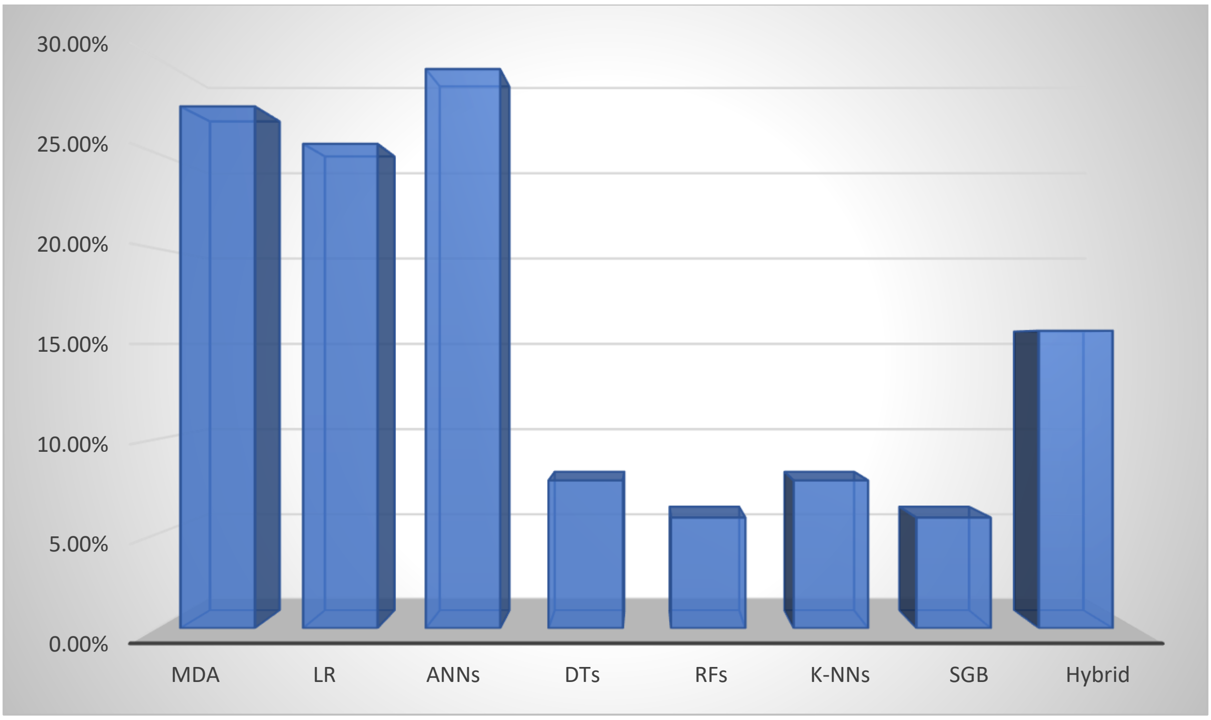
Chapter 2: Literature Review

The first chapter introduced the concept of FDP, outlined the gaps in the literature, the motivations behind picking this topic, the potential gains of exploring the area of FDP further, and highlighted the research questions and hypotheses to be investigated in this dissertation. This chapter provides an overarching survey of the available literature, introducing seminal and contemporary research alike. Each subsequent chapter will provide a specific literature review dealing with the topic introduced in each respective chapter.

In this thesis, when referring to FDP *techniques*, the meaning refers to the overarching, generic algorithms and procedures for dealing with a set of issues, this includes both traditional statistical techniques, such as LR and MDA, as well as machine learning techniques, such as ANNs and SGB. On the other hand, FDP *models*, are the particular models constructed using any FDP technique based on specific data-sets and explanatory variables. For example, researchers might adopt seminal statistical techniques, such as LR or MDA, but when applying them in their FDP research, they create models based on the aforementioned techniques. Examples of such FDP models will be presented throughout this thesis.

Numerous models have been developed over the years that deal with FDP using various techniques. They vary in the methodologies they utilise to achieve their results; however, their core aims tend to be similar, that is, analysing variables or achieving the most possible accurate predictions – refer to Figure 2.1 below for a visual comparison of FDP techniques used in prior studies. As is evident in the figure, MDA, LR, and ANNs make up the lion's share of techniques used in the literature. Burgeoning machine learning techniques like RFs and SGB are used in fewer studies, however, due to their superior performance vis-à-vis traditional statistical techniques, they are likely to become more popular in the coming years. The percentages were calculated by reviewing 220 peer-reviewed journal articles, books, conference papers, and other publications from the literature pertaining to FDP, and subsequently classifying them as per the technique(s) used. The total is more than 100% since some studies use multiple techniques in their research. The various studies were extracted from different portals, such as: Google Scholar and Bond University's Online Library. The research were selected based on reviewing a wide variety of both seminal and contemporary works published in reputable journals, and by following trails within each study.

Figure 2. 1 Percentage Comparison of FDP Techniques in the Literature



In the literature, the accuracy of a model's prediction is generally determined by the Type I and Type II error rates. Type I error refers to misclassifying a failing business as successful, whereas Type II error refers to misclassifying a successful business as a failing one. Type I error results in a realised financial loss caused by participation with a business that is doomed to fail, for example: losing money or shares invested in a potentially failed company. Whereas, Type II error results in a lost opportunity cost from participating with a successful business, for example: missed investment gains from not investing in a potentially successful company. However, it is important to note that the weights of each error type are not necessarily equal, that is, these costs may vary according to the stakeholder or circumstance (Gepp & Kumar, 2012). For instance, a risk-averse person might assign a higher weight to Type I errors, as they are more concerned with a realised financial loss vis-à-vis missed opportunities; whereas, risk-seeking people might assign higher importance to Type II errors, as they are more concerned with potential gains from their investments. From a statistical analysis point of view, the actual amount is not important, but rather the ratio of the two costs. Type I and Type II errors were introduced here as they will be mentioned throughout the thesis.

2.1 Univariate Analysis

The prediction of financial distress for businesses has been extensively researched ever since the early 1930s, pioneered by FitzPatrick (1932), followed by Winakor and Smith (1935) who found that trends in certain financial ratios can lead to bankruptcy. These studies were furthered by Beaver (1966) through establishing the first statistical model – Univariate Analysis, which used financial ratios individually for FDP. Beaver used 30 financial ratios in his research. A classification model was conducted separately for each ratio to determine an optimal cut-off point with the goal of minimising misclassification. He tested his models on 158 large businesses for the time-period 1954-1964, half of which were successful and the other half failed. Beaver adopted paired sampling for determining the accuracy of ratios and developing his models. Beaver considered a business to be failed if it had gone into bankruptcy, there was an overdrawn bank account, a miss out on preferred stock dividends, or a defaulted debt. He established a set of ratios with the greatest predictive power, namely:

- Cash Flow to Total Debt;
- Net Income to Total Assets;
- Total Debt to Total Assets;
- Working Capital to Total Assets;
- Current Ratio (Current Assets to Current Liabilities);
- No Credit Interval (Defensive Assets minus Current Liabilities to Fund Expenditures for Operations).

Beaver's model had approximately 22% Type I error and 5% Type II error. However, this was not time-constant, that is, the amount of error increased as the length of prediction increased, which is problematic for long-term predictions. Another issue faced by Beaver's model was that various ratios could result in conflicting predictions, and so the models would cease to be feasible (Gepp & Kumar, 2012).

2.2 Multivariate Discriminant Analysis

After Beaver's univariate analysis, Altman (1968) founded the first multivariate statistical approach pertaining to FDP – Multivariate Discriminant Analysis (MDA). Altman's model was designed to address the main issue faced by Beaver's models, that being, different ratios could result in conflicting predictions. Altman devised a single weighted score (Z) for each business based on five variables. The variables were financial ratios but excluded cash flow ratios as they were not found to be statistically significant, hence contrasting Beaver's model. The ratios used in Altman's (1968) paper are as follows:

- x_1: Working capital divided by total assets,
- x_2: Retained earnings divided by total assets,
- x_3: Earnings before interest and tax divided by total assets,
- x_4: Market value of equity divided by book value of total liabilities,
- x_5: Sales divided by total assets.

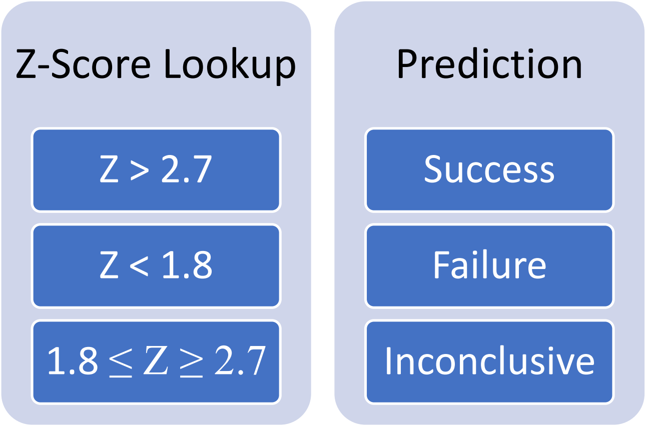
The single weighted score (Z) was calculated according to the following equation:

$$Z = 1.2x_1 + 1.4x_2 + 3.3x_3 + 0.6x_4 + 1.0x_5 \qquad [Equation\ 2.1]$$

- Z = Discriminant Score of a Company
- x_i = Independent Variables (the five abovementioned financial ratios)

Altman (1968) analysed how well financial ratios performed in predicting financial distress of manufacturing firms whose assets ranged from $0.7 million to $25.9 million. His sample included 66 businesses (33 bankrupted and 33 non-bankrupted). Each company's Z-score was referenced with cut-off scores that determined the financial health of the company – this is presented in Figure 2.2. Altman's model outperformed that of Beaver's, as the short-term accuracy of the model was 95%; however, that drops down to 72% when it is predicting bankruptcies two or more years in advance. Therefore, the long-term issues persisted, that is, Altman's model was only viable for short-term predictions.

Figure 2. 2 Altman's Z-score Model



As was shown in Figure 2.1, MDA is one of the most popular techniques in the literature for analysing financial distress – this claim was also issued by Perez (2006). MDA has been used in many FDP studies, including: Altman, Iwanicz-Drozdowska, Laitinen, and Suvas (2017); Chung, Tan, and Holdsworth (2008); Grice and Ingram (2001); Le and Viviani (2018); Lee and Choi (2013). The main benefit of the MDA technique for predicting financial distress is its capability to reduce a multidimensional problem to a single score with a fairly high level of accuracy, thus overcomes the problem identified with the Beaver's univariate model.

However, MDA has a few disadvantages in the form of being subjected to various restrictive assumptions. Firstly, MDA requires the decision set that is used for differentiating between bankrupt and non-bankrupt businesses be linearly separable. Secondly, unless an interaction term is introduced, MDA does not allow a ratio's signal to fluctuate based on its relationship with another ratio, or set of ratios in the model (Veal, 2005). Although, in practice, a ratio can signal financial distress if it is below or above the normal value. These problems, along with issues such as the multivariate assumption of normality, multicollinearity, bias of extreme data points and equal group variance-covariance matrix, might confirm that MDA is unfitted to the complex nature, interrelationships, and boundaries of financial ratios (Coats & Fant, 1993). It remains, however, widely used and a good benchmark (Altman, Iwanicz-Drozdowska, Laitinen, & Suvas, 2014). There are other forms of MDA, such as quadratic discriminant analysis that can overcome some of the drawbacks mentioned. The form of MDA discussed earlier and commonly used is linear discriminant analysis.

Li (2012) examined corporate failures in the United States between 2008-2011. Three models were created, namely: Altman's original Z-Score model, a re-estimated Z-Score model and a re-estimated model with an added variable. The ratio with the highest predictive power was found to be 'Market Value of Equity/Total Liabilities'. To address the failure of the Altman's (1968) model to include a measure of asset volatility, a new variable was added to the re-estimated model, namely: 'Total Assets One Year Prior to Bankruptcy – Total Assets Two Years Prior to Bankruptcy)/Total Assets Two Years Prior to Bankruptcy.' Li's results indicated that Altman's original model performed with predictive accuracy rates ranging from 80% -94%. The re-estimated model accurately predicts 70% of bankrupt firms for one year prior to bankruptcy. Using data from two years prior to bankruptcy, the re-estimated model accurately predicted 92% of bankrupt companies. The third model's results were the most accurate, correctly classifying 96% of companies. However, all three models yielded unencouraging Type II results. The added variable did not add value to the model.

Chung et al. (2008) applied FDP modelling on firms based in New Zealand using MDA. Their results showed that prior to failing, companies had low profitability, higher leverage ratios, less liquidity, and lower asset quality. Their findings also showed that financial ratios have different predictive abilities for detecting financial distress in New Zealand finance companies, and the ratios of failed versus non-failed companies vary substantially. Altman et al. (2014) applied Altman's (1968) Z-score model to multinational firms, as well as using additional variables, re-estimation, and using another statistical method to test for the effect of classification performance. Their results showed that the original Z-score model performed well in an international context, the re-estimation of the coefficients using MDA marginally improved classification performance, and the use of additional variables generally improved classification accuracy of the original model. However, the results vary by country, hence implying that a country-specific model will be more accurate – this justifies developing Australian-specific models like the ones used in this research.

2.3 Logistic Regression

2.3.1 Standard Logistic Regression

As was shown in Figure 2.1, LR is one of the most popular models for forecasting financial distress, some of the prominent studies using LR pertaining to FDP include: Chen (2011); Collins and Green (1982); Daniel and Ionuţ (2013); Hall (1994); Hua, Wang, Xu, Zhang, and Liang (2007); Laitinen and Laitinen (2001); Laitinen and Kankaanpaa (1999); Le and Viviani (2018); Min and Lee (2005).

Analogous to MDA, LR devises a score for each company, but unlike MDA, it is not affected when assumptions of equal variance-covariance and normality of the variables are violated (Altman & Hotchkiss, 2010). Ohlson (1980) pioneered the application of LR to forecast business financial distress. Comparable to the Z-Score devised by Altman (1968), Ohlson's O-score can be labelled as a statistical financial distress indicator produced from a predefined set of variables. In his ground-breaking study, three distinct logistic regression models were produced to predict financial distress for one, two, and three years in advance. The variables selected in the study comprised standard financial ratios, dummy variables based on comparisons of balance sheet numbers, and a variable demonstrating the change in net income over the past year. He devised a probabilistic model of bankruptcy, where the logarithm of the likelihood of any specific outcome, as reflected by the binary sample space of financial health vis-à-vis financial distress, is shown by the following equation:

$$l(\beta) = \sum_{i \in S_1} logP(X_i, \beta) + \sum_{i \in S_2} \log(1 - P(X_i, \beta)) \qquad [Equation\ 2.2]$$

- X_i = Vector of Predictors for observation i
- β = Vector of Unknown Parameters
- $P(X_i, \beta)$ = Probability of Bankruptcy for X_i and β
- S_1 = Set of Bankrupt Companies
- S_2 = Set of Healthy Companies

To remedy for the problem of selecting appropriate class functions of *P*, Ohlson developed the following logistic function, presented in Equation 2.3 below.

$$P = 1 + e^{(-\gamma_i)^{-1}} \qquad [Equation\ 2.3]$$

The implications of the above logistic function are twofold: firstly, P is increasing in γ; and secondly, $\gamma = \log(\frac{P}{1-P})$, hence making the model more statistically valid and easily interpreted (Ohlson, 1980).

Ohlson's developed his model using a much bigger sample than that of Altman's. Ohlson's sample included 2,058 successful businesses and 105 failed businesses. Ohlson's empirical results were not encouraging, for example, his first model yielded a Type I error of 63% at the 0.50 cut-off mark.

Despite Ohlson's empirical results being unencouraging, later studies used LR to developed FDP models. Collins and Green (1982) compared forecasting results by using an LR model, an MDA model, and a linear probability model. Their results demonstrate that the logistic model performs better. Hall (1994), created a logistic model with nonfinancial variables and the model could differentiate bankrupt businesses from non-bankrupt ones with an impressive accuracy rate of 95%. Also, various later studies on logistic regression have shown that it is typically marginally empirically superior to discriminant analysis in both prediction and classification accuracy, for example: Laitinen and Kankaanpaa (1999); Min and Lee (2005). In Chen's (2011) study, LR was found to have better prediction accuracy for long run predictions (more than one and a half years) when compared to decision trees – to be discussed is Section 2.6.1. Daniel and Ionuț (2013) conducted FDP tests using LR on companies in Romania; their results yielded 70% accuracy in predicting bankruptcy over a five-year period.

2.3.2 Bayesian Logistic Regression

According to Tsai (2005), statistical inferences are generally based on Maximum Likelihood Estimation (MLE). MLE picks the parameters that maximize the likelihood of the data. In MLE, parameters are presumed to be unknown but fixed, therefore, can be estimated with a degree of confidence. However, in Bayesian statistics, the uncertainty about the unknown parameters is quantified by the means of probability in order for those parameters to be considered random variables. Bayesian inference is the manner of analysing statistical models with the inclusion of prior knowledge about the model or its parameters; the root of such inference is the Bayes' theorem, which is presented in Equation 2.4 below:

$$P(parameters|data) = \frac{P(data|parameters) \times P(parameters)}{P(data)} \propto likelihood \times prior$$

$$[Equation\ 2.4]$$

Bayes' theorem suggests that an update to the knowledge regarding the distribution of an unknown parameter is achievable if its prior information is known. Bayesian statistics assumes that there are precise distributions for the unknown parameters. It fits the probability model of interest through incorporating prior information relating to the unknown parameters and the likelihood function of the observed data to generate a posterior probability. Bayesian model is especially beneficial when there is limited amount of data available (Tsai, 2005). Bayesian logistic regression is not used in this thesis due to the abundance of data available.

Two similar studies, Chaudhuri (2013); He and Trabelsi (2013) used Bayes' theorom to examine the effect of cut-off points, business cycle, and sampling procudure on the accuracy of FDP. Four models were created and different cut-off points selected to find the optimal FDP model. The study was conducted on U.S. firms. The results show that the Hazard logit model had the highest predictive power when ratio of costs is equal, however the Bayesian and Rough Bayesian models have higher predictive powers when the ratio of cost of Type I error to Type II error is high. This makes the Bayesian models a preferable option due to consistency across all of the sampling methods.

A recent paper by Shrivastava, Kumar, and Kumar (2018) applied LR and Bayesian techniques on a panel data-set comprising 628 Indian companies (341 financially healthy, 287 distressed) for the 2006-2015 time-period. 15 variables were used in their study. The Bayesian model's predictive accuracy outperformed that of LR by a marginal amount – 98.9% versus 98.6%, both being very accurate models.

2.3.3 Dynamic Panel Data Logistic Regression

Unlike cross-sectional data, where studies are conducted at a particular point in time, panel data uses both cross-sectional and time-series data for the study, which is arguably a more realistic way of conducting research, especially pertaining to FDP modelling (Bond, 2002). Dynamic panel data tests using lagged dependent variables for past periods. The dynamic panel data model is presented Equation 2.5:

$$\delta_{it} = \alpha\delta_{i,t-1} + (\theta_i + \mu_{it}) \qquad |\alpha| < 1; \qquad i = 1,2, \dots , N; \qquad t = 2,3, \dots , T$$

$$[Equation\ 2.5]$$

- δ_{it} is an observation for individual i in period t;
- $\delta_{i,t-1}$ is δ_{it} in the previous period;
- θ_i is an unobserved individual-specific, time-invariant effect that allows for heterogeneity in the means of δ_{it} series amongst individuals;
- μ_{it} is a disturbance term.

2.4 Support Vector Machines

Support Vector Machines (SVMs) are supervised learning models used for classification and regression analysis. SVMs are based on Statistical Learning Theory (Boser, Guyon, & Vapnik, 1992). Basically, the way SVMs function is that input vectors are mapped in a nonlinear fashion to a high dimension feature space. SVMs can change complex issues into simpler ones that are able to use linear discriminant functions, through creating a linear decision surface in the feature space. The SVM technique does not concentrate on all of the training data, rather, it concentrates on the data points that are extremely difficult to identify, this is because when it identifies those points, the others are easily seen. The vectors that are the hardest to identify and can be easily misclassified are found close to the hyperplane (in the case of FDP, separating healthy and distressed companies) – these are called support vectors. The margin is the distance from the closest data points in each particular class to the hyperplane. SVMs try to maximise these margins, so that the hyperplane is at an identical distance from both groups (healthy and distressed companies). The advantage of SVMs is that they combine the strengths of traditional statistical and machine learning techniques. SVMs are applied in numerous fields, including: FDP, image recognition, and bioinformatics (Le & Viviani, 2018; Min & Lee, 2005).

A recent study applying the SVM technique to FDP is by Le and Viviani (2018). They compared the FDP accuracy of statistical techniques, namely: LR and MDA; vis-à-vis machine learning techniques, namely: SVMs, K-NNs, and ANNs. 31 financial ratios were used as variables to model on a data-set consisting of 3000 banks (1562 operational and 1438 failed) in the United States between 2011-2016. Their results indicated that the ANN model had the superior predictive power in determining banks' financial distress with an accuracy of 75.7%. The SVM model performed the worst with an accuracy of 71.6%. The difference between the best and the worst models is very close (less than four percentage points), thus indicating it was a close call amongst all models.

2.5 Artificial Neural Networks

Artificial Neural Networks (ANNs) have been used in many FDP studies, including: Ciampi and Gordini (2013); Coats and Fant (1993); Le and Viviani (2018); Lee and Choi (2013); Tan (2001). ANNs are computerised techniques that can be trained to mimic the cellular connections in the brains of human beings (Hertz, Krogh, & Palmer, 1991). It is made up of interconnected units that process and evaluate the interactions between the units in a complex set of existing data – ANNs can also be used for non-complex data, but their ability to evaluate complex interactions is what sets them apart. ANNs assign weights to the respective inputs to enable the precise deduction of the ultimate outcome (Dorsey, Edmister, & Johnson, 1995). This overcomes the issue of prespecifying interactions between independent variables, because ANNs will model them.

According to Dorsey et al. (1995), there are steps involved in the prediction process of ANNs, these include:

1. Define network typology/structure;
2. Select input variables and determine learning parameters;
3. Train network
4. Optionally test new variables and forecast.

Odom and Sharda (1990) employed the same financial ratios used by Altman (1968) and applied ANNs to a sample of 129 firms – 65 bankrupt and 64 non-bankrupt businesses. Their training set contained 74 firms (38 bankrupt and 36 non-bankrupt), whereas their testing set contained 55 firms (27 bankrupt and 28 non-bankrupt). In their study, three-layer feed-forward networks are employed and the results are compared to those of MDA. They tested the effects of different levels on the predictive ability of ANNs and MDA. Their model correctly identified all bankrupt and existing businesses in the training sample, as opposed to 86.8% accuracy by the MDA model. As for the performance with holdout samples, ANNs had an accuracy rate of above 77%, whereas MDA's accuracy rate was between 59% - 70%. Thus, ANNs were much more accurate in both training and test results.

Following Odom and Sharda (1990), a multitude of studies further investigated the use of ANNs in FDP. For example, Salchenberger, Cinar, and Lash (1992) presented an ANN approach to predict bankrupt loans and save businesses from financial distress. The results found ANNs to be as good as or better than the LR models across three different lead times of 6, 12 and 18 months. A paper by Lee and Choi (2013) is one of few that talks about industry-specificity pertaining to FDP – they tested their hypotheses on 229 Korean companies (91 failed and 138 operating). They used MDA and Back-propagation Neural Networks (BNNs) to developed their FDP models on construction, retail, and manufacturing industries – refer to Chapter 3 for elaboration on BNNs. Their results found that the prediction accuracy is improved for industry-specific prediction modelling vis-à-vis an industry-wide model for all models, and that the BNN models outperformed all MDA models. Their models also indicated the most important variables for each industry, and the variables differed amongst industries, thus entailing a need to account for industry-specificity when modelling for multi-industry FDP. Another study by Coats and Fant (1993) examined 282 firms between 1970-1989. Their results suggested that BNNs outperformed MDA, correctly predicting 80% of failed companies with a lead time of up to four years. Ciampi and Gordini (2013) applied their financial distress prediction study on 7,000 Italian small enterprises. Their results showed an ANN predictive superiority when compared with logistic regression and MDA.

ANNs have many advantages, including: they do not need the pre-specification of a functional form, nor do they require the adoption of restrictive assumptions regarding the characteristics of statistical distributions of the variables and errors existing in the model; ANNs are able to function with inexact variables as well as with changes to the model over time; they have an adaptability feature to the presence of new cases that signify changes in the situation. On the other hand, some of the limitations of ANNs include: creating oscillating behaviour in the learning stage, the learning stage can be very prolonged and tedious, and ANNs may not attain a steady absolute minimum cost, but might lock on local minimums without the capability to move to the global optimum (Altman, Marco, & Varetto, 1994).

2.6 K-Nearest Neighbours

There are only a handful of studies that deal with business financial distress prediction using K-Nearest Neighbours (K-NNs), some of these studies include: Chen et al. (2011); Le and Viviani (2018); Park and Han (2002). K-NN is a versatile and simple machine learning and data mining technique that is a non-parametric learning method that may be applied for regression and classification modelling. The model development method is comprised of the K-nearest training instances in the feature space – refer to Chapter 5 for an elaboration on the feature space concept. For both classification and regression, weighting the contributions of the neighbours is essential, in order for the nearer neighbours to contribute more to the average than the distant ones. In classification, the output is a class member, however, in regression, the output is the property value of the entity (Altman, 1992).

Park and Han (2002) used financial and non-financial ratios and proposed a weighting approach on the K-NN algorithm to predict financial distress. Their model outperformed the traditional K-NN algorithm through showcasing enhanced modelling in FDP, as their results indicated increased classification accuracy and a justification to incorporate qualitative criteria alongside the quantitative.

Chen et al. (2011) devised a novel model for FDP; in their study, an adaptive fuzzy K-NN method was applied to FDP. Fuzzy K-NN allocates degrees of membership to various classes while considering the distance of its k-nearest neighbours. This means that all the instances are assigned a membership value in each class rather than binary decision of ‘failed’ or ‘non-failed’. Their model outperformed five counterpart cutting-edge classifiers in terms of Type I and Type II errors and Area Under the Receiver Operating Characteristic (AUROC) criteria. They also pointed out the best discriminative ratios pertinent for FDP.

2.7 Recursive Partitioning Techniques

Recursive partitioning refers to a set of machine learning techniques for multivariate analysis. They are intelligent, nonparametric classification or regression that evolved to lessen or remove the distribution assumptions associated with parametric techniques, such as MDA, LR, and others (Breiman, Friedman, Stone, & Olshen, 1984). These models are more versatile and have a wider scope than traditional models, since they can handle nominal variables, outliers, nonlinear relationships, interactions, missing values, and qualitative variables, hence making them more broadly applicable than traditional parametric techniques (Zhang & Singer, 2010). On the downside, there is no formal test for assessing the statistical significance of variables (Altman & Hotchkiss, 2010). Some examples of recursive partitioning techniques include: Decision Tree (DTs), Random Forest (RFs), and Stochastic Gradient Boosting (SGB). These techniques will be elaborated upon in the following subsections. Due to their recent invention, relative to parametric models, they are naturally less occurrent in the literature, however, they are slowly gaining traction because of their superior predictive capabilities (Gepp, 2015).

2.7.1 Decision Trees

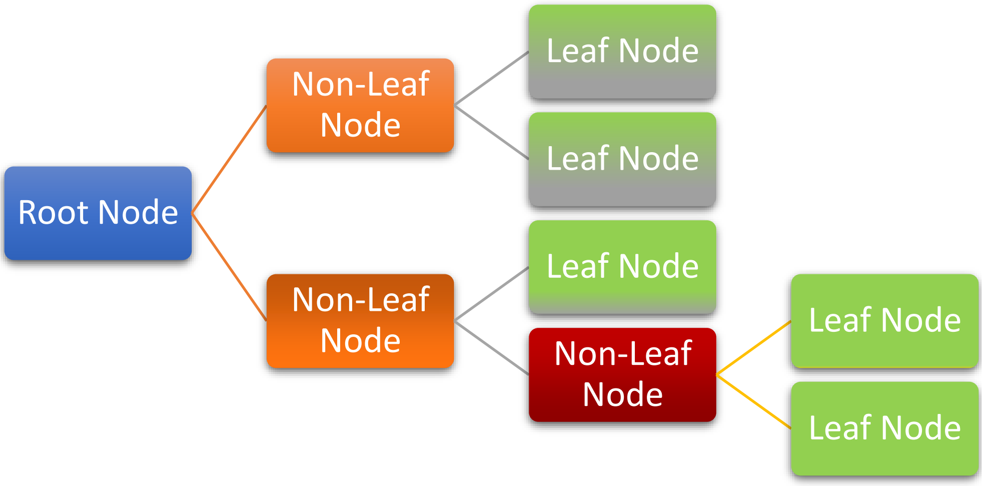
Decision Trees (DTs) have not been as extensively used in FDP studies vis-à-vis their parametric counterparts. Some of the studies that apply DTs to FDP include: Chen (2011); Geng, Bose, and Chen (2015); Gepp, Kumar, and Bhattacharya (2010); Hung and Chen (2009); Sun and Li (2008).

Decision Trees (DTs) are models that construct a set of tree-based classification rules that recursively break down a data-set into smaller and smaller subsets (partitions). The tree is generated in a recursive process that splits the data from a higher level to a lower level of the tree, ending with leaf nodes that characterise classification groups (distressed or successful). When applied to FDP, DTs commonly assign businesses to either the successful or distressed group. The splitting at each node is determined by comparing an expression that is assessed for each company with a cut-off point. There are two main tasks for the algorithms that generate DTs. First, to choose the optimal splitting rule at each non-leaf node to differentiate between distressed and successful companies, and secondly, to determine the number of nodes in the decision tree (Gepp & Kumar, 2012). A sample DT can be seen in Figure 2.3 below.

DTs consist of the following:

- A root node: Topmost decision node that corresponds to the best predictor
- Non-leaf nodes (non-leaf nodes project 2 branches leading to 2 distinct nodes)
- Leaf nodes: Represents a classification or decision
- Connecting branches: connecting nodes

Figure 2. 3 Decision Tree



A drawback of DTs is that they do not provide precise probabilities of group membership, that is, financial distress – except for a whole node (group of businesses). However, DTs are beneficial for many reasons, including: invariance to monotonic alterations of input variables, handling outliers in the data effectively as well as mixed variables, and being able to deal with a data set that contains missing data. There are different algorithms that can be used to generate DTs. These algorithms all create similar tree structures but selecting the correct algorithm for a particular circumstance can have a huge impact on the predictive power of the generated model. Popular implementations of decision trees include Classification and Regression Trees (CART) and See5 (Gepp et al., 2010). In a 2005 pioneering study, Huarng, Yu, and Chen (2005) compared the accuracy of CART and See5; their results showed CART to be empirically superior to See5. However, it is crucial to note that the data-sets encompassed less than 12 businesses and five variables, that is, the sample is too small to obtain reliable results. However, Gepp et al. (2010) confirmed that CART empirically superior to See5, thus solidifying Huarng et al.'s (2005) claim.

According to Gepp et al. (2010), DTs are empirically found to be superior predictors vis-à-vis MDA when it pertains to forecasting companies' financial distress. Studies that solidify this claim include: Chen (2011); Frydman, Altman, and Kao (1985); Kumar and Ravi (2007). When comparing DTs to LR, Chen (2011) found that DTs classification approach yields superior FDP accuracy in the short-run (less than one year), hence implying that ANNs are better predictors in the short-term.

Chen (2011) applied his study on 100 listed Taiwanese companies – 50 distressed vis-à-vis 50 healthy companies – using 37 financial and non-financial ratios that are common in the literature. He used Principal Component Analysis (PCA) to extract suitable variables – PCA will be explored further in Chapter 6. Three DT classification methods were used to create the FDP model; a logistic regression model was also developed for comparison purposes. Chen's FDP model using DTs outperformed his LR model by yielded around 97% accuracy for identifying distressed firms in the short-term (two seasons prior to actual financial distress); however, the LR model marginally outperformed the DT model in the long-term (over one and a half years) by almost three percentage points (91.7% versus 88.8%). Chen concluded that Artificial Intelligence (AI) techniques are superior to traditional statistical techniques in predicting financial distress in the short-term.

Geng et al. (2015) employed data mining techniques to construct three main models for three time-periods preceding the companies' financial distress, using DTs, neural networks, and Support Vector Machines (SVMs). Their study was based on 107 Chinese "Special Treatment" companies, that is, implying financial distress, and the same number of financially healthy companies, for the time-period 2008-2011. They incorporated 31 financial variables in all of their FDP models. Their results showed that the neural network model was the most accurate at predicting financial distress, closely followed by the DT model. 'Net Profit Margin of Total Assets, "Return on Total Assets", "Earnings per Share", and "Cashflow per Share" were the financial indicators with the highest predictive capability in pointing out financial distress.

Gepp et al. (2010) provided a classic case of the Occam's razor philosophical principle, that being, the most parsimonious models are better than more complex ones. They employed 20 financial variables and applied it on the original data-set used by Frydman et al. (1985), comprising 200 businesses, and conducted a cross-sectional analysis. They devised DT models using different implementations of DT, including: CART, See5, and Recursive Partitioning Analysis (RPA). See5 yielded the best in-sample classification capability, but the poorest predictions. CART and RPA were the best overall predictors. The three DT models were compared with MDA and they outperformed it. Profitability and liquidity ratios were the most important variables at predicting financial distress.

Hung and Chen (2009) used 30 financial ratios that are common in the literature on a data-set consisting of 56 bankrupt companies and 64 healthy companies, for the time-period 1997-2001. They proposed an ensemble method of three classifiers, namely: DTs, BNNs, and SVMs in an attempt to harness their pooled advantages, all the while mitigating the individual disadvantages of each technique. Their selective ensembles outperform weighting and voting ensembles for FDP by around 2.5 percentage points.

Sun and Li (2008) incorporated 35 financial ratios and applied them on 198 listed Chinese companies, of which 92 are financially distressed and 106 are financially healthy, for the time-period 2000-2005. They present a data mining method which includes attribute-oriented induction, information gain, and DT. Adopting entropy-based method, their model achieved a prediction accuracy rate of 95.33%.

2.7.2 Random Forests

Random Forests (RFs) is an ensemble learning method for regression and classification that consists of creating many decision trees. In classification, the output is the mode of the classifications of the individual trees. In regression, the output is the mean from every generated tree. As part of their intrinsic structure, RF predictors lead to a dissimilarity measure between the observations. One can also define a RF dissimilarity measure between unlabelled data. The idea is to build a RF predictor that distinguishes the observed data from suitably created synthetic data. RF has similar advantages to single trees, such as: handling mixed variables effectively, invariance to monotonic transformations of input variables, robustness to outlying observations, and accommodation to different strategies for dealing with missing data (Chandra, Ravi, & Bose, 2009).

Only a handful of studies apply RF to FDP throughout the literature, but it is generally found to be highly accurate because of the multiple trees generated. A study by Fantazzini and Figini (2009) compared a variant of RF, namely Random Survival Forests (RSF) with a standard logistic model. Their findings showed that RF outperforms the logit model for the in-sample, but the opposite is true for the out-of-sample. A pioneering study by Nanni and Lumini (2009) investigated the performance of several systems based on ensemble of classifiers for FDP. Their results showed that Random Subspace, a method were each stand-alone classifier uses only a subset of all features for training and testing, outperformed other ensemble methods.

2.7.3 Stochastic Gradient Boosting

Stochastic Gradient Boosting (SGB) is a dynamic and adaptable data driven tool that creates numerous small decision trees in an incremental error–correcting process. SGB’s versatility enables it to deal with data contaminated with erroneous target labels. Such data are usually extremely problematic for conventional boosting and are a challenge to handle using traditional data mining tools; au contraire, SGB is less affected by such errors. SGB also has a degree of accuracy that is typically not achievable by a single model or ensembles like bagging or conventional boosting. SGB has advantages on ANNs of not being sensitive to erroneous data and requires minimal data preparation time, imputation of missing values, or pre-processing (Mukkamala, Vieira, & Sung, 2008).

As with RF, there are a handful of studies that apply SGB to FDP. Ravi, Kumar, Srinivas, and Kasabov (2007) presented a research on predicting financial distress in financial engineering . They used an alogirthm to train radial basis function neural networks in a semi-online fashion. It incoporated online and evolving clustering alogirthms and the traditional least squeares estimation. Their results showed that their algorithm outperformed other neural netowrk techniques, however, SGB outperformed their alogrithm in both data-sets. Another study by Ravi, Kurniawan, Thai, and Kumar (2008) presented an ensemble system with a multi-faceted statistical technique constituency to predict financial distress of banks. They adopted a novelty method to use SGB for feature selection (selecting the top five predictor variables), and then added them to the fuzzy rule based classifier. Their results yielded lower Type I and Type II errors vis-à-vis the constituent models in stand-alone mode.

2.8 Hybrid Models

Hybrid models pool various individual statistical techniques in order to maximise their advantages, all the while minimising the combined model's disadvantages. The idea is, the advances achieved by certainty and precision in more traditional methods, such as: MDA and LR, are not justified by their costs (Kumar & Ravi, 2007).

A study by McKee and Lensberg (2002) presented a hybrid financial diagnosis model combining rough sets and genetic programming. Their sample comprised 291 businesses from the U.S. for the time-period 1991 to 1997 using 11 variables to describe the cases. They concluded that the hybrid model reaches a Type I and Type II error rates of 20%, that is, average predictive accuracy rate of 80% on the validation set, whereas the simple rough-set performs significantly lower on the same data-set achieving an average accuracy rate of 67%).

Another study by Ahn, Cho, and Kim (2000) worked on combining neural networks and rough sets for business financial distress prediction. They used Korean data for the time-period between 1994 and 1997 and compared their results to different standard neural network techniques. Their model's predictive accuracy rates exceeded 80% in many instances.

Lee, Han, and Kwon (1996) developed hybrid neural network models for predicting financial distress on Korean firms. Their results showed that integrating unsupervised with supervised learning yields more accurate predictions.

Other studies applying hybrid models to FDP include: Chandra et al.'s (2009) study which presented a novelty study to predict the financial distress of 240 dotcom companies using hybrid intelligent systems, which included RFs, LR, and CART, to name a few. Their results yielded high accuracies for all the techniques, even superseding previous studies' accuracy rates on the same data-set; Tinoco and Wilson's (2013) study which tested 23,218 company-year observations for the time-period 1980-2011. They combined accounting, market-based, and macroeconomic data to predict financial distress. When benchmarked against Altman (1968) Z-score model and neural network models, their results were more accurate in terms of both Type I and Type II errors. The macroeconomic variables contributed only marginally to the overall classification accuracy of the model; and finally, an extensive review carried out by Kumar and Ravi (2007) investigated papers cenetred around FDP of banks and firms for the time-period 1968-2005. They categorised the research based on the techniques used in each study. Their results showed that statistical techniques in stand-alone mode are no longer used, and among the stand-alone intelligent techniques, ANNs were most ofen adopted. However, they found a trend emerging to build hybrid intelligent systems to predict financial distress, and that ensemble classifiers outerpform individual techniques.

Refer to Table 2.1 in the following section for a consise summary of the most important points of the various statistical and machine learning techniques that were presented in this section.

2.9 Statistical and Machine Learning Technique Summary

A summary of various statistical and machine learning techniques in the context of FDP, along with their relative advantages and disadvantages, is given below in Table 2.1.

Table 2. 1: Statistical Technique Comparison

Technique	Advantages	Disadvantages
Univariate Analysis	1. Simple to use;	1. Various ratios results in conflicting predictions (Gepp & Kumar, 2012);
	2. High short-term predictive accuracy.	2. Predictive accuracy declines for long-term predictions.
MDA	1. Extensively used throughout literature;	1. Multicollinearity problem;
	2. Simple to use;	2. Predictive accuracy declines for long-term predictions;
	3. High short-term predictive accuracy;	3. Decision set needs to be linearly separable;
	4. Reduces multidimensional problems to an accurate single score.	4. Affected when basic assumptions are violated.
LR	1. Less affected than MDA when basic assumptions are violated;	1. Predictive accuracy declines for long-term predictions (Altman & Hotchkiss, 2010);
	2. Extensively used throughout literature.	2. It may require more data than MDA to achieve reliable results.
ANNs	1. They do not need the pre-specification of a functional form;	1. The learning stage can be very long;
	2. They are able to function with imprecise variables.	2. A steady absolute minimum cost may not be attained, but may lock on local minimums without moving to the global optimum.
K-NNs	1. No assumptions about the concepts' characteristics to learn need to be executed.	1. It is computationally expensive to find the K-NNs when the data-set is large.
Recursive Partitioning: DTs, RFs, SGB	1. Eliminates some problems faced in parametric techniques, e.g.: distribution assumptions with variables;	1. Harder to interpret than parametric techniques, DTs excepted;
	2. They can handle qualitative variables and are immune to outliers and irrelevant variables.	2. No formal test of variable significance.
Hybrid	1. Combines advantages of various models & minimises disadvantages.	1. Can be complex, less user-friendly, and difficult to interpret.

2.10 Conclusion

This chapter has expanded on the topic of FDP and presented an overarching review of the literature pertaining to the topic at hand. More than 200 studies that use statistical and machine learning techniques were investigated from their inception to contemporary research. There are gaps found in the literature that justify the research conducted in this thesis, including: FDP research mostly on non-industry-specific basis; machine learning techniques are not used extensively, despite their tendency to yield more accurate results vis-à-vis traditional statistical techniques; and scarcity of FDP studies centred around Australia. The literature overwhelmingly show that machine learning techniques tend to outperform traditional statistical techniques in terms of predictive accuracy. This can be explained due to a number of factors, including:

- Traditional statistical techniques generally use a default cut-off of value of 0.5 when classifying companies as healthy or distressed – this is not always an accurate representation of reality, especially when data-sets have a class imbalance issue; whereas, machine learning techniques are usually impervious to this issue and can provide optimised cut-off points for each model – this will be explored in later chapters;

- Predictive accuracy of statistical techniques are generally measured by the Type I and Type II errors in the model and/or simple averages of classification accuracy – again, in some instances this does not reflect real-life situations, therefore other methods of measurement available in machine learning techniques can offer a more accurate representation of reality, including the Receiver Operating Characteristic (ROC) graph – to be explored in later chapters;

- The machine learning techniques' algorithms are intrinsically far more complex than their statistical counterparts, thus enabling them to utilise computing power to analyse data in ways that are virtually impossible for the statistical techniques to do, for example, RF and SGB techniques can generate thousands of trees to find out the most accurate result; and,

- The machine learning techniques are generally not constrained by many of the restrictive assumptions of the statistical technique, thus rendering them an overall more versatile and effective predictive tool.

As explained earlier, the literature survey presented in this chapter is not comprehensive, as this thesis is designed to address various topics in separate chapters. Therefore, each consecutive chapter will include its own literature review section that will be relevant to each chapter’s topic.

Chapter 3: Industry-Specificity*

*This chapter is based on a published paper in a peer-reviewed Journal, namely: Halteh, K. (2015). Bankruptcy Prediction of Industry-Specific Businesses Using Logistic Regression. *Journal of Global Academic Institute Business & Economics, 1*(2), 151-163.

This chapter investigates the predictive accuracy of industry-wide and industry-specific FDP models, outlines the variables that are most important in predicting financial distress in each industry, and experiments on varying the cut-off point pertaining to the LR models in order to showcase how Type I and Type II errors can change in accordance with objective of the user. For example, a lower Type I error may be preferred for a risk-averse person, whereas a lower Type II error may be preferred by a risk-seeking person. This chapter does not compare the predictive accuracies between statistical vis-à-vis machine learning techniques (*Hypothesis 3*) amongst the created models, as this will be done in later chapters.

3.1 Introduction

As mentioned in Chapter 2, only a fraction of the FDP literature is concerned with industry-specificity. Aligning with *Hypotheses 1 and 2* stated in the Chapter 1, namely:

> ***H_1:*** Industry-specific models have a greater ability to predict financial distress when compared to a *one-size-fits-all* industry-wide model.
>
> ***H_2*:** Independent variables differ in predictive importance across the models mentioned in ***RQ_1/H_1***.

This chapter will investigate the effect industry-specificity poses on FDP modelling. There are two main aims to this chapter, namely: to ascertain whether industry-specific models – these are FDP models based on segregating the companies as per each industry they subscribe to – can outperform an industry-wide *one-size-fits-all* model; as well as, to investigate whether the variables most useful to FDP models differ by industry – this is done by checking the statistical significance or variable importance in the developed models. This is done by utilising three techniques to develop aforesaid models, namely: LR, MDA, and ANNs.

Many FDP models test their hypotheses by using a specific set of variables, such as Altman's (1968) five financial ratios, or by using the same variables across various industries in the economy, that is, paying little or no attention to industry-specificity (Gepp & Kumar, 2012). Very few studies paid attention to industry-specificity pertaining to FDP modelling, as will be explored later in the Literature Review section. This chapter's findings contribute to the literature by recommending the construction of tailored industry-specific models which include variables with the highest predictive power for each respective industry.

3.2 Literature Review

As was mentioned in Chapter 2, there are many different techniques that can be applied to create FDP models. This chapter surveys the literature pertaining to studies that apply FDP modelling with an industry-centric focus. Therefore, in order to limit repetition, if the studies mention statistical and/or machine learning techniques whose mechanics were already mentioned in Chapter 2, they will be only briefly explained.

It might come intuitively that variables should have varying effects on different industries. For example, the balance sheet figure 'Total Assets' or 'Enterprise Value', will generally be significantly higher for a company operating in the mining industry, as opposed to, say, a firm in the service or retail industries. Therefore, variables or ratios that include 'Total Assets' may be more informative about companies' financial health that operate in an asset-intensive industry vis-à-vis companies operating in low-asset industries. Despite this intuition, surprisingly, there is a scarcity in the literature of empirical studies that tests for industry-specificity pertaining to FDP.

Most studies concerning FDP modelling have been concentrating on a single industry, or, if many industries are involved, no investigation is undertaken to highlight the differences between the industries. This presented a clear gap in the literature that this study contributes towards. In addition to most of the studies mentioned in Chapter 2, what follows are some examples to add to the long list of studies not paying attention to industry-specificity. He and Kamath (2005) assessed the efficacy of two successful FDP models used by Ohlson (1980) and Shumway (2001) with the aid of a multi-industry sample in discerning between healthy and distressed businesses from a single industry – the equipment and machinery manufacturing industry. Another study by Dewaelheyns and Van Hulle (2006) indicated that models involving financial distress variables defined at both subsidiary and at group levels, provide a significantly improved fit and classification performance. The studies aforementioned did not examine the differences in industries, in terms of prediction accuracy, independent variables, and practically working models. Therefore, the difference in FDP accuracy of industry-specific models vis-à-vis a *one-size-fits-all* model is unclear.

Lee and Choi (2013) is a rare study that investigates industry-specificity pertaining to FDP. They tested their hypotheses on 229 Korean companies, 91 of which were bankrupt, for the time-period 2000-2009. Starting from an initial list of 100 variables, they were later cut down as per statistical significance to each industry. Some of the variables used included: a set of growth, profitability, productivity, liquidity, and asset quality ratios. They used MDA and Back-propagation Neural Networks (BNN) to developed their FDP models on construction, retail, and manufacturing industries. BNNs are supervised learning models that generally have a single input layer, one or more hidden layers, and a single output layer. Every layer of an ANN structure has many neurons, and the output units of a layer serve as input units of its following layer. BNNs mimic the way human brains learn – the main idea behind BNN training is to create the weight of the connection between neurons, so that the squared error sum concerning the actual and predicted values is minimised (Lee & Choi, 2013).

Lee and Choi (2013) used the t-test method for different means between the groups to determine the statistically significant variables for each industry at the 5% level, and then only incorporated those variables in their models. Their results indicated that there are in fact differences between variables pertaining to each industry, for example: for the construction industry – the growth, productivity, and liquidity variables were found significant; for the retail industry – the stability and liquidity variables were found to be significant; and finally, for the manufacturing industry, the growth, productivity, and stability variables were found to be significant. This shows the importance of net profit ratio, operating income, and turnover rate of assets in the manufacturing and construction industry; whereas, retained earnings, and operating cash flow are important in the retail industry.

Lee and Choi's (2013) results indicated that their BNN models outperformed the MDA models across all models, and the prediction accuracy is improved for industry-specific prediction modelling vis-à-vis an industry-wide model across all models by a margin ranging between 6-12%, thus empirically proving the necessity of industry-specificity.

Given the limited literature available with regards to industry-specificity of FDP models, this presented the motivation to further investigate this area. Although industry differences were found in Lee and Choi's (2013) FDP study, however, due to the limited data used in their study, location of companies (Korea), limited number of industries, and limited number of modelling techniques used, further investigation on this area is warranted. To the best of the author's knowledge, there are no studies investigating the effect of industry-specificity in an Australian context.

3.3 Data

MorningStar database has been used to collect data on 803 operating and delisted companies from a number of different industries in Australia – energy, industrials, financial, health, and Information Technology (IT). The financial data collected from the Australian companies is used to conduct a cross-sectional study for the time period 2013-2014. Using a larger data-set than Lee and Choi's (2013) study, increases the validity of the study by improving the chances of representing the population in a fair and unbiased manner, and reduce the chances of falling into sampling error by using a small set of data.

This research uses all available data from the MorningStar database for 'failed' and 'successful' Australian businesses, that is, as per classification by database for company status – listed or delisted, respectively. According to the Australian Securities Exchange (ASX) – which is MorningStar's primary source for obtaining Australian company data – a company is 'listed' if its currently operational, whereas a company is 'delisted' for a number of reasons, including: insolvency, merger, or take-over – hence, collectively implying an element of financial distress leading to delisting of the company (MorningStar, 2016).

A dichotomous variable – coded 1 if the company is healthy and 0 if the company is distressed, was used to refer as the dependent variable for each company. 18 variables were used in the study as predictors – refer to Table 3.1 for a complete list of the variables used in this study. The variables in this study are standard accounting and financial variables that were selected based on use in prior empirical research and literature, as per availability of data, and default classification by the database as variables that are industry-specific.

Table 3. 1 Complete List of Variables

Variable	Description
TR	Total Revenue excluding interest – measured in $
EBIT	Earnings Before Interest and Tax – measured in $
Working Capital	Measured In $
Retained Earnings	Measured In $
Total Equity	Measured In $
NPM	Net Profit Margin = Net Profit / Revenue
ROE	Return on Equity = Net Profit After Tax / (Shareholders Equity – Outside Equity Interests)
ROA	Return on Assets = Earnings before interest / (Total Assets Less Outside Equity Interests)
Enterprise Value	Monetary value of the enterprise – measured in $
Current Ratio	Current Assets / Current Liabilities
Quick Ratio	(Cash + Securities + Accounts Receivable) / Current Liabilities
Cash per Share	Cash / Share
Gross Gearing	Total Debt / Total Equity
Price/CF	Share Price / Gross Cash Flow
Net Gearing	(Total Debt - Cash) / Book Value of Equity
PER	Price per Earnings = Market Value per Share / Earnings per Share
Debt/CF	Gross Debt per Cash Flow
EV/EBITDA	Enterprise Value / Earnings Before Interest, Tax, Deprecation, and Amortisation

Some of the variables presented above did not yield any outcome for companies operating in certain industries when extracting data from the database. This is due to the fact that there are inherent differences across various industries in the economy, for example, a dotcom company may not have physical assets or plants as a mining company would. Naturally, this produces different variables that are only pertinent to the specific industry the company subscribes to. Table 3.2 presents the variables used when constructing each model. The ✓ symbol indicates that the variable presented in the first column was used when creating FDP models for its respective industry (presented in the first row). Whereas, the ✗ symbol indicates that the variable was not used when creating FDP models for its respective industry.

Table 3. 2 Ratios Used in All Models

Variables / Industry	Energy	Financials	Industrials	Health	IT	Industry-Wide
TR	✓	✗	✓	✓	✓	✗
EBIT	✓	✗	✓	✓	✓	✗
Working Capital	✓	✓	✓	✓	✓	✓
Retained Earnings	✓	✓	✓	✓	✓	✓
Total Equity	✓	✓	✓	✓	✓	✓
NPM	✓	✓	✓	✓	✓	✓
ROE	✓	✓	✓	✓	✓	✓
ROA	✓	✓	✓	✓	✓	✓
Enterprise Value	✓	✓	✓	✓	✓	✓
Current Ratio	✓	✓	✓	✓	✓	✓
Quick Ratio	✓	✓	✓	✓	✓	✓
Cash per Share	✓	✓	✓	✓	✓	✓
Gross Gearing	✓	✓	✓	✓	✓	✓
Price/CF	✓	✓	✓	✓	✓	✓
Net Gearing	✓	✓	✓	✓	✓	✓
PER	✓	✓	✓	✓	✓	✓
Debt/CF	✓	✓	✓	✓	✓	✓
EV/EBITDA	✗	✓	✓	✗	✗	✗

3.4 Methodology

Using the SPSS statistical software package, the models were built using all of the extracted company data. Three statistical techniques were used to build the models, namely: LR, MDA, and ANNs. 18 models were constructed (three for each industry) for each of the following industries, using the variables mentioned in Table 3.2.

- Industry-Wide model containing all 803 companies – 15 variables;
- Energy Sector model containing 148 companies – 17 variables;
- Financial Sector model containing 166 companies – 16 variables;
- Industrial Sector model containing 188 companies – 18 variables;
- Health Sector model containing 149 companies – 17 variables,
- IT Sector model containing 152 companies – 17 variables.

3.4.1 Models Created

Both the industry-wide and industry-specific models were created using the LR, MDA, and ANN techniques, as shown below. To limit repetition, the mechanics of each technique will not be restated – refer to Chapter 2 for an elaboration on each technique. Since the objectives of this chapter are to check whether industry-specific models are superior to industry-wide models, and whether variable importance differ by industry; the training and testing methods differed amongst techniques when constructing the models. This is because no comparison between technique superiority is undertaken in this chapter – as this is done in later chapters in the thesis. The models constructed had the following properties:

- **LR Models:** Standard settings were used when creating all LR models, such as, probability for Stepwise: entry = 0.05, removal = 0.1; maximum iterations = 20; cut-off point = 0.5.
- **MDA Models:** Standard settings for classification were used, such as, testing method: tenfold cross validation, all groups count equally towards the prior probabilities, and covariance matrix was used within groups.
- **ANN Models:** Standard settings were used, such as: training the model was based on randomly selecting 70% of cases, and testing on the remaining 30%; automatic architecture selection: minimum and maximum number of units in hidden layer, 1 and 50, respectively; and finally, optimisation algorithm used: scaled conjugate gradient.

3.5 Results

This section showcases the results achieved for all the different models constructed, for both the industry-wide data-set and industry-specific data-sets using two traditional statistical techniques (LR and MDA) and a machine learning technique (ANN). Due to the large number of models constructed, the classification tables and figures will only be shown for the industry-wide models; however, Table 3.6 in Section 3.5.3 provides the empirical results and variable importance of each created model.

3.5.1 Industry-Wide Models

3.5.1.1 LR Model

The model initially contained 15 independent variables*.* Statistical level of significance (α) was chosen to be 10%, this is because exit was set at 10%, therefore what remains in the model is significant at the 10% level. Only three of the independent variables made a unique statistically significant contribution to the model, namely: Working Capital, Current Ratio, and Quick Ratio. The results of the model are based on in-sample testing. The full model containing all predictors was statistically significant, χ^2 (15, N = 803) = 35.62, $p < .003$, indicating that the model was able to distinguish between companies that are listed as distressed or healthy. The model as a whole explained between 43% (Cox and Snell R Square) and 59% (Nagelkerke R Squared) of the variance in company status. As for classification of cases accuracy, the model's overall correct classification was 61.8%.

Table 3.3 below showcases the classification table for LR. As can be seen, due to a default cut-off of 0.5, the model correctly classified 98.6% (1.4% Type II error) of the healthy companies, but only correctly classified 4.5% (95.5% Type I error) of the distressed companies. As is evident, there is a high level of Type I error, this is due to the default cut-off point assigned by the technique (0.5). Experimentation on varying the cut-off points will be explored in Section 3.5.4 to check the effect that poses on Type I and Type II errors .

Table 3. 3 Classification Results for Industry-Wide Model using LR

Observed			Predicted		
			Status		% Correct
			Distressed (0)	Healthy (1)	
Step 1	Status	Distressed (0)	14	300	4.5
		Healthy (1)	7	482	98.6
	Overall %				61.8

3.5.1.2 MDA Model

The Industry-Wide MDA model yielded an unencouraging result for the correctly classifying the classes. Table 3.4 shows the cross-validated results of the MDA model. 49.8% of original grouped cases were correctly classified, and after cross-validation that result fell to 47.9%. This model is not better than a coin flip in discerning whether a company is failed or successful. The top three independent variables that made a unique statistically significant contribution to the model, were: ‘Total Equity’, ‘Enterprise Value’, and ‘Retained Earnings’.

Table 3. 4 Classification Results for Industry-Wide Model using MDA

Observed			Predicted		
			Status		% Correct
			Distressed (0)	Healthy (1)	
Step 1	Status	Distressed (0)	237	77	75.4
		Healthy (1)	341	148	30.3
	Overall %				47.9

3.5.1.3 ANN Model

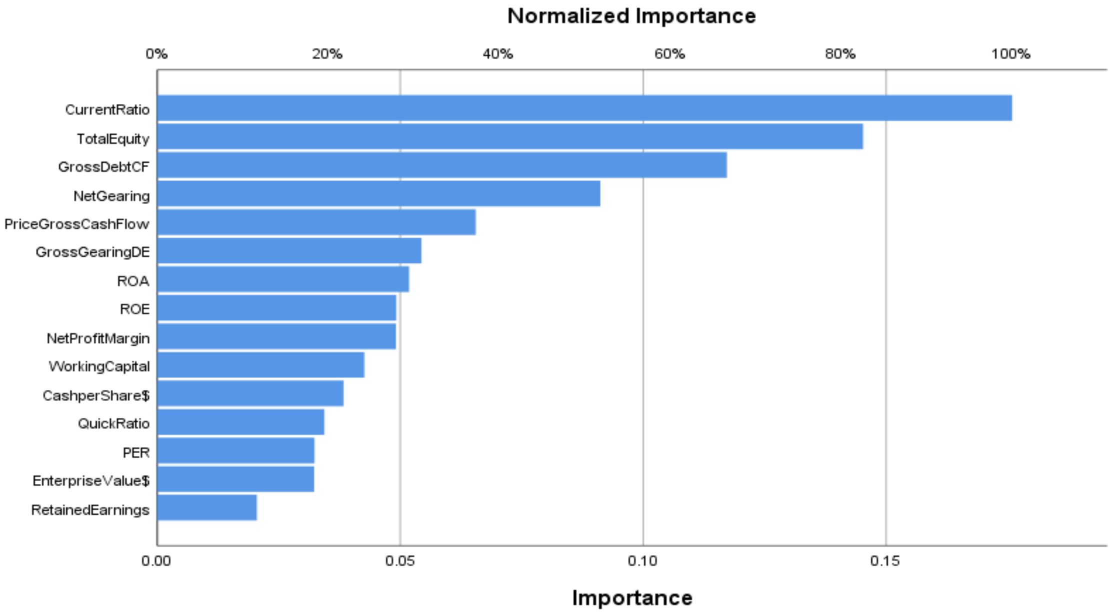
The Industry-Wide ANN model yielded a much better classification result vis-à-vis the LR and MDA models. However, as mentioned in the Methodology section, comparison between techniques cannot be drawn due to the differences when constructing the models. As seen in Table 3.5, 65.3% of cases in the testing group were correctly classified.

Table 3. 5 Classification Results for Industry-Wide Model using ANN

Observed			Predicted		
			Status		% Correct
			Distressed (0)	Healthy (1)	
Step 1	Status	Distressed (0)	1	80	1.2
		Healthy (1)	3	155	98.1
	Overall %				65.3

As for variable importance, the top three variables that had the greatest predictive power in shaping this model were ‘Current Ratio’, ‘Total Equity’, and ‘Gross Debt / Cash Flow’. Figure 3.1 below shows the ‘Independent Variable Importance Analysis’ – they are based on a sensitivity analysis, which calculates the importance of each predictor in determining the neural network.

Figure 3. 1 Industry-Wide’s ANN Model Variable Importance



3.5.2 Industry-Specific Models

This subsection presents the results for the Industry-Specific models for each of the five industries, using LR, MDA, and ANNs. The Energy industry's results are showcased first, followed by the Financials industry, the Health industry, the Industrials industry, and finally, the IT industry. As mentioned earlier, the classification tables and figures are only presented for the industry-wide models, Table 3.6 will showcase the empirical results for all constructed models.

3.5.2.1 Energy

- **LR Model:** The model initially contained 17 independent variables – refer to the Methodology section for list of variables. Six independent variables made a statistically significant contribution to the model (Total Revenue, EBIT, Total Equity, ROE, Enterprise Value, and Cash per Share). The full model containing all predictors was statistically significant, χ^2 (17, N = 148) = 53.33, *p* < .001, indicating that the model was able to distinguish between companies that are listed as failed or successful. The model as a whole explained between 30.3% (Cox and Snell R Square) and 41.2% (Nagelkerke R Squared) of the variance in company status, and correctly classified 77.7% of cases.
- **MDA Model:** The Industry-Specific MDA model for the Energy industry yielded a result of 72.2% for the original grouped cases that were correctly classified, and after cross-validation that result fell to 66.5%. This result is better than all the results out of all the Industry-Wide models. Only one independent variable made a unique statistically significant contribution to the model, namely: 'Cash per Share'.
- **ANN Model:** The Industry-Specific ANN model for the Energy industry yielded an overall classification accuracy result of 82.7%. As for variable importance, the top three variables that had the greatest predictive power in shaping this model were 'Cash per Share, 'PER', and 'Share Price/Cash Flow'. Refer to Discussion section for rationale of variable importance.

3.5.2.2 Financials

- **LR Model:** The model initially contained 16 independent variables. Three independent variables made a unique statistically significant contribution to the model (ROA, EV/EBITDA, and Current Ratio). The full model containing all predictors was statistically significant, χ^2 (16, N = 166) = 30.86, p < .02, indicating that the model was able to distinguish between companies that are listed as failed or successful. The model as a whole explained between 17% (Cox and Snell R Square) and 23% (Nagelkerke R Squared) of the variance in company status, and correctly classified 66.3% of cases.
- **MDA Model:** The Industry-Specific MDA model for the Financials industry yielded a result of 54.9% for the original grouped cases that were correctly classified, and after cross-validation that result fell to 49.2%. This result is only slightly better than the industry-wide MDA model, but it is still largely an unencouraging result. Only two independent variables made a unique statistically significant contribution to the model, namely: ‘PER’ and ‘Gross Gearing’.
- **ANN Model:** The Industry-Specific ANN model for the Financials industry yielded a classification accuracy average of 63.8%. As for variable importance, the top three variables that had the greatest predictive power in shaping this model were ‘Gross Debt per Cash Flow’, ‘Cash per Share’, and ‘Current Ratio’. Refer to Discussion section for rationale of variable importance.

3.5.2.3 Health

- **LR Model:** The model contained 17 independent variables. Two independent variables made a unique statistically significant contribution to the model (EBIT and Quick Ratio). The full model containing all predictors was statistically significant, χ^2 (16, N = 166) = 30.86, p < .01, indicating that the full model containing all predictors was statistically significant, χ^2 (17, N = 149) = 22.8, p < .05, indicating that the model was able to distinguish between companies that are listed as failed or successful. The model as a whole explained between

14.2% (Cox and Snell R Square) and 19.6% (Nagelkerke R Squared) of the variance in company status, and correctly classified 69.8% of cases.

- **MDA Model:** The Industry-Specific MDA model for the Health industry yielded a result of 65.6% for the original grouped cases that were correctly classified, and after cross-validation that result fell to 56.4%. This result is slightly better than the industry-wide MDA model, but it is still largely an unencouraging result. Only two independent variables made a unique statistically significant contribution to the model, namely: ‘Cash/Share’ and ‘Current Ratio’.
- **ANN Model:** The Industry-Specific ANN model for the Health industry yielded a classification accuracy of 75.50%. As for variable importance, the top three variables that had the greatest predictive power in shaping this model were ‘Gross Debt per Cash Flow’, ‘Current Ratio’, and ‘Gross Gearing’. Refer to Discussion section for rationale of variable importance.

3.5.2.4 Industrials

- **LR Model:** The model contained 18 independent variables. One independent variable made a unique statistically significant contribution to the model (Current Ratio). The full model containing all predictors was statistically significant, χ^2 (18, N = 188) = 21.07, $p < .02$, indicating that the model was able to distinguish between companies that are listed as failed or successful. The model as a whole explained between 10.3% (Cox and Snell R Square) and 14% (Nagelkerke R Squared) of the variance in company status, and correctly classified 66% of cases.
- **MDA Model:** The Industry-Specific MDA model for the Industrials industry yielded a result of 60.2% for the original grouped cases that were correctly classified, and after cross-validation that result fell to 53.8%. This result is slightly better than the industry-wide MDA model, but it is still largely an unencouraging result. Three independent variables made a unique statistically significant contribution to the model, namely: ‘Enterprise Value’, ‘PER, and ‘Cash/Share’.

- **ANN Model:** The Industry-Specific ANN model for the Industrials industry yielded an average classification accuracy result of 69.10%. As for variable importance, the top three variables that had the greatest predictive power in shaping this model were ‘Gross Debt per Cash Flow’, ‘Gross Gearing’, and ‘Current Ratio’. Refer to Discussion section for rationale of variable importance.

3.5.2.5 IT

- **LR Model:** The model contained 17 independent variables. Six independent variables made a unique statistically significant contribution to the model (Total Equity, ROE, Enterprise Value, Gross Gearing, PER, and Debt/CF). The full model containing all predictors was statistically significant, χ^2 (17, N = 152) = 48.48, $p < .001$, indicating that the model was able to distinguish between companies that are listed as failed or successful. The model as a whole explained between 27.3% (Cox and Snell R Square) and 36.5% (Nagelkerke R Squared) of the variance in company status, and correctly classified 75% of cases.
- **MDA Model:** The industry-specific MDA model for the IT industry yielded a result of 59.7% for the original grouped cases that were correctly classified, and after cross-validation that result fell to 50.3%. This result is slightly better than the industry-wide MDA model, but it is still largely an unencouraging result. Only one independent variable made a unique statistically significant contribution to the model, namely: ‘Enterprise Value’.
- **ANN Model:** The Industry-Specific ANN model for the IT industry an average classification accuracy result of 58.00%. As for variable importance, the top three variables that had the greatest predictive power in shaping this model were ‘Gross Gearing’, ‘Cash/Share’, and ‘Current Ratio’. Refer to Discussion section for rationale of variable importance.

3.5.3 Models and Variables Comparison

Table 3.6 presents the overall classification accuracies of all 18 constructed models, as well as the most important variables in determining companies' financial distress in each model. As is evident in Table 3.6, only two out of the 15 (≈13%) industry-specific models constructed had slightly worse results than their respective industry-wide model using the same technique (the IT and Financials industries using the ANN technique – these are highlighted in yellow in Table 3.6). In other words, approximately 87% of the time, using an industry-specific model is the superior choice. This is an important finding to FDP of industry-specificity in general, and to the Australian marketplace, in particular. These findings are in concert with Lee and Choi's (2013) findings regarding the superior predictive importance of industry-specific FDP models. These findings are more inclusive due to the use of a larger data-set and more industries.

As for the variable differences amongst the constructed models, as is evident in Table 3.6, differences exist. This an important finding since it demonstrates that each industry is more so affected by different variables, thus management should keep a close eye on the variables that are most important to the industry their company operates in. The Discussion section elaborates on the variable differences amongst the industries investigated in this chapter.

Table 3. 6 Models and Variables Comparison

Technique	Models	Overall % Classification Accuracy	Most Important Variables
ANN	Industry-Wide	65.30%	Current Ratio; Total Equity; Debt/Cash Flow
	Energy	82.70%	Cash/Share; PER; Price/CF
	Financials	63.80%	Debt/Cash Flow; Cash/Share; Current Ratio
	Health	75.50%	Debt/Cash Flow; Current Ratio; Gearing
	Industrials	69.10%	Debt/Cash Flow; Gearing; Current Ratio
	IT	58%	Gearing; Cash/Share; Current Ratio
MDA	Industry-Wide	47.90%	Total Equity; Enterprise Value; Retained Earnings
	Energy	66.50%	Cash/Share
	Financials	49.20%	PER; Gross Gearing
	Health	56.40%	Cash/Share; Current Ratio
	Industrials	53.80%	Enterprise Value; PER; Cash/Share
	IT	50.30%	Enterprise Value
LR	Industry-Wide	61.80%	Working Capital; Current Ratio; Quick Ratio
	Energy	77.70%	Total Revenue; EBIT; Total Equity
	Financials	66.30%	ROA; EV/EBITDA; Current Ratio
	Health	69.80%	EBIT; Quick Ratio
	Industrials	66%	Current Ratio
	IT	75%	Total Equity; ROE; Enterprise Value

3.5.4 Varying Cut-Offs Experimentation

The average classification scores shown in Table 3.6 can be misleading. As was shown in Table 3.3, despite the average of the classification accuracy being 61.8%, only 14 out of the 314 distressed companies were correctly classified by the model (4.5%), whereas 482 out of the 489 healthy companies were correctly classified by the model (98.6%). It is clear that model is more biased towards the healthy companies – this results in a high Type I error vis-à-vis Type II error. As mentioned in Chapter 2, the weighting of the errors can differ amongst users. Therefore, this section experiments with varying the cut-off points to check their effects on the Type I and Type II errors – the LR models were chosen for the experiments due to the large range between Type I and Type II errors. Table 3.7 below presents the original results for all the constructed LR models at the default 0.5 cut-off point. Table 3.8 presents the results after the cut-off points were experimentally changed to check their effects on the classification accuracies of the models. An elaboration on the meaning of the terms used in the table is presented below:

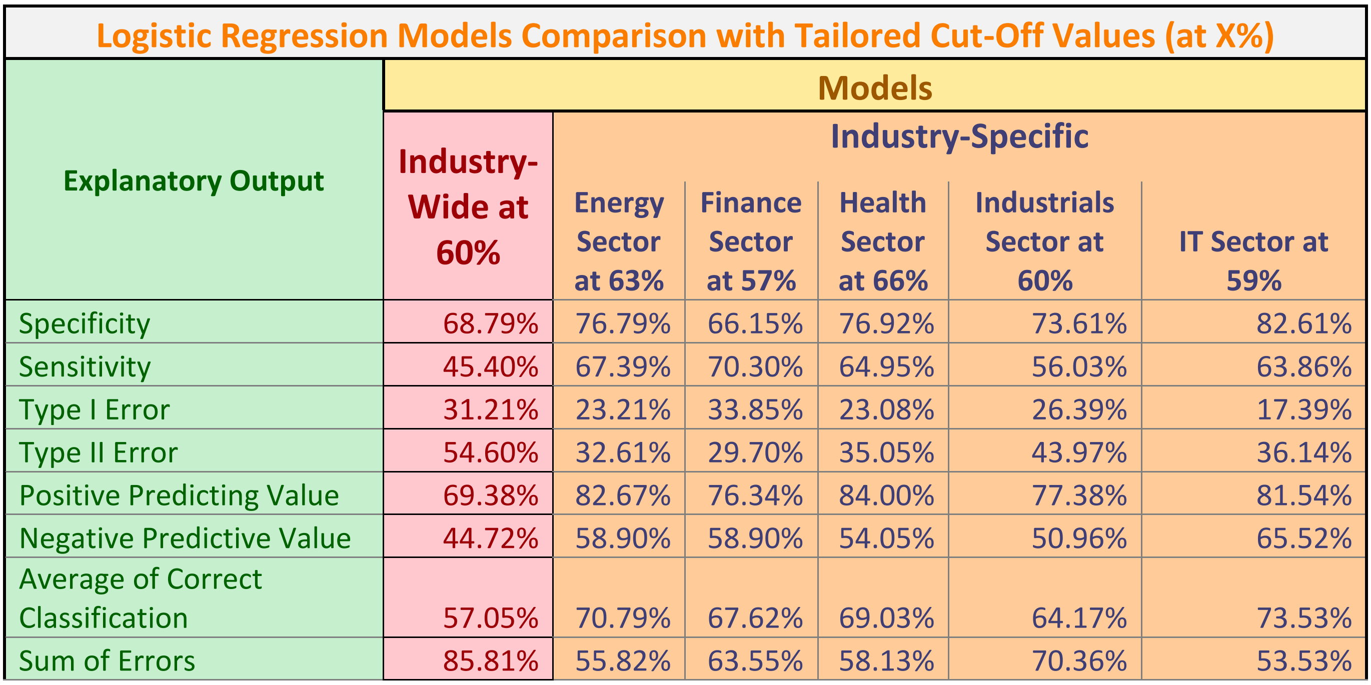
- **True Positives (Sensitivity):** Percentage of successful firms correctly predicted by model as such.
- **True Negatives (Specificity):** Percentage of failed firms correctly predicted by model as such.
- **Type I Error:** =100%- Specificity (percentage of actually failed but predicted as successful).
- **Type II Error:** =100%- Sensitivity (percentage of actually successful but predicted as failed).
- **Positive Predicting Value:** Percentage of predicted as successful that are actually successful.
- **Negative Predicting Value:** Percentage of predicted as failed that are actually failed.
- **Sum of Errors:** Type I + Type II Error.

Table 3. 7 Logistic Regression Models Comparison with Default Cut-Offs

Logistic Regression Models Comparison at Default 50% Cut-Off						
Explanatory Output	Models					
	Industry-Wide	Industry-Specific				
		Energy Sector	Finance Sector	Health Sector	Industrials Sector	IT Sector
Sensitivity	4.46%	57.14%	36.92%	19.23%	31.94%	71.01%
Specificity	98.57%	90.22%	85.15%	95.88%	85.34%	78.31%
Type I Error	95.54%	42.86%	63.08%	80.77%	68.06%	28.99%
Type II Error	1.43%	9.78%	14.85%	4.12%	14.66%	21.69%
Positive Predicting Value	61.64%	77.57%	67.72%	68.89%	69.89%	76.47%
Negative Predictive Value	66.67%	78.05%	61.54%	71.43%	58.50%	73.13%
Average of Correct Classification	64.12%	77.81%	64.63%	70.16%	64.2%	74.8%
Sum of Errors	96.97%	52.64%	77.93%	84.89%	82.72%	50.68%

Table 3.7 shows the results comparison of all constructed models (one industry-wide model and five industry-specific models) using LR using the default cut-off value of 0.5 for each model. The 'Sum of Errors' column shows the total percentage of companies that were misclassified by the model. As is evident in Table 3.7, the industry-wide combined error rate was 96.97%. None of the models using the industry-specific method have a combined error rate that exceeds that of the industry-wide model. This shows that the models using industry-specific LR are a more accurate choice.

Table 3. 8 Logistic Regression Models Comparison with New Cut-Offs



Logistic Regression Models Comparison with Tailored Cut-Off Values (at X%)						
Explanatory Output	Models					
	Industry-Wide at 60%	Industry-Specific				
		Energy Sector at 63%	Finance Sector at 57%	Health Sector at 66%	Industrials Sector at 60%	IT Sector at 59%
Specificity	68.79%	76.79%	66.15%	76.92%	73.61%	82.61%
Sensitivity	45.40%	67.39%	70.30%	64.95%	56.03%	63.86%
Type I Error	31.21%	23.21%	33.85%	23.08%	26.39%	17.39%
Type II Error	54.60%	32.61%	29.70%	35.05%	43.97%	36.14%
Positive Predicting Value	69.38%	82.67%	76.34%	84.00%	77.38%	81.54%
Negative Predictive Value	44.72%	58.90%	58.90%	54.05%	50.96%	65.52%
Average of Correct Classification	57.05%	70.79%	67.62%	69.03%	64.17%	73.53%
Sum of Errors	85.81%	55.82%	63.55%	58.13%	70.36%	53.53%

Table 3.8 shows the results comparison of all constructed models (one industry-wide model and five industry-specific models) using LR after applying new cut-off values for each model – the new cut-off values are presented underneath each model's name. These experimentally new cut-off values were chosen as they reduced the value of Type I error, all the while minimising the increase in Type II error. For example, the Type I error of the industry-wide model using the default cut-off point was 95.54%, whereas after applying the new cut-off point this dropped to 31.21%. The Sum of Errors was 96.97% for the industry-wide model using the default cut-off point, however this dropped to 85.81% after applying the new cut-off point.

This empirical experimentation showcases that varying the cut-off points for each FDP model can lead to a superior and more balanced model. Chapters 4, 5, and 8 present more robust methods for dealing with class imbalance and cut-off optimisation techniques that can present a more fair representation and alter the classification accuracy of the models.

3.6 Discussion

It is important to try and understand why each industry yielded different variables that are most pertinent for each industry's FDP model. Understanding the differences has the potential to yield to tailor-made industry-specific models with a high predictive accuracy. This section attempts to rationalise the reasoning behind those differences. Due to a lack of previous studies in this area that are able to provide justifications for the variable importance differences amongst industries, the following rationales are based on discussions with an expert in accountancy. It is important to note that these rationales are up for discussion and further studies should be done to cement those claims. The variables explained here are as per the model that yielded the highest overall accuracy, as was shown in Table 3.6.

- **Energy Industry Rationale:** One reason to explain why 'Cash per Share' came out as the most important predictive variable for the Australian energy industry may be because the choice of capital structure involves considering different costs and different risks – firms in the Energy industry are considered less risky (as they are assured a steady flow of cash payments from customers), therefore have access to higher risk funding (debt financing).
- **Financials Industry Rationale:** One reason to explain why 'ROA' – a profitability ratio that compares income to total assets – came out as the most important predictive variable in the Australian financials industry may be because if the company is not able to convert its investments in assets into profits, it is doomed to fail in such a liquid-driven industry.
- **Health Industry Rationale:** One reason to explain why 'Debt/Cash Flow' came out as the most important predictive variable in the Australian health industry may be due to the fact that the Australian healthcare system is largely

subsidized and government-funded, therefore a decrease or increase in the 'Debt/Cash Flow' level can have a direct impact on whether the entity will succeed or fail.

- **Industrials Industry Rationale:** One reason to explain why 'Gross Debt per Cash Flow', which is one of the most important indicators of cash flows, came out as the most important predictive variable in the Australian industrials industry may be due to the behemoths that operate within that industry that are able to accumulate high levels of debt, therefore, this ratio is indicative of a company's success or failure within this industry.
- **IT Industry Rationale:** One reason to explain why 'Total Equity', came out as the most important predictive variable in the Australian IT industry may be due to the fact that IT is an unforgiving volatile and fast-paced industry were technological obsolescence is always looming, therefore a company's equity is of utmost importance in determining its success or failure.

3.7 Conclusion

To conclude, this chapter showed how the literature is limited pertaining to the effects of industry-specificity on FDP, applied three techniques – two statistical techniques, namely: LR and MDA, and a machine learning technique, namely: ANN, on a large data sample comprising hundreds of Australian companies operating across five different industries. 18 models were created (three for each industry and three for the industry-wide model). The results indicate that using industry-specific models will lead to an increase in the predictive accuracy vis-à-vis an industry-wide model. Also, the most important variables pertaining to each industry were outlined and elaborated upon. The FDP models in this chapter have the potential to momentously aid various parties in the economy – from shareholders to government agencies; thus, leading to the improvement of the economy in general. This chapter has validated *Hypotheses 1 and 2,* and contributed is in the form of presenting adequate evidence to prove that financial distress in companies can be more accurately predicted by allocating companies to their respective industry, as opposed to a *one-size-fits-all* approach, which is still commonly used throughout the literature.

Chapter 4: Australian Mining Case Study*

*This chapter is based on a published paper in a peer-reviewed Journal, namely: Halteh, K., Kumar, K., & Gepp, A. (2018). Using Cutting-Edge Tree-Based Stochastic Models to Predict Credit Risk. *Risks*, 6(2), 55. doi:10.3390/risks6020055

4.1 Introduction

The previous chapter explored whether industry-specificity has an effect on the accuracy of predicting financial distress. The empirical results revealed that they actually do, and that the variables affecting financial distress differ by industry.

This chapter builds on the information gained from the previous chapter through exploring the Australian mining industry and applying parametric and nonparametric statistical models to evaluate which model has the superior predictive capabilities pertaining to financial distress. Aligning with *Hypotheses 3 and 4* stated in the Chapter 1, namely:

> ***H*$_3$*:*** Using cutting-edge recursive partitioning techniques will yield empirically superior results compared to traditional statistical techniques.
>
> ***H*$_4$*:*** Class imbalance does affect the detection accuracy of FDP models, and it can be enhanced by optimising the cut-off points or using SMOTE vis-à-vis a model that is built on a standard imbalanced data-set.

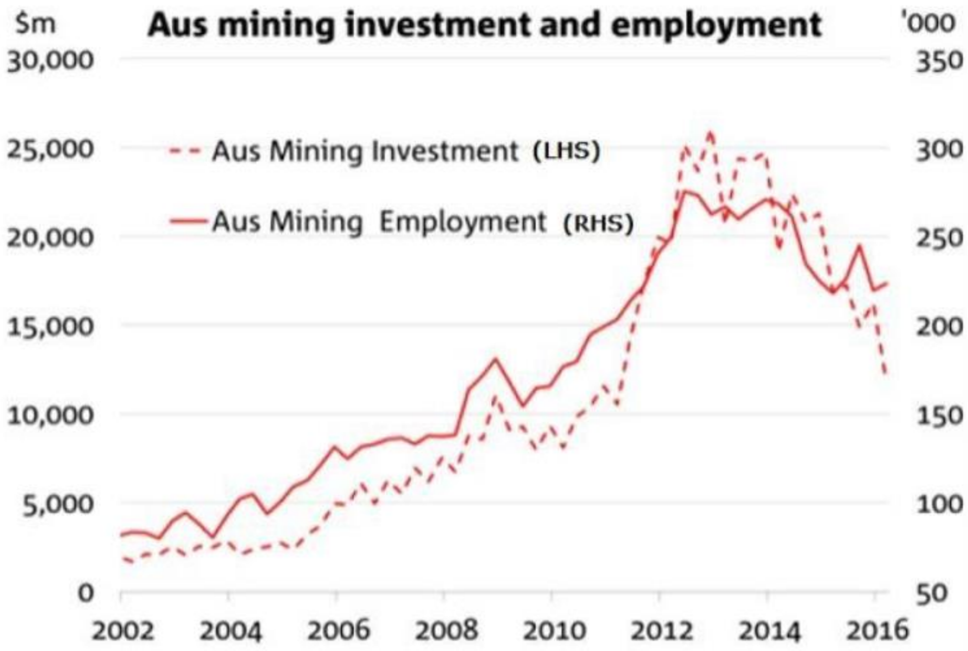
This chapter aims at verifying the aforementioned hypothesis by utilising cutting-edge machine learning techniques and comparing them with a standard LR model. Also, this chapter presents a method of dealing with an imbalanced data-set – as this was the case for the Australian mining industry, and showcases the most important variables for determining financial distress for each developed model and the predictive accuracy of each model – presented in terms of specificity and sensitivity, as well as the Area Under the Receiver Operating Characteristic Curve (AUROC) method – the mechanics of this metric will be discussed in detail in Chapter 5.

The Australian mining industry was chosen for several reasons, including:

- Mining helped cushion the Australian economy during from the 2008 Global Financial Crisis (GFC), as the mining sector was enjoying a boom during that period (Shah, 2014);
- Mining is a major contributor to Australia's economy, generating around $140 billion annually, hence making up more than half of the total goods and services (Shah, 2014);
- Mining is an important part of the Australian workforce – during 2007-2012, the mining sector set the highest employment growth nationwide, increasing by a record-breaking 94.3% to reach almost 270,000 workers, a record high (Shah, 2014);
- Mining makes up around 8% of the national GDP, 38% of all foreign direct investment, and approximately 60% of all exports (Frydenberg, 2015);
- Australia is the global leader when it comes to iron ore exports, making up more than half of the world's trade in 2014 (Frydenberg, 2015);
- Australia is also one of the world's leading exporters of coal, aluminium, copper, uranium, gold, and zinc (Frydenberg, 2015).

During the mining boom in Australia, between 2011-2012, mining contributed towards the national economic growth by approximately 66% and towards the GDP by about 8%. On the downside, however, a report by the National Australia Bank – as was shown in Letts (2016) – suggested that the contribution towards the GDP fell to around 4% in 2016, and is projected to fall to around 1%. Petroleum and mineral mining expeditions have also been falling, with a drop of around 8% from 2015. Adding insult to injury, tens of thousands of jobs are going to be lost, and investment is going to fall by as much as 70% in the coming years as the mining boom draws to an end – refer to Figure 4.1 for a visual representation of the Australian mining investment and employment for the time-period 2002-2016.

Figure 4. 1 Australian Mining Investment & Employment 2002-2016 (Letts, 2016)



Given the above statistics, this presents a gloomy future for the mining industry in Australia, at least for the short term. This might have dire consequences on not only the mining companies which may start declaring bankruptcies, but also on the whole economy due to the massive influence the mining sector has on it. This is already starting to materialise by an increase in individual bankruptcy rates as a result of a decline in the mining sector. Personal bankruptcies increased by around 5% in 2016 from the previous year, and around 6% in 2017 from the previous year (Butler, 2018). Other effects of the end of the mining boom are experienced in the construction and real estate industries, in which the value of construction work has been falling, and the house prices in western Australia has plummeted drastically near mining towns (Scutt, 2017; Wahlquist, 2017).

The points presented in the aforementioned paragraphs provide the justification to conduct an FDP analysis concentrating on the Australian mining industry to empirically determine the model(s) best suited for forecasting financial distress, as well as outlining the most important variables that effect a mining company's financial health. As was presented in Chapter 1, FDP modelling provides many advantages not only to decision makers and shareholders, but due to the enormous influence of the mining industry, will have far-reaching economic implications.

4.2 Literature Review

Chapter 2 presents a literature survey on the statistical techniques that are used in this chapter, namely: LR, DT, RF, and SGB. Therefore, to avoid repetition, this section will solely concentrate on the limited studies available in the literature pertaining to the FDP of the Australian mining industry.

There are limited FDP studies in the literature concentrating on companies operating in the mining industry. This section presents some of these studies that applied FDP modelling to mining sectors in Indonesia; as for studies using FDP modelling concentrating on Australian mining companies, the studies are extremely rare.

A recent study by Syamni, Majid, and Siregar (2018), applied FDP modelling to 19 coal mining companies operating in Indonesia between 2013-2015. Their study generated five models using a unique technique for each, which will generate scores for each model. Following this, a multiple panel regression model is estimated to investigate the effects the FDP models have on the stock prices of the coal mining companies – refer to Equation 4.1. The scores of the models generated earlier were used as predictors to predict the stock prices.

$$lnSP_{it} = \alpha + \beta_1 OS_{it} + \beta_2 ZM_{it} + \beta_3 GS_{it} + \beta_4 SS_{it} + \beta_5 ZS_{it} + \varepsilon_{it}$$

$$[Equation\ 4.1]$$

- *lnSP* = Natural logarithm of stock prices
- *OS* = Ohlson (1980) Score
- *ZM* = Modified Altman (1968) Z-Score – removed the fifth variable, different cut-off points for classification of company's financial health status
- *GS* = Grover and Lavin (2001) Score
- *SS* = Springate (1978) Score
- *ZS* = Zmijewski (1983) Score
- *ε* = Error term
- *i & t* = Company i for year t.

Syamni et al. (2018) found that the Grover and Lavin (2001) model identified most of the healthy companies, whereas the Ohlson (1980) model identified most of the distressed companies. Both models were also found to directly and negatively affect stock prices of the coal mining companies, that is, the higher the prediction scores, the lower the stock prices.

Another study by Nindita and Indrawati (2014), applied FDP modelling in the form of LR using five financial and two nonfinancial variables on 13 publicly listed mining companies in Indonesia for the time-period 2008-2010. Their findings indicate that Current Ratio, Cash Ratio, and Debt Ratio have a significant and negative effect on predicting financial distress, that is, the higher the ratio, the lower probability of financial distress; whereas nonfinancial variables were not found to be statistically significant.

As for Australian studies, a paper by Ferguson, Clinch, and Kean (2011), applied FDP modelling in the form of LR to determine the success or distress to a sample of 85 single-project gold mining companies following disclosure of a feasibility study for the time-period 1990-2007. Their results indicate that nonfinancial information had a direct effect on financial distress, whereas the Altman (1968) Z-score financial predictors were not useful in explaining financial distress.

The findings by Ferguson et al. (2011) are in direct contradiction to the results in the study presented above by Nindita and Indrawati (2014) in terms of predictive effect of nonfinancial variables on mining companies. However, there is an important distinction between the two studies that must be noted, which may offer an explanation towards the disparity in results. Ferguson et al. (2011) defined failure not in terms of a mining company's closure, but in terms of four development projects criteria outlined in their paper. Whereas, Nindita and Indrawati (2014) did not focus on projects, but on the company's overall financial status (healthy/distressed).

Another paper by Shah (2014) applied FDP modelling on the Australian mining industry during the 2012-2013 time-period. Shah selected 20 independent variables made up of standard financial ratios for the FDP modelling. Shah's data-set consisted of 351 and 44 financially healthy and distressed mining companies, respectively. Shah used various parametric, nonparametric, and hybrid statistical techniques to create the FDP models. Shah's models are presented below, and the most significant/important variables are presented in Table 4.1 below.

- **LR model:** Six statistically significant variables were found significant at the 5% level. The model's accuracy in predicting financially healthy companies (specificity) was an impressive 99.1%, however the accuracy in predicting financial distressed companies (sensitivity) was a modest 34.1%.

- **MDA model:** Nine statistically significant variables were found significant at the 5% level. The model's specificity was 86.9%, whereas the sensitivity was 13.1%.

- **ANN model:** Two important variables were presented. The testing sample's accuracy results were specificity = 98%, sensitivity = 36.4%.

- **DT model:** The DT model was built using the CHAID growing method. The model's accuracy ratings for the testing sample correctly classified all the financially healthy companies but none of the distressed ones. The testing sample's accuracy results were specificity = 100%, sensitivity = 0%.

- **Hybrid model 1:** Two important variables were presented. The model's specificity was 98.2%, whereas the sensitivity was 29.4%.

- **Hybrid model 2:** Two important variables were presented. The model's specificity was 100%, whereas the sensitivity was 18.8%

Table 4. 1 Most Significant/Important Variables in Shah's (2014) FDP Models

	Models					
	LR	MDA	ANN	DT	Hybrid 1 (ANN and LR)	Hybrid 2 (ANN and MDA)
Significant Variables	Depreciation / PPE	Asset Turnover	Gross Gearing	*Root Node:* Price / Book Value	PER	Price / Gross Cash Flow
	Gross Debt / Cash Flow	Current Ratio	Price / Gross Cash Flow		Price / Gross Cash Flow	PER
	Price / Gross Cash Flow	Invested Capital Turnover				
	ROA	Long-term Asset Turnover				
	ROIC	Net Gearing				
		PPE Turnover				
		Price / Book Value				
		Price / Gross Cash Flow				
		Quick Ratio				

The enormous gap between sensitivity and specificity in Shah's paper is due to the class imbalance problem in the data-set used to create the models (351 vis-à-vis 44), as well as the default 0.5 cut-off used in the models. Shah did not give attention to these important factors, thus resulting mostly in impractical models that predict the majority class by default, thus giving a deceiving average accuracy rating (by averaging the true positives and negatives). These issues were faced in the data-set adopted for this study, but were addressed carefully, as is shown later in this chapter.

4.3 Data

Archival data were extracted from MorningStar database pertaining to the Australian mining companies used in the research. The MorningStar database has been used previously in the literature across different disciplines, some of these studies include: Halteh (2015); Halteh et al. (2018b); Shah (2014); Smith et al. (2011).

This chapter used all available data from the database for listed and delisted mining companies. Time-series data were then chosen for the years 2011-2015. The company status variable in MorningStar was used to determine the listed or delisted status. According to the Australian Securities Exchange (ASX), the source of much of the data from MorningStar, an Australian company is 'listed' if it is currently operational, whereas a company is 'delisted' for a number of reasons including insolvency, merger, or take-over. All of these collectively imply an element of financial distress leading to delisting of the company (MorningStar, 2016). This chapter refers to listed companies as healthy, and delisted companies as distressed.

As was the case in Chapter 3, the variables were chosen based upon several factors, including standard accounting and financial variables, use in prior empirical research and literature, endorsement by theorists, and as per availability of data. It is important to note that since this study uses the companies' actual financial status (distressed/healthy), no nonfinancial variables were included in the study – this is due to the findings presented in Nindita and Indrawati's (2014) study, which found that nonfinancial variables were not statistically significant.

The extracted data yielded 632 healthy companies and 118 distressed companies. The data were then downloaded to a spreadsheet for cleaning. The initial count was 590 observations (118 companies multiplied by 5 years) for distressed companies and 3160 observations (632 companies multiplied by 5 years) for healthy companies, a total of 3750 observations incorporating data for 29 explanatory variables. After examining the data, some observations needed to be deleted due to insufficient data. Variables that had 50% or more missing data were deleted. Following this removal, companies that had 50% or more missing data were also deleted. Such a high percentage of missing data were deemed to be insufficient to build a credible model. This resulted in omitting ten variables; as for companies' financial data, the final sample contained 19 variables with 3375 observations – 339 observations for distressed companies and 3036 for healthy companies. All 29 variables are shown in Table 4.2, with the ones omitted being followed by an asterisk.

Table 4. 2 Complete List of Variables

Variable	Description
Net Profit Margin*	Net Profit / Revenue
EBIT Margin*	Earnings Before Interest and Tax (EBIT) / Net Revenue
ROE	Return on Equity = Net Profit After Tax / (Shareholders Equity – Outside Equity Interests)
ROA	Return on Assets = Earnings before interest / (Total Assets Less Outside Equity Interests)
ROIC	Return on Invested Capital = Net Operating Profit Less Adjusted Tax / Operating Invested Capital
NOPLAT Margin*	Net Operating Profit Less Adjusted Tax (NOPLAT) / Revenue
Inventory Turnover*	Net Sales / Inventory
Asset Turnover	Operating Revenue / Total Assets
PPE Turnover	Revenue / (Property, Plant & Equipment (PPE) – Accumulated Depreciation)
Depreciation/PPE	Depreciation / Gross PPE
Depreciation/Revenue*	Depreciation / Revenue
Working Capital/Revenue*	Working Capital / Revenue
Working Capital Turnover	Operating Revenue / Operating Working Capital
Gross Gearing	(Short-Term Debt + Long-Term Debt) / Shareholders Equity
Financial Leverage	Total Debt / Total Equity
Current Ratio	Current Assets / Current Liabilities
Quick Ratio	(Current Assets - Current Inventory) / Current Liabilities
Gross Debt/CF	(Short-Term Debt + Long-Term Debt) / Cash Flow
Cash per Share	Cash Flow / Shares Outstanding
Invested Capital Turnover	Operating Revenue / Operating Invested Capital before Goodwill
Net Gearing	(Short-Term Debt + Long-Term Debt - Cash) / Shareholders Equity
NTA per Share	Net Tangible Assets / Number of Shares on Issue
BV (Book Value) per Share	(Total Shareholder Equity - Preferred Equity) / Total Outstanding Shares
Receivables/Operating Revenue*	Debtors / Operating Revenue
Inventory/Trading Revenue*	Inventory / Trading Revenue
Creditors/Operating Revenue*	Creditors / Operating Revenue
Sales per Share	Total Revenue / Weighted Average of Shares Outstanding
EV/EBITDA*	Enterprise Value (EV) / Earnings Before Interest, Tax, Depreciation & Amortisation (EBITDA)
PER	Price per Earnings = Market Value of Share / Earnings per Share

: *Cells with a red background/asterisk indicate the variables that were later excluded from the model due to missing data – refer to the last paragraph on the previous page for explanation.

4.4 Methodology

Following the data collection and cleaning (as mentioned in the Data section), a dichotomous binary variable was used to refer to the status of each company—coded ‘1’ if the company is healthy and ‘0’ if the company is distressed. The data were then partitioned by randomly selecting 80% healthy and 80% distressed companies for a training set used to develop statistical models, with the remaining 20% of the healthy and distressed companies being used for testing and evaluating models. Having a separate data-set is necessary to obtain representative estimates of real-world performance for fair comparisons between models. This process and the resulting data-sets are summarised in Table 4.3.

Table 4. 3 Data Overview

Sample Partition	Number of Observations	Percentage	Healthy Companies	Distressed Companies	Class Imbalance %
Train	2,700	80.00%	2,419	281	89.59% Healthy – 10.41% Distressed
Test	675	20.00%	617	58	91.41% Healthy – 8.59% Distressed
Total	3,375	100.00%	3036	339	89.96% Healthy – 10.04% Distressed

As is evident in Table 4.3, there is class imbalance in the data-set, meaning that there are much more healthy companies than distressed ones. When creating the testing/holdout sample, the class imbalance percentage was ensured to be kept very similar to that of the training sample to enable a fair representation of the data-set.

The class imbalance is particularly problematic when the difference is extreme, as the models will tend to automatically overlook the minority class and predict everything as the majority class. In this case, the overall results will appear good overall, but they will be unusable as all predictions are the same.

The model building methods that are used in the study are logistic regression (as a well-established benchmark), decision trees, random forests, and stochastic gradient boosting. In line with *Hypothesis 3* presented in Chapter 1, the results of the state-of-the-art recursive partitioning models are expected to outperform the parametric logistic regression model. This would provide confirmatory evidence from a larger data-set of similar results in the limited existing literature.

The following subsections outline the methodologies used in the study and expand upon the three aforementioned models, as well as the optimised cut-off value approach used to deal with the class imbalance problem in this data-set. With the exception of the optimised cut-off value approach – which will be discussed below in Section 4.4.5, the mechanics of the techniques used will not be discussed, as they were previously mentioned in Chapter 2.

4.4.1 Logistic Regression Model

The logistic regression model was estimated with all 19 variables used as covariates to explain the companies' status (healthy or distressed). SPSS statistical software was used to develop the model, but as the model is deterministic, the same results would be obtained using other software packages.

4.4.2 Decision Tree Model

Classification and Regression Trees (CART) using Salford Predictive Modeller (SPM) have been used to generate the FDP tree. All 19 variables were selected as predictors in the model. The Gini splitting rule was used because of its popularity and widespread use. The minimum data points in a non-leaf node was set to 10 to avoid the tree becoming too large. This setting assists in avoiding over-fitting, that is, looking for patterns in very small subsamples that are likely not to generalise to future data.

4.4.3 Random Forests Model

SPM has again been used and all 19 variables were used as predictors. There are two main parameters to set for a RF model: the number of trees to be generated, and the number of variables to be considered at each node. A model was developed for each of 200, 500, and 1000 trees to empirically determine the best choice for this parameter. The number of variables considered at each node was set to the square root of the total number of predictors: $\sqrt{19} \approx 4.36 \approx 4$. The square root heuristic was chosen as it has been recommended by and used in prior literature, including: Bhattacharyya, Jha, Tharakunnel, and Westland (2011); Gepp (2015); Whiting, Hansen, McDonald, Albrecht, and Albrecht (2012).

4.4.4 Stochastic Gradient Boosting Model

Once again, SPM has been used and all 19 variables were used as predictors. Models were developed based on 200, 500, and 1000 trees, to empirically determine the best choice for this parameter. As mentioned in the literature review, SGB relies on incremental improvements and therefore, it is important that no individual tree is too complex (large). Consequently, individual trees are kept small by setting the maximum nodes per tree to six (a standard setting) with a minimum number of data points of ten in each node. The criterion to determine the optimal number of trees, that is, how much incremental improvement to perform, was chosen based on the default of cross entropy.

4.4.5 Cut-Off Values for Classification

All four models can estimate the probability of being healthy (1). Often, a default value of 0.5 is used such that if a company has a value greater than 0.5 it will be classified as healthy, else as distressed. However, this is commonly unsuitable when there is a substantial class imbalance, as will be demonstrated in this case in the Results section. Consequently, the cut-off values are empirically optimised using the train sample. This approach involves experimentally and empirically determining the optimal cut-off value for each constructed model. Since cut-off values are optimised based on the same training data used to construct models, the existence of any sample selection bias will be common to both processes, hence not worrying (Gepp, 2015). Because of this, and for consistency with prior research in the field, cut-off values have been empirically optimised in this chapter – this approach has been used successfully in the literature, studies include: Bayley and Taylor (2007); Beneish (1997); Gepp (2015); Perols (2011).

Since this optimisation will be completed for each model, it is possible that the cut-off values will vary between the models. The cut-off value must be between zero and one, as they are the limits of any probability figure. The optimised cut-off value is chosen as the value that produces the most balanced accuracy on the train sample. The most balanced is defined by minimising the difference between prediction accuracy for healthy companies and prediction accuracy for distressed companies. It is important to highlight that the cut-off values were optimised exclusively on the train sample, so that model evaluation on the test sample still represents performance on data that is completely new to the model.

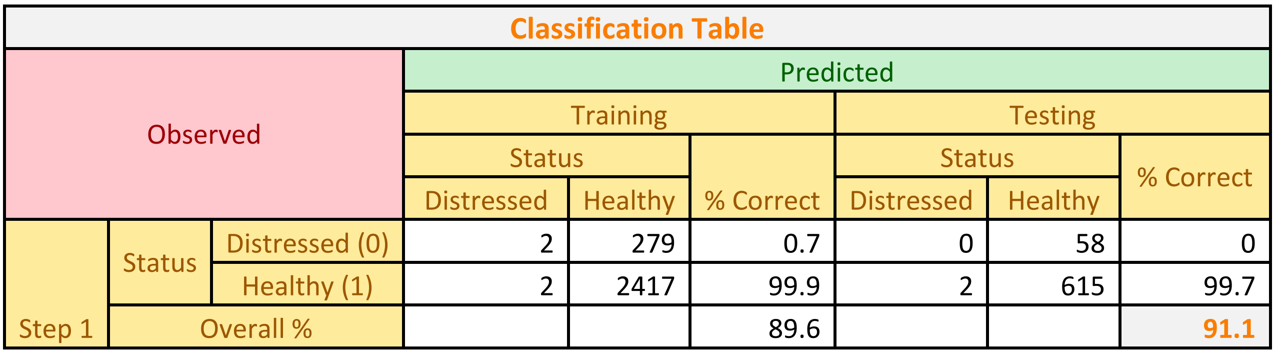
4.5 Results

The following subsections explore and analyse, in detail, the results achieved and performances of the various models used in the study. Specificity represents the accuracy at classifying healthy companies, while sensitivity represents the accuracy at classifying distressed companies.

4.5.1 Logistic Regression Model

As shown in Table 4.4 below, the default logistic regression model yielded an average accuracy of 91.1% on the test sample. However, as mentioned in the Methodology section above, this model is not practically useful because of the class imbalance. When using the default 0.5 cut-off value, the model predicts almost all the companies as healthy (1), which results in a mirage of high predictive accuracy. Even though their overall accuracy is high, the model is useless because it cannot successfully predict distressed companies: 0.7% on the training data and 0% on the testing data.

Table 4. 4 LR Classification Table at Default 0.5 Cut-Off Value



Classification Table

Observed			Predicted					
			Training			Testing		
			Status		% Correct	Status		% Correct
			Distressed	Healthy		Distressed	Healthy	
Step 1	Status	Distressed (0)	2	279	0.7	0	58	0
		Healthy (1)	2	2417	99.9	2	615	99.7
	Overall %				89.6			91.1

To remedy this class imbalance problem, the cut-off values in the training sample were empirically optimised to give the most accurate balanced rates, as explained in the Methodology section (Section 4.4.5). Results for both the training and testing samples are shown in Tables 4.5A and 4.5B, respectively. As shown in Table 4.5C, the overall model's accuracy dropped to an average of 56.71%. However, the accuracy is now more balanced between distressed (0) and healthy (1) companies. Therefore, this model is of more practical use and its assessment is more indicative of a logistic regression model. As for the variable importance, 'PER', 'Sales per Share', and 'Gross Debt / Cash Flow' were found to be the most statistically significant variables, all having *p*-values less than 10%.

Table 4. 5A Optimised LR – Train Sample

Train Sample			
Class	Cases	Misclassified	% Error
Distressed (0)	281	111	39.50%
Healthy (1)	2419	1,422	58.78%

Table 4. 5B Optimised LR – Test Sample

Test Sample			
Class	Cases	Misclassified	% Error
Distressed (0)	58	16	27.59%
Healthy (1)	617	364	59.00%

Table 4. 5C Model Accuracy (Test Sample) with Optimised Cut-Off Values

Accuracy at Predicting Healthy Companies (Specificity)	41.00%
Accuracy at Predicting Distressed *Companies* (Sensitivity)	72.41%
Simple Average	56.71%

4.5.2 Decision Tree Model

The empirical optimisation of the cut-off value on the training sample resulted in a cut-off value of 0.9. As shown in Table 4.6C, the decision tree yielded an average accuracy of 71.72% on the test sample. This is already a better outcome vis-à-vis LR, both for the specificity and sensitivity measures. This is consistent with existing literature that recursive partitioning models outperform traditional models. More detailed results for the train and test samples are shown in Tables 4.6A and 4.6B, respectively. As for the variable importance, ‘Invested Capital Turnover’, ‘Book Value per Share’, and ‘NTA per Share’ were found to be the most important variables for predicting financial distress in this model.

Table 4. 6A Optimised DT – Train Sample

Train Sample			
Class	Cases	Misclassified	% Error
Distressed (0)	281	66	23.49%
Healthy (1)	2419	788	32.58%

Table 4. 6B Optimised DT – Test Sample

Test Sample			
Class	Cases	Misclassified	% Error
Distressed (0)	58	14	24.14%
Healthy (1)	617	200	32.41%

Table 4. 6C Model Accuracy (Test Sample)

Accuracy at Predicting Healthy Companies (Specificity)	67.59%
Accuracy at Predicting Distressed *Companies* (Sensitivity)	75.86%
Simple Average	71.72%

4.5.3 Random Forests Model

Experimentation was conducted on generating 200, 500, and 1000 trees. Using 1000 trees yielded the most accurate results, which have been reported below. The empirical optimisation of the cut-off value on the training sample resulted in a value of 0.47, which was close to the default 0.5, and so the default cut-off value. Results for both the training and testing samples are shown in Tables 4.7A and 4.7B, respectively. As shown in Table 4.7C, the RF model yielded an average accuracy of 72.26% on the test data. However, compared to a single decision tree, this model is better at predicting distressed companies, but slightly worse at predicting healthy companies. As for variable importance, 'Invested Capital Turnover, 'Book Value per Share', and 'NTA per Share' were found to be the most important variables for predicting financial distress in this model.

Table 4. 7A Optimised RF – Train Sample

Train Sample			
Class	Cases	Misclassified	% Error
Distressed (0)	281	81	23.49%
Healthy (1)	2419	806	32.58%

Table 4. 7B Optimised RF – Test Sample

Test Sample			
Class	Cases	Misclassified	% Error
Distressed (0)	58	13	22.41%
Healthy (1)	617	204	33.06%

Table 4. 7C RF Model Accuracy (Test Sample)

Accuracy at Predicting Healthy Companies (Specificity)	66.94%
Accuracy at Predicting Distressed *Companies* (Sensitivity)	77.59%
Simple Average	72.26%

4.5.4 Stochastic Gradient Boosting Model

Experimentation was conducted across 200, 500, and 1000 trees – the model with 1000 trees yielded the most accurate results. The empirical optimisation of the cut-off value on the training sample resulted in a value of 0.91, which was close to the 0.9 mark, hence 0.9 was chosen. Both the training and testing samples results are shown in Tables 4.8A and 4.8B, respectively. As shown in Table 4.8C, stochastic gradient boosting yielded an average accuracy of 73.70% on the test data. On average, and as per the specificity score, this model outperforms all other models in the study. However, DT and RF yielded slightly better sensitivity accuracy. As for the variable importance, ‘Property, Plant, & Equipment (PPE) turnover’, ‘invested capital turnover’, and ‘PER’ were found to be the most important variables for predicting financial distress in this model.

Table 4. 8A Optimised SGB – Train Sample

Train Sample			
Class	Cases	Misclassified	% Error
Distressed (0)	281	49	17.44%
Healthy (1)	2419	596	24.64%

Table 4. 8B Optimised SGB – Test Sample

Test Sample			
Class	Cases	Misclassified	% Error
Distressed (0)	58	15	25.86%
Healthy (1)	617	165	26.74%

Table 4. 8C SGB Model Accuracy (Test Sample)

Accuracy at Predicting Healthy Companies (Specificity)	73.26%
Accuracy at Predicting Distressed *Companies* (Sensitivity)	74.14%
Simple Average	73.70%

4.5.5 Model Comparison

Table 4.9 summarises the performance of all four models. The rightmost column of the table represents the AUROC – a measure that is widely used in the literature, studies include: Burez and Van den Poel (2009); Chawla (2009); Chawla, Bowyer, Hall, and Kegelmeyer (2002); Duda, Hart, and Stork (2001).The AUROC measure was added in order to solidify the findings as to which model has the highest predictive accuracy. The closer the percentage is to 100%, the more accurate the model is in classifying the distressed and healthy companies. The presented percentages represent the AUROC for the test samples.

Table 4. 9 Model Comparison Table using Test Data

Model	Overall Model Accuracy	Most Important Variables	AUROC %
Logistic Regression	Specificity: 41.00% Sensitivity: 72.41% Average: 56.71%	PER, Sales per Share, Gross Debt / CF	59.00%
Decision Tree	Specificity: 67.59% Sensitivity: 75.86% Average: 71.72%	Invested Capital Turnover, BV per Share, NTA per Share	74.00%
Random Forest	Specificity: 66.94% Sensitivity: 77.59% Average: 72.26%	Invested Capital Turnover, BV per Share, NTA per Share	78.99%
Stochastic Gradient Boosting	Specificity: 73.26% Sensitivity: 74.14% Average: 73.70%	PPE Turnover, Invested Capital Turnover, PER	88.98%

As can be seen in Table 4.9, all of the machine learning techniques outperform the LR technique in terms of predictive accuracy. The SGB model outperformed all others as per the Overall Model Accuracy and the AUROC criteria. This empirically demonstrates the predictive superiority of machine learning models vis-à-vis the traditional parametric LR model, thus verifying *Hypothesis 3,* and that SGB is the most accuracy machine learning model compared with DT and RF.

As for the most important variables affecting the financial standing of a company, 'Invested capital turnover' is of utmost importance when trying to work out the level of financial distress, because it constantly appeared in all three non-parametric recursive partitioning models. This ratio measures the revenue generated from working capital investments. This enables the company to realise the tie between invested capital to fund normal operations, and the amount of sales created through these operations. It is meaningful that said variable is important to mining companies, since, the higher the capital turnover, the more efficient the company is at using current assets and liabilities to sustain its revenues. Inversely, a low capital turnover may lead to bad debts and obsolescence of inventory (Kenton, 2019). Therefore, due to the inventory-intensive nature of mining companies, it is crucial that mining company executives maintain a high capital turnover ratio in order to prevent their companies from failing. The results are in concert with findings found in the FDP of mining sector literature, such as Nindita and Indrawati (2014); Shah (2014); that is, the statistical significance of cash and debt ratios to mining companies.

4.6 Conclusion

To conclude, this chapter has showcased a real-world problem that needs to be addressed, that is, a high number of business failures in Australia in general, and an impending financial distress of mining companies, in particular. FDP can be utilised to forecast impending distress to enable the decision makers to take the preventive measures to hold-off financial distress or mitigate its effect. LR and recursive partitioning models were employed to test for the most accurate model at predicting financial distress. These models are not exclusive to the mining industry; they can be used in any industry worldwide.

The results indicated that 'Invested Capital Turnover' was the variable most occurring amongst the recursive partitioning models. In terms of the best model overall, SGB yielded the most accurate results in predicting financial distress in the Australian mining industry, as per the AUROC and averages of the sensitivity and specificity criteria. However, the random forests model yielded the best results at predicting the distressed companies (sensitivity). All in all, the analysis has shown that tree-based models are more accurate, versatile and have a wider scope than traditional models, such as logistic regression – this verifies *Hypotheses 3 and 4* presented in Chapter 1.

The main takeaway from this chapter is that modern models, such as the recursive partitioning models, can offer substantial accuracy improvements and should be considered in future research and in practice, especially in conjunction with qualitative measures and managerial decision-making. The models analysed in this chapter can be algorithmically automated to input new data as soon as they become available, for example through interim or annual reports, thus saving time to reconstruct the models manually and ensuring up-to-date models. It is imperative to address the class imbalance problem; in this chapter, ‘empirically optimised cut-off scores’ were used. There are other approaches in the literature to handle class imbalance that can change the overall data set, such as the Synthetic Minority Oversampling Technique (SMOTE), which is investigated in Chapter 5.

Chapter 5: Class Imbalance: Synthetic Minority Oversampling Technique (SMOTE) in the Context of FDP

Aligning with *Hypothesis 4* stated in the Chapter 1, namely:

> ***H*$_4$*:*** Class imbalance does affect the detection accuracy of FDP models, and it can be enhanced by optimising the cut-off points or using SMOTE vis-à-vis a model that is built on a standard imbalanced data-set.

This chapter will verify the aforementioned hypothesis by applying the SMOTE technique to an imbalanced data-set.5.1 Introduction

As mentioned in the Conclusion section of the previous chapter, another approach to dealing with class imbalance – other than the 'empirically optimised cut-off scores' method – is Synthetic Minority Oversampling Technique (SMOTE), which is investigated in this chapter. Class imbalance is present when there is a substantial difference in the ratio between the classes in a data-set, for example, a large number of healthy companies, vis-à-vis a small number of distress companies.

The presence of class imbalance is problematic, as it can lead to suboptimal and/or deceptive prediction accuracy levels in traditional data driven models. This is due to the algorithms that are used to construct the models being biased towards the majority class, hence resulting in a mirage of high predictive accuracy. For example, a data-set containing 90% healthy companies and 10% distressed companies, will yield to a deceptive predictive accuracy result of 90% if the model simply classifies all companies as healthy. For further reading on this topic, refer to Chapter 4 for an application of a LR model on an imbalanced data-set, that yielded a fallaciously high predictive accuracy result. The class imbalance prevalence can be found across many fields, some of these include: FDP studies, including: Kim, Kang, and Kim (2015); Zhou (2013); fraud detection studies, including: Gepp (2015); Perols (2011); Provost and Fawcett (2001); detection of oil spills in satellite radar imaging, such as in the research by Kubat, Holte, and Matwin (1998); diagnosis of rare medical conditions, as was shown by Murphy and Aha (1994); and lastly, monitoring of helicopter gearbox failure, as was shown by Japkowicz, Myers, and Gluck (1995).

The evaluation of the predictive accuracy of statistical and machine learning models generally occurs through inspecting the confusion matrix table, which presents the number and percentages of cases correctly and incorrectly identified in the developed model. However, if one simply looks at the overall percentage accuracy of the developed model, that will lead to a deceptive result if the data-set was imbalanced; hence, rendering this approach only reflective of true model performance when the classes are balanced and when the weights of the errors are equal. For example, a mammography test contains around 98% normal pixels vis-à-vis 2% abnormal ones (Woods et al., 1993). Creating a model that solely predicts the majority class will yield a high prima facie result of 98% predictive accuracy; however, as explained earlier, this result is illusory. This happens due to not emphasising the presence of the minority class. Therefore, only looking at the overall percentage accuracy of the model is not prudent (Chawla et al., 2002). A better single measure is the Receiver Operating Characteristic (ROC) curve, which visualises all possible thresholds, that is, the true positive and false positive error rates (Type I and Type II errors). It is plotted with the sensitivity on the y-axis, and the specificity on the x-axis. The Area Under the Curve (AUC) is a performance measure for the ROC, often referred to as Area Under Receiver Operating Characteristics (AUROC), and is widely-used across various disciplines in the literature, studies include: Burez and Van den Poel (2009); Chawla (2009); Chawla et al. (2002); Duda et al. (2001).

Figure 5.1 below shows an ROC graph in which the AUC is 1, that is, a perfect model in its distinguishing ability to separate between classes – in the area of FDP, those classes would be healthy and distressed businesses. As is clear, the red line runs along the y-axis, then veers to the right at the '1' mark, then runs parallel to the x-axis, thus encompassing the total AUC, yielding an AUROC score of 1.

Figure 5. 1 ROC Graph Representing a Perfect Model

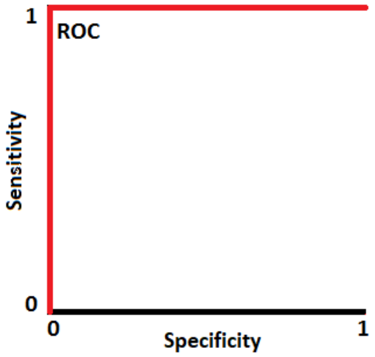
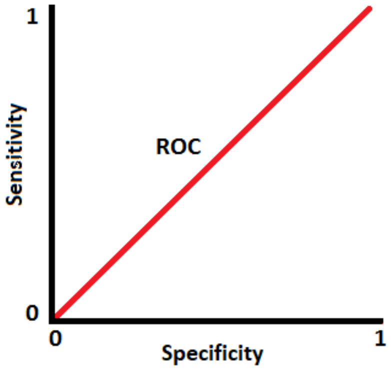


Figure 5.2, on the other hand, showcases a model that has no discernment or distinguishing ability between classes, thus yielding an AUROC score of 0.5. The graph runs diagonally from the 0 mark and cuts the graph in half, thus encompassing 50% of the total AUC.

Figure 5. 2 ROC Graph Representing an Undiscerning Model



Therefore, the aim is to have the ROC graph look as much like the one showcased in Figure 5.1, although, in reality, errors are always present, but the idea is to try and have a model with the least amount of error, which is presented both graphically – by a line that steeps vertically upwards and as close as possible to the y-axis, veers to the right as it approaches the ‘1’ mark, and then runs parallel to the x-axis. Empirically, the aim is to have the AUROC score that is close to 1. When comparing models, the model with the higher AUROC score is superior.

This chapter has introduced the concept of class imbalance, then proceeds to survey the literature on the various methods of dealing of class imbalance problem, followed by presenting various FDP models created using an imbalanced data-set comprised of companies operating in the Australian mining industry, and then compares those results with results achieved after applying SMOTE to the same data-set, followed by concluding remarks. This chapter's contribution is in the form of creating and comparing various machine learning FDP models built using a standard data-set and a data-set that has been *SMOTEd*. This furthers the understanding of the class imbalance pertaining to FDP through an empirical analysis of SMOTE on machine learning techniques. The presumed resilience of said techniques towards data-sets that are imbalanced is also checked. Therefore, this chapter verifies *Hypothesis 4* presented in Chapter 2.

5.2 Literature Review

This section presents other techniques that can deal with the class imbalance problem, some of these include:

- Empirical cut-off optimisation – as was shown in Chapter 4
- Random resampling of the original data-set
- Bagging
- Boosting
- Synthetic Minority Oversampling Technique (SMOTE)

In classification models, a default value of 0.5 is used as a cut-off value of classifying the predicted variable, whereby a predicted value equal to or greater than 0.5 will result in a classification of (1) – in the case of FDP analysis in this thesis, that is reflective of successful/operating/non-distressed companies (1). On the other hand, a predicted value less than 0.5 will result in a classification of (0), that is, financially distressed companies. This is suitable so long as there is no substantial difference between the (1's) and (0's) in the data-set, that is no class imbalance problem.

As mentioned earlier, there are many methods to deal with the class imbalance problem. One approach is the empirical optimisation of the cut-off values. In brief, the cut-off value is determined as the value that optimises a chosen accuracy metric on the training sample – an example metric is a weighted average of Type I and Type II errors. This approach has been used successfully used in the literature, studies include: Bayley and Taylor (2007); Beneish (1997); Gepp (2015); Halteh et al. (2018b); Perols (2011) – for more on this topic, refer to Chapter 4.

Another approach to the class imbalance problem is through random resampling of the original data-set, by either under-sampling the majority class or over-sampling the minority class – this is often referred to as bootstrapping (Tibshirani & Efron, 1993). This results in a more balanced data-set, therefore standard statistical techniques can then be used. Some of the studies incorporating these techniques include: Drummond and Holte (2003); Japkowicz (2000); Ling and Li (1998). Under-sampling is a method whose purpose is to balance the classes in a data-set by randomly eliminating from the majority class. The main problem with under-sampling is loss of invaluable data that would have been included in the model. As for over-sampling, similar to under-sampling, it attempts to balance the class distribution, but this is done through replicating data from the minority class. The main problem with over-sampling is the non-value-adding repetitiveness of data which may lead to over-fitting (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012).

Another approach is bagging – which was pioneered by Breiman (1996). Bagging combines bootstrapping and aggregating, hence the name bagging. It is a hybrid ensemble method which is usually applied to classification cases in order to enhance the classification accuracy through combining single classifications. Bagging trains various classifiers on bootstrapped copies of the original training data-set – this results in achieving diversity with the resampling procedure through the use of different data-set. When predicting new cases, each training data is created a classification tree and the majority vote (mode) or weighted vote is utilised to deduce the class (Galar et al., 2012). Models constructed using the bagging technique generally outperform those built using sample random sampling (Hakim, Sartono, & Saefuddin, 2017). The RF technique is an example of bagging that uses DTs.

Another approach is boosting, formally known as ARCing (Adaptive Resampling and Combining). Boosting was pioneered by Schapire (1990) – he showed how a weak learner – which is marginally superior to random guessing – can be turned into a strong learner. Boosting is an ensemble method that aims at minimising variants due to the average refractive effect of the ensemble. The classification power of decision trees is "boosted" through applying the classification function repeatedly and combining, including weights, the results in order to minimise the classification error. Dissimilar to bagging that builds models which are independent of one another, boosting is repetitive since the inaccurate predictions from the existing model are provided higher probability of being selected in the data that will be used to grow the successive tree. Therefore, the classification accuracy is improved through repetition, hence is immune to the problem of reduced performance on holdout data. Boosting has the advantage of being simpler than bagging by using simpler classifiers, that is, small trees. The SGB technique is an example of boosting (Gepp, 2015; Sutton, 2005).

5.2.1 SMOTE

Since SMOTE is going to be used in this chapter, this subsection thoroughly investigates SMOTE's mechanics and its advantages vis-à-vis other class imbalance approaches, like the ones mentioned earlier. This provides the justification to use SMOTE in this study.

Chawla et al. (2002) presented a breakthrough study that coined SMOTE. They argued that although in an imbalanced data-set, under-sampling the majority class may be used as a method to increase sensitivity of the classifier towards the minority class. However, combining under-sampling of the majority class with over-sampling of the minority class leads to an improved classifier performance, as per the ROC curve. The over-sampling involves the creation of synthetic data, which can mitigate the effect of over-fitting – this will be explained in the following paragraphs.

Given a positive training document, its *k*-nearest-neighbours among other positive training documents are first identified. Let $\overrightarrow{\beta_i}$ be the feature vector of document α_i, and $\overrightarrow{\beta_m}$ be the feature vector of one of the *k*-nearest-neighbours of α_i. The feature vector of a synthetic document is created by $(\overrightarrow{\beta_i} + \mu(\overrightarrow{\beta_m} - \overrightarrow{\beta_i}))$ where *μ* is a random value between 0 and 1 (Sun, Lim, & Liu, 2009).

This method requires the decision region of the minority (rare) class to become more general. In other words, the main merit of SMOTE is to generate new rare class instances by interpolating between numerous rare class instances that lie together. Therefore, the problem of over-fitting can be eliminated, as no non-value-adding or repetitive data will be created. This causes the decision boundaries for the rare class to spread further into the prevalent class space (Lin, Chang, & Hsu, 2013).

The mechanics of SMOTE are as follows:

1. Every data point is plotted,
2. The feature vector and its nearest neighbours are identified,
3. The difference between the two data points is calculated,
4. The difference between the two data points is multiplied by a random number between 0 and 1,
5. A new point on the line segment is identified by adding a random number to the feature vector, and then
6. The process is repeated for identified feature vectors.

Figure 5.3 below shows a two-dimensional illustration of a feature space, that is points in a data-set. Each dot in the feature space represents a point in the data-set. The blue points represent the majority class, whereas the orange points within the rectangle represent the minority class. Figure 5.4 represents a zoomed-in view of said minority class and showcases how SMOTE synthesises data. First, the SMOTE algorithm identifies the feature vector and its nearest neighbours, this is illustrated with the orange arrows. After this, the linear distance between the two points is calculated – the feature vectors in the feature space, which is represented by the white dotted lines. The algorithm then multiplies this distance by a random number between 0 and 1, then plots a new data point on the line with the achieved result (green points). The feature vector for this new point (green arrow) is the new synthetic data point. This process is repeated as many times as required to obtain a new synthesised training sample.

Figure 5. 3 Two-Dimensional Illustration of Points in a Data-Set (Feature Space)

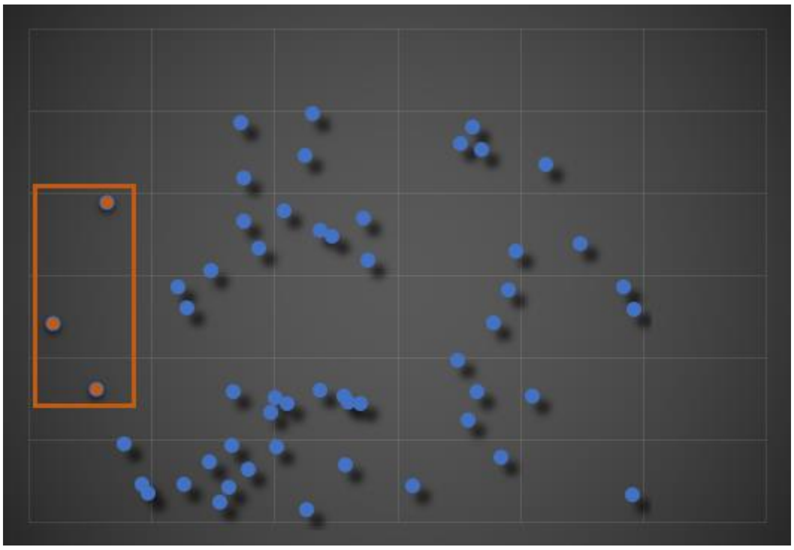
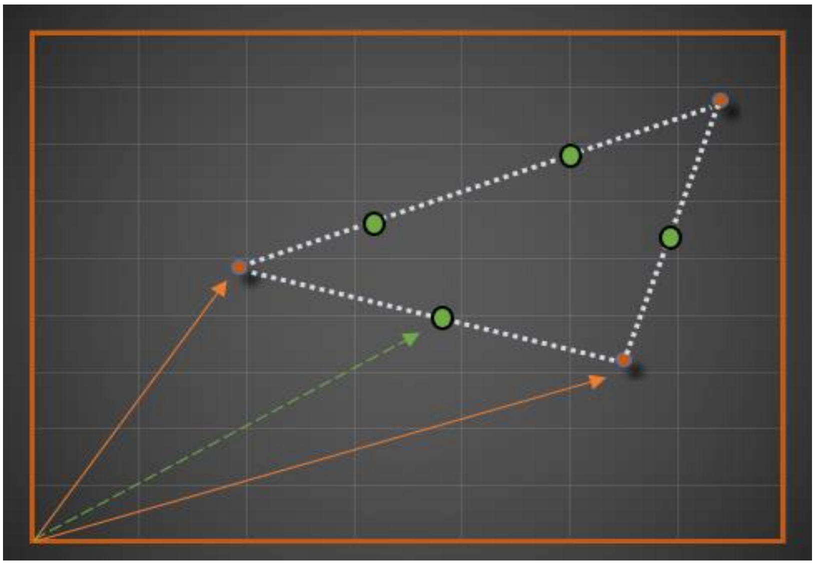


Figure 5. 4 A Zoomed-In View of the Minority Class Showcased in Figure 5.3 with Synthetic Data Points Synthetically Generated in the Feature Space



5.3 Data

The same explanatory variables that were used in Chapter 4 were used in this study. Table 5.1 presents the 19 variables used. The data collected for the companies within the Australian mining sector were extracted from the official portal of MorningStar. Time-series data were then chosen for the years 2011-2015. The outcome was 632 healthy companies and 118 distressed companies. The data were then downloaded to a spreadsheet for cleaning. The initial count was 590 rows (118 companies multiplied by 5 years) for distressed companies and 3160 rows (632 companies multiplied by 5 years) for healthy companies, a total of 3750 rows incorporating data for 29 explanatory variables – 10 were later omitted due to insufficient data. The data cleaning process entailed using a criterion that deletes company information that had 50% or more missing data. This resulted in the final sample containing 19 variables with 3375 rows; 339 rows for distressed companies and 3036 for healthy companies. The companies count was reduced to 631 for healthy companies and 117 for distressed companies. Refer to Table 5.2 for a breakdown of the data used in this study.

Table 5. 1 List of Variables Used in Study

Variable	Description
ROE	Return on Equity = Net Profit After Tax / (Shareholders Equity – Outside Equity Interests)
ROA	Return on Assets = Earnings before interest / (Total Assets Less Outside Equity Interests)
ROIC	Return on Invested Capital = Net Operating Profit Less Adjusted Tax / Operating Invested Capital
Asset Turnover	Operating Revenue / Total Assets
PPE Turnover	Revenue / (Property, Plant & Equipment – Accumulated Depreciation)
Depreciation/PPE	Depreciation / Gross Property, Plant & Equipment
Working Capital Turnover	Operating Revenue / Operating Working Capital
Gross Gearing	(Short-Term Debt + Long-Term Debt) / Shareholders Equity
Financial Leverage	Total Debt / Total Equity
Current Ratio	Current Assets / Current Liabilities
Quick Ratio	(Current Assets - Current Inventory) / Current Liabilities
Gross Debt/Cash Flow	(Short-Term Debt + Long-Term Debt) / Gross Cash Flow
Cash per Share	Cash Flow / Shares Outstanding
Invested Capital Turnover	Operating Revenue / Operating Invested Capital before Goodwill
Net Gearing	(Short-Term Debt + Long-Term Debt - Cash) / Shareholders Equity
NTA per Share	Net Tangible Assets / Number of Shares on Issue
Book Value per Share	(Total Shareholder Equity - Preferred Equity) / Total Outstanding Shares
Sales per Share	Total Revenue / Weighted Average of Shares Outstanding
PER	Price per Earnings = Market Value of Share / Earnings per Share

Following this, a dichotomous binary variable was used to refer to the status of each company – coded ‘1’ if the company is healthy and ‘0’ if the company is distressed. For creating the training sample, the data were split in half by randomly selecting 50% of the observations (3035 ÷ 2 = 1688 rows for training sample). The other half of the observations were used to construct the testing/holdout sample. When creating both the training and testing samples, it is imperative to retain ratio of percentage imbalance in the original observations, as otherwise the generated model will not have a fair representation of the original data – this process and the resulting data sets are summarised in Table 5.2.

Table 5. 2 Data Overview

Sample Partition	Number of Observations	Percentage	Number of Observations of Healthy Companies	Number of Observations of Distressed Companies	Class Imbalance %
Train	1,688	50.00%	1,512	176	89.57% Healthy – 10.43% Distressed
Test/Holdout	1,687	50.00%	1,524	163	90.34% Healthy – 9.66% Distressed
Total	3,375	100.00%	3036	339	89.96% Healthy – 10.04% Distressed

As is evident in Table 5.2, there is an issue of class imbalance in the data-set, meaning that there are much more healthy companies than there is distressed – 89.96% to 10.04%, respectively. As shown in the Train row, half of the observations were split to generate the training sample, and the class imbalance ratio was kept very similar to that of the original data-set's. This is also true for the testing sample, as shown in the Test row. Therefore, using this data-set will be a good representation of the class imbalance problem since the difference between healthy and distressed companies is extreme.

5.4 Methodology

There are three subsections in the Methodology section. The first subsection presents the evaluation methods used in this study for assessing detection accuracy of the created models. The second subsection explains the data-sets used in this study and how the training and testing samples were constructed. The third subsection showcases the models that were created for this study using the following techniques: DT, treebag, RF, and SGB. With the exception of treebag, the mechanics of the aforementioned techniques were already presented in Chapter 2. Therefore, to limit repetition only the mechanics of the treebag technique is presented in this chapter.

5.4.1 Evaluation Methods

The evaluation methods used in this chapter incorporates both visual – as per the ROC graph, and empirical – as per the AUROC score, sensitivity and sensitivity aspects. Combining both aspects reinforces the validity of the results. These evaluation methods are used for all constructed models. As mentioned earlier in this chapter, the justification for using these methods is due to the fact that if one simply observes the overall model's accuracy, one can never know whether the model is considering a 50-50 split or otherwise. Since the holdout sample has a class imbalance of 90.34%, if the model simply classifies all companies as healthy, it will yield a default accuracy of 90.34%. This is a high result at face value, but a deceptive one nonetheless, as it does not take the distressed companies into consideration.

5.4.2 Data-sets

Two data-sets were used in this study's analysis, the original and the *SMOTEd* data-sets, as shown:

- **Original Data-set:** As explained the Data section above, the original data-set was split evenly to create training and holdout samples. This training sample is then tested on the holdout to create the models pertaining to the original data-set – refer to Table 5.3 for samples used in this study. Since the holdout sample contains real-life data, it is also used as the holdout sample for the *SMOTEd* data-set. This ensures unbiasedness when testing for the effectiveness of SMOTE – this is because if the *SMOTEd* data-set was split and the same processes performed as in the case of original data-set, the *SMOTEd* test sample will comprise fictitious/synthetic data. Therefore, testing all of the *SMOTEd* data-set on a holdout sample containing real-life data increases the validity of the results achieved in this study.

- ***SMOTEd* Data-set:** The same training sample from the original set was used to create the *SMOTEd* data-set. The parameters for creating the *SMOTEd* data-set are as follows: Let *k* be the over-sampling ratio, where one synthetic positive training example is generated from each of the its *k*-nearest-neighbours of a positive training example. The rare event needed to be oversampled, therefore *k* was set to 1 – this oversamples the rare events by 100% (doubles them). As for the majority class, it needed be undersampled. Let *j* be the under-sampling ratio – therefore *j* was set to 2 – this undersamples the negative target by twice the amount oversampled, through randomly removing observations from the negative target (successful companies) – as was explained in the Literature Review section. After *SMOTEing*, results yielded a *SMOTEd* data-set with 704 observations – 352 distressed (50%) and 352 healthy companies (50%), thus eliminating the class imbalance problem that existed in the original data-set. The *SMOTEing* process has oversampled the healthy companies from 176 to 352 (100% increase) and has undersampled the distressed companies from 1,688 to 352 (randomly removed 1,336 observations). The *SMOTEd* data-set is more than two times smaller (≈40%) of the size of the original training data-set. Table 5.3 for below presents the two samples used in this study. All of this *SMOTEd* data-set is used to train the various models constructed in this chapter, and is tested on the holdout sample from the original data containing 1,687 observations, as was shown in Table 5.2.

Table 5. 3 Original and *SMOTEd* Samples

Data-set	Sample Partition	Number of Observations	Percentage
Train (2 Options)	Original	1,688	50%
	SMOTEd	704	100%
Holdout Sample for All Models	Total	1,687	100%

5.4.3 Created Models

This subsection explains the models built for this study. Two software packages were used to aid with the analysis, namely: ‘Salford Predictive Modeler’ and ‘R’ software package. ‘R’ is a programming language commonly used for statistical and machine learning modelling. It provides both empirical and graphical outcomes that aids statisticians in their analyses (R, 2019). The ‘R’ software package has been used in many studies across various disciplines throughout the literature, some of these include: Calenge (2006); Knezevic, Streibig, and Ritz (2007); Noguchi, Gel, Brunner, and Konietschke (2012). Whereas, ‘Salford Predictive Modeler’ is a platform that is used for developing both statistical and cutting-edge tree-based models that can deal with complex data – this software has been used previously in the literature (Gepp & Kumar, 2012; Gepp et al., 2010). ‘R’ was used to develop the treebag model, whereas ‘Salford Predictive Modeler’ was used to develop the DT, RF, and SGB models. To minimise repetition, the mechanics of these techniques will not be presented, refer to Chapter 2 for in-depth analysis of the aforementioned techniques.

5.4.3.1 Decision Tree Models

Two models were created using the DT technique, one using the original data-set and the other using the *SMOTEd* data-set. Building the DT models had the following properties – all are commonly used metrics:

- Testing method to determine optimal size was based on random selection of 50% of the cases;
- The parameters influencing the selection of the best tree were based on commonly used criteria:
  - a) standard error rule: minimum cost tree regardless of size,
  - b) variable importance formula: all surrogates count equally;
- The splitting method for the classification trees was the popular Gini criterion.

5.4.3.2 Treebag Models

Treebag is R language denoting an ensemble of machine learning algorithms for creating a bagging framework that can be used for classification or regression modelling. In brief, it is a recursive partitioning technique that constructs many individual tree models from disconnected subsections of training data, then builds an aggregated and superior model (Brownlee, 2016). The model was trained using the "caret" package on the training sample using the commonly used fivefold cross validation. Whether the company is healthy or distressed, was set as the response variable, whereas everything else were set as predictors.

Two models were created using treebag – one using the original data-set, and the other using the *SMOTEd* data-set.

5.4.3.3 Random Forests Models

Two models were constructed using the RF technique. The same training process in terms of data-sets was used. Construction of the data-sets had the properties shown below – they are all commonly used criteria. The other parameters influencing the model were kept as per default criteria.

- Number of trees built: 1,000
- Number of predictors: Square root (√19 ≈ 4)
- Testing method was based on the random selection of 50% of the cases

5.4.3.4 Stochastic Gradient Boosting Models

As with the aforementioned models, two models were built using the SGB technique. The same training process in terms of data-sets was used. Testing of the data-sets had the properties shown below – they are all commonly used criteria. The other parameters influencing the model were kept as per default criteria.

- Number of trees built: 1,000
- Testing method was based on the random selection of 50% of the cases
- Maximum nodes per tree: 6
- Criterion for selection optimal number of trees for model: AUROC

5.5 Results

This section presents the results in this study for the four techniques used after they have been tested on the holdout sample containing 704 observations, as was shown in Table 5.3. The results are in terms of ROC graphs, AUROC scores, as well as sensitivity and specificity scores for the recursive partitioning models. Refer to the Appendices section (Appendix 1) for the raw R-code and data summary.

5.5.1 Decision Tree Models

5.5.1.1 AUROC Results

The AUROC scores of the models using DT are as follows:

- **Original:** The treebag model using the original data-set yielded an AUROC result of 0.5794.

- ***SMOTEd:*** The treebag model using the *SMOTEd* data-set yielded an AUROC result of 0.6179 – hence the superior model.

5.5.1.2 ROC Results

As for the ROC graphs, as is evident in Figures 5.5 and 5.6, the model's line (blue) of the model using the *SMOTEd* data-set, runs closer to the Y-axis, thus encompassing a larger area beneath it. If the visual representation is not clear, then refer to the AUROC score.

Figure 5. 5 ROC of Original Model

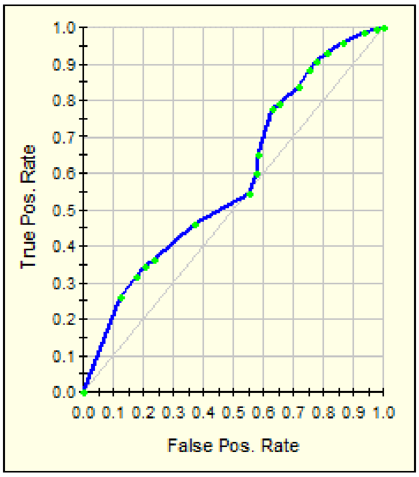
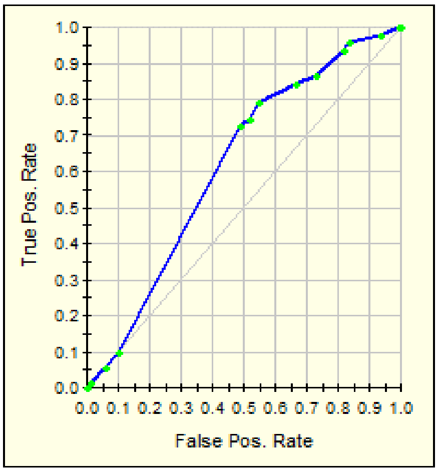


Figure 5. 6 ROC of *SMOTEd* Model



5.5.2 Treebag Models

5.5.2.1 AUROC Results

The AUROC scores of the models using treebag are as follows:

- **Original:** The treebag model using the original data-set yielded an AUROC result of 0.5736.

- ***SMOTEd:*** The treebag model using the *SMOTEd* data-set yielded an AUROC result of 0.6388 – hence the superior model.

What is remarkable here, is that a data-set that is not only much smaller than the original one, but also contains synthesised or fictitious data, was able to outperform the predictive accuracy of a model that is much larger and contains real data. This seems to be at odds with generic statistical rules which state that the larger the sample size is, the more accurate the representation of the population is, but since class imbalance exists, the results are sensible.

5.5.2.2 ROC Results

As for the ROC graphs, Figures 5.7 and 5.8 below present the ROC graphs for the original and *SMOTEd* models, respectively. The black lines represent the models' predictive performance. The grey lines are there just for illustrative purposes of a model with no discerning or distinguishing capabilities between the classes. As explained in the Introduction section, the closer the model's line (the black line in this example) runs to the Y-axis, and then veers right parallel to the X-axis, the more area it encompasses – thus indicating a model with superior predictive power. As is evident in the graphs, the black line of the *SMOTEd* model, runs closer to the Y-axis, thus encompassing a larger area beneath it, which is reflected in the higher AUROC score of the *SMOTEd* Model vis-à-vis the Original Model.

Figure 5. 7 ROC of Original

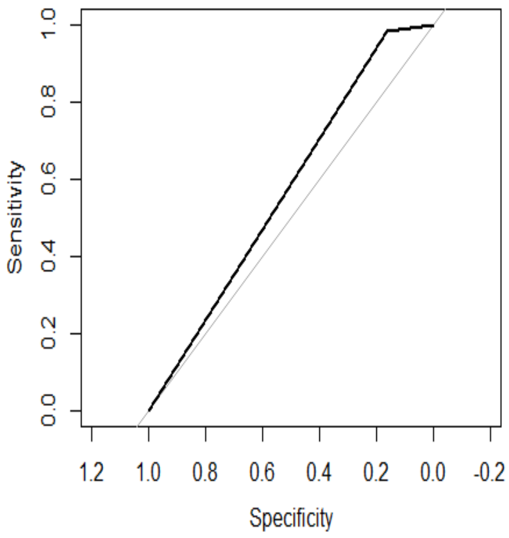
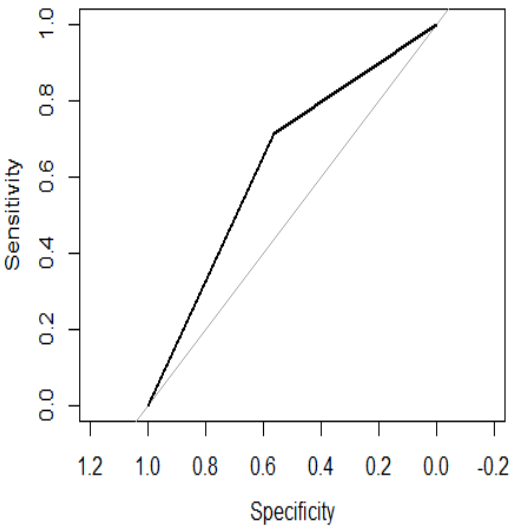


Figure 5. 8 ROC of *SMOTEd*



These results visually verify that applying SMOTE to an imbalanced data-set yields a higher predictive accuracy, tested using a treebag model.

5.5.3 Random Forests Models

5.5.3.1 AUROC Results

The AUROC scores of the models using RF are as follows:

- **Original:** The treebag model using the original data-set yielded an AUROC result of 0.7045 – hence is slightly the superior model (0.89% greater than the *SMOTEd* model stated below).
- ***SMOTEd:*** The treebag model using the *SMOTEd* data-set yielded an AUROC result of 0.6983.

5.5.3.2 ROC Results

As for the ROC graphs, the model's lines in Figures 5.9 and 5.10, look very similar in terms of area encompassed under the curve, therefore, it is prudent to check the AUROC score to make an empirical determination to as which is the superior model. The AUROC score of the *SMOTEd* Model was 0.6983 vis-à-vis the 0.7045 for the Original Model. These AUROC scores are very similar, as they are only about 0.89% apart. So, despite, the Original Model having an ever so slightly higher AUROC score, both model's detection accuracies are essentially the same.

Figure 5. 9 ROC of Original

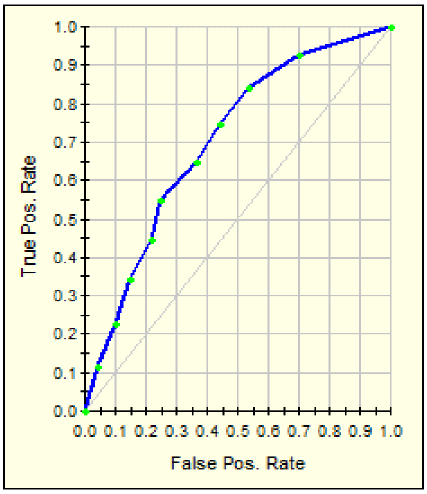
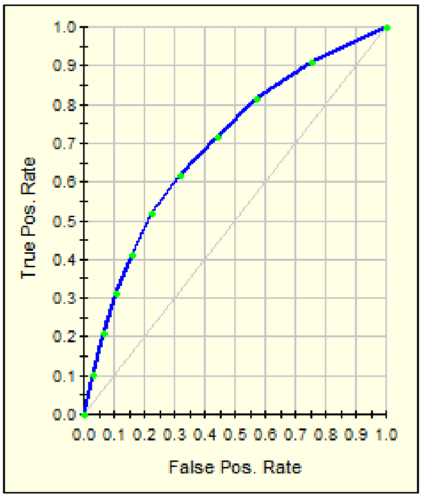


Figure 5. 10 ROC of *SMOTEd*



5.5.4 Random Forests Models

5.5.4.1 AUROC Results

The AUROC scores of the models using SGB are as follows:

- **Original:** The treebag model using the original data-set yielded an AUROC result of 0.6730.

- ***SMOTEd:*** The treebag model using the *SMOTEd* data-set yielded an AUROC result of 0.7103.

5.5.4.2 ROC Results

As for the ROC graphs, as is evident in Figures 5.11 and 5.12, the model's line (blue) of the *SMOTEd* model, runs closer to the Y-axis, thus encompassing a larger area beneath it.

Figure 5. 11 ROC of Original

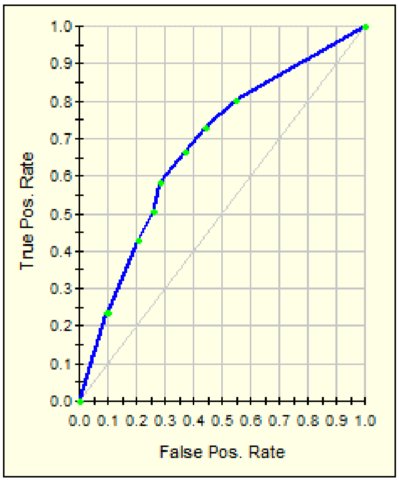
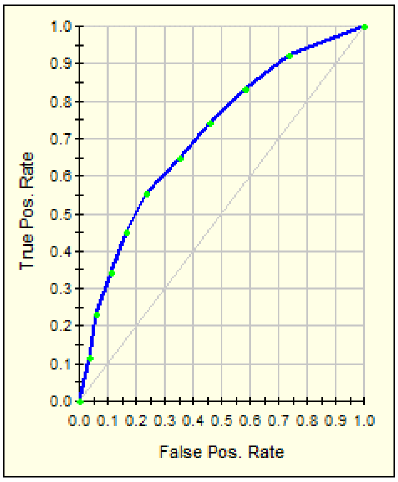


Figure 5. 12 ROC of *SMOTEd*



5.5.5 Model Comparison

This subsection presents the models' AUROC, specificity, and sensitivity in a tabulated fashion. The tables presented below allow for convenient comparisons to be made in order to deduce whether using SMOTE yielded empirically superior models vis-à-vis models created using the original data-set.

Table 5.4 below shows the simple averages of the sensitivity and specificity scores for all of the models created. As is evident, all the models using the *SMOTEd* data-set have outperformed the models using the original data-set.

Table 5. 4 Models' Specificity & Sensitivity Average Result Comparison

Data/Model	**DT**	**Treebag**	**RF**	**SGB**
Original Data	57.15%	57.36%	63.81%	50.00%
***SMOTEd* Data**	61.54%	63.88%	64.10%	63.80%

Table 5.5 below shows all of the models' AUROC scores. As is evident, the models using the *SMOTEd* data-set have yielded a higher AUROC score for all the data-sets, except for the RF model, which is only 0.89% greater, thus essentially the same score. Again, this empirically proves SMOTE's superiority vis-à-vis the original data-set.

Table 5. 5 Models AUROC Result Comparison

Data/Model	**DT**	**Treebag**	**RF**	**SGB**
Original Data	0.5794	0.5736	0.7045	0.6730
***SMOTEd* Data**	0.6179	0.6388	0.6983	0.7103

These results clearly indicate that building models using the *SMOTEd* data-set yields empirically superior results to those using real data. This is showcased in two areas, firstly, both the AUROC, and the specificity and sensitivity averages, yielded higher scores for the models using the *SMOTEd* data-set (except for the RF model, as they are almost the same); and secondly, the increase in accuracy across the various models conform more so with the literature when using the *SMOTEd* data-set as opposed to real data. This is in terms of the predictive accuracy of tree ensembles over single tree techniques (RFs/SGB>treebag>DTs). As is clear in Table 5.5, all tree ensemble models using the *SMOTEd* data-set outperformed the models using the original data-set, as measured by the AUROC criterion. The results also show that even with the recursive partitioning models' resilience to class imbalance, using a *SMOTEd* data-set yields more accurate detection accuracy scores. This is an important finding that contributes towards the literature through recommending the use of SMOTE even when using machine learning techniques due to empirically superior results, as was shown in this chapter.

5.6 Conclusion

This chapter presented the application of Synthetic Minority Oversampling Technique (SMOTE) to an imbalanced data-set comprising 748 Australian mining – 631 of which are financially healthy and 117 are distressed. Four machine learning tree-based techniques were used to create the models for this study. For comparison purposes, the models were trained on two data-sets, the original imbalanced data-set and a balanced *SMOTEd* data-set, in order to empirically deduce the detection accuracy of SMOTE. A holdout sample using real-life data were used to test the accuracy of the aforementioned trained models using both data-sets. The results indicated that despite the *SMOTEd* data-set being around 80% smaller than the original, it resulted in superior detection accuracy. This was measured by AUROC, specificity, and sensitivity results. The AUROC results showed the superiority of SMOTE for the DT and SGB models, as for RF, the scores were almost identical pre and post SMOTE. This study has showcased that using SMOTE is not only easier to handle due to the smaller data-set, but is also empirically superior to the original class imbalanced data-set. This research has contributed towards the literature by investigating the detection accuracy of SMOTE using a multi-approach system and recommending the use of SMOTE even when using machine learning techniques due to empirically superior results. This chapter has verified *Hypothesis 4*.

Chapter 6: Financial Distress Prediction Index (FDPI)

Aligning with *Hypothesis 6* stated in the Chapter 1, namely:

> ***H*₆:** Creating an FDP index is more accurate, informative, and user-friendly than solely relying on standard FDP models.

This chapter will verify the aforementioned hypothesis by creating FDP indices and comparing them to a standard FDP model constructed using LR.

6.1 Introduction

Indices provide a quick and user-friendly way of relaying relevant information to the user. Developing indices is increasingly becoming a popular method of relaying information in a quick and effective manner that is easily interpreted by the general public (Nardo et al., 2005). Rating mechanisms are usually used to rank or rate the performance of companies, countries, sports teams, and medicines, to name a few. The Council on Foreign Relations (2015) outlines the most internationally well-known indices pertaining to companies' financial ratings, namely: Moody's, S&P, and Fitch – they are known as the *Big Three* and encompass around 95% of the global credit ratings' market share.

The advantages of using an index to make decisions relating to companies include:

- Enables ease of interpretation, understandability, and user friendliness;

- Enables banks and lenders to easily assess a company's financial distress probability before determining whether a loan is suitable, and if so, how much interest to charge;

- Allows governments and watchdog institutions to utilise the models to focus on companies with high financial distress probabilities;

- Allows existing and potential stockholders to use the indexes to make more informed investment decisions for best Return on Investment (ROI) opportunities;

- Provides conciseness through reducing the number of variables, that is, a solitary index can showcase the ranking of the desired data-set, which paves the way for prompt decision making processes and easy comparisons;

- Enables other stakeholders and potential merger companies to assess the likelihood of a company's failure or success, as an indicator of whether there will be sustainable benefits gained from continuous operation with the company at hand (Gepp & Kumar, 2012; Krishnan, 2010).

One might query why a Financial Distress Prediction Index (FDPI) is needed when you can already view the ratings from one of the *Big Three* rating agencies? Some of the disadvantages of relying solely on these ratings include:

- Lack of rating information for many companies – the rating agencies do not provide ratings for all companies worldwide;

- Subscription costs – credit rating agencies get paid either by the entity that requests the rating and/or by the subscribers wishing to view the ratings;

- No intra-rating information provided – the agencies group the companies into different categories, such as: AAA, AA, and AA+, but do not provide a ranking for the companies within each category;

- Not to be used for investment – the rating agencies confess that their ratings reflect their opinion and not to be used as recommendations for investing or divesting (Moody's, 2009; The Telegraph, 2012).

Another valid question regarding the use of FDPI is why not only use FDP modelling (as was used throughout this thesis) to gain information about the prospective company? Some of the disadvantages of relying solely on FDP modelling include:

- Classification and cut-off point problem – as seen in previous chapters, cut-off points were varied experimentally to decide optimal cut-off point, which can be a tedious task. However, the cut-off point does not have to be decided when creating an index;

- Matching problem – there tends to be subjectivity when selecting samples for the model – for example, the problem with determining which successful and bankrupt companies to add or omit from the sample. However, the index is built on all companies;

- Class imbalance problems – validity of the results is in question when the data-set has a big difference in the ratio between successful and distressed companies, since the accuracy of the model will be misleading due to solely classifying by the majority class. This is not an issue in an index, as ranking is done on a case-by-case basis.

FDPI amalgamates the concepts of index construction and FDP modelling – in the sense that it can be used as a tool to gain a prompt indication of a company's financial health. This concept pools the advantages of both approaches, hence increasing the validity of the results achieved. This chapter presents the construction of the indices, then compares the superior index with the LR model to ascertain which is more in line with commonly used performance metrics, namely: market capitalisation and share price. Even though shares can be split, thus affecting price; share price is still commonly considered a performance metric as it is indicative of a company's financial health – generally, a positive correlation exists between share price and company performance (Murphy, 2018). Similarly, market capitalisation is indicative of company size, the higher its value, the more established the company is. On their own, these metrics do not provide a holistic perspective of company performance, since they offer a myopic perspective; whereas the index uses many variables, therefore the results are more robust and comprehensive. This chapter's methodology can be applied to any field across any industry. The premise is that the FDPI index provides a ranking of companies that is more consistent with common performance metrics vis-à-vis the LR model.

6.2 Literature Review

This section covers some key techniques used for constructing indices, such as Principal Component Analysis (PCA) and Factor Analysis, including three different approaches of presenting the indices. The literature pertaining to FDP was presented in Chapter 2.

The PCA technique transforms numerous variables in a data-set into a reduced set of uncorrelated/orthogonal factors, known as the principal components. These principal components account for the lion's share of the variance amongst the set of original variables used. Every component is a linearly-weighted amalgamation of the original variables; the weights for every component are shown by the eigenvectors in the correlation matrix, or the covariance matrix, should the data be standardised. Every principal component’s variance is characterised by the eigenvalue of the matching eigenvector. The order of these principal components places the component which accounts for the largest amount of variation in the original variables on top. The second component is totally uncorrelated with the first one and accounts for the maximum variation that is not accounted to by the first component; this pattern is followed for each component (Krishnan, 2010).

PCA was pioneered by Pearson (1901), this was followed by Hotelling (1933). Future studies include that of Pomeroy, Pollnac, Katon, and Predo (1997), which applied PCA on a survey of 200 houses in the Philippines. The subjects were asked to score ten indicators on a scale from 1-15, to present their opinions on recent community-based coastal resource management projects in their communities. Their results yielded 3 principal components – the first component dealt with the behaviour or community members, whereas the second component dealt with fisheries resource, and the third component was in relation to the well-being of the household. Their principal components explained 66% of the total variance in the model. Further details and various applications of PCA can be seen in Jolliffe (1990).

Factor analysis, also known as ‘spectral decomposition,’ reduces the number of variables used in the model, all the while, capturing most of the information based upon eigenvalues of the covariance matrix. Its major advantage is reducing the number of original variables in the models to a set of factors with no problem of multicollinearity. This technique has been vastly used in the literature pertaining to indicators or constructing indices (Dialga, 2017; Helmes, Goffin, & Chrisjohn, 1998; Pasimeni, 2013).

Factor analysis incorporates PCA and principal factors analysis – PCA being an estimate to the principal factor analysis, especially if the components are rotated. The common rotational approaches are: *quartimax, varimax,* and *equamax*. The aim in adopting a rotational approach is to achieve a clear pattern of loadings for variables, high for some and low for others, in order to help with interpretation. The notion of factor loadings refers to the correlations between the factors and the variables. Varimax rotation is a variance-maximizing approach aiming at maximising the variance of the factor. The main difference between PCA and factor analysis, is that in PCA it is assumed that all variability in a variable must be used in the analysis, whereas in factor analysis, the variability is only used in a variable that is common with the other variables (Krishnan, 2010).

There are many different types of indices and index-construction methods in the literature, they vary in the way they portray their scores, but their core aim is similar – to relay a clear and user-friendly message to the viewer, which enables efficient and effective decision making. Abeyasekera (2005) presents various multivariate approaches found in the literature, mainly using PCA, to construct indices. These approaches are advantageous in a number of ways, including: presenting a complex model in a simple manner, enabling graphical representations, and explores patterns across the variables. Nardo et al. (2005) also provides an invaluable repository on several methods used to construct indices, of which three will be explored in this chapter, namely: the Factor Weighted Index (FWI) approach, the Weighted Factor Loading Index (WFLI) approach, and finally, the Non-Standardised Index (NSI) approach.

The first approach, the Factor Weighted Index (FWI), is constructed using both the original data from each of the variables and the percentage values of the variance explained by each factor in the model using PCA. The data under each variable for each factor is summed to form an aggregated factor. After this, the variance percentage contributions that the first factor contributes towards the model after rotation is divided by the overall percentage explained to yield a weighted score. This weighted score is multiplied by the aggregated factor found earlier, which results in a weighted first factor. This process is done for all factors in the model. Finally, these weighted factors are summed and then numerically sorted to create the index (Abeyasekera, 2005; Nardo et al., 2005; Pomeroy et al., 1997).

The second approach, the Weighted Factor Loading Index (WFLI), is a more complex index-construction method, it involves the following steps: after performing PCA on the data-set, the user checks the variables that make up each factor from the *Component Score Coefficient Matrix* based on their higher loadings. The score associated with each variable is divided by the total score of the variables that make up each factor, thus resulting in a weighted value for each variable. Subsequently, these weighted values are multiplied by the actual values of their respective variables. Following this, the results are aggregated and then multiplied by the weighted variance percentage contribution of each factor. Finally, the results are summed and then numerically sorted to create the index (Nardo et al., 2005).

Lastly, the Non-Standardised Index (NSI) approach, uses PCA on the data-set. Following this, the percentage of variance explained by each factor is divided by the total variance explained by the model. That is then multiplied by each factor score, and is finally aggregated. This yields a single score for each data point which are used to create the NSI (Nardo et al., 2005).

The index-construction approaches can be standardised, that is, having a score for each case in the data-set ranging from 0 to 100 to be more presentable and easily understood. The Standardised Index (SI) is used extensively in the literature, some of the studies include: Antony and Rao (2007); Hightower (1978); Krishnan (2010); Sekhar, Indrayan, and Gupta (1991). Further details of the aforementioned index-construction approaches are provided in Methodology section.

As is evident, there are different methods of constructing and presenting indices, however, there are no studies that have combined the concepts of indices and FDP modelling. This research spearheads this initiative in the hope of encouraging further research to be done in this area in the future.

6.3 Data

The data-set used in the study was extracted from the Capital IQ database. Financial data were collected for 779 mining companies listed on the Australian Stock Exchange (ASX) for the financial year of 2015. The study incorporated 27 explanatory variables – refer to Table 6.1 for a comprehensive and explanatory list of the variables used in the study. The variables were comprised of standard accounting and financial information, chosen based upon several factors, including use in the literature and as per availability of data.

Table 6. 1 Variables Used in this Chapter

Variable	Description
ln Total Assets	Natural Logarithm value of Total Assets
ln Current Liabilities	Natural Logarithm value of Current Liabilities
ln Current Assets	Natural Logarithm value of Current Assets
ln Cash & Equivalents	Natural Logarithm of Cash and Equivalents
Net Working Capital	Net Working Capital
ln Market Capitalisation	Natural Logarithm value of Market Capitalisation
Cash per Share	Cash / Share
Net Income	Net Income – measured in $ (millions)
Operating Income	Operating Income – measured in $ (millions)
Gross Profit	Gross Profit – measured in $ (millions)
Retained Earnings	Retained Earnings – measured in $ (millions)
Accounts Receivable	Accounts Receivables – measured in $ (millions)
Inventory	Inventory – measured in $ (millions)
Long-Term Debt	Long-Term Debt – measured in $ (millions)
Current Ratio	Current Assets / Current Liabilities
Quick Ratio	(Total Cash & Short-Term Inventory + Accounts Receivables) / Current Liabilities
ROA	Return on Assets = Income / Total Assets
ROC	Return on Capital = Income / Average Total Capital
ROE	Return on Equity = Earnings from operations / Average Total Equity
ROIC	Return on Investment Capital = (Net Income - Tot Dividends Paid) / Capital
SGA Margin	Selling, General, & Administration Expenses Margin = (SGA Expense/Total Revenue)
Total Assets Turnover	Total Revenue / Average Total Assets
Fixed Assets Turnover	Total Revenue / Average Net Property, Plant & Equipment
Accounts Receivables Turnover	Total Revenue / Average Accounts Receivables
TD/TC	Total Debt / Total Capital
TL/TA	Total Liabilities / Total Assets
Altman Z-Score	Z = 1.2*(Working Capital/TA) + 1.4*(Retained Earnings/TA) + 3.3*(EBIT/TA) + 0.6*(Market Value of Equity/Book Value of TL) + 1.0*(Sales/TA)

6.4 Methodology

Before developing the index, it is prudent to check the Kaiser-Meyer-Olkin (KMO) – a measure of sampling adequacy – that is used to check for multicollinearity in the data-set, in order to determine the suitability of carrying out a factor analysis. The sampling adequacy forecasts whether the data is likely to factor properly based on correlations and partial correlations. If the variables do have common factors, the partial correlation coefficients should be marginal in relation to the total correlation coefficient. The maximum score for the KMO statistic is 1. Following this, a test of the strength of the relationship among variables was executed using the Bartlett (1954) *test of sphericity*. This test tests the null hypothesis that the variables in the population correlation matrix are not correlated with the alternative that they are correlated.

In this study, factor analysis was executed by including all 27 variables and financial data for the 779 companies. Factor analysis was chosen to lessen the number of dimensions and provide a concise set of factors with no problem of multicollinearity. Principal Component Analysis (PCA) was chosen as the extraction method and Varimax with Kaiser Normalisation as the rotation method, since this is a prevalent method with success in the literature. The commonly used Kaiser's criterion, or the eigenvalue rule, retains only the factors with an eigenvalue of 1.0 or more.

A graphical method, known as the Cattell (1966) scree test – shown later in Figure 6.1 – was produced to showcase the plots of each of the eigenvalues of the factors. The user can visually inspect the plot to pinpoint where the smooth decrease of eigenvalues seems to plateau. After this point, what is found is only 'factorial scree,' that is, debris that accumulates on the lower part of a rocky slope. This means that the marginal value from additional factors is minimal and is likely outweighed by the negative of the additional complexity.

After conducting factor analysis, the factor scores for each factor were tabulated. Following this, constructing the FDP index was initiated. Each approach outlined above is discussed in more detail in separate sections below.

6.4.1 Factor Weighted Index

The FWI was constructed using the original data from the variables after the performing factor analysis. All the variables in each corresponding factor were aggregated to form factors for each company in the data-set. Following this, the percentage contributions of each factor towards the model were multiplied by the preceding sums to produce weighted scores for each company – this was done for all eight factors. After this, the weighted scores for each company under each factor were aggregated to produce a single aggregate score for each company. For ease of interpretation, the scores from each company were then standardised to provide a score falling between the 0 to 100 range. See Equations 6.1-6.4 below.

$$w_i = \frac{v_i}{\sum v} \qquad [Equation\ 6.1]$$

- w_i = Weight of ith factor
- v_i = Percentage value of variance explained by ith factor
- v = Total variance explained in the model

$$F_i = (\sum x_i).w_i \qquad [Equation\ 6.2]$$

- F_i = Weighted factor
- x_i = Variables pertaining to its respective factor

$$FWI = \sum F_i \qquad [Equation\ 6.3]$$

$$Standardised\ FWI = \frac{FWI\ value\ of\ each\ case - Min\ FWI}{Max\ FWI - Min\ FWI} * 100$$

$$[Equation\ 6.4]$$

- $Min\ FWI$ = Minimum FWI value
- $Max\ FWI$ = Maximum FWI value

6.4.2 Weighted Factor Loading Index

The WFLI was constructed by firstly inspecting the percentage contribution of each factor towards the model, as shown in the *Total Variance Explained Table*, under the *% of Variance* – this is illustrated in Table 6.3 in the Results section. Each factor's variance contribution percentage was divided by the total variance of the factors to achieve a percentage contribution out of 100 for each factor. Subsequently, the *Component Score Coefficient Matrix* was inspected – this is shown in Table 6.7 in the Results section. This table shows the contribution score of each variable towards its associated factor. Firstly, the scores of the variables that make up each factor were summed to create an aggregate score for each factor.

After this, each variable's individual score was divided by the aggregate score for its associated factor to yield a weighted score, that is, a percentage contribution of each variable towards the factor. Following this, the weighted score of each variable was multiplied by the actual data for each company to yield a weighted value of each variable to each factor. Then, these weighted values were aggregated to achieve a sum of the weighted variables for each factor. This sum for each company was then multiplied by the percentage contribution of each factor, as explained earlier; which yielded a weighted score for each company for each factor. Subsequently, these values for each factor were summed to yield a final score for each company. Finally, the scores were standardised to perform an index for companies with a range of 0 to 100. See equations 6.5-6.8 below.

$$\varphi_i = \frac{\gamma_i}{\sum \gamma} \qquad [Equation\ 6.5]$$

- φ_i = Weight of ith factor
- γ_i = Score of ith variable
- γ = Total score of variables

$$\theta_i = \sum \varphi_i . x_j \qquad [Equation\ 6.6]$$

- θ_i = Weighted factor
- x_j = Variable pertaining to its respective factor

$$WFLI = \sum \theta_i . v_i \qquad [Equation\ 6.7]$$

- v_i = Percentage value of variance explained by ith factor

$$Standardised\ WFLI = \frac{WFLI\ value\ of\ each\ case - Min\ WFLI}{Max\ WFLI - Min\ WFLI} * 100$$

$$[Equation\ 6.8]$$

- $Min\ WFLI$ = Minimum WFLI value
- $Max\ WFLI$ = Maximum WFLI value

6.4.3 Non-Standardised and Standardised Indices

The first step is estimating the component scores by adopting the Non-Standardised Index (NSI) method used by Krishnan (2010). Equation 6.9 shows the methodology for computing the NSI. The percentage of variance explained by each factor was divided by the total variance explained by the model, then multiplied by each factor score before being summed. This yields a single score for each company which holistically generates an NSI.

$$NSI = \left(\frac{V_1}{V_t} * F_1\right) + \left(\frac{V_2}{V_t} * F_2\right) + \cdots \left(\frac{V_i}{V_t} * F_i\right) \qquad [Equation\ 6.9]$$

- V_i = Proportion of variance explained by ith factor
- V_t = Total variance explained by the model
- F_i = Factor score of ith factor

To convert this NSI to a Standardised Index (SI), the methodology used by Krishnan (2010) was adopted. Equation 6.10 shows this study's methodology for computing the SI. The minimum and maximum values within the generated NSI were retrieved. Next, the minimum value from the NSI is deducted from the component score for each company within the NSI, then divided by the maximum minus the minimum values, before multiplying by 100 to achieve the SI that ranges from 0 to 100. The results are then ordered from largest to smallest – with companies having the highest scores being the least financially distressed, and companies with the lowest scores being the most financially distressed.

$$SI = \frac{NSI\ value\ of\ each\ case - Minimum\ NSI}{Maximum\ NSI - Minimum\ NSI} * 100 \qquad [Equation\ 6.10]$$

- $Min\ NSI$ = Minimum NSI value
- $Max\ NSI$ = Maximum NSI value

6.4.4 Index Comparison

As a means to compare and contrast the performance of the optimal index, a well-established technique, namely, Logistic Regression (LR), was used, to test whether the results achieved when creating the FDPI are similar to that in FDP. Model and index construction were done on a training data-set comprising randomly chosen 80% of the data with the remaining 20% being used for testing. The results from the LR model produce probability scores for each company. These scores can then be used for other purposes – in this case for ranking. The scores were tabulated and organised from largest to smallest, with the highest score indicating the company with the least financial distress, and inversely, the companies with the lowest scores are the most financially distressed.

6.5 Results

6.5.1 Pre-Index Validation Checks

As shown in Table 6.2, KMO result was 0.73. According to Antony and Rao (2007), a value of 0.9 is considered *marvellous*, 0.80, *meritorious*; 0.70, *middling*; 0.60, *mediocre*; 0.50, *miserable*. Therefore, the score lies between the *meritorious* and *marvellous* rankings, which indicates the suitability of using factor analysis for the study. The result of Bartlett's Test of Sphericity shows a significance level of 0.00, thus indicating that the null hypothesis can be rejected as it is less than the level of significance of 0.05. Therefore, it was certain that the correlation matrix is not an identity matrix, or the relationship strength amongst the variables is strong, as is essential by factor analysis to be effective. All in all, the aforementioned diagnostic tests validate that factor analysis is fitting for this analysis.

Table 6. 2 KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**0.733561131**
Bartlett's Test of Sphericity	Approx. Chi-Square	**21477.86683**
	df	**351**
	Sig.	**0.00**

The scree plot – presented in Figure 6.1 below – shows a downward sloping curve with the eigenvalues on the Y-axis and factor numbers on the X-axis. The point where the slope of the curve is levelling-off indicates the most efficient number of factors that should be generated by the model. As is clear in the graph, the decision as to where the line plateaus is not clear-cut and can be subjective. Therefore, the scree plot should be used in conjunction with the empirical results showcased in Table 6.3 under the "Rotation Sums of Squares Loadings" section, which indicate that the optimal number of factors in the models is eight, explaining 81.32% of variation in the data.

Figure 6. 1 SPSS Factor Analysis Scree Plot

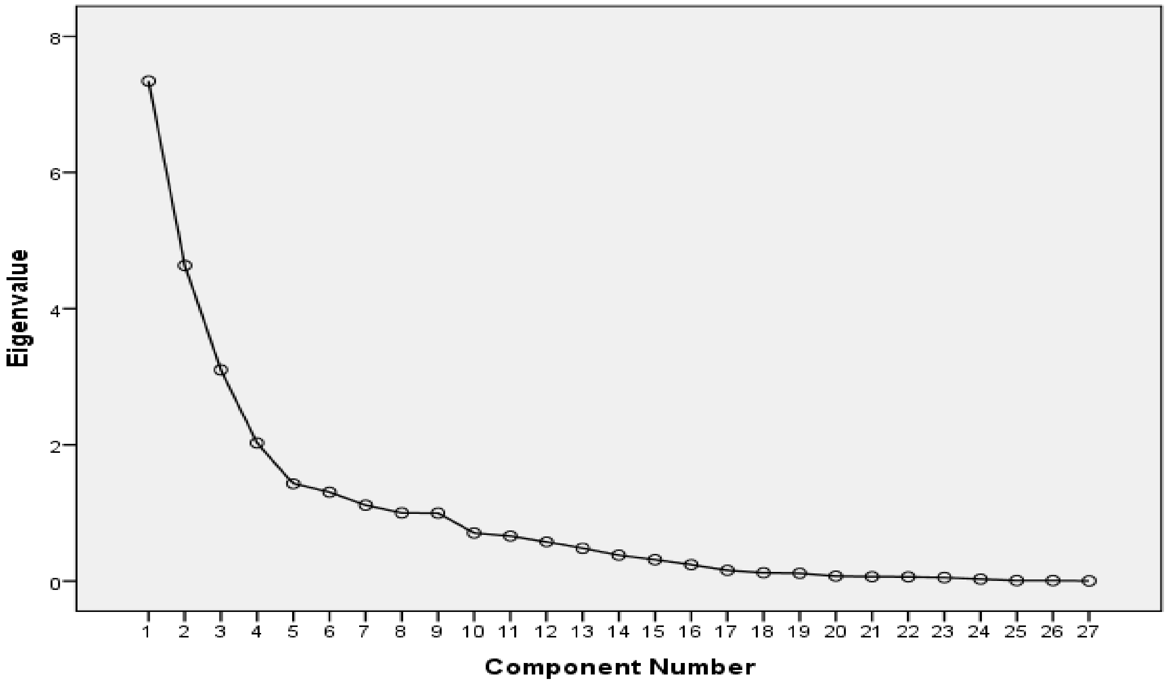


Table 6.3 below presents the contribution that each of the eight factors provides towards the total variance explained by the model – this is shown in the 'Rotation of Sums of Squared Loadings' section. The overall variance explained by the model equals 81.328%.

Table 6. 3 Total Variance Explained – Extraction Method: PCA

Total Variance Explained									
	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.345	27.204	27.204	7.345	27.204	27.204	5.821	21.559	21.559
2	4.636	17.172	44.377	4.636	17.172	44.377	4.647	17.213	38.772
3	3.100	11.482	55.859	3.100	11.482	55.859	3.660	13.554	52.326
4	2.028	7.513	63.371	2.028	7.513	63.371	2.193	8.123	60.448
5	1.430	5.295	68.666	1.430	5.295	68.666	1.882	6.970	67.419
6	1.304	4.829	73.495	1.304	4.829	73.495	1.444	5.350	72.768
7	1.114	4.127	77.623	1.114	4.127	77.623	1.302	4.822	77.590
8	1.000	3.705	81.328	1.000	3.705	81.328	1.009	3.738	**81.328**
9	0.998	3.695	85.023						
10	0.708	2.624	87.647						
11	0.664	2.458	90.105						
12	0.576	2.133	92.238						
13	0.479	1.772	94.011						
14	0.378	1.399	95.410						
15	0.311	1.153	96.563						
16	0.239	0.887	97.450						
17	0.157	0.580	98.030						
18	0.121	0.449	98.479						
19	0.114	0.421	98.900						
20	0.072	0.265	99.166						
21	0.064	0.238	99.404						
22	0.062	0.230	99.634						
23	0.052	0.191	99.825						
24	0.029	0.106	99.932						
25	0.008	0.030	99.962						
26	0.008	0.029	99.991						
27	0.002	0.009	100.000						

The results of PCA using varimax rotation are presented in Table 6.4. As shown in the table, each variable contributes a certain loading towards the overall model. The group of variables that offer strong loadings towards the factor, be it positive or negative, have been highlighted. These loadings are correlation coefficients of each variable with the factor, therefore range from -1 to +1. The eight factors in the model were subsequently named according to the variables they are comprised of; they are shown in Table 6.5. For example, Factor 1 (F1) was dubbed the ‘Balance Sheet and Income Statement’ factor – this is due to the variables that it represents being found in the aforementioned financial statements.

Table 6. 4 Rotated Component Matrix with PCA Extraction and Varimax Rotation

Rotated Component Matrix

	Component							
	1	2	3	4	5	6	7	8
Operating Income	0.991							
Retained Earnings	0.978							
Gross Profit	0.978							
Accounts Receivables	0.923	0.206						
Inventory	0.890	0.262				0.263		
Long-Term Debt	0.854	0.285				0.263		
Net Income	0.524	-0.441	0.112			-0.223	0.260	
ln Current Assets	0.149	0.901	0.173	0.164			0.115	
ln Market Capitalisation	0.205	0.875	0.101					
ln Total Assets	0.170	0.851	0.385					
ln Cash & Equivalents	0.139	0.835	0.169	0.228	-0.103			
ln Current Liabilities	0.165	0.819		-0.324	0.152		0.161	
Return on Capital		0.181	0.945					
Return on Investment Capital		0.147	0.942					
Return on Assets %		0.217	0.898					
Return on Equity			0.754		-0.179			
Quick Ratio				0.984				
Current Ratio				0.983				
Total Liabilities / Total Assets			-0.390		0.797			
Total Debt / Total Capital					0.795		-0.105	
Altman Z-Score			0.287	0.174	-0.709			
Net Working Capital	-0.112	0.177				-0.882		
Cash per Share	0.349	0.414				0.672		
Accounts Receivables Turnover		0.263					0.734	
Total Assets Turnover		0.423	0.115				0.671	
Fixed Assets Turnover							0.374	
Selling & Admin. Expenses Margin								0.996

Table 6. 5 Factor Names

Factor	Name
F1	Balance Sheet and Income Statement
F2	Monetary Figures
F3	Investment Ratios
F4	Liquidity Ratios
F5	Credit Default Ratios
F6	Efficiency Ratios
F7	Revenue Ratios
F8	Short-term Ratio

6.5.2 Factor Weighted Index Construction

Table 6.6 presents the ranking of the top ten and bottom ten companies according to the standardised FWI. Laneway Resources Limited was the topmost ranked company with an index value of 100, whereas Atlas Iron Limited was the lowest ranked company with an index value of 0.

Table 6. 6 Top 10 and Bottom 10 Mining Companies according to the FWI

Index Value	Australian Mining Company Name
100	1. Laneway Resources Limited (ASX: LNY)
52.14	2. G8 Communications Limited (ASX: G8C)
5.51	3. 4DS Memory Limited (ASX: 4DS)
2.19	4. BHP Billiton Limited (ASX: BHP)
1.20	5. Dourado Resources Limited
1.00	6. China Waste Corporation Limited (ASX: CWC)
0.85	7. Rio Tinto Limited (ASX: RIO)
0.60	8. Corizon Limited (ASX: CIZ)
0.58	9. Genesis Resources Limited (ASX: GES)
0.44	10. Pawnee Energy Limited
0.05	770. Sundance Energy Australia Limited (ASX: SEA)
0.05	771. Energy Resources of Australia Limited (ASX: ERA)
0.05	772. Aurelia Metals Limited (ASX: AMI)
0.05	773. Silver Lake Resources Limited (ASX: SLR)
0.05	774. Coal of Africa Limited (ASX: CZA)
0.05	775. Resolute Mining Limited (ASX: RSG)
0.04	776. Wollongong Coal Limited (ASX: WLC)
0.03	777. Paladin Energy Limited (ASX: PDN)
0.03	778. Mount Gibson Iron Limited (ASX: MGX)
0	779. Atlas Iron Limited (ASX: AGO)

6.5.3 Weighted Factor Loading Index Construction

Table 6.7 presents the component score coefficient matrix which showcases all the variables and their respective component score. The scores highlighted with the same colour correspond to the variable(s) that make up each respective factor.

Table 6.8 presents the ranking of the top ten and bottom ten companies according to the standardised WFLI. Laneway Resources Limited was the topmost ranked company with an index value of 100, whereas Atlas Iron Limited was the lowest ranked company with an index value of 0.

Table 6. 7 Component Score Coefficient Matrix

Component Score Coefficient Matrix								
	Component							
	1	2	3	4	5	6	7	8
Ln of TA	-0.015	0.191	0.052	-0.040	0.020	-0.042	-0.103	0.014
Ln of CL	-0.015	0.190	-0.028	-0.126	0.047	-0.043	0.011	0.048
Ln of CA	-0.019	0.219	-0.048	0.083	-0.029	-0.056	-0.003	-0.044
Ln of Cash & Equivalent	-0.021	0.204	-0.050	0.108	-0.048	-0.022	-0.008	-0.051
Net Work Cap	0.046	0.098	-0.032	0.020	0.004	-0.659	-0.119	0.015
Ln of Market Cap	-0.011	0.220	-0.056	-0.012	-0.035	-0.037	-0.064	0.028
Cash/Share	-0.014	0.067	-0.015	0.024	0.013	0.456	-0.010	0.010
Net Income	0.151	-0.188	0.061	0.050	-0.001	-0.189	0.270	-0.033
Operating Income	0.188	-0.060	0.010	0.008	-0.003	-0.031	0.040	-0.005
Gross Profit	0.194	-0.029	-0.005	0.006	-0.007	-0.141	-0.008	-0.001
Retained Earnings	0.182	-0.039	0.000	0.001	-0.007	-0.028	-0.005	-0.004
Acc Rev	0.172	0.003	-0.015	0.002	-0.004	-0.101	-0.017	0.006
Inventory	0.137	0.017	-0.012	-0.015	-0.001	0.107	-0.057	0.007
Long-Term Debt	0.128	0.030	-0.012	-0.017	0.008	0.109	-0.095	0.009
Current Ratio = CA/CL	0.002	0.011	-0.024	0.464	0.045	-0.012	0.028	0.027
Quick Ratio = (Tot Cash and Short term Inv + Acc Rec) / TCL	0.001	0.011	-0.025	0.464	0.046	-0.009	0.028	0.028
Return On Assets % (EBIT * (1-.375)/ Avg TA	-0.004	-0.035	0.279	-0.014	0.085	0.005	0.009	0.020
Return On Capital % (EBIT * (1-.375)/ Avg Tot Capital)	-0.004	-0.046	0.298	-0.020	0.082	0.012	-0.013	0.009
Return On Equity % (Earnings from Cont Operations/ AVG Tot Equity)	-0.002	-0.052	0.230	-0.035	0.002	0.024	-0.007	-0.018
Return on Investment Capital % (Net Income - Tot Dividends Paid)/ Tot Capital	-0.004	-0.052	0.302	-0.014	0.089	0.014	-0.032	-0.017
Selling, Gerneral, & Admin Expenses Margin % (SG&A Expense/Tot Revenue)	0.002	0.004	-0.005	0.026	-0.014	-0.007	-0.004	0.989
TA Turnover = TR/ Avg TA	-0.012	0.027	-0.015	0.000	0.008	-0.035	0.506	0.010
FA Turnover = TR / Avg Net PP&E	0.002	-0.068	-0.012	0.022	-0.034	0.052	0.332	-0.016
Acc Rec Turnover = TR / Avg Acc Rec	-0.018	-0.018	-0.030	0.028	-0.017	0.030	0.588	0.025
Tot Debt / Tot Cap %	-0.003	-0.017	0.148	0.068	0.511	-0.003	-0.131	-0.050
TL/TA %	-0.002	0.017	-0.022	0.056	0.424	-0.011	0.033	0.016
Altman Z-Score	0.001	0.026	-0.018	0.013	-0.379	-0.019	-0.052	-0.013

Table 6. 8 Top 10 and Bottom 10 Mining Companies according to the WFLI

Index Value	Australian Mining Company Name
100	1. Laneway Resources Limited (ASX: LNY)
52.11	2. G8 Communications Limited (ASX: G8C)
6.65	3. Hexagon Resources Limited (ASX:HXG)
5.47	4. 4DS Memory Limited (ASX: 4DS)
3.58	5. IPB Petroleum Limited (ASX:IPB)
2.92	6. Bulletin Resources Limited (ASX: BNR)
2.07	7. Breaker Resources NL (ASX: BRB)
1.16	8. Dourado Resources Limited
0.95	9. China Waster Corporation Limited (ASX: CWC)
0.58	10. Mount Burgess Mining NL (ASX: MTB)
0.01	770. Coal of Africa Limited (ASX: CZA)
0.01	771. Callabonna Resources Limited
0.01	772. Wollongong Coal Limited (ASX: WLC)
0.01	773. Star Striker Limited (ASX: SRT)
0.01	774. Cougar Metals NL (ASX: CGM)
0.01	775. LWP Technologies Limited (ASX: LWP)
0.01	776. Paladin Energy Limited (ASX: PDN)
0.01	777. Gulf Manganese Corporation Limited (ASX: GMC)
0.01	778. Mount Gibson Iron Limited (ASX: MGX)
0	779. Atlas Iron Limited (ASX: AGO)

6.5.4 Standardised Index Construction

When constructing the NSI, the percentage variance explained by each factor was multiplied by the factor score for each company, which was then divided by the total variance explained by the model (81.328%). The results were then summed for all the companies in the sample. Subsequently, as was explained in the Methodology section, the values for each company were standardised using the formula shown in Equation 6.10. This index was dubbed the K-Index – presented in Table 6.9 below. Table 6.9 shows the standardised K-Index values for the top 10 and bottom 10 companies according to the index, as shown BHP Limited and Rio Tinto Limited are on top of the list indicating they have the least financial distress, whereas Image Resources NL and Magnis Resources Limited are at the bottom, indicating they are the most financially distress companies.

Table 6. 9 Top 10 and Bottom 10 Mining Companies according to the K-Index

K-Index Value	Australian Mining Company Name
100	1. BHP Billiton Limited (ASX: BHP)
68.9132	2. Rio Tinto Limited (ASX: RIO)
39.80434	3. Resource Mining Corp. Ltd. (ASX: RMI)
35.19202	4. Fortescue Metals Group Limited (ASX: FMG)
33.73464	5. Neon Capital Limited (ASX: NEN)
32.19876	6. WorleyParsons Limited (ASX: WOR)
30.18073	7. Molopo Energy Limited (ASX: MPO)
28.74564	8. Laneway Resources Limited (ASX: LNY)
27.87678	9. Woodside Petroleum Ltd. (ASX:WPL)
27.24218	10. Northern Star Resources Limited (ASX: NST)
5.304597	770. Castillo Copper Limited (ASX: CCZ)
4.522382	771. Capital Mining Limited (ASX: CMY)
4.241543	772. Estrella Resources Limited (ASX: ESR)
4.038806	773. Image Resources NL (ASX: IMA)
3.738755	774. Magnis Resources Limited (ASX: MNS)
3.712362	775. Mount Ridley Mines Limited (ASX: MRD)
3.697516	776. Oro Verde Limited (ASX: OVL)
1.997241	777. Genesis Minerals Limited (ASX: GMD)
0.371579	778. Lithium Australia NL (ASX: LIT)
0	779. Empire Resources Limited (ASX: ERL)

6.5.5 Comparison to Performance Metrics

To check whether the results of the created indices aligned with commonly used metrics for determining financial standing of companies, namely: “ordinary shares market capitalisation” and “share price,” a comparison of the ranking of companies in the indices with their respective ordinary shares market capitalisation and share price figures was carried out. The K-Index was found to be the one that is in parallel the most with the aforementioned metrics. Due to this, analyses were carried out solely on the K-Index. Table 6.10 showcases the top five mining companies with the highest share price, and Table 6.11 presents the top five companies with the highest ordinary shares market capitalisation. As shown in Table 6.10, the top five mining companies with the highest share price all fall in the top ten companies in the K-Index. This is also true for Table 6.11 – the top five mining companies with the highest ordinary shares market capitalisation all fall in the top ten companies in the K-Index. Therefore, having the K-Index results align with both of the abovementioned figures empirically substantiates the validity of the FDPI in general, and the K-Index in particular.

Table 6. 10 Share Price for Top 5 Companies in the Australian Mining Industry

Australian Mining Company Name	Share Price ($)
2. Rio Tinto Limited (ASX: RIO)	53.75
9. Woodside Petroleum Ltd. (ASX: WPL)	34.23
1. BHP Billiton Limited (ASX: BHP)	27.05
6. WorleyParsons Limited (ASX: WOR)	10.41
10. Northern Star Resources Limited (ASX: NST)	2.21

Table 6. 11 Ordinary Shares Market Capitalisation for Top 5 Companies in the Australian Mining Industry

Australian Mining Company Name	Ordinary Shares Market Capitalisation ($million)
1. BHP Billiton Limited (ASX: BHP)	143,942.7764
2. Rio Tinto Limited (ASX: RIO)	98,203.6967
9. Woodside Petroleum Ltd. (ASX: WPL)	28,202.46179
4. Fortescue Metals Group Limited (ASX: FMG)	5,947.35447
6. WorleyParsons Limited (ASX: WOR)	2,576.79267

6.5.6 K-Index Comparison to a Logistic Regression Model

As mentioned earlier, an FDP model was created for comparative purposes with the K-Index. The results from using LR yielded 24.10% Type I error and 21.74% Type II error for the holdout sample. This provides a model with an average predictive accuracy rate of about 77%. The probability scores for the top ten companies are shown in Table 6.12; despite it being a relatively accurate model, only three out of the top-ten companies are found in the top-ten section of the K-Index. Also, as is evident in the same table, the difference between each company is not even measurable in some cases, such as BHP and WPL, thus making it near-impossible and impractical to make informed and affirmed decisions about the financial ranking of each company.

Hence, these results show that solely relying on the LR FDP model is not sufficient to gain a clear understanding of company ranking. This indicates a legitimate need for an index such as the one developed in this study.

Table 6. 12 Top 10 Probability Scores using Logistic Regression

Australian Mining Company Name	Probability Score
BHP Billiton Limited (ASX: BHP)	1
Woodside Petroleum Ltd. (ASX:WPL)	1
New Hope Corporation Limited (ASX:NHC)	1
Catalyst Metals Ltd (ASX: CYL)	1
Broken Hill Prospecting Limited (ASX: BPL)	1
OZ Minerals Limited (ASX: OZL)	0.999984
Beacon Minerals Limited (ASX: BCN)	0.998615
Tribune Resources Limited (ASX: TBR)	0.997757
Rio Tinto Limited (ASX: RIO)	0.997162
Emu NL (ASX: EMU)	0.996859

6.6 Conclusion

This chapter presented various methods to develop FDP indices. It also presented a novel, user-friendly, standardised index pertaining to companies' financial distress. This chapter also explained why the novel developed index outperforms popular repositories, such as the *Big Three* credit agencies, commonly-used single value metrics, and the Logistic Regression (LR) model. Factor analysis was used to concise the number of variables in the original data-set and subsequently generate the index according to the weighted score of each component. This was tested on the Australian mining sector, by using financial data from 779 companies to develop an index that best describes the financial position of listed Australian mining companies. Three indices were created and the SI index was found to be the optimal through comparing the ranking of companies in the index vis-à-vis established performance metrics – this industry-specific FDPI was coined the K-Index. Subsequently, an LR model was created to showcase the downfalls of relying solely on FDP, as well as the ease of using the K-Index as opposed to FDP. This chapter has verified *Hypothesis 6.*

Chapter 7: Islamic Banking*

* This chapter is based on a published paper in a peer-reviewed Journal, namely: Halteh, K., Kumar, K., & Gepp, A. (2018). Financial distress prediction of Islamic banks using tree-based stochastic techniques. *Managerial Finance*. doi: https://doi.org/10.1108/MF-12-2016-0372

Aligning with *Hypothesis 7* stated in the Chapter 1, namely:

> ***H₇:*** The most important variables in FDP models for Islamic banks vary according to the measure of financial distress used.

This chapter will verify the aforementioned hypothesis by applying the creating FDP models using a data-set comprised of international Islamic banks and then comparing the most important variables in the constructed FDP models for each measure of financial distress used.

This chapter highlights some key differences and similarities between Islamic and conventional banks, surveys the literature on the topic, presents a methodology that identifies the most important predictors pertaining to Islamic banks' financial distress, and discusses key findings before providing the concluding remarks. Unlike other chapters in this thesis, which conduct FDP analyses based on the classification method, this chapter conducts regression analyses. Three measures are employed for assigning financial distress scores for each Islamic bank in the data-set; these scores are subsequently used in the regression analyses to present the most important variables in predicting Islamic banks' financial distress according to each measure – the Literature Review section introduces these measures, whereas the Methodology section presents how they were applied to this study.

7.1 Introduction

The banking industry is extremely crucial not only to local economies, but to the global economy as well, so much so, that when multinational behemoths, such as, Citigroup and Lehman Brothers, experienced extreme financial difficulties mainly due to holding huge derivative portfolios in subprime mortgages, essentially meaning that the borrowers had weak credit-ratings, that is, their capability to repay the loan is dubious. This eventually led to Lehman Brothers going bankrupt, and Citigroup receiving a multi-hundred billion dollar bailout from the United States' government in order to rescue it from insolvency (Wilchins & Stempel, 2008). This, in effect, was the catalyst that led to a domino effect, resulting in plummeting consumer confidence worldwide, thus leading to a stock market crash. In Australia, stimulus packages were announced to try and resuscitate the fragile economy and increase consumer confidence (Davies, 2017).

As is evident, the banking system is directly proportional to the condition of the economy, therefore, for an economy to develop sustainably, an effective banking system needs to be in place (Jan & Marimuthu, 2015b). Measures of sustainability include:

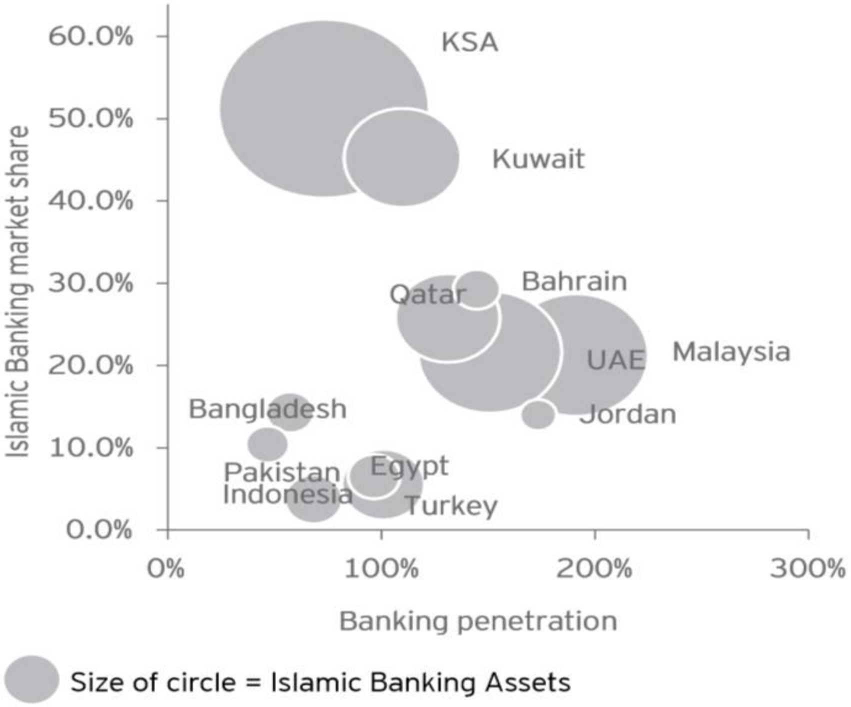
- The internationally recognised CAMELS rating system, which ranks banks with respect to six variables, as the acronym suggests, namely: Capital adequacy, Asset quality, Management, Earnings, Liquidity, and Sensitivity – this rating system allows managers to assess performance and allow for informed decision making;

- The Financial Stability Board, which is an international body that was established post the GFC, to monitor and make recommendations to financial institutions globally (Jan & Marimuthu, 2015b);

- The Basel Accords deliver recommendations on banking regulations pertaining to different types of risk – refer to the Discussion section for further elaboration on the Basel Accords.

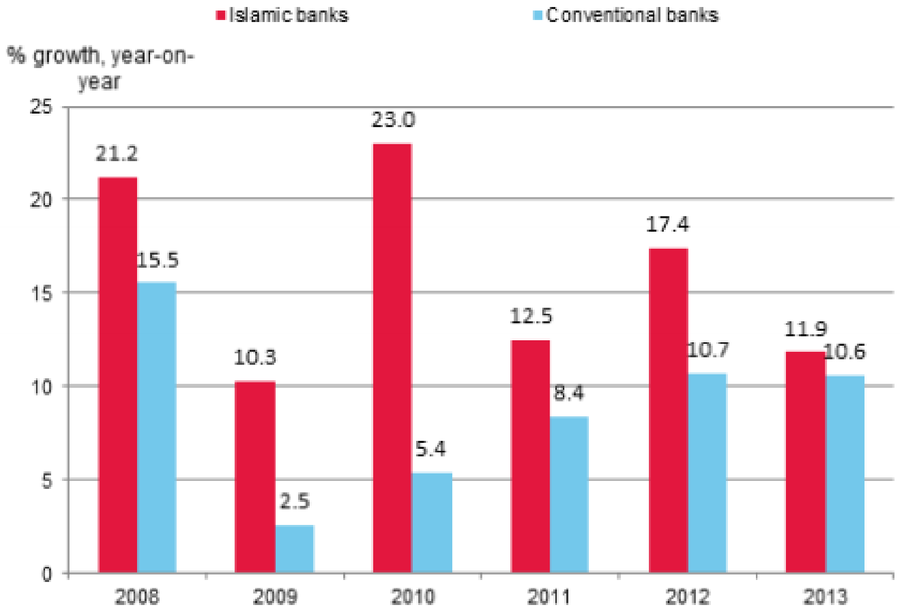
Islamic banks dominate the banking market share in predominantly Muslim nations, especially in the Middle East region with 80% market share vis-à-vis 20% in the rest of the world. Their presence has also expanded on a global scale and they can be found in more than 50 countries (Hanif, 2011). Figure 7.1 below shows the banking penetration and participation asset market share for Islamic banks. The graph clearly shows that the countries with the highest market share of Islamic banks are Middle Eastern nations with predominantly Muslim population, while banking penetration is higher amongst nations with a greater number of conventional banks.

Figure 7. 1 Banking Penetration and Participation Asset Market Share – Source: (EY, 2016)



Shariah-compliant financial assets are predicted to reach $3 trillion in the next decade – an increase from approximately $2 trillion in the year 2016, as well as sales of Islamic bonds, called *sukuk*, increased by 24% to $44 billion in 2016 (Liau, 2017). According to Standard & Poor's (2014), Islamic banking asset-growth has been overtaking conventional banks for a number of years – as shown in Figure 7.2 below. This demonstrates the importance for expanding the currently limited literature available on Islamic banks, and even more so, financial distress prediction pertaining to Islamic banks. Applying FDP modelling to banks can showcase important variables that have a direct effect on a banks' financial distress levels. Another use for applying FDP to banks is that it enables the banks to assess a person's/firm's financial distress probability before determining whether a loan is suitable, and if so, how much excess and premium to charge – in the case of Islamic banks, a *murabaha* contract, where the bank purchases a good then on-sells it to the buyer at a premium price (Beck, Demirgüç-Kunt, & Merrouche, 2013).

Figure 7. 2 Asset Growth Comparison: Islamic and Conventional Banks – Source: Standard & Poor's (2014)



In theory, Islamic banks differ substantially from conventional banks, most notably through the absence of interest charges, as it is considered usury (*riba*) – which is religiously forbidden (*haram*). This is in accordance with Shariah law's dicta that forbids charging interests; speculation (*gharar*); and funding of illicit products – such as: pork, weaponry, and alcohol; as well as, requiring prices to be placed on goods and services only. Islamic finance also requires transactions to be backed by a a pecuniary transaction involving a tangible asset, this is due to the concept of risk/profit-loss sharing (*mudaraba and musharaka*) – for both assets and liabilities – inherent in Islamic finance. *Mudaraba* are partnership loans between the bank and the borrowers, where profits are shared, but the bank bears the losses. Under *Musharaka*, the bank is one of many investors, and both the profits and losses are shared amongst all investors. Therefore, the key differences here between Islamic and conventional banks are the nature of interests and the risk and reward aspects. In terms of interest, conventional banks can offer fixed and predetermined interests to consumers; in terms of risk and reward, the bank bears all the risk and reward after servicing the consumers. On the other hand, in the case of Islamic banks – due to *muskaraka and mudaraba* – the both risk and reward are shared by the bank and the consumers (Hanif, 2011).

In practice, however, these striking differences are not very apparent, as the products are similar to those of conventional banks, but executed differently. For instance, interest rates and discounts are replaced with fees and conditional payment plans (Beck et al., 2013). An example of this apparent difference but practical similarity can be shown in the following scenario of buying a car from a conventional bank vis-à-vis an Islamic bank:

- **Conventional bank:**

The customers do not have the funds in full to pay for the vehicle, therefore, they approach a conventional bank asking for a loan to buy the car. The loan is granted on either a fixed or variable interest rate outlined by the bank. Repayments are done accordingly to pay-off the principal and interest amounts over a designated time period. Let’s assume the interest rate was 10% over a period of one year, and the loan is $10,000. Assuming all else being equal, the amount to be repaid is $11,000 over the course of one year.

- **Islamic bank:**

The customers approach the Islamic bank for funds to pay for the car. Since, there needs to be an actual transaction with a tangible asset involved, the bank offers to purchase the car, and then resells it to the consumers at a premium. So, the bank pays $10,000 for the dealership and purchases the car, then offers to resell it to the consumers at $11,000 to be repaid over the course of one year.

This simple example goes to show that despite fundamental theoretical differences in the methods of conducting financial transactions between Islamic and conventional banks, the practical implications are very similar. This notion of theoretical dissimilarity, but practical similarity, is presented in various studies in the literature that outline other similarities, such as: the Islamic banks' method of calculating the premium price is by pegging it to the interest rates of conventional banks, and that the risk/profit-loss sharing only plays a small role in Islamic banks (Beck et al., 2013; Chong & Liu, 2009; Khan, 2010).

Given the aforementioned similarities and differences between conventional and Islamic banks, and due to the limited literature available on the topic of FDP of Islamic banks. This presents a gap in the literature that this study contributes towards, through utilising machine learning techniques to create FDP models that present the most important predictors of Islamic banks' financial distress. This aids bank managers in their strategic and financial decision-making processes to detect early sings of financial distress, and hence implement preventive measures. This chapter verifies *Hypothesis 7*, presented in Chapter 1.

7.2 Literature Review

Despite there being a large number of papers that use statistical models to predict financial distress of companies, only a fraction deal with the banking industry, and of those, merely a handful pertain to Islamic banking. This section explores some of the literature pertaining to the prediction of financial distress of the banking industry in general, and Islamic banking in particular.

Furthering his work in the seminal paper of 1968, Altman (2000) devised a model specifically for predicting financial distress of service firms. He retained the same financial variables, which he deemed having the strongest predictive power, as was presented in his 1968 paper, with the sole exception of excluding the fifth variable (Sales/Total Assets) – for an elaboration on the equation used in Altman's (1968) paper, refer to Chapter 2. This exclusion was done in order to mitigate the industry effect, which is likely to occur when such an industry-sensitive value is incorporated. Altman's model accuracy was around 90% one year prior to failure, and up to 70% five years prior to failure. His new model is as follows:

$$Z = 6.56x_1 + 3.26x_2 + 6.72x_3 + 1.05x_4 \qquad [Equation\ 7.1]$$

This model was used in later research, some of these studies include: Jan and Marimuthu (2015b); Jan, Marimuthu, Shad, Zahid, and Jan (2019); Kyriazopoulos Georgios (2014); Mamo (2011); Sharma (2013).

In Kyriazopoulos Georgios's (2014) study, the financial distress of six Greek banks was predicted using data from 2001-2009. His research outlined that the reason for failure was mainly due to direct burrowing from the financial market. In Sharma's (2013) study, an application of Altman's (2000) model was conducted on 36 Indian banks. Sharma's model achieved an FDP accuracy level of 70%. In Mamo's (2011) study, an application of Altman's (2000) was conducted on a model containing data pertaining to 43 Kenyan banks. Mamo's model yielded an accuracy level of 90% in identifying non-distressed banks, and 80% pertaining to financially distressed banks. In Jan and Marimuthu's (2015b) study, they used financial distress as a proxy to argue for Islamic banks' sustainability. They applied Altman's (2000) model on Islamic banks from the top five Islamic banking countries. Their aims were threefold: examining financial distress, finding performance indicators that affect the banks' financial health, and perform a comparative analysis on said performance indicators. Their results indicate that the performance indicators in Islamic banking were declining with an average of 79% across liquidity, profitability, insolvency, and productivity. And finally, a recent study by Jan et al. (2019) applied Altman's (2000) model on a data-set comprising 14 Islamic and 14 conventional banks in Malaysia for the economic-postapocalyptic time-period of 2009-2013. Their results indicated that six out of the 14 conventional banks were in distress, compared to ten out of the 14 Islamic banks – which is contradictory to other research that claim superior resilience of Islamic banks vis-à-vis conventional banks. The profitability ratio (Retained Earnings/Total Assets) was found to be the most important variable in predicting banks' financial distress.

Kumar and Ravi (2007) presented an invaluable comprehensive review of studies between 1968-2005, which used both parametric and nonparametric techniques to predict financial distress of firms and banks. His review showed that, while the majority of papers used various financial ratios, there were a few that still used Altman's (1968) original variables. The standard statistical techniques were outperformed by the nonparametric techniques, such as: ANNs and DTs. The paper ends by recommending extra research to be done on machine learning methods, as well as, the use of ensemble and hybrid techniques, as they have the superior predictive capabilities, as well as pooling the advantages and mitigating the drawbacks of individual models.

Both Olson and Zoubi's (2008) and Beck et al.'s (2013) studies investigated the key differences between Islamic and conventional banks. Olson and Zoubi's (2008) study was centred around banks operating in the Gulf Cooperation Council (GCC). They used 26 variables for their study, and developed logit, ANN, and K-NN models.

conforming with the FDP literature, the nonparametric techniques – the ANN and K-NN models, outperformed the logit model. The models were able to differentiate Islamic vis-à-vis conventional banks in the out-of-sample tests with a success percentage rate of 92%. In Beck et al.'s (2013) study, they used a sample of 88 Islamic and 422 conventional banks across 22 countries, for the 1995-2009 time period. They also use another sample of 209 listed banks to check the effect the GFC had on the stock market condition of both types of banks. Their results indicated that there are no major differences in business orientation, and although Islamic banks are less efficient and cost-effective, they have higher intermediation ratios, asset quality and are better capitalised, which led them to outperform conventional banks during the GFC. Two ratios were used to achieve a standardised (z) score for each bank – the formula and ratios are presented in Equation 7.2 below:

$$z = \frac{(ROA + CAR)}{\sigma(ROA)} \qquad [Equation\ 7.2]$$

- z**:** Indicates the distance from insolvency, combining accounting measures of profitability, leverage, and volatility.
- σ**:** Standard Deviation
- ROA**:** (Return on Assets) **=** Profits/Total Assets
- CAR**:** (Capital Asset Ratio) = Total Equity/Total Assets

Al-Shayea, El-Refae, and El-Itter (2010) used ANNs to predict financial distress of Spanish banks using a sample of 66 banks, of which 37 were insolvent. Nine variables were used in the study, comprising various financial statements ratios. They developed two ANN models using different supervised and unsupervised learning algorithms. Their results indicated that their models were able to learn patterns that led to financial distress of the banks, yielding a predictive accuracy rate between 92%-94%.

Al Zaabi (2011) presented a study on applying Altman's (2000) Z-score model on Islamic banks in the United Arab Emirates (UAE) for the 2004-2007 time period. He measured the banks' Z-score for the three years prior to his publication, and then compared it with the then current Z-score of the banks, in order to establish an FDP model. Banks with a Z-score of less than 1.1 were deemed to be financially stressed, above 1.6 were financially healthy, and between 1.1 and 1.6 were uncertain. His results indicated that the Islamic banks in the UAE are overwhelmingly financially healthy

Anwar and Mikami (2011) developed multiple models to predict the *mudaraba* time-deposit return in Islamic banks, including: ANN, LR, and a generalised autoregressive conditional heteroskedasticity model. They used ten years' worth of data and six macroeconomic variables. Their results indicated that ANN outperformed the other models in predicting the average rate of return of one-month *mudaraba* time deposit.

A recent study by Le and Viviani (2018) compared the FDP accuracy of statistical techniques, namely: LR and MDA; vis-à-vis machine learning techniques, namely: SVMs, ANNs, and K-NNs. They incorporated 31 financial ratios to be tested on a data-set consisting of 3000 banks in the United States – 1562 operating and 1438 failed, for the time-period 2011-2016. Their results indicated that ANNs were the superior model with a predictive accuracy of 75.7%, followed by K-NNs (74.1%), LR (73.9), MDA (72%), and finally SVM (71.6%). In terms of variable importance, all 31 ratios were found to be statistically significant, but ratio groups such as: operation efficiency, profitability, and liquidity ratios were found to be the most important – these groups include ratios such as: Impaired Loans divided by Gross Loans, Capital Ratio, Operation Income divided by Average Assets, ROA, and others.

As is evident, there are various studies in the literature regarding banks' financial distress, however, there are no studies that have combined three measures of financial distress to determine the most important variables in determining Islamic banks' financial distress. This research spearheads this initiative in the hope of encouraging further research to be done in this area in the future.

7.3 Data

This research extracted financial data for the year 2014 using a data-set of 101 Islamic banks that operate on a global scale. Due to difficulty and limited availability of extracting failed Islamic banks' data, this study used three measures – outlined later in the Methodology section, that assigns scores to each Islamic bank, which is used to determine the financial distress level of the each bank. The number of independent variables used is 18 – comprising financial ratios, actual figures, margins, and rates, as shown in Table 7.1.

As explained in the Introduction section, Islamic banks may refer to certain financial terms by Arabic terms, such as *mudaraba* and *musharaka* pertaining to assets and liabilities. However, in this study's data-set, the Islamic banks referred to their financials by standard English terms in their statement. This is why the variables used in this study are not referred to by Arabic terms. The data for the companies used in the research were extracted from the Capital IQ database, which, as mentioned in Chapter 1, is a web-based data repository that provides ubiquitous financial information and company data (Phillips, 2012). Capital IQ has been used various interdisciplinary research, including: Feldman and Zoller (2012); Halteh et al. (2018a); Kahle and Stulz (2013).

Table 7. 1 Variables used in this study

Variable	Description
Total Assets (TA)	Actual Balance Sheet Figure
Dividends / Shares	The Number of Dividends that the Shareholders Receive on a Per-Share Basis
ROE (Return on Equity)	Net Income / (Shareholders' Equity - Outside Equity Interests)
ROA (Return on Assets)	Earnings Before Interest / (Total Assets - Outside Equity Interests)
Operating Income / TA	Financial Ratio
Working Capital / TA	Financial Ratio
Retained Earnings / TA	Financial Ratio
Earnings Before Income & Tax (EBIT) / TA	Financial Ratio
Market Value of Equity/Total Liabilities (MVE / TL)	Financial Ratio
Revenue / TA	Financial Ratio
Debt Ratio	Total Liabilities / Total Assets
Current Ratio	Current Assets / Current Liabilities
ROR (Return on Revenue)	Net Income / Total Revenue
Asset Turnover	Total Revenue / Total Assets
Efficiency Ratio	Total Expenses / Total Revenue
Total Equity / Total Assets	Financial Ratio
Equity Ratio	Total Equity / Total Assets
Total Debt / Total Equity	Financial Ratio

7.4 Methodology

Three measures, namely: Altman Z-Score, Altman Z-Score for Service Firms, and the Standardised Profits, were utilised to extract a score that is used to measure each bank's financial distress. The software package 'Salford Predictive Modeler' was used to develop and test the models built using three machine learning techniques, namely: DTs, RFs, and SGB. This software package has been used previously in the literature (Gepp & Kumar, 2012; Gepp et al., 2010). The aforementioned techniques were chosen to construct models due to previous research presenting the empirical superiority of said techniques vis-à-vis traditional statistical techniques, such as MDA and LR – some of these studies include: Berg (2007); Gepp and Kumar (2012); Gepp et al. (2010); Kumar and Ravi (2007). Additionally, these techniques do not make any distributional assumptions, which is prudent in this case because of the limited literature pertaining to Islamic banking. The techniques used and the way they were developed will be further explained in the following subsections.

Three regression analyses were conducted. Each one of the measures stated above were used as a continuous dependent variable in each technique. This yielded three models for each of the three techniques used, that is, nine models altogether. The results of each model were subsequently compared and contrasted with one another in order to deduce the most important predictors at identifying Islamic banks' financial distress. Since this study only focuses on variable importance, there is no need for a test/holdout sample.

The independent variables are based on data for the year 2014, whereas the dependent variables (for all three measures), use 2015 data. That is, one-year lagged independent variables have been used. The dependent variable changed based on the measure used, that is, Altman Z-Score, Altman Z-Score for Service Firms, or the Standardised Profits measure. For each of the measures of financial distress, and using the 18 variables each time, the models were built using the techniques mentioned above. The identification of important variables that affect Islamic banks' financial distress enables various stakeholders, including shareholders and government bodies, as well as regulatory influences, such as the Basel Accords, to monitor those variables and install measures to prevent possible distress – these implications will be discussed in detail in the Discussion section later on in the chapter.

As mentioned above, three measures of financial distress were used in this study. The rationales behind choosing said measures are explained below:

- Firstly, the "Altman (1968) Z-Score" measure has been chosen because of its extensive use in the FDP literature. Five ratios were used to achieve a Z-score for each bank using data from the year 2015; the five ratios used were outlined in Chapter 2. However, the results for each Islamic bank were not classified as per Altman (1968) classification, as this research has conducted a regression analysis, not a classification/logistic binary analysis.

- Secondly, the "Altman Z-Score for Service Firms" measure has been chosen – which was discussed earlier in the Literature Review section. This measure was chosen since this study is concerned with Islamic banks – a service industry, and so this approach is arguable more appropriate and accurate. This measure has been applied by various researchers to the banking industries in a number of countries worldwide, including Greece, India, and Kenya, and they have achieved high FDP accuracy rates (Jan & Marimuthu, 2015a; Kyriazopoulos Georgios, 2014; Mamo, 2011; Sharma, 2013). Four ratios were used to achieve a Z-score for each bank using data for the year 2015.

- Thirdly, the "Standardized Profits" measure that was utilised by Beck et al. (2013). This measure was chosen as it is a novel approach that can be applied to FDP of banks, both conventional and Islamic. As discussed in the Literature Review section, their model measures a standardised 'z' score, which is indicative of bank stability. This includes accounting measures of profitability, volatility, and leverage.

The techniques presented below provide information that are specific in constructing the models presented in this chapter. For an elaboration of the mechanics of each technique, refer to Chapter 2.

7.4.1 Decision Tree Model

In this research, regression trees have been used with the standard Gini criterion to determine the best splitting rule at each point. All 18 variables were used as predictors (independent variables), and the target variable (dependent variable) was selected as one of Altman Z-Score, Altman Z-Score for service firms, or the Standardised Profits measure, to achieve results for all three models. The standard V-fold cross validation using 10 folds was used for the testing component of the model. This helps to ensure that the model is not over-trained, meaning that it can detect patterns that appear in the data-set given, but will not generalise well to new data.

7.4.2 Random Forests Model

The same variables were used as for DTs. Testing of the model was based on out-of-bag data, which is also used for testing and avoiding over-training to increase the generalisability of the findings. The number of variables considered at each node was set to the square root of the total number of predictors: $\sqrt{18} \approx 4.24 \approx 4$. Different numbers of trees were tested (200, 500, and 1000), but 500 trees were determined to be sufficient.

7.4.3 Stochastic Gradient Boosting Model

The standard V-fold cross validation using 10 folds was used for the testing component of the model. Individual trees were kept small by setting the maximum nodes per tree to six (a standard setting) with a minimum number of data points of ten in each node. The criterion to determine the optimal number of trees, that is, how much incremental improvement to perform, was chosen based on the default of cross entropy. Different numbers of trees were tested, but for similar reasons as stated previously, 200 stochastic random boosting trees were finally determined to be sufficient.

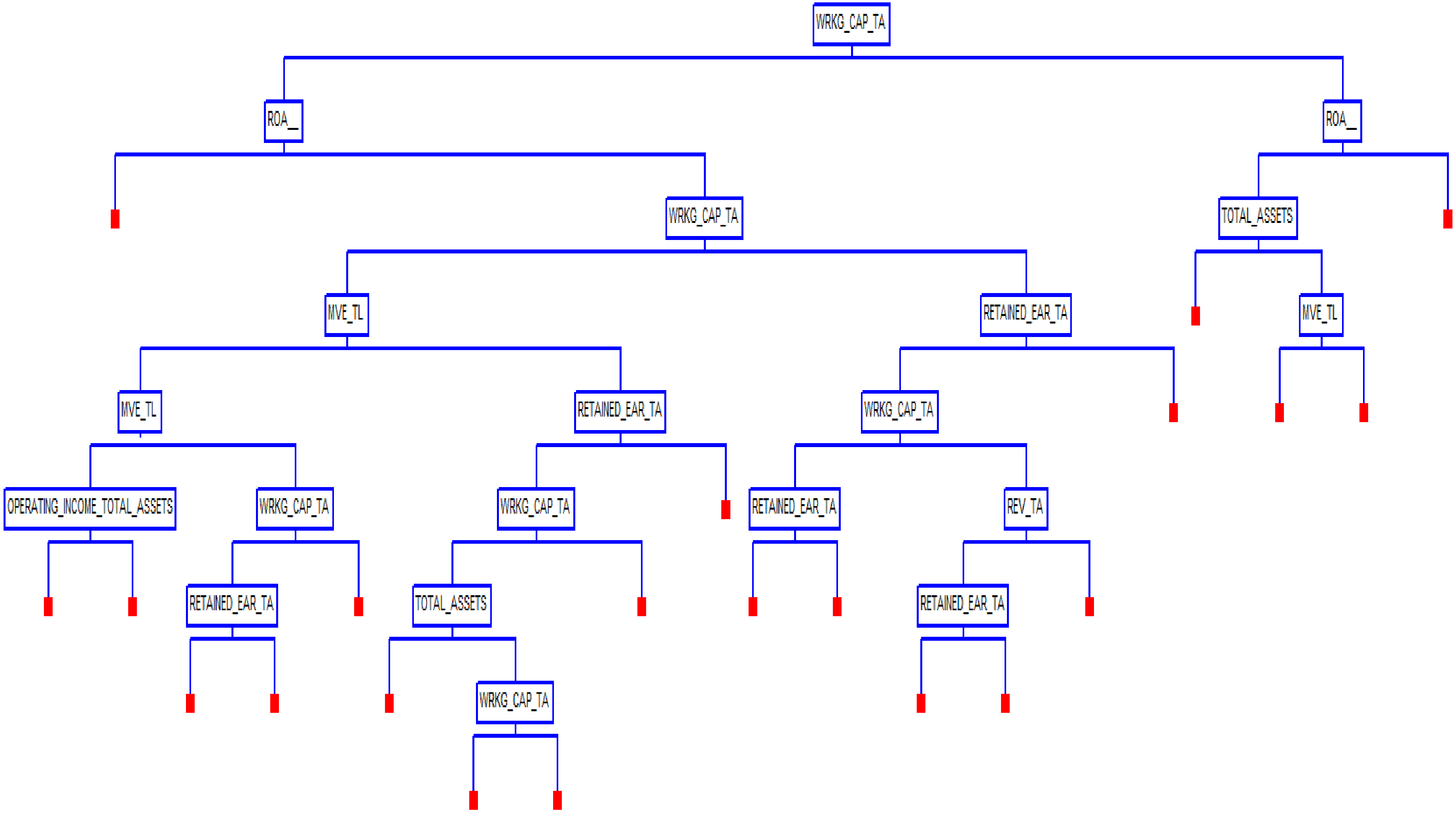
7.5 Results

The models for each of the three definitions of financial distress are analysed separately below. Table 7.2 at the end of this section, provides a summary of the most important variables in each model, according to both technique and definition of financial distress. A sample DT is shown for each measure. For RF and SGB, a similar visualisation is unattainable because they are an ensemble of many trees, which is one of their disadvantages, but they are likely to be more accurate and better at handling inaccuracies in the data.

7.5.1 Altman Z-Score Measure

For the DT model, the results yielded 'Working Capital/Total Assets' as the root node, the most important variable, followed by ROA as the next non-leaf node, leading through connecting branches to multiple consecutive non-leaf nodes and finally ending with leaf nodes – refer to Figure 7.3 for an illustration. The ratio of Working Capital to Total Assets was also the most important variable in both the RF and SGB models. Current Ratio appeared as the second most important variable using DT and RFs, whereas the Debt Ratio was the second most important variable using SGB – refer to Table 7.2 for more detail.

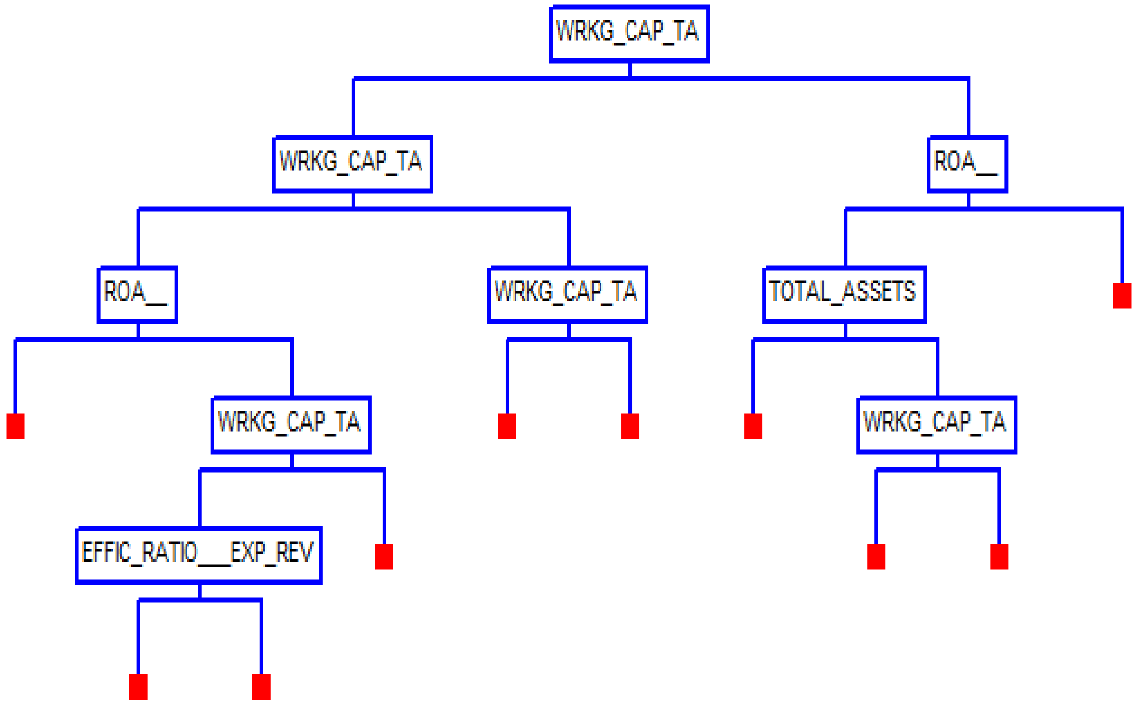
Figure 7. 3 Decision Tree Model using Altman's Z-Score as the measure of Financial Distress



7.5.2 Altman Z-Score for Service Firms Measure

Figure 7.4 provides an illustration of the single DT model for Altman’s Z-Score for Service Firms measure. Again, ‘Working Capital/Total Assets’ shows up as the most important variable in FDP. RF and SGB models confirmed this as the most important variable. The second most important variable was Current Ratio for both DT and RF, whereas it appeared as the third most important using SGB. Refer to Table 7.2 for more information.

Figure 7. 4 Altman Z-Score for Service Firms Decision Tree



7.5.3 Standardised Profits Measure

A decision tree with 'ROR' as the root node was developed for this measure. See Figure 7.5 for illustration. 'ROR' is the most important variable in this model, but RF found Total Debt/Total Equity. SGB agreed with the single tree that ROR is the most important variable. Total Debt/Total Equity and Retained Earnings/Total Assets are clearly also important across all models. Refer to Table 7.2 for more detail.

Figure 7. 5 Standardised Profits Decision Tree

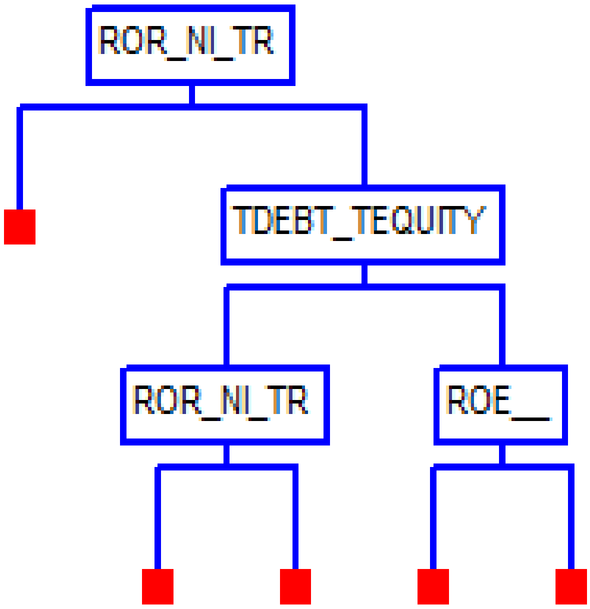


Table 7. 2 Model Comparison Table

Measure	Model	Most Significant Variables (in order of significance)
Altman Z-Score	Decision Tree (CART)	1. Working Capital/Total Assets 2. Current Ratio 3. Debt Ratio 4. Retained Earnings/Total Assets
	Random Forest	1. Working Capital/Total Assets 2. Current Ratio 3. Total Assets 4. Equity Ratio
	Stochastic Gradient Boosting (TREENET)	1. Working Capital/Total Assets 2. Debt Ratio 3. Retained Earnings/Total Assets 4. Market Value of Equity/Total Liabilities
Altman Z-Score for Service Firms	Decision Tree (CART)	1. Working Capital/Total Assets 2. Current Ratio 3. Debt Ratio 4. Total Assets
	Random Forest	1. Working Capital/Total Assets 2. Current Ratio 3. Total Assets 4. Equity Ratio
	Stochastic Gradient Boosting (TREENET)	1. Working Capital/Total Assets 2. Debt Ratio 3. Current Ratio 4. Retained Earnings
Standardised Profits	Decision Tree (CART)	1. ROR 2. Total Debt/Total Equity 3. Market Value of Equity/Total Liabilities 4. Retained Earnings/Total Assets
	Random Forest	1. Total Debt/Total Equity 2. ROR 3. Total Revenue/Total Assets 4. Retained Earnings/Total Assets
	Stochastic Gradient Boosting (TREENET)	1. ROR 2. Retained Earnings/Total Assets 3. ROA 4. Market Value of Equity/Total Liabilities

Table 7.2 shows the most predictive variables achieved for each model constructed using the three tree-based techniques. 'Working Capital divided by Total Assets' is the most important variable in determining Islamic banks' financial distress using the Altman Z-Score and the Altman Z-Score for Service Firms measures across all techniques. As for the Standardised Profits measure, 'Return on Revenue' is the most

important variable for both the DT and SGB techniques. ‘Total Debt divided by Total Equity’ is the most important variable using RF – for further analysis of the results tabulated above, refer to the Discussion section below.

7.6 Discussion

As was evident in the Results section, there are significant variable differences between the measures used, especially between the Altman (1968) and the Beck et al. (2013) approaches. The results above showcase the similarities between the Altman (1968) Z-Score measure and the Altman Z-Score for Service Firms measure, as they have ‘Working Capital/Total Assets’ as the most predictive variable. The Current and Debt Ratios appear frequently as the next most predictive variables. As for the Standardised Profits measure, ‘ROR’ (Return on Revenue) = Net Income/Total Revenue, was the most important predictive variable, using the DT and SGB techniques, and the second most predictive using RFs. These results contribute to the literature and further the understanding of Islamic banks financial distress. The results are meaningful since the banks are service firms, and the results achieved comprehensively deal with the capital/monetary aspects of the bank. This explanation is in concert with previous Islamic banking literature, including: Jan and Marimuthu (2015a, 2015b); Jan et al. (2019), as well as the Basel Accords (explained below), that recommends focusing on the capital risks of the banks.

By using lagged variables to predict the future state of Islamic banks, this gives rise to the potential of implementing proactive measurements by senior management to deviate the bank from the road to bankruptcy. It can also provide governmental watchdog institutions an alert to notify the bank of the impending dangers ahead should they perpetuate the status quo. Managers will benefit from the findings in this study by having a clear picture of what to look for when assessing their bank’s financial distress levels. Other stakeholders like investors also benefit as they can make informed decisions about whether to stay with the bank or go elsewhere due to a forecasted danger the following year.

Moving on to regulatory implications, the Basel Accords play a key role in reforming the banks' operations. The Basel Accords are three sets of banking regulations (Basel I, II and III) set by the Basel Committee on Bank Supervision (BCBS). They provide recommendations on banking regulations in regard to capital risk, market risk, and operational risk. The function of the accords is to make sure that financial institutions have sufficient capital on hand to meet obligations and withstand unforeseen losses. Basel I was issued in 1988 and it focuses on the capital adequacy risk of financial institutions – international banks should have a risk weight of 8% or less. Basel II is an updated version of the original accord; it coined the 3 pillars: minimum capital requirements, supervisory review of an institution's capital adequacy and internal assessment process, and effective use of disclosure (Federal Reserve, 2003). Basel III was established in the wake of the GFC, it is a continuation of the three pillars, as well as extra requirements and safeguards (Bank for International Settlements, 2016).

Even though, as Beck et al. (2013) found, Islamic banks are better capitalised, hence can withstand unforeseen losses better vis-à-vis conventional banks, by using lagged variables (identified earlier in this chapter and in the predictive models) in conjunction with the Basel Accords, management can determine whether the company is in the 'danger zone' or whether their risk is marginal. This will enhance the longevity of banks in the marketplace.

7.7 Conclusion

This study has focused on cutting-edge financial distress prediction models and applied them to Islamic banks. These models can be used to forecast impending risks to enable the decision makers to take the preventive measures to hold-off such risks or mitigate their effect. Recursive partitioning techniques were employed to test for the most accurate measure in predicting financial distress. The results indicated that there is a need for a specific financial distress mechanism for Islamic banks, as variables that are indicative of a bank's status differ between the old Altman (1968) standard and novel approaches. 'Working Capital/Total Assets' was the most predictive variable for forecasting financial distress in Islamic banks using all three models used in this study across both measures: Altman Z-Score and Altman Z-Score for Service Firms. As for the Standardised Profits measures, ‘Return on Revenue’ was the most influential variable. Therefore, the aforementioned two variables can be used in conjunction with the recommendations made by the Basel Accords, when making decisions pertaining to FDP of Islamic banks. This presents an opportunity for future research to investigate the differences in the results achieved, which will contribute towards further understanding of variables affecting Islamic banks’ financial status. This chapter has verified *Hypothesis 7.*

Chapter 8: FDP of Large Companies and Small & Medium Enterprises

Aligning with *Hypotheses 4 and 5* stated in the Chapter 1, namely:

> ***H$_4$:*** Class imbalance does affect the detection accuracy of FDP models, and it can be enhanced by optimising the cut-off points or using SMOTE vis-à-vis a model that is built on a standard imbalanced data-set.
>
> ***H$_5$:*** Independent variables' importance vary between FDP models for SMEs vis-à-vis LCs.

This chapter will verify the aforementioned hypotheses by applying the SMOTE technique to imbalanced data-sets comprised of SMEs and LCs, and investigating the differences in independent variable importance between them.Introduction

Small and Medium Enterprises (SMEs) make up the majority of businesses on a global scale, employ many more people vis-à-vis Large Companies (LCs), and are considered to be the main drivers of economic growth by entrepreneurs. These factors enable SMEs to be a mighty force in the fight against poverty worldwide – which explains why they are widely viewed as the cornerstone of businesses globally, and have led governments around the world to encourage SME development, be it through grants, subsidies, and/or limiting red tape (Gupta, Gregoriou, & Healy, 2015; Koshy & Prasad, 2007).

In Australia, SMEs account for around 56% of the total Gross Domestic Product (GDP) – around 577 billion dollars; constitute around a whopping 99.8% of all businesses – thus leaving only around 0.2% for; and employ around 7.3 million people – that makes up around 68% of all employees in the country (ASBFEO, 2016). The prevalence of SMEs can vary by industry – for example, industries such as: health care, professional services, accommodation and food services, real estate, construction, forestry, fishing, and agriculture are predominantly run by SMEs; whereas, gas, electricity, water, telecommunications, transport, manufacturing, and mining are mainly run by LCs (ASBFEO, 2016).

A worrying statistic, however, indicates that 44 Australian small businesses close doors every day (Cornish & Landy, 2013). Survival rates of businesses seem to increase with size – starting with about 56% for sole proprietorships, and ending with 83% for companies employing more than 200 people, tested over a four-year period, from 2011-2015 (ASBFEO, 2016). Given the predominance of SMEs, these statistics are troubling for the Australian economy. This presents an opportunity to develop FDP models that will aid in understanding the variables affecting business failure for both SMEs and LCs.

The definition of what makes up an SME differs from country to country, or region to region, and is even often nonbinding – such as in Australia, thus making it difficult and subjective to ascertain which definition to use when conducting studies. For example, according to the Australian Bureau of Statistics (ABS), a business is classified as a "non-employing business," if it is a sole proprietorship or partnership without any employees; a "micro business," if it has less than five employees; a "small business," if it has at least 5, but less than 20 employees; a "medium business," if it has at least 20, but less than 200 employees; and a "large business," if it has 200 or more employees ABS (2001). However, according to the Australian Securities and Investments Commission (ASIC), for a company to be classified as a "large proprietary company," it must satisfy two of the following three criteria:

1. The consolidated annual revenue of the company is $25 million or more,
2. The annual consolidated gross assets the company owns is $12.5 million or more, and/or
3. The company employs 50 or more employees (ASIC, 2014).

This chapter outlines the inherent differences between SMEs and LCs and the statistics relating to a large number of small business failures in Australia. The chapter proceeds to survey the literature on the various studies that have developed FDP models pertaining to SMEs. After this, the data-set used in this study is presented – comprising Australian LCs and SMEs. Similar to Chapter 5, the data-set used in this study has a class imbalance problem. To remedy this, the same methodology used in

Chapter 5 was used again here due to its empirical substantiation, that is, using SMOTE and creating FDP models using machine learning techniques. Following this, the results of the aforementioned models are presented, followed by concluding remarks. This chapter's contribution is in the form of using SMOTE and applying and comparing various machine learning techniques to create FDP models pertaining to Australian LCs and SMEs – this has not been done before to the best of the author's knowledge. This furthers the understanding of variables affecting the financial health of SMEs and LCs, and offers invaluable insight to decision makers to install proactive measures to alleviate possible bankruptcies.

8.2 Literature Review

The literature is inundated with studies that deal with FDP of LCs, many of which are based on Altman's (1968) seminal paper. The papers tend to use historical data to predict financial distress of firms (Gupta et al., 2015). Only a fraction of the FDP studies are applied to SMEs. This may be because the definition of what constitutes an SME varies across different countries or regions, or due to the fact that it is much easier to obtain data pertaining to LCs, as they tend to be more publicly listed in comparison with SMEs. Being a publicly listed company entails providing public access to their archival data, and such data tends to be readily available across many databases. On the other hand, SMEs – especially micro and small companies – tend to be privately owned, hence are under no obligation to disclose their financial statements, which in turn, makes it much more difficult to retrieve the required data to perform an FDP study (Edmister, 1972).

As mentioned above, the definition of SMEs can be regional and nonbinding, which has led researchers to use different definitions for their studies. For example, Freel (2000); Spithoven, Vanhaverbeke, and Roijakkers (2013) used the European Commission and the Organisation for Economic Cooperation and Development's (OECD) definition of SMEs for splitting their data-set, that is, by using a 250-employee cut-off; whereas, Narula (2004); Van de Vrande, De Jong, Vanhaverbeke, and De

Rochemont (2009) used a 500-employee cut-off; while, Bianchi, Campodall'Orto, Frattini, and Vercesi (2010) used a 50-employee cut-off.

The seminal paper that spearheaded the application of FDP modelling on an SME data-set was by Edmister (1972). His study incorporated 19 financial ratios as predictors. Those variables were selected as they were supported by theorists or found to be significant in previous empirical research. and employed MDA on data-sets that were based on restrictive assumptions, ranging from 42-562 small businesses. His results yielded a discriminant function with seven variables that can be used to infer whether a business is going to be fail or not with 93% accuracy.

Altman and Sabato (2007) applied LR techniques on a sample of SMEs in the United States. Their findings indicate that their FDP model outperforms generic credit scoring models, and it leads to lower capital requirements for banks. However, they acknowledge that their model’s performance could be improved by addition of qualitative data.

Altman, Sabato, and Wilson (2010) heeded Altman and Sabato’s (2007) recommendation regarding qualitative data and incorporated both nonfinancial, regulatory compliance, and event data when developed FDP models using a sample of 5.8 million unlisted SMEs in the UK, of which over 66,000 failed between 2000-2007. Their findings showed a 13% improvement in their model’s performance when qualitative information are added alongside traditional financial ratios.

Spithoven et al. (2013) used a sample consisting of 792 SMEs and 175 LCs in their study. They considered a company to be an SME if it had fewer than 250 employees. They used independent variables that consist of control, open innovation – breadth, and open innovation – intensity, in their study. They investigated how open innovation affects the innovative performance of SMEs vis-à-vis LCs. Their findings indicate that SMEs are more dependent on open innovation compared to LCs, and are more effective in using different open innovation practices concurrently when they present new products to the marketplace. Intellectual property protection mechanism drives revenues from new products in, however, in the case of LCs, they gain from search strategies.

Camacho-Miñano, Segovia-Vargas, and Pascual-Ezama (2015) used Artificial Intelligence (AI) techniques, namely: rough set and PART methods (rule-learning algorithm based on partial DTs) to model for FDP of SMEs in Spain. Their sample included 235 bankrupt companies. They started with an initial set of 23 variables but they were later reduced to nine. The objective of the study was to identify the characteristics of bankrupt firms. The AI models' results indicated that there are five important FDP variables, namely: Sector, Size, Number of Shareholdings, Return on Assets, and Liquidity.

Keasey, Pindado, and Rodrigues (2015) used a sample of 18,580 firms from five European countries for the time-period 1999-2006. 74.4% of the companies were healthy and 25.6% were distressed. They considered a firm to be distressed if it had two consecutive years of having an Earnings Before Interest, Tax, Depreciation, and Amortisation (EBITDA) value less than financial expenses, (Net Worth / Total Debt) being less than 1, and if the company's net worth falls between the two periods. The used five variables in their study. The study's objective was to identify the most important FDP variables. They showed that the expected costs of financial distress, can be estimated by an innovative model that allows for an interaction between the possibility of financial distress and its costs when it happens. Their results indicated that forecasted financial distress costs depend on the likelihood of financial distress and on the variables that effect the period of time and costs incurred during the bankruptcy process. Particularly, financial costs are lesser where the capability to use tangible assets as collateral and short-term debt is larger; they are larger the more the use of long-term secured debt. Also, the effect of these variables can be controlled by the firm's ownership and bankruptcy laws.

Gupta et al. (2015) used a sample consisting of 8,162 distressed and 385,733 healthy companies in the United Kingdom for the time-period 2000-2009. 20 financial and nonfinancial variables used to predict a firm’s failure hazard. They estimate separate hazard models for each sub-category of SMEs, and compare their accuracy with an SMEs hazard model that include all three sub-categories. They test their hypotheses using discrete-time duration dependent hazard rate modelling techniques, which control for both survival time and macro-economic circumstances. Their results present the differences in the financial distress attributes of micro firms and SMEs, and showcase that there is no need to segregate small and medium firms when creating FDP models, since almost all explanatory variables affect the failure hazard of SMEs, small, and medium firms.

Calabrese, Andreeva, and Ansell (2019) used 92 predictors and extracted data pertaining to 27,533 companies in London from an anonymous database for their study. They used the European definition of what constitutes an SME (less than 250 employees and annual turnover below 50 million euros). They studied the effects of incorporating the interdependence among SME bankruptcies into a risk analysis framework using data prior to the GFC. Their findings indicate that the interdependence or contagion component defined based upon spatial and demographic characteristics is significant, and it enhances the ability to predict defaults of non-start-ups in London.

As is evident in the literature survey presented above, there are many studies applying FDP to SMEs. However, there are no studies that combine SMOTE with FDP modelling pertaining to Australian SMEs and LCs. This presents an opportunity to fill an existing gap in the literature. Therefore, this chapter contributes towards this gap by applying four machine learning techniques on ASX and *SMOTEd* data-sets, in order to create FDP models pertaining to Australian LCs and SMEs. The results of the models using the aforementioned data-sets will be compared to ascertain the most effective model at predicting financial distress of SMEs and LCs. To add, the most important variables that directly affect SME and LC financial distress will be presented and discussed.

8.3 Data

In order to develop FDP models for LCs and SMEs, this study adopts ASIC's definition of businesses – as was outlined earlier in the Introduction section. This is because ABS's definition purely focuses on the number of employees in a business, and this is a myopic and simplistic category to classify by. Whereas, ASIC's definition is more holistic and provides extra dimensions that is more in-touch with real-world situations. Therefore, classifying businesses as SMEs will be according to whether they satisfy any two of the following three criteria: less than 50 employees, less than $12.5 million in assets, and/or less than $25 million in annual revenue. Hence, this definition of SMEs will encompass micro, small, and medium businesses. Businesses that do not meet two of the three criteria are classified as large.

The Capital IQ database was used to extract financial data for all companies listed on the Australian Stock Exchange (ASX) as at 28th of May, 2018 – their latest financial statements were used (30th of June, 2017). After classifying the companies as per ASIC's criteria and cleaning the data, the final data-set pertaining to healthy (listed) companies was as follows: 1,233 SMEs and 260 LCs. Some of the companies in the data-set had missing information – the literature presents a number of ways for dealing with this issue, including: deletion, replacement with mean, replacement with mean for a given class, replacement with median for a given class, replacement with mode, to name a few. Due to the presence of outliers in the data-set, replacing the missing values with median for the given class was chosen, as the median is immune to outliers (Kantardzic, 2011). This methodology has been previously used across various disciplines in the literature, including: Gromski et al. (2014); Kaiser (2014).

Capital IQ was also used to collect data pertaining to delisted (distressed) ASX companies from 1/7/2016 – 30/5/2018. Financial data were extracted for the latest annual financial statement prior to delisting, for example, if a company delisted on the 3rd of July, 2018, the financial statement for the 2017 financial year (as at 30th of June, 2017) was used. After classifying the companies as per ASIC's criteria and cleaning the data, the final distressed companies’ data-set was as follows: 42 SMEs and 32 LCs. Some companies in the data-set had missing information, so, as with the listed companies above, replacement of missing values by the median was conducted.

The data for both SMEs and LCs are imbalanced in terms of healthy and distressed companies, that is, the ratio of healthy companies vastly outweighs that of distressed companies – refer to Table 8.1 below for a breakdown of the data-sets. As is evident in Table 8.1, there is an issue of class imbalance for both the SME and LC data-sets, meaning that there are much more healthy companies than there are distressed – for SMEs: 96.71% to 3.29%, respectively; for LCs: 89.04% to 10.96%, respectively.

Table 8. 1 Final Data-Set

Companies	Healthy	Distressed	Total	Class Imbalance
SMEs	1,233	42	1,275	96.71% Listed – 3.29% Delisted
Large	260	32	292	89.04% Listed – 10.96% Delisted
Total	1,493	74	1,567	95.28% Listed – 4.72% Delisted

The 24 variables selected for the study are given in Table 8.2. The variables used in this research were chosen in line with prior studies dealing with the SMEs, including: Altman et al. (2010); Camacho-Miñano et al. (2015); Gepp (2015); Gupta et al. (2015); Keasey et al. (2015); Spithoven et al. (2013).

Table 8. 2 Variables used in the study

Variable	Description
ln TA	Natural Logarithm value of Total Assets (TA)
ln TR	Natural Logarithm value of Total Revenue (TR)
ln Employees	Natural Logarithm value of Total Employees
ROE	Return on Equity = Net Income/ (Shareholders' Equity - Outside Equity Interests)
ROA	Return on Assets = Earnings Before Interest (EBIT) / Total Assets Less Outside equity interests
ROC	Return on Capital = Earnings Before Interest and Tax (EBIT) * (1-0.375) / Average Total Capital
Gross Margin	Gross Profit / Total Revenue
ROCE	Return on Capital Employed = EBIT / (Total Assets - Current Liabilities)
SG&A Margin	Selling, General, and Administration Costs / Net Sales
TD/TE	Total Debt (TD) / Total Equity (TE) * 100
TD/TC	Total Debt (TD) / Total Capital (TC) * 100
TL/TA	Total Liabilities (TL) / Total Assets (TA)
Cash/CL	Cash / Current Liabilities
Cash/TA	Cash / Total Assets
CA/TA	Current Assets / Total Assets
NWC/TA	Net Working Capital / Total Assets
NI/TA	Net Income / Total Assets
EBITDA/TA	Earnings Before Interest, Tax, Depreciation, & Amortisation / Total Assets
RE/TA	Retained Earnings / Total Assets
CFO/CL	Cash from Operations / Current Liabilities
Current Ratio	Current Assets / Current Liabilities
Quick Ratio	(Current Assets - Current Inventory) / Current Liabilities
Asset Turnover	Total Revenue / Total Assets
Altman Z-Score	$Z = 1.2x_1 + 1.4x_2 + 3.3x_3 + 0.6x_4 + 1.0x_5$ – refer to Chapter 2 for variables

Following this, a dichotomous binary variable was used to refer to the status of each company – coded '1' if the company is listed (healthy) and '0' if the company is delisted (distressed). For example, when creating the SMEs training sample, the data were split in half by randomly selecting 50% of the observations (1,275 ÷ 2 = 638). The other half of the observations were used to construct the holdout validation sample. Same process was repeated for the LCs. When creating both the training and testing samples for both data-sets, it is was ensured that the class imbalance ratio did not vary significantly from the overall data-set, as otherwise the generated model will not have a fair representation of the original data – this process and the results sets are summarised in Table 8.3.

Table 8. 3 Original Training and Testing Data-Sets for SMEs and LCs

Companies	Training	Holdout	Total	Class Imbalance for Training Samples	Class Imbalance for Holdout Samples
SMEs	638 (613 Healthy – 25 Distressed)	637 (620 Healthy – 17 Distressed)	1,275	96.08% Healthy – 3.92% Distressed	97.33% Healthy – 2.67% Distressed
Large	146 (130 Healthy – 16 Distressed)	146 (130 Healthy – 16 Distressed)	292	89.04% Healthy – 10.96% Distressed	89.04% Healthy – 10.96% Distressed

8.4 Methodology

There are two subsections within the Methodology section. The first subsection explains the data-sets used in this study and how the training and testing samples were constructed. The second subsection showcases the models that were created for this study using the following techniques: DT, treebag, RF, and SGB. The evaluation methods used in this study for assessing detection accuracy of the created models incorporates both visual (as per the ROC graph) and empirical (as per the AUROC score) aspects. Combining both aspects reinforces the validity of the results.

8.4.1 Data-sets

Four data-sets were used in this study's analysis, the original and the *SMOTEd* data-sets for both LCs and SMEs, as shown:

8.4.1.1 Original SME and LC Data-sets

As explained the Data section above, the original data-sets of both SMEs and LCs were split evenly to create training and holdout samples for each. These training samples of each data-set are then tested on their respective holdout sample to create the models pertaining to the original data-sets – which will be explored later in this section. Since the holdout samples for both LCs and SMEs contain real-life data, they will also be used to as the holdout samples for the *SMOTEd* data-sets.

8.4.1.2 *SMOTEd* Data-set for SMEs

The same process and parameters that were used in Chapter 5 for creating the *SMOTEd* data-sets were used in this chapter. After *SMOTEing* the SMEs original training data-set, the results yielded a *SMOTEd* data-set with 100 observations – 50 healthy and 50 distressed observations, thus eliminating the prevailing class imbalance problem that existed in the original SMEs data-set. The *SMOTEing* process has oversampled the healthy companies by doubling their amount from 25 to 50, and has undersampled the distressed companies from 638 to 50 (removed 588 observations). The *SMOTEd* data-set is more than six times smaller than the original data-set – refer to Table 8.4. This balanced *SMOTEd* data-set is used to train the various models constructed in this chapter, before being tested on the holdout sample from the original data – as shown in Table 8.5.

Table 8. 4 Original versus *SMOTEd* Data for SMEs

Data-set	Number of Companies	Class Imbalance %
Original	638	96.08% Healthy – 3.92% Distressed
SMOTEd	100	50.00% Healthy – 50% Distressed

Table 8. 5 Training and Holdout Samples for SMEs

Data-set	Sample Partition	Number of Observations	Percentage
Train (2 Options)	**Original**	638	50%
	SMOTEd	100	100%
Holdout Sample for All SME Models		637	50%

8.4.1.3 *SMOTEd* Data-set for LCs

After *SMOTEing* the LCs original data-set, the results yielded a *SMOTEd* data-set with 64 observations – 32 healthy and 32 distressed observations, thus also eliminating the prevailing class imbalance problem that existed in the original LCs data-set. The *SMOTEing* process has oversampled the healthy companies by doubling their amount from 32 to 64, and undersampled the distressed companies from 146 to 32 (removed 114 observations). The *SMOTEd* data-set is less than half the size of the original training data-set – refer to Table 8.6 for a comparison of the original and *SMOTEd* data-sets for LCs. This balanced *SMOTEd* data-set is used to train the various models constructed in this chapter, and is then tested on the holdout sample from the original data – as shown in Table 8.7.

Table 8. 6 Original versus *SMOTEd* Data for LCs

Data-set	Number of Companies	Class Imbalance %
Original	146	89.04% Listed – 10.96% Delisted
SMOTEd	64	50.00% Listed – 50% Delisted

Table 8. 7 Training and Holdout Samples for LCs

Data-set	Sample Partition	Number of Observations	Percentage
Train (2 Options)	**Original**	146	50%
	SMOTEd	64	100%
Holdout Sample for All LC Models		146	50%

8.4.2 Models Constructed

This subsection explains the models built for this study. Two software packages were used to aid with the analysis, namely: 'Salford Predictive Modeler' and 'R' software package. 'R' was used to develop the treebag models, whereas 'Salford Predictive Modeler' was used to develop the DT, RF, and SGB models. To minimise repetition, the mechanics of these techniques will not be presented, refer to Chapter 2 for in-depth analysis of the DT, RF, and SGB techniques, and to Chapter 5 for an analysis of the treebag technique.

8.4.2.1 Decision Tree Models

Four models were created using DTs – two for the SMEs, one using the original data-set, and the other using the *SMOTEd* data-set; and the other two for the LCs, again, one for each data-set. Building the DT models had following properties – all are commonly used metrics:

- Testing method to determine optimal size was based on the commonly used tenfold cross validation
- The parameters influencing the selection of the best tree were based on commonly used criteria:
  - Standard error rule: Minimum cost tree regardless of size,
  - Variable importance formula: All surrogates count equally
- The splitting method for the classification trees was the popular Gini criterion.

The SME models were then tested on the SMEs holdout sample. Similarly, the LC models were tested on the LCs holdout sample.

8.4.2.2 Treebag Models

The models for both SMEs and LCs were trained using the “caret” package on the training samples using the commonly used tenfold cross validation. Whether the company is healthy or distressed was set as the response variable, whereas all of the variables shown in Table 8.2 were set as predictors. Standard parameters were used when developing the treebag model.

As mentioned earlier, to check how the treebag models performed, the ROC and AUROC measures were used to provide both visual and empirical results. Four models were created using treebag – two for the SMEs, one using the original data-set, and the other using the *SMOTEd* data-set; and the other two for the LCs, again, one for each data-set. The SME models were then tested on the SMEs holdout sample. Similarly, the LC models were tested on the LCs holdout sample.

8.4.2.3 Random Forests Models

Four models were created using RFs – two for the SMEs, one using the original data-set, and the other using the *SMOTEd* data-set; and the other two for the LCs, again, one for each data-set. Building the RF models had following properties – all are commonly used metrics.

- Testing method was based on the commonly used out of bag method
- Number of trees built: 1,000
- Number of predictors: Square root (√24 ≈ 5)

The SME models were then tested on the SMEs holdout sample. Similarly, the LC models were tested on the LCs holdout sample.

8.4.2.4 Stochastic Gradient Boosting Models

As with all the other models, four models were created using SGB – two for the SMEs, one using the original data-set, and the other using the *SMOTEd* data-set; and the other two for the LCs, again, one for each data-set. Building the SGB models had following properties – all are commonly used metrics.

- Testing method was based on the popular tenfold cross validation
- Number of trees built: 1000
- Maximum nodes per tree: 6
- Criterion for determining optimal number of trees for model: AUROC

The SME models were then tested on the SMEs holdout sample. Similarly, the LC models were tested on the LCs holdout sample.

8.5 Results

This section presents the results in this study for both SMEs and LCs using the four aforesaid techniques after they have been tested on their respective holdout samples – SMEs holdout sample size: 637 observations; LCs holdout sample size: 146 observations. Refer to the Appendices section (Appendix 2 and Appendix 3) for the raw R-code and data summary.

8.5.1 Treebag Models

8.5.1.1 AUROC Results

- **SMEs:** The treebag model using the original data-set yielded an AUROC result of 0.76. On the other hand, the treebag model using the *SMOTEd* data-set yielded an AUROC result of 0.82.
- **LCs:** The treebag model using the original data-set yielded an AUROC result of 0.89. On the other hand, the treebag model using the *SMOTEd* data-set yielded an AUROC score of 0.89.

What is notable in these results, is that the models using the *SMOTEd* data-sets outperformed the models using the original data-sets for both SMEs and LCs.

8.5.1.2 ROC Results

The black lines in the figures below represent the models' predictive performance. The grey lines are there purely for illustrative purposes showcasing a hypothetical model with no distinguishing capabilities between the classes. Refer to Chapter 5 for an explanation of the mechanics of interpreting the ROC graphs. The results are as follows:

- **SMEs:** Figures 8.1 and 8.2 below present the ROC graphs for both the original and *SMOTEd* models. As is evident when comparing both graphs, the black line of the *SMOTEd* model runs closer to the Y-axis, thus encompassing a larger area beneath it, which is reflected in the higher AUROC score of the *SMOTEd* Model vis-à-vis the Original Model.

Figure 8. 1 Original SMEs Treebag ROC

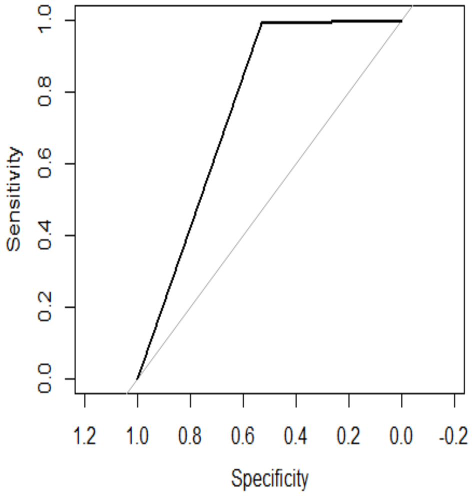
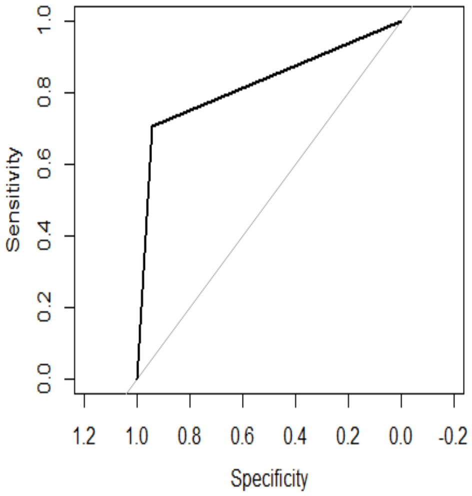


Figure 8. 2 *SMOTEd* SMEs Treebag ROC



- **LCs:** Figures 8.3 and 8.4 below present the ROC graphs for both the original and *SMOTEd* models. As is evident when comparing both graphs, the black lines of the both models look very similar. This is reflected in the same AUROC score of either model, thus indicating no empirical superiority of SMOTE here. However, due to the much smaller data-set, using the *SMOTEd* model is preferable as it is much easier to deal with.

Figure 8. 3 Original LCs Treebag ROC

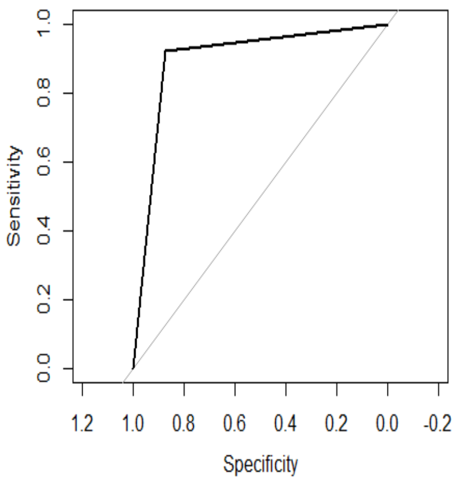
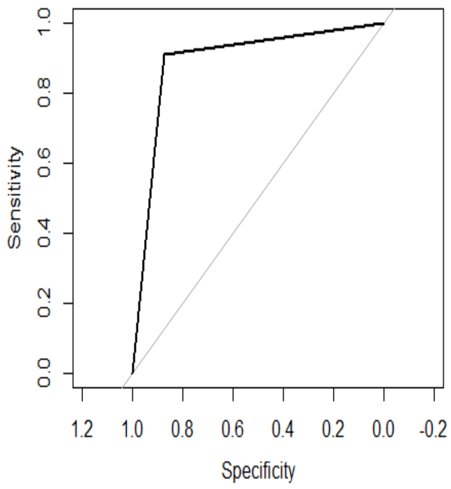


Figure 8. 4 *SMOTEd* LCs Treebag ROC



8.5.2 Decision Tree Models

8.5.2.1 AUROC Results

- **SMEs:** The DT model using the original data-set yielded an AUROC result of 0.76. On the other hand, the DT model using the *SMOTEd* data-set yielded an AUROC result of 0.78.
- **LCs:** The DT model using the original data-set yielded an AUROC result of 0.76. On the other hand, the DT model using the *SMOTEd* data-set yielded an AUROC result of 0.86.

8.5.2.2 ROC Results

- **SMEs:** Figures 8.5 and 8.6 below present the ROC graphs for both the original and *SMOTEd* models. As is evident when comparing both graphs, the blue line of the *SMOTEd* model runs closer to the Y-axis, thus encompassing a larger area beneath it, which is reflected in the higher AUROC score of the *SMOTEd* Model vis-à-vis the Original Model.

Figure 8. 5 Original SMEs DT ROC

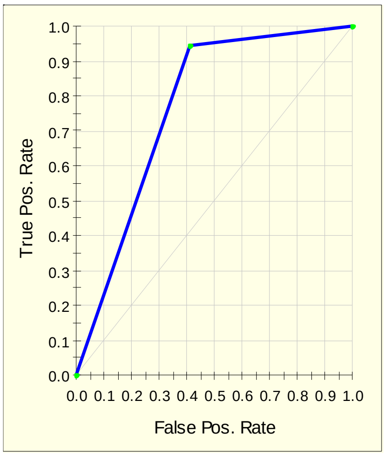
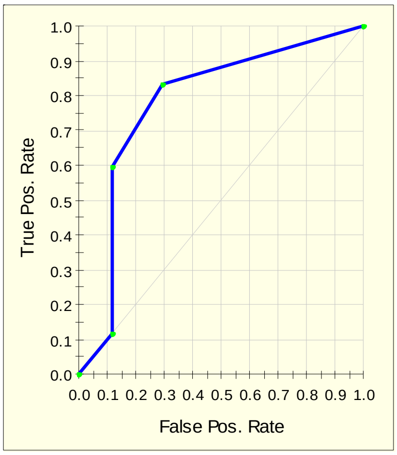


Figure 8. 6 *SMOTEd* SMEs DT ROC



- **LCs:** Figures 8.7 and 8.8 below present the ROC graphs for both the original and *SMOTEd* models. As is evident when comparing both graphs, the blue line of the *SMOTEd* model runs closer to the Y-axis, thus encompassing a larger area beneath it, which is reflected in the higher AUROC score of the *SMOTEd* Model vis-à-vis the Original Model.

Figure 8. 7 Original LCs DT ROC

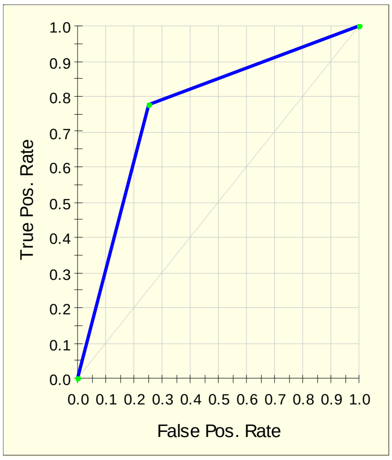
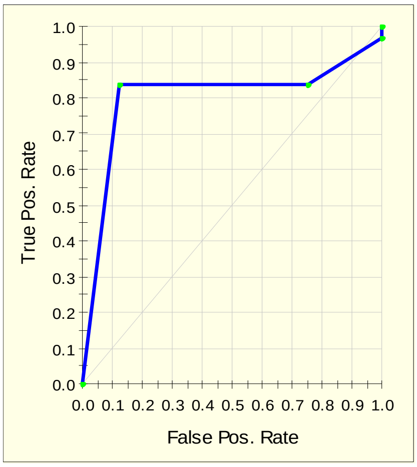


Figure 8. 8 *SMOTEd* LCs DT ROC



8.5.3 Random Forests Models

8.5.3.1 AUROC Results

- **SMEs:** The RF model using the original data-set yielded an AUROC result of 0.89. On the other hand, the RF model using the *SMOTEd* data-set yielded an AUROC result of 0.9.
- **LCs:** The RF model using the original data-set yielded an AUROC result of 0.88. On the other hand, the RF model using the *SMOTEd* data-set yielded an AUROC result of 0.9.

8.5.3.2 ROC Results

- **SMEs:** Figures 8.9 and 8.10 below present the ROC graphs for both the original and *SMOTEd* models. As is evident when comparing both graphs, the blue lines of the both models look very similar. Therefore, it is imperative to check the AUROC score in order determine which model is empirically superior. As presented above, the model using *SMOTEd* data is empirically superior.

Figure 8. 9 Original SMEs RF ROC

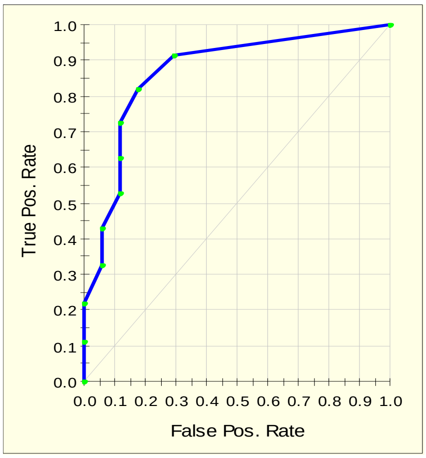
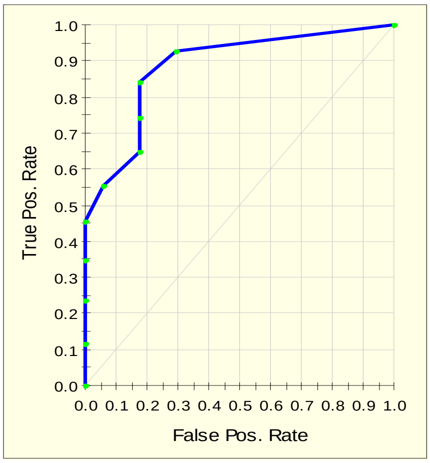


Figure 8. 10 *SMOTEd* SMEs RF ROC



- **LCs:** Figures 8.11 and 8.12 below present the ROC graphs for both the original and *SMOTEd* models. Similar to the SMEs models, the blue lines of the both models look very similar. Therefore, after referring to the AUROC score, it is clear that the model using *SMOTEd* data is empirically superior.

Figure 8. 11 Original LCs RF ROC

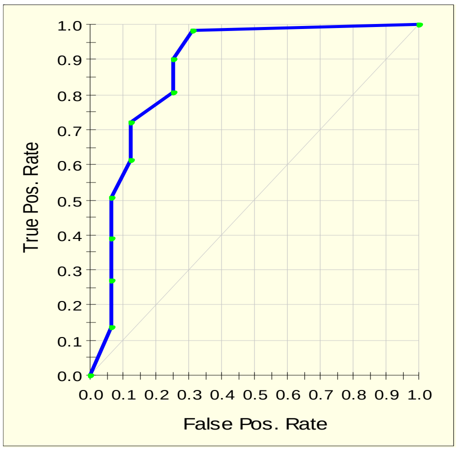
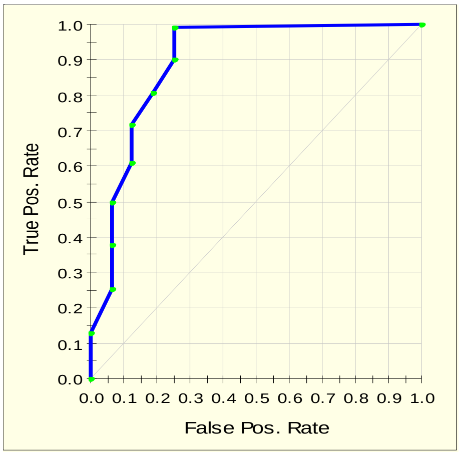


Figure 8. 12 *SMOTEd* LCs RF ROC



8.5.4 Stochastic Gradient Boosting Models

8.5.4.1 AUROC Results

- **SMEs:** The SGB model using the original data-set yielded an AUROC result of 0.86. On the other hand, the SGB model using the *SMOTEd* data-set yielded an AUROC result of 0.9.
- **LCs:** The SGB model using the original data-set yielded an AUROC result of 0.89. On the other hand, the SGB model using the *SMOTEd* data-set yielded an AUROC result of 0.91.

8.5.4.2 ROC Results

- **SMEs:** Figures 8.13 and 8.14 below present the ROC graphs for both the original and *SMOTEd* models. As is evident when comparing both graphs, the blue line of the *SMOTEd* model runs closer to the Y-axis, thus encompassing a larger area beneath it, which is reflected in the higher AUROC score of the *SMOTEd* Model vis-à-vis the Original Model.

Figure 8. 13 Original SMEs SGB ROC

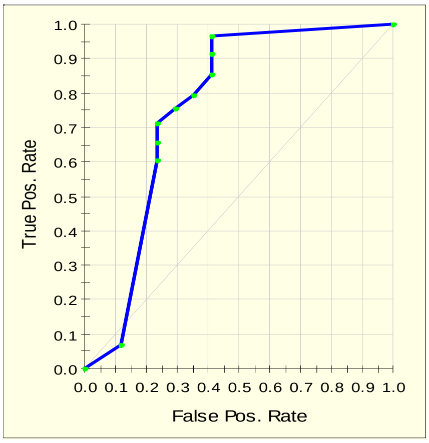
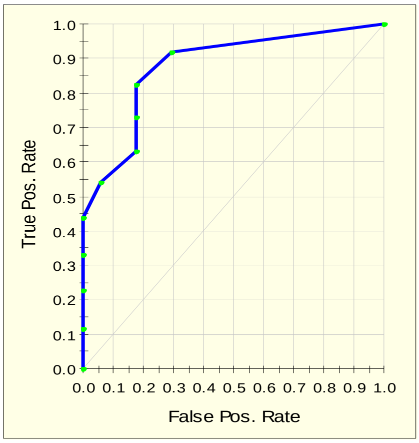


Figure 8. 14 *SMOTEd* SMEs SGB ROC



- **LCs:** Figures 8.15 and 8.16 below present the ROC graphs for both the original and *SMOTEd* models. Similar to the SMEs models, the blue lines of the both models look very similar. Therefore, after referring to the AUROC score, it is clear that the model using *SMOTEd* data is empirically superior.

Figure 8. 15 Original LCs SGB ROC

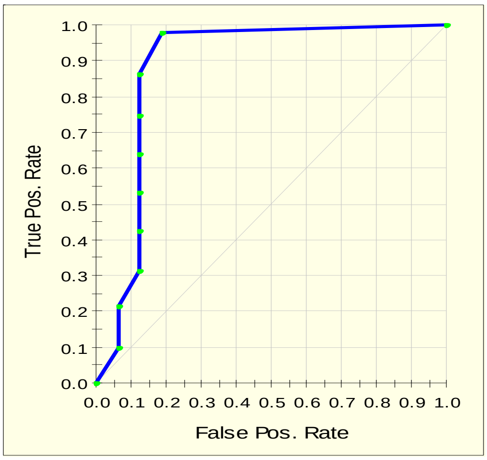
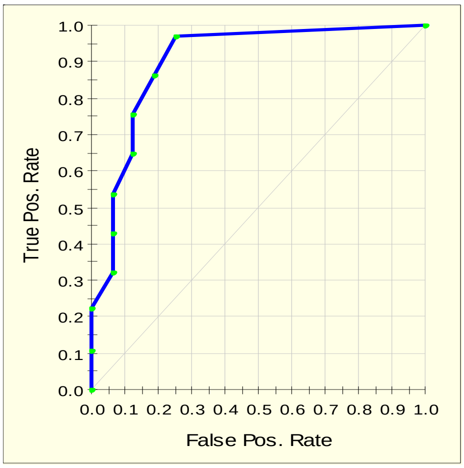


Figure 8. 16 *SMOTEd* LCs SGB ROC



8.5.5 Model Comparison

The following subsection presents the AUROC results and the most important variables in detection of financial distress for both SMEs and LCs in a tabulated fashion for ease of comparison. The top five variables, in order of importance, are presented for each developed model.

8.5.5.1 SMEs

Table 8. 8 SMEs Most Important Variables using Original Data

Model	AUC	IMPORTANT VARIABLES
DT	0.76	(1) QR; (2) Cash/CL; (3) CR; (4) TL/TA; (5) Cash/TA
Treebag	0.76	(1) CR; (2) ROA; (3) QR; (4) ROC; (5) ROE
RF	0.89	(1) Cash/TA; (2) CA/TA; (3) TL/TA; (4) LnTA; (5) QR
SGB	0.86	(1) QR; (2) CR; (3) Cash/CL; (4) EBITDA/TA; (5) RE/TA

Table 8. 9 SMEs AUROC Results Using *SMOTEd* Data

Model	AUC	IMPORTANT VARIABLES
DT	0.78	(1) ROC; (2) ROA; (3) ROE; (4) ROCE; (5) Cash/TA
Treebag	0.82	(1) ROC; (2) ROA; (3) ROE; (4) Cash/TA; (5) LnTA
RF	0.9	(1) Cash/TA; (2) ROC; (3) ROE; (4) ROA; (5) QR
SGB	0.9	(1) ROC; (2) Cash/TA; (3) TL/TA; (4) LnTA; (5) RE/TA

As is evident in Tables 8.8 and 8.9, the AUROC scores of the models using *SMOTEd* data are higher than those using original data. This indicates that using SMOTE provides empirically superior results. The increase in accuracy across the various models conform with the literature – in terms of the predictive accuracy of DT<treebag<RF/SGB. The DT uses builds a single tree, whereas treebag, RF, and SGB build an ensemble of trees, therefore, the accuracy in such models tend to outweigh DT.

As for the most important variables in detecting financial distress, the results from both data-sets showcase liquidity-driven variables, as shown:

- **Original:** In the models using original data, variables such as Quick Ratio (QR), Current Ratio (CR), Cash divided by Total Assets (CASH/TA), and Cash divided by Current Liabilities (CASH/CL) appear frequently across the models.
- ***SMOTEd*:** As for the models using *SMOTEd* data, there is a little more consistency of variables across the models. Variables such as Return on Capital (ROC), Cash divided by Total Assets (CASH/TA), Return on Equity (ROE), and Return on Assets (ROA) appear frequently across the models.

The SME models present an important finding to showcase that SMEs are liquidity driven, unlike large corporations, since they cannot access debt as easily, so liquidity is key to their success.

8.5.5.2 LCs

Table 8. 10 LCs Most Important Variables Using Original Data

Model	AUC	Important Variables
DT	0.76	(1) ROE; (2) ROCE; (3) NI/TA; (4) ROA; (5) CFO/CL
Treebag	0.89	(1) LnEmp; (2) TL/TA; (3) CA/TA; (4) Cash/CL; (5) Cash/TA
RF	0.88	(1) ROE; (2) ROCE; (3) ROA; (4) LnEmp; (5) NI/TA
SGB	0.89	(1) LnEmp; (2) ROE; (3) CA/TA; (4) Asset Turnover; (5) CR

Table 8. 11 LCs AUROC Results Using *SMOTEd* Data

Model	AUC	Important Variables
DT	0.86	(1) LnEmp; (2) TL/TA; (3) CA/TA; (4) Cash/CL; (5) Cash/TA
Treebag	0.89	(1) LnEmp; (2) NI/TA; (3) ROE; (4) CA/TA; (5) Altman Z-Score
RF	0.9	(1) LnEmp; (2) Gross Margin; (3) CA/TA; (4) NI/TA; (5) EBITDA/TA
SGB	0.91	(1) LnEmp; (2) Gross Margin; (3) EBITDA/TA; (4) ROE; (5) NI/TA

As is evident in Tables 8.10 and 8.11, the AUROC scores of the models using *SMOTEd* data are higher than those using original data, expect for the treebag models which yielded similar scores. This indicates that using SMOTE provides empirically superior results. The increase in accuracy across the various models conform with the literature – in terms of the predictive accuracy of tree ensembles over single tree techniques.

As for the most important variables in detecting financial distress, the results from both data-sets showcase variables that are asset and employment-driven, as shown:

- **Original:** In the models using original data, variables such as Return on Equity (ROE), Return on Capital Enterprise (ROCE), Net Income divided by Total Assets (NI/TA), and the natural logarithm of employees (LnEMP) appear frequently across the models.
- ***SMOTEd*:** In the models using *SMOTEd* data, there is a little more consistency of variables across the models. Variables such as the natural logarithm of employees (LnEMP), Gross Margin, Current Assets divided by Total Assets

(CA/TA), and Earnings Before Interest, Tax, Depreciation, and Amortisation divided by Total Assets (EBITDA/TA) appear frequently across the models.

The LC models present important findings, namely: LCs are more asset-driven than SMEs, and the number of employees a company has is an important determinant of a company's financial health, that is, the more the employees, the more likely the company is financially healthy. This is in concert with the ASBFEO (2016) statistics presented in the Introduction section regarding the survival rates of companies, as well as the studies presented in the Literature Review section.

8.6 Conclusion

In this chapter, FDP models were created using data pertaining to SMEs and LCs that are listed on the ASX. Another set of FDP models were created on data that was *SMOTEd*. Both visual (as per the ROC graph) and empirical (as per the AUROC scores) results were presented for all the models created in this study. The empirical AUROC results – which takes into consideration specificity and sensitivity – indicated that the models using SMOTE outperformed the models using the original data, with the SGB model being the superior model. These results cement Chapter 5's findings in terms of superiority of SMOTE pertaining to FDP modelling.

In terms of variable importance, this chapter's findings indicate that variables affecting financial distress differ substantially for SMEs and LCs, as was shown by the variable importance analysis. Most notably, SMEs are liquidity-driven, with the most important variable that appeared frequently across all models being 'Return on Capital'. A rationale to explain the liquidity-driven nature of SMEs is that it is much harder for SMEs to access funds from creditors due to the limited collateral on offer – hence, creditors can be more cautious providing a loan to SMEs. Therefore, if an SME is presenting low liquidity ratios, that ought to raise red flags (indicative of possible financial distress).

On the other hand, LCs are more employee and asset-driven, with the most important variable that appeared frequently across all models being 'natural logarithm of employees.' One rationale to explain the asset-driven nature of LCs is due to the company size, that is, generally, the bigger the company is, the more employees it has and the more assets it acquires. Therefore, this opens the doors for easy access to creditors, lobbying power, and influence on stakeholders. These factors can explain the positive correlation between LC financial health, employment numbers, and high asset ratios. Therefore, if an LC is presenting low asset ratios or has low employment figures, that ought to raise red flags (indicative of possible financial distress).

This study is a novel way of combining SMOTE with machine learning techniques to create FDP models pertaining to SMEs and LCs, in order to present the most accurate models, and present the most important variables in determining their financial distress. The results indicated that the 'Return on Capital' variable was the most important variable in determining the success or failure of a SME; as for LCs, the number of employees was directly proportional with a firm's success or failure. The findings present a need for distinctly separate FDP models to be created when modelling for SMEs or LCs. This chapter has verified *Hypotheses 4 and 5.*

Chapter 9: Conclusions, Study Limitations, & Future Works

This thesis has explored and tackled the issue of Financial Distress Prediction (FDP) from numerous angles. This was done by firstly introducing the concept of FDP, the research questions and hypotheses – as was outlined in Chapter 1. Following this, an extensive review of the literature was conducted – as was shown primarily in Chapter 2, but also in each subsequent chapter, as per each chapter's specific topic. Later chapters utilised various traditional and machine learning statistical techniques were utilised to create models that were used to:

- Test the efficacy of industry-specificity vis-à-vis a *one-size-fits-all* model on FDP – this was done by creating separate industry-specific models through segregating companies in the Australian marketplace as per each industry they subscribe to. After this, various techniques were utilised to create FDP models for each industry. The results indicated the superiority of industry-specific models vis-à-vis industry-wide models. Also, results indicated that variable importance differ per industry – refer to Chapter 3;

- Outline the empirically superior FDP model when applied to the Australian mining industry, as well as, presenting the most important variables that showcased a mining company's success or failure – this was done through creating four models, each using a different modelling technique, namely: LR, DT, RF, and SGB. Since the data-set was imbalanced, the models' cut-off points were optimised – refer to Chapter 4;

- Apply the Synthetic Minority Oversampling Technique (SMOTE) to remedy for the class imbalance problem – this was done by comparing the empirical predictive accuracy of a model using a standard data-set suffering from class imbalance vis-à-vis a model developed using a data-set that has been *SMOTEd*. The predictive accuracies of both approaches were compared using machine learning models, namely: DT, treebag, RF, and SGB. The results conclusively established the superiority of the SMOTE technique – refer to Chapter 5;

- Explore the differences in FDP between Large Companies (LCs) versus Small and Medium Enterprises (SMEs) – since both data-sets suffered from the class imbalance problem, SMOTE was used in the same manner it was used in Chapter 5. Four techniques were used: DT, treebag, RF, and SGB. The results indicated a superior predictive accuracy for the models that used the *SMOTEd* data-set, as well as outlining the different variables with the highest predictive power for both SMEs and LCs – refer to Chapter 8;

- Apply FDP modelling to Islamic banking, as well as outline the differences and similarities vis-à-vis conventional banking – this was done through using three different measures of financial distress/success pertaining to Islamic banks, namely: Altman Z-Score, Altman Z-Score for Service Firms, and the Standardised Profits, to measure the banks' financial distress. DT, RF, and SGB techniques were used to build the models for each measure. The results indicated that 'Working Capital/Total Assets' was the most predictive variable for predicting financial distress in Islamic banks using all three models, for both the Altman Z-Score and Altman Z-Score for Service Firms methods. On the other hand, the Standardised Profits method, yielded 'Return on Revenue' as the most important variable – refer to Chapter 7;

- Develop FDP indices – a novel method of presenting the financial health of companies was presented – this was done by using factor analysis and Principal Component Analysis (PCA). Following this, Factor Weighted Index (FWI), Weighted Factor Loading Index (WFLI), and a Standardised Index (SI) approaches were used to create three indices. The SI was the optimal choice, as demonstrated by a comparison with a standard LR model, and an evaluation with established performance metrics, namely: share price and ordinary shares market capitalisation. A particular index for the Australian mining companies was created and dubbed the 'K-Index'. The K-Index showcased the top 10 and bottom 10 mining companies. – refer to Chapter 6.

9.1 Contributions

This dissertation shows that there are real-world problems that need to be addressed, namely: high business failure rates in Australia, and a lack of FDP research focusing on the Australian marketplace. This research has investigated past literature to demonstrate how FDP models may aid in addressing the aforementioned problems, unravelled gaps in the literature that fail to show: the differences amongst different industries, variable differences amongst SMEs and large companies, and the benefits of creating an FDP index.

This thesis explored statistical and machine learning techniques, including: MDA, LR, DT, RF, and SGB. These techniques were used to create models that contributed to the literature through showcasing that differences exist in terms of FDP modelling amongst SMEs and LCs; hence it is important to analyse SMEs/LCs separately. Differences also exist across various industries; thus, using industry-specific models vis-à-vis an industry-wide model yields empirically superior business failure predictions. The research also empirically showcased the predictive pre-eminence of machine learning techniques when compared with traditional statistical techniques – this was done by creating models using data from sectors that are rarely studied, such as various industries in the Australian marketplace and Islamic banks. To add, this treatise added to the limited literature available on Islamic banking by applying FDP modelling and outlining the most important variables in identifying an Islamic bank's success or failure. Finally, a novel concept of creating FDP indices was introduced, which pools the benefits of both, the empirical findings of FDP modelling and the user-friendliness of indices.

Although this research was primarily focused on the Australian marketplace, the methodologies presented can be applied to companies operating in any industry on a global scale, that is, the implications and applicability of this study is not confined to Australia. Thus, this research has the potential to greatly benefit various stakeholders, from investors to governmental agencies, which, if applied alongside competent managerial decisions, may lead reduced business failure, which will, in-effect, have a positive consequence on the global economy. This can be done by monitoring and assigning red flags to certain variables that have been shown to have a direct effect on companies operating in particular industries – thus leading to managerial reactive measures to act accordingly.

Limitations do exist, therefore one of the main objectives of future work is to find ways to eliminate them, or at least alleviate their effects on the results. The next sections outline some of the limitations of this research, followed by some of the prospects for future works.

9.2 Study Limitations

- **Scope of Study:** With the exception of Chapter 7, the thesis was largely centred around Australian data. The models which were developed in Australia may not be applicable to other countries or regions, as each country is unique in terms of its laws, accounting standards, and micro and macroeconomics. The methods used can be theoretically applied on a global scale, however, using international data would widen the scope of the study.

- **Solely Conducting Quantitative Research:** This thesis has solely been quantitative-based. Although there are many advantages for the quantitative method, limitations do exist. The limitations regarding this point are generic to the quantitative method, hence do not necessarily reflect the research conducted in this study. Some of these limitations include: the possibility of presenting a myopic perspective due to: the results solely offering numerical explanations, lack of thorough narrative and elaborate accounts of human perception, and an unconscious bias when presenting results that might not accurately showcase real-life occurrences (Kruger, 2003).

- **Use of Archival Data:** Even though utmost diligence was carried out when obtaining data for the studies, databases may sometimes output data that is erroneous, missing, static (not interactive/dynamic), is not enough for a comprehensive analysis, or corrupted when extracted or downloaded on to a spreadsheet. To add, according to Shultz et al. (2005), limitations to archival data include: appropriateness of the data, detection of errors can be extremely difficult, and the lure of dustbowl empiricism (collection of data and creation of empirical observations, as opposed to producing a theoretical framework). Therefore, in order to capture the full picture, employing qualitative aspects, such as, regulatory measures, stakeholders' pressures, board members influence, could offer a more comprehensive perspective towards FDP – see Future Works section below.

- **Rationale of Variable Importance:** As mentioned in Chapter 3's Discussion section, due to a lack of studies offering insight as to why the variables that were deemed important by the industry-specific FDP models are in fact a crucial determinant of company financial distress, the rationales were provided by after discussing them with an expert in accountancy. This method is not watertight, therefore, further studies should be undertaken to provide a more valid justification for the differences amongst variables pertaining to different industries.

- **Use of Delisted Companies:** Delisted companies were regarded as financially distressed in this thesis; this is not necessarily always the case. Some databases include merged, withdrawn, suspended, or acquired companies under the delisted category. Therefore, this might result in unreliable results.

- **Private Company Data:** This was especially relevant for Chapter 8, since many SMEs are private companies, hence, are not mandated by law to surrender their financials. Therefore, the data used was not exhaustive and a huge chunk of the marketplace was overlooked, as access to private company data is extremely difficult. Refer to the Future Works section below for possible data-gathering options.

9.3 Future Works

There are a number of areas that could be explored in the future to help cement the claims presented in this dissertation, some of these areas include:

- **Focus on Family Businesses:** Family businesses are generally privately-owned, and in Australia, they account for 70% of all businesses; of those, 64% are small businesses (Clark, Eaton, Meek, Pye, & Tuhin, 2012; FBA, 2014). As was presented in the thesis, on average, 44 small businesses close doors every day (Cornish & Landy, 2013), therefore, this presents a legitimate cause of concern for the Australian economy. Given that the findings in this thesis empirically showed that there are differences amongst industries (Chapter 3) and company size (Chapter 8), it can be hypothesised that FDP modelling tailored to family businesses will yield better models and more accurate results.

- **Fraud, Neglect, and Disaster Variables:** The leading causes of business failure can be classified according to financial, economic, neglect, disaster, or fraud aspects (Anderson, 2006; Gepp, 2015). This thesis researched the financial component of business failure; this presents room to investigate the effects fraud, neglect, and disaster have on FDP. This can be done by quantifying the aforementioned components, in order to use them as predictors in FDP modelling.

- **Cross-Regional/International:** As mentioned earlier, this thesis mainly focused on Australian companies' data. It would be interesting to apply the statistical techniques and FDP models on international data.

- **Bayesian Model:** Many studies show that Bayesian models' accuracy supersedes other statistical techniques, including: Chaudhuri (2013); Shrivastava et al. (2018); Tsai (2005). This presents an opportunity to investigate in the future.

- **Dynamic Panel Logistic Modelling:** There are no studies applying dynamic panel logistic modelling to FDP thus far, this provides a pioneering research opportunity which may be investigated in future studies.

- **Qualitative and Corporate Governance (CG) Variables**: Adding qualitative variables adds another dimension to the study – examples include: ratio of males versus females in a company, role of females in a company, and board members decisions. CG is the set of rules, processes, and practices that direct and control companies. In Australia, following the collapse of major corporations such as HIH, Ansett, and OneTel, there has been an increasing concern about the quality of corporate governance. In 2002, the Horwath CG Report was introduced which provided an objective analysis of the governance structures in Australia’s top 250 listed companies by market capitalisation. The rankings are based on information about the board and its principal committees that is found in the companies’ annual reports and disclosures. The index is calculated similarly for all companies, irrespective of size. The report provides companies’ rankings and a five-star-scale analysis of how well the company’s CG standards are – one-star indicating CG structures are lacking in several areas, whereas five-stars indicated outstanding CG practices (Psaros & Seamer, 2002). The report also found that 30% of the companies had inferior CG structures. It pointed out the significance of having independent directors in the board as it will lead to better CG practices. In 2008, an updated version was released, namely the WHK Horwath Corporate Governance Report, this included five-star-scales, ranking information, and comparisons for the top 250 listed companies for the past three years, that is, 2006-2008 (Horwath, 2008). Previous studies have used the Horwath Report as an indicator for a company’s CG standards (Beekes & Brown, 2006; Lama, 2012). Inclusion of a CG index may aid the research by adding a new dimension to financial distress prediction.

- **SMEs:** Since many SMEs are private companies, they are not lawfully obliged to make their financials public. To overcome this, one option may be to ask them for a confidential or incognito raw data; another option may be surveying SMEs and collecting information. This will serve as an extension to the research conducted on SMEs in Chapter 8, which will, in-effect, widen the scope of the research and increase its credibility.

- **Macroeconomic Variables:** It might be intuitive that poor economic conditions might increase the rate of bankrupt firms. For the economy to thrive there needs to be a healthy rate of demand and supply, therefore when people's spending concentrate on necessities, not materialistic goods, firms will start incurring losses, which may lead to bankruptcy. Therefore, including macroeconomic variables may aid the research by providing a wider scope when forecasting financial distress. Duffie, Saita, and Wang (2007) solidifies this claim, as one of the factors their model was significantly dependent on was the state of the economy. Some of the macroeconomic variables the research will incorporate are: percentage change in annual GDP, interest rates, aggregate default rates, and unemployment rates.

Bibliography

ABC. (2011). Moody's affirms Australia's AAA rating. Retrieved from http://www.abc.net.au/news/2011-12-22/moody27s-affirms-australia-rating/3743656

Abeyasekera, S. (2005). Chapter XVIII Multivariate methods for index construction. *Household Sample Surveys in Developing and Transition Countries: Design, Implementation and Analysis*, 367-387.

ABS. (2001, 23/10/2002). Small Business in Australia. Retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/mf/1321.0

ABS. (2017). *Key Economic Indicators* Retrieved from http://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/1345.0?opendocument?opendocument#from-banner=LN

Ahn, B., Cho, S., & Kim, C. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Applications, 18*(2), 65-74.

Al-Shayea, Q. K., El-Refae, G. A., & El-Itter, S. F. (2010). Neural Networks in Bank Insolvency Prediction. *International Journal of Computer Science and Network Security, 10*(5), 240-245.

Al Zaabi, O. S. H. (2011). Potential for the application of emerging market Z-score in UAE Islamic banks. *International Journal of Islamic and Middle Eastern Finance and Management, 4*(2), 158-173.

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

Altman, E. I. (2000). Predicting financial distress of companies: revisiting the Z-score and ZETA models. *Stern School of Business, New York University*, 9-12.

Altman, E. I., & Hotchkiss, E. (2010). *Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt* (Vol. 289): John Wiley & Sons.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2014). Distressed Firm and Bankruptcy Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model. *Available at SSRN 2536340*.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model. *Journal of International Financial Management & Accounting, 28*(2), 131-171.

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance, 18*(3), 505-529.

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus, 43*(3), 332-357.

Altman, E. I., Sabato, G., & Wilson, N. (2010). The value of non-financial information in small and medium-sized enterprise risk management. *Journal of Credit Risk, 6*(2), 95-127.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175-185.

Anderson, S. (2006). *Investment management and mismanagement: History, findings, and analysis* (Vol. 17): Springer Science & Business Media.

Antony, G., & Rao, K. V. (2007). A composite index to explain variations in poverty, health, nutritional status and standard of living: Use of multivariate statistical methods. *Public Health, 121*(8), 578-587.

Anwar, S., & Mikami, Y. (2011). Comparing accuracy performance of ANN, MLR, and GARCH model in predicting time deposit return of Islamic bank. *International Journal of Trade, Economics and Finance, 2*(1), 44-51.

ASBFEO. (2016). *Small Business Counts: Small Business in the Australian Economy*. Australia Retrieved from https://www.asbfeo.gov.au/sites/default/files/Small_Business_Statistical_Report-Final.pdf.

ASIC. (2014). Are you a large or small proprietary company. Retrieved from https://asic.gov.au/regulatory-resources/financial-reporting-and-audit/preparers-of-financial-reports/are-you-a-large-or-small-proprietary-company/

ASIC. (2015). *Corporate insolvencies: September quarter 2015*. Australia Retrieved from http://download.asic.gov.au/media/3446712/insolvency-statistics-summary-september-quarter-2015-published-10-november-2015.pdf.

ASIC. (2018). *Corporate insolvencies: September quarter 2018*. Australia Retrieved from https://download.asic.gov.au/media/4952377/201809-sep-qtr-2018-summary-analysis.pdf.

Austrade. (2015). Investment – creating more opportunities for all Australians. Retrieved from https://www.austrade.gov.au/international/invest/benefits-of-foreign-direct-investment

Bank for International Settlements. (2016). Basel III: international regulatory framework for banks. Retrieved from https://www.bis.org/bcbs/basel3.htm

Bartlett, M. S. (1954). A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 296-298.

Bayley, L., & Taylor, S. L. (2007). Identifying earnings overstatements: A practical test. *Available at SSRN 995957*.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 71-111.

Beck, T., Demirgüç-Kunt, A., & Merrouche, O. (2013). Islamic vs. conventional banking: Business model, efficiency and stability. *Journal of Banking & Finance, 37*(2), 433-447.

Beekes, W., & Brown, P. (2006). Do Better-Governed Australian Firms Make More Informative Disclosures? *Journal of Business Finance & Accounting, 33*(3-4), 422-450.

Beneish, M. D. (1997). Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy, 16*(3), 271-309.

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry, 23*(2), 129-143.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602-613.

Bianchi, M., Campodall'Orto, S., Frattini, F., & Vercesi, P. (2010). Enabling open innovation in small- and medium-sized enterprises: how to find alternative applications for your technologies. *R&D Management, 40*(4), 414-431.

Bloomberg. (2011, 9/8/2011). Dow plunges as $2.5tn erased from equities. Retrieved from https://www.irishtimes.com/business/markets/equities/dow-plunges-as-2-5tn-erased-from-equities-1.593127

Bond, S. (2002). Dynamic panel data models: a guide to micro data methods and practice. *Portuguese Economic Journal, 1*(2), 141-162.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers.* Paper presented at the Proceedings of the 5th Annual Workshop on Computational Learning Theory, 144-152

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*: CRC press.

Brownlee, J. (2016). How to Build an Ensemble Of Machine Learning Algorithms in R (ready to use boosting, bagging and stacking). Retrieved from https://machinelearningmastery.com/machine-learning-ensembles-with-r/

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626-4636.

Butler, B. (2018). Mining decline spurs bankruptcies. Retrieved from https://www.theaustralian.com.au/business/economics/mining-decline-spurs-bankruptcies/news-story/b5597ffdd2724031bc236bb0109da560

Calabrese, R., Andreeva, G., & Ansell, J. (2019). “Birds of a Feather” Fail Together: Exploring the Nature of Dependency in SME Defaults. *Risk Analysis, 39*(1), 71-84.

Calenge, C. (2006). The package “adehabitat” for the R software: a tool for the analysis of space and habitat use by animals. *Ecological Modelling, 197*(3-4), 516-519.

Camacho-Miñano, M. D. M., Segovia-Vargas, M. J., & Pascual-Ezama, D. (2015). Which Characteristics Predict the Survival of Insolvent Firms? An SME Reorganization Prediction Model. *Journal of Small Business Management, 53*(2), 340-354.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Chandra, D. K., Ravi, V., & Bose, I. (2009). Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications, 36*(3), 4830-4837.

Chaudhuri, A. (2013). Bankruptcy prediction using Bayesian, hazard, mixed logit and rough Bayesian models: A comparative analysis. *Computer and Information Science, 6*(2), 103-125.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview *Data mining and knowledge discovery handbook*. *Springer*, 875-886.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chen, H.-L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S.-J., & Liu, D.-Y. (2011). A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. *Knowledge-Based Systems, 24*(8), 1348-1359.

Chen, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications, 38*(9), 11261-11272.

Chong, B. S., & Liu, M.-H. (2009). Islamic banking: interest-free or interest-based? *Pacific-Basin Finance Journal, 17*(1), 125-144.

Chung, K. C., Tan, S. S., & Holdsworth, D. K. (2008). Insolvency prediction model using multivariate discriminant analysis and artificial neural network for the finance industry in New Zealand. *International Journal of Business and Management*, *39*(1), 19-28.

Ciampi, F., & Gordini, N. (2013). Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of Italian small enterprises. *Journal of Small Business Management, 51*(1), 23-45.

Clark, M., Eaton, M., Meek, D., Pye, E., & Tuhin, R. (2012). *Australian Small Business Key Statistics and Analysis*. Canberra: Department of Industry, Innovation, Science, Research and Tertiary Education Retrieved from http://www.treasury.gov.au/~/media/Treasury/Publications%20and%20Media/Publications/2012/Australian%20Small%20Business%20-%20Key%20Statistics%20and%20Analysis/downloads/PDF/AustralianSmallBusinessKeyStatisticsAndAnalysis.ashx.

Coats, P. K., & Fant, L. F. (1993). Recognizing financial distress patterns using a neural network tool. *Financial Management*, 142-155.

Collins, R. A., & Green, R. D. (1982). Statistical methods for bankruptcy forecasting. *Journal of Economics and Business, 34*(4), 349-354.

Cornish, L., & Landy, S. (2013, 7/8/2013). Average of 44 small businesses closing their doors each day, according to Australian Bureau of Statistics data. Retrieved from http://www.heraldsun.com.au/news/victoria/average-of-44-small-businesses-closing-their-doors-each-day-according-to-australian-bureau-of-statistics-data/story-fni0fit3-1226692393716?nk=e7843220d5bb04e01c4f152f4cbfaaa7

Council on Foreign Relations. (2015). The Credit Rating Controversy. Retrieved from https://www.cfr.org/backgrounder/credit-rating-controversy

Cybinski, P. (2001). Description, explanation, prediction–the evolution of bankruptcy studies? *Managerial Finance, 27*(4), 29-44.

Daniel, B.-O., & Ionuț, G. (2013). Prediction of corporate bankruptcy in Romania through the use of logistic regression. *The Annals of the University of Oradea*, 976-986.

Davies, J. (2017). Global Financial Crisis - What caused it and how the world responded. Retrieved from https://www.canstar.com.au/home-loans/global-financial-crisis

Dewaelheyns, N., & Van Hulle, C. (2006). Corporate failure prediction modeling: Distorted by business groups' internal capital markets? *Journal of Business Finance & Accounting, 33*(5-6), 909-931.

DFAT. (2018). The benefits of foreign investment. Retrieved from https://dfat.gov.au/trade/investment/pages/the-benefits-of-foreign-investment.aspx

Dialga, I. (2017). Highlighting Methodological Limitations in the Steps of Composite Indicators Construction. *Social Indicators Research, 131*(2), 441-465.

Dorsey, R. E., Edmister, R. O., & Johnson, J. D. (1995). Bankruptcy prediction using artificial neural systems, *Research Foundation of The Institute of Chartered Financial Analysts,* 1-68.

Drummond, C., & Holte, R. C. (2003). *C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling.* Paper presented at the Workshop on Learning from Imbalanced Datasets II (Vol. 17), 1-8.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification 2nd Edition*. New York*. *John Wiley&Sons*.

Duffie, D., Saita, L., & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics, 83*(3), 635-665.

Edmister, R. O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative analysis, 7*(2), 1477-1493.

Elliott, L., Treanor, J., & Rushe, D. (2011, 6/8/2011). US credit rating downgraded to AA+ by Standard & Poor's. Retrieved from https://www.theguardian.com/business/2011/aug/05/ftse-slumps-us-jobs-data

EY. (2016). *World Islamic Banking Competitiveness Report*. Retrieved from http://www.ey.com/Publication/vwLUAssets/ey-world-islamic-banking-competitiveness-report-2016/$FILE/ey-world-islamic-banking-competitiveness-report-2016.pdf

Fantazzini, D., & Figini, S. (2009). Random survival forests models for SME credit risk measurement. *Methodology and Computing in Applied Probability, 11*(1), 29-45.

FBA. (2014). Australian Family Business Sector Statistics. Retrieved from http://fambiz.org.au/documents/AustralianFamilyBusinessSectorStatistics.pdf

Federal Reserve. (2003). Capital Standards for Banks: The Evolving Basel Accords. Retrieved from https://www.federalreserve.gov/pubs/bulletin/2003/0903lead.pdf

Feldman, M., & Zoller, T. D. (2012). Dealmakers in place: Social capital connections in regional entrepreneurial economies. *Regional Studies, 46*(1), 23-37.

Ferguson, A., Clinch, G., & Kean, S. (2011). Predicting the failure of developmental gold mining projects. *Australian Accounting Review, 21*(1), 44-53.

FitzPatrick, P. J. (1932). *A comparison of the ratios of successful industrial enterprises with those of failed companies*. Washington.

Freel, M. S. (2000). Barriers to product innovation in small manufacturing firms. *International Small Business Journal, 18*(2), 60-80.

Frydenberg, J. (2015). Mining and the Australian economy: the Australian Government's priorities for the mining sector. Retrieved from http://www.minister.industry.gov.au/ministers/frydenberg/speeches/mining-and-australian-economy-australian-governments-priorities-mining

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

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 463-484.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research, 241*(1), 236-247.

Gepp, A. (2015). *Financial statement frraud detection using supervised learning methods.* (PhD Dissertation). Bond, Business School, 1-319.

Gepp, A., & Kumar, K. (2012). Business failure prediction using statistical techniques: A Review. *Some Recent Developments in Statistical Theory and Applications*. Boca Raton, Florida, USA: Brown Walker Press, 1-25.

Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting, 29*(6), 536-555.

Grice, J. S., & Ingram, R. W. (2001). Tests of the generalizability of Altman's bankruptcy prediction model. *Journal of Business Research, 54*(1), 53-61.

Gromski, P., Xu, Y., Kotze, H., Correa, E., Ellis, D., Armitage, E., . . . Goodacre, R. (2014). Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. *Metabolites, 4*(2), 433-452.

Grover, J., & Lavin, A. (2001). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy: A Service Industry Extension of Altman’s Z-Score Model of Bankruptcy Prediction. *Southern Finance Association*.

Gupta, J., Gregoriou, A., & Healy, J. (2015). Forecasting bankruptcy for SMEs using hazard function: To what extent does size matter? *Review of Quantitative Finance and Accounting, 45*(4), 845-869.

Hakim, L., Sartono, B., & Saefuddin, A. (2017). Bagging Based Ensemble Classification Method on Imbalance Datasets. *International Journal of Computer Science and Network, 6*(6), 670-676.

Hall, G. (1994). Factors distinguishing survivors from failures amongst small firms in the UK construction sector. *Journal of Management Studies, 31*(5), 737-760.

Halteh, K. (2015). Bankruptcy Prediction of Industry-Specific Businesses Using Logistic Regression. *Journal of Global Academic Institute Business & Economics, 1*(2), 151-163.

Halteh, K., Kumar, K., & Gepp, A. (2018a). Financial distress prediction of Islamic banks using tree-based stochastic techniques. *Managerial Finance*, *44*(6), 759-773.

Halteh, K., Kumar, K., & Gepp, A. (2018b). Using Cutting-Edge Tree-Based Stochastic Models to Predict Credit Risk. *Risks, 6*(2), 55.

Hanif, M. (2011). Differences and Similarities in Islamic and Conventional Banking. *IntInternational Journal of Business and Social Science, 2*(2), 166-175.

He, Y., & Kamath, R. (2005). Bankruptcy prediction of small firms in an individual industry with the help of mixed industry models. *Asia-Pacific Journal of Accounting & Economics, 12*(1), 19-36.

He, Z. L., & Trabelsi, S. (2013). Using a Bayesian model for bankruptcy prediction: a comparative approach. *Available at SSRN 2201224*.

Helmes, E., Goffin, R. D., & Chrisjohn, R. D. (1998). Confirmatory factor analysis of the life satisfaction index. *Social Indicators Research, 45*(1), 371-390.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computing*. New York: Addison Wesley.

Hightower, W. L. (1978). Development of an index of health utilizing factor analysis. *Medical care*, 245-255.

Horwath, W. (2008). WHK Horwath Corporate Governance Report. *University of Newcastle, Newcastle, Australia*.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417-441.

Hua, Z., Wang, Y., Xu, X., Zhang, B., & Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications, 33*(2), 434-440.

Huarng, K., Yu, H., & Chen, C. (2005). *The application of decision trees to forecast financial distressed companies.* Paper presented in the Proceeding of the 2005 International Conference on Intelligent Technologies and Applied Statistics.

Hung, C., & Chen, J.-H. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert Systems with Applications, 36*(3), 5297-5303.

Jaikengkit, A.-o. (2004). *Corporate Governance and Financial Distress: An Empirical Analysis. The Case of Thai Financial Institutions.* Case Western Reserve University.

Jan, A., & Marimuthu, M. (2015a). Altman model and bankruptcy profile of islamic banking industry: A comparative analysis on financial performance. *International Journal of Business and Management, 10*(7), 110-119.

Jan, A., & Marimuthu, M. (2015b). Bankruptcy and Sustainability: A Conceptual Review on Islamic Banking Industry. *Global Business & Management Research, 7*(1), 109-138.

Jan, A., Marimuthu, M., Shad, M. K., Zahid, M., & Jan, A. A. (2019). Bankruptcy profile of the Islamic and conventional banks in Malaysia: a post-crisis period analysis. *Economic Change and Restructuring, 52*(1), 67-87.

Japkowicz, N. (2000). *The class imbalance problem: Significance and strategies.* Paper presented in the Proceedings of the International Conference on Artificial Intelligence. Retreived from https://pdfs.semanticscholar.org/907b/02c6322d0e7dff6b0201b03e3d2c6bc1d38f.pdf

Japkowicz, N., Myers, C., & Gluck, M. (1995). *A novelty detection approach to classification* (Vol. 1), 518-523.

Jolliffe, I. T. (1990). Principal component analysis: a beginner's guide—I. Introduction and application. *Weather, 45*(10), 375-382.

Kahle, K. M., & Stulz, R. M. (2013). Access to capital, investment, and the financial crisis. *Journal of Financial Economics, 110*(2), 280-299.

Kaiser, J. (2014). Dealing with missing values in data. *Journal of systems integration, 5*(1), 42-51.

Kantardzic, M. (2011). *Data Mining – Concepts, Models, Methods, and Algorithms* (2nd ed.). New Jersey: John Wiley & Sons.

Keasey, K., Pindado, J., & Rodrigues, L. (2015). The determinants of the costs of financial distress in SMEs. *International Small Business Journal, 33*(8), 862-881.

Kenton, W. (2019). Working Capital Turnover Definition.   Retrieved from https://www.investopedia.com/terms/w/workingcapitalturnover.asp

Khan, F. (2010). How 'Islamic'is Islamic banking? *Journal of Economic Behavior & Organization, 76*(3), 805-820.

Kim, M.-J., Kang, D.-K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications, 42*(3), 1074-1082.

Knezevic, S. Z., Streibig, J. C., & Ritz, C. (2007). Utilizing R software package for dose-response studies: the concept and data analysis. *Weed Technology, 21*(3), 840-848.

Koshy, P., & Prasad, V. (2007). Small and Micro Enterprises: A tool in the fight against poverty. Retrieved from https://mpra.ub.uni-muenchen.de/22827/

Krishnan, V. (2010). Constructing an area-based socioeconomic index: A principal components analysis approach. *Edmonton, Alberta: Early Child Development Mapping Project*.

Kruger, D. (2003). Integrating quantitative and qualitative methods in community research. *The Community Psychologist, 36*(2), 18-19.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning, 30*(2-3), 195-215.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European Journal of Operational Research, 180*(1), 1-28.

Kyriazopoulos Georgios, M. (2014). *The Edward I. Altman's Model of Bankruptcy and the implementation of it on the Greek cooperative banks.* Paper presented at the 9th Annual MIBES International Conference, 423-436.

Laitinen, E. K., & Laitinen, T. (2001). Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International Review of Financial Analysis, 9*(4), 327-349.

Laitinen, T., & Kankaanpaa, M. (1999). Comparative analysis of failure prediction methods: the Finnish case. *European Accounting Review, 8*(1), 67-92.

Lama, T. (2012). Empirical evidence on the link between compliance with governance of best practice and firms' operating results. *Australasian Accounting Business & Finance Journal, 6*(5), 63-80.

Le, H. H., & Viviani, J.-L. (2018). Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. *Research in International Business and Finance, 44*, 16-25.

Lee, K. C., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems, 18*(1), 63-72.

Lee, S., & Choi, W. S. (2013). A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. *Expert Systems with Applications, 40*(8), 2941-2946.

Letts, S. (2016, 11/6/2016). Mining industry to lose 50,000 more jobs as boom comes to an end: NAB. Retrieved from http://www.abc.net.au/news/2016-06-10/mining-boom-halfway-down-the-mining-cliff/7500700

Li, J. (2012). Prediction of corporate bankruptcy from 2008 through 2011. *Journal of Accounting and Finance, 12*(1), 31-41.

Liau, Y.-S. (2017). Islamic Finance. Retrieved from https://www.bloomberg.com/quicktake/islamic-finance

Lin, S.-J., Chang, C., & Hsu, M.-F. (2013). Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction. *Knowledge-Based Systems, 39*, 214-223.

Ling, C. X., & Li, C. (1998). *Data mining for direct marketing: Problems and solutions.* Paper presented at the KDD (Vol. 98), 73-79.

Mamo, A. Q. (2011). Applicability of Altman (1968) model in predicting financial distress of commercial banks in Kenya. (PhD Dissertation).

McKee, T. E., & Lensberg, T. (2002). Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European Journal of Operational Research, 138*(2), 436-451.

Min, J. H., & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), 603-614.

Moody's. (2009). Moody's Rating System in Brief. Retrieved from https://www.moodys.com/sites/products/ProductAttachments/Moody's%20Rating%20System.pdf

MorningStar. (2015). Independent. Insightful. Trusted. Retrieved from http://corporate.morningstar.com/au/asp/subject.aspx?xmlfile=5677.xml

MorningStar. (2016). About DatAnalysis. Retrieved from http://datanalysis.morningstar.com.au/af/help

Mukkamala, S., Vieira, A., & Sung, A. (2008). *Model selection and feature ranking for financial distress classification.* Paper presented in the Proceedings of the 8th International Conference on Enterprise Information Systems (ICEIS 2006).

Murphy, C. (2018). Why Do Companies Care About Their Stock Prices? Retrieved from https://www.investopedia.com/investing/why-do-companies-care-about-their-stock-prices/

Murphy, P., & Aha, D. (1994). UCI repository of machine learning databases, University of California, Department of Information and Computer Science, Irvine, CA: US., Tech. Rep.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 36*(2), 3028-3033.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). Handbook on constructing composite indicators. OECD Statistics Working Paper.

Narula, R. (2004). R&D collaboration by SMEs: new opportunities and limitations in the face of globalisation. *Technovation, 24*(2), 153-161.

Nindita, K., & Indrawati, M. N. K. (2014). Prediction on Financial distress of Mining Companies Listed in BEI using Financial Variables and Non-Financial Variables. *European Journal of Business and Management, 6*(34).

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software, 50*(12).

Odom, M. D., & Sharda, R. (1990). *A neural network model for bankruptcy prediction.* Paper presented at the 1990 IJCNN International Joint Conference on Neural Networks, IEEE, 163-168.

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

Olson, D., & Zoubi, T. A. (2008). Using accounting ratios to distinguish between Islamic and conventional banks in the GCC region. *The International Journal of Accounting, 43*(1), 45-65.

Park, C.-S., & Han, I. (2002). A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications, 23*(3), 255-264.

Pasimeni, P. (2013). the Europe 2020 index. *Social Indicators Research, 110*(2), 613-635.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559-572.

Perez, M. (2006). Artificial neural networks and bankruptcy forecasting: a state of the art. *Neural Computing & Applications, 15*(2), 154-163.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19-50.

Phillips, C. H. (2012). S&P Capital IQ. *Journal of Business & Finance Librarianship, 17*(3), 279-286.

Pomeroy, R. S., Pollnac, R. B., Katon, B. M., & Predo, C. D. (1997). Evaluating factors contributing to the success of community-based coastal resource management: the Central Visayas Regional Project-1, Philippines. *Ocean & Coastal Management, 36*(1-3), 97-120.

Provost, F., & Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning, 42*(3), 203-231.

Psaros, J., & Seamer, M. (2002). Horwath Corporate Governance Report. Retrieved from http://www.ecgi.org/codes/documents/horwath_cg_02.pdf

R. (2019). What is R? Retrieved from https://www.r-project.org/about.html

Ravi, V., Kumar, P. R., Srinivas, E. R., & Kasabov, N. K. (2007). A Semi-Online Training Algorithm for the Radial Basis Function Neural Networks: Applications to Bankruptcy. *Advances in Banking Technology and Management: Impacts of ICT and CRM: Impacts of ICT and CRM*, 243-260.

Ravi, V., Kurniawan, H., Thai, P. N. K., & Kumar, P. R. (2008). Soft computing system for bank performance prediction. *Applied Soft Computing, 8*(1), 305-315.

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

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197-227.

Scutt, D. (2017). The mining boom's end is still dragging down the value of Australian construction. Retrieved from https://www.businessinsider.com.au/the-mining-booms-end-is-still-shrinking-the-value-of-australian-construction-2017-2

Sekhar, C. C., Indrayan, A., & Gupta, S. (1991). Development of an index of need for health resources for Indian states using factor analysis. *International Journal of Epidemiology, 20*(1), 246-250.

Shah, N. (2014). Developing financial distress prediction models using cutting edge recursive partitioning techniques: a study of Australian mining performance. *Review of Integrative Business and Economics Research, 3*(2), 103.

Sharma, N. (2013). Altman model and financial soundness of Indian banks. *International Journal of Accounting and Finance, 3*(2), 55-60.

Shrivastava, A., Kumar, K., & Kumar, N. (2018). Business Distress Prediction Using Bayesian Logistic Model for Indian Firms. *Risks, 6*(4), 113.

Shultz, K. S., Hoffman, C. C., & Reiter-Palmon, R. (2005). Using archival data for IO research: Advantages, pitfalls, sources, and examples. *The Industrial-Organizational Psychologist, 42*(3), 31.

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

Smith, M., Ren, Y., & Dong, Y. (2011). The predictive ability of "conservatism" and "governance" variables in corporate financial disclosures. *Asian Review of Accounting, 19*(2), 171-185.

Spithoven, A., Vanhaverbeke, W., & Roijakkers, N. (2013). Open innovation practices in SMEs and large enterprises. *An Entrepreneurship Journal, 41*(3), 537-562.

Springate, G. L. (1978). *Predicting the possibility of failure in a Canadian firm: A discriminant analysis.* Simon Fraser University. (PhD Dissertation).

Standard & Poor's. (2014). *Asset Growth Comparison: Islamic and Conventional Banks*. Retrieved from http://www.ft.com/fastft/files/2014/10/13-Capture(2).PNG

Sun, A., Lim, E.-P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems, 48*(1), 191-201.

Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems, 21*(1), 1-5.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of Statistics, 24*, 303-329.

Syamni, G., Majid, M. S. A., & Siregar, W. V. (2018). Bankruptcy Prediction Models and Stock Prices of the Coal Mining Industry in Indonesia. *Etikonomi: Jurnal Ekonomi, 17*, 57-68.

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

The Telegraph. (2012). How do credit ratings work? Retrieved from https://www.telegraph.co.uk/finance/financialcrisis/9081354/QandA-How-do-credit-ratings-work.html

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability, 57*, 1-436.

Tinoco, M. H., & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis, 30*, 394-419.

Tsai, C.-h. (2005). *Bayesian Inference in Binomial Logistic Regression: A Case Study of the 2002 Taipei Mayoral Election*: Chia-hung Tsai, 103-123.

Van de Vrande, V., De Jong, J. P., Vanhaverbeke, W., & De Rochemont, M. (2009). Open innovation in SMEs: Trends, motives and management challenges. *Technovation, 29*(6-7), 423-437.

Veal, A. J. (2005). *Business research methods: A managerial approach*: Pearson Education Australia/Addison Wesley.

Wahlquist, C. (2017). 'They've lost the lot': how the Australian mining boom blew up in property owners' faces. Retrieved from https://www.theguardian.com/australia-news/2017/may/12/theyve-lost-the-lot-how-the-australian-mining-boom-blew-up-in-property-owners-faces

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & Operations Research, 32*(10), 2543-2559.

Whiting, D. G., Hansen, J. V., McDonald, J. B., Albrecht, C., & Albrecht, W. S. (2012). Machine learning methods for detecting patterns of management fraud. *Computational Intelligence, 28*(4), 505-527.

Wilchins, D., & Stempel, J. (2008). Citigroup gets massive government bailout. Retrieved from https://www.reuters.com/article/us-citigroup/citigroup-gets-massive-government-bailout-idUSTRE4AJ45G20081125

Winakor, A., & Smith, R. (1935). Changes in the Financial Structure of Unsuccessful Industrial Corporations. *Bulletin, 51*, 1-41.

Woods, K. S., Doss, C. C., Bowyer, K. W., Solka, J. L., Priebe, C. E., & Kegelmeyer Jr, W. P. (1993). Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence, 7*(06), 1417-1436.

Yu, Q., Miche, Y., Séverin, E., & Lendasse, A. (2014). Bankruptcy prediction using extreme learning machine and financial expertise. *Neurocomputing, 128*, 296-302.

Zhang, H., & Singer, B. (2010). *Recursive partitioning and applications*: Springer Science & Business Media.

Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems, 41*, 16-25.

Zmijewski, M. (1983). Predicting corporate bankruptcy: An empirical comparison of the extant financial distress models. *Document de travail. State University of New York at Buffalo*.

Appendices

Appendix 1: Chapter 5’s R Report

```
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(pROC)

## Type 'citation("pROC")' for a citation.

## 
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
## 
##     cov, smooth, var

library(DMwR)

## Loading required package: grid

setwd("C:/Users/Khaled/Downloads")
mydata <- read.csv("CRAll.csv",header=TRUE)
summary(mydata)

##                          Company.Name      Status            ROE
##  Rift Valley Resources Limited:   6   Min.   :0.0000   Min.   :-3
95.8600
##  3D Resources Limited         :   5   1st Qu.:1.0000   1st Qu.:
-0.5300
##  A1 Consolidated Gold Limited :   5   Median :1.0000   Median :
-0.1700
##  ABM Resources NL             :   5   Mean   :0.8996   Mean   :
-0.8621
##  Accent Resources N.L.        :   5   3rd Qu.:1.0000   3rd Qu.:
-0.0500
##  Activex Limited              :   5   Max.   :1.0000   Max.   : 2
49.6800
##  (Other)                      :3344
##       ROA                 ROIC             Asset.Turnover
##  Min.   :-2127.290   Min.   :-8826.290   Min.   : -0.5400
##  1st Qu.:   -0.480   1st Qu.:   -1.620   1st Qu.:  0.0000
##  Median :   -0.170   Median :   -0.310   Median :  0.0000
##  Mean   :   -1.654   Mean   :   -7.876   Mean   :  0.3108
##  3rd Qu.:   -0.060   3rd Qu.:   -0.090   3rd Qu.:  0.0000
##  Max.   :   49.140   Max.   : 1118.220   Max.   :673.9000
```

```
##
##   PPE.Turnover      Depreciation.PP.E Working.Cap.Turnover
##  Min.   :  -0.460   Min.   :-1.8000   Min.   :-1239.890
##  1st Qu.:   0.000   1st Qu.: 0.0300   1st Qu.:    0.000
##  Median :   0.000   Median : 0.0900   Median :    0.000
##  Mean   :   2.392   Mean   : 0.1342   Mean   :   -1.319
##  3rd Qu.:   0.590   3rd Qu.: 0.1600   3rd Qu.:    0.000
##  Max.   :1569.090   Max.   :11.2100   Max.   :  615.300
##
##  Gross.Gearing..D.E. Financial.Leverage Current.Ratio
##  Min.   :-60.1900    Min.   :-155.580   Min.   :   0.00
##  1st Qu.: 0.0000    1st Qu.:   1.030   1st Qu.:   1.26
##  Median : 0.0000    Median :   1.080   Median :   3.59
##  Mean   : 0.1125    Mean   :   1.313   Mean   :  14.23
##  3rd Qu.: 0.0300    3rd Qu.:   1.320   3rd Qu.:  10.21
##  Max.   : 74.4600    Max.   :  78.410   Max.   :7073.96
##
##   Quick.Ratio      Gross.Debt.CF       Cash.per.Share....
##  Min.   :   0.00   Min.   :-353.0900   Min.   : 0.0000
##  1st Qu.:   0.99   1st Qu.:  -0.0350   1st Qu.: 0.0000
##  Median :   3.46   Median :   0.0000   Median : 0.0100
##  Mean   :  14.13   Mean   :  -0.0029   Mean   : 0.3667
##  3rd Qu.:  10.16   3rd Qu.:   0.0000   3rd Qu.: 0.0400
##  Max.   :7073.96   Max.   : 512.5700   Max.   :616.3700
##
##  Invested.Capital.Turnover   Net.Gearing       NTA.per.Share....
##  Min.   :-0.2900           Min.   :-57.3200   Min.   :  -4.30
##  1st Qu.: 0.0000           1st Qu.: -0.5000   1st Qu.:   0.01
##  Median : 0.0000           Median : -0.1700   Median :   0.06
##  Mean   : 0.2704           Mean   : -0.2406   Mean   :   1.25
##  3rd Qu.: 0.1800           3rd Qu.: -0.0200   3rd Qu.:   0.15
##  Max.   :36.3500           Max.   : 61.9000   Max.   :1388.35
##
##  BV.per.Share....    Sales.per.Share....      PER
##  Min.   :  -4.300   Min.   :  0.0000   Min.   :-721.740
##  1st Qu.:   0.020   1st Qu.:  0.0000   1st Qu.:  -8.035
##  Median :   0.060   Median :  0.0000   Median :  -3.000
##  Mean   :   1.289   Mean   :  0.4184   Mean   :  -5.788
##  3rd Qu.:   0.160   3rd Qu.:  0.0000   3rd Qu.:  -0.850
##  Max.   :1396.100   Max.   :108.7200   Max.   : 725.000
##
```

```
str(mydata)
```

```
## 'data.frame':    3375 obs. of   21 variables:
##  $ Company.Name             : Factor w/ 747 levels "3D Resources
Limited",..: 19 19 19 19 19 58 58 58 58 58 ...
##  $ Status                   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ROE                      : num  0.06 0.02 0.02 0.09 0.04 -0.44
-0.56 -0.54 -0.46 -1.8 ...
```

```
##  $ ROA                      : num  0.03 0.02 0.02 0.04 0.02 -0.42
-0.52 -0.48 -0.43 -1.53 ...
##  $ ROIC                     : num  0.08 0.04 0.04 0.1 0.07 -4.4 -
3.66 -1.12 -0.74 -19.8 ...
##  $ Asset.Turnover           : num  0.62 0.328 0.328 -0.54 -0.48 .
..
##  $ PPE.Turnover             : num  1.29 2.31 2.31 -0.46 -0.43 ...
##  $ Depreciation.PP.E        : num  0 0.136 0.136 -1.8 -1.53 ...
##  $ Working.Cap.Turnover     : num  50.71 -1.17 -1.17 -0.27 -0.25
...
##  $ Gross.Gearing..D.E.      : num  0.55 0.117 0.117 -0.13 -0.11 .
..
##  $ Financial.Leverage       : num  2.33 1.35 1.35 -4.17 -2.68 ...
##  $ Current.Ratio            : num  1.28 1.3 1.14 1.49 1.53 ...
##  $ Quick.Ratio              : num  0.8 0.82 0.7 0.94 0.87 ...
##  $ Gross.Debt.CF            : num  7.24 4.27 3.75 3.1 3.46 0 0 0
0 0 ...
##  $ Cash.per.Share....       : num  1.79 1.68 1.5 1.88 2 0.05 0.03
0.01 0.01 0.01 ...
##  $ Invested.Capital.Turnover: num  1.41 1.47 1.55 1.64 1.7 0.09 0
0 0 0 ...
##  $ Net.Gearing              : num  0.43 0.42 0.51 0.47 0.51 -0.89
-0.84 -0.5 -0.34 -0.91 ...
##  $ NTA.per.Share....        : num  7.49 6.88 7.07 6.34 5.73 0.05
0.04 0.03 0.02 0.01 ...
##  $ BV.per.Share....         : num  12.8 11.9 11.1 12.3 12.6 ...
##  $ Sales.per.Share....      : num  23.1 21.4 24.1 25.1 24.5 ...
##  $ PER                      : num  13.3 31.9 44.8 16.6 25 ...
```

```
mydata <- mydata[,-1] # remove company name
print(table(mydata$Status))
```

```
## 
##    0    1 
##  339 3036
```

```
print(prop.table(table(mydata$Status)))
```

```
## 
##         0         1 
## 0.1004444 0.8995556
```

```
set.seed(1234)
splitIndex <- createDataPartition(mydata$Status, p = .50, list = FAL
SE, times = 1)
trainSplit <- mydata[ splitIndex,]
testSplit <- mydata[-splitIndex,]
ctrl <- trainControl(method = "cv", number = 5)

train <- trainSplit
```

```
tbmodel1 <- train(factor(Status) ~ ., data = trainSplit, method = "t
reebag", trControl = ctrl)
```

```
## Loading required package: ipred
```

```
## Loading required package: plyr
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:DMwR':
##
##     join
```

```
## Loading required package: e1071
```

```
predictors <- names(trainSplit)[names(trainSplit) != 'Status']
pred1 <- predict(tbmodel1$finalModel, testSplit[,predictors])

mean(testSplit$Status == as.numeric(pred1)-1)
```

```
## [1] 0.902786
```

```
auc <- roc(testSplit$Status, as.numeric(pred1)-1)
print(auc)
```

```
##
## Call:
## roc.default(response = testSplit$Status, predictor = as.numeric(p
red1) -     1)
##
## Data: as.numeric(pred1) - 1 in 163 controls (testSplit$Status 0)
< 1524 cases (testSplit$Status 1).
## Area under the curve: 0.5736
```

```
plot.roc(testSplit$Status,as.numeric(pred1)-1)
```

```
trainsplit1 <- train
trainsplit1$Status <- as.factor(trainsplit1$Status)
smotedata <- SMOTE(Status ~ ., trainsplit1, perc.over = 100, perc.un
der=200)
smotedata$Status <- as.numeric(smotedata$Status) - 1
print(prop.table(table(smotedata$Status)))
```

```
##
##   0   1
## 0.5 0.5
```

```
tbmodel2 <- train(factor(Status) ~ ., data = smotedata, method = "tr
eebag", trControl = ctrl)

pred2 <- predict(tbmodel2$finalModel, testSplit[,predictors])
```

```
auc <- roc(testSplit$Status, as.numeric(pred2)-1)

print(auc)

##
## Call:
## roc.default(response = testSplit$Status, predictor = as.numeric(p
red2) -     1)
##
## Data: as.numeric(pred2) - 1 in 163 controls (testSplit$Status 0)
< 1524 cases (testSplit$Status 1).
## Area under the curve: 0.6388

mean(testSplit$Status == as.numeric(pred2)-1)

## [1] 0.6988737

plot.roc(testSplit$Status,as.numeric(pred2)-1)
```

Appendix 2: Chapter 8’s SMEs’ R Report

```
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(DMwR)

## Loading required package: grid

setwd("C:/Users/Khaled/Downloads")
mydata <- read.csv("SMEs3.csv",header=TRUE)
summary(mydata)

##     Excel.Company.ID Listed.Delisted       ROA
##  IQ46235974 :  23    Min.   :0.0000   Min.   :-25.00445
##  IQ108954538:   2    1st Qu.:1.0000   1st Qu.: -0.36585
##  IQ327168517:   2    Median :1.0000   Median : -0.13134
##  IQ4481685  :   2    Mean   :0.9671   Mean   : -0.34989
##  IQ100315307:   1    3rd Qu.:1.0000   3rd Qu.: -0.03917
##  IQ100718430:   1    Max.   :1.0000   Max.   :  0.46402
##  (Other)    :1244
##       ROC                  ROE                Gross.Margin
##  Min.   :-67.73667   Min.   :-127.15055   Min.   :-2.9293
##  1st Qu.: -0.40436   1st Qu.:   -0.80402   1st Qu.: 0.0000
##  Median : -0.13444   Median :   -0.24126   Median : 0.3817
##  Mean   : -0.56746   Mean   :   -1.03846   Mean   : 0.3902
##  3rd Qu.: -0.04563   3rd Qu.:   -0.07771   3rd Qu.: 1.0000
##  Max.   :  0.55729   Max.   :    4.71552   Max.   : 1.0000
##
##       ROCE                SGA.Margin            TD.TE
##  Min.   :-305.58210   Min.   :      0.0   Min.   : 0.00000
##  1st Qu.:   -0.79900   1st Qu.:      0.0   1st Qu.: 0.00000
##  Median :   -0.29117   Median :      0.7   Median : 0.00000
##  Mean   :   -1.39875   Mean   :    798.5   Mean   : 0.32646
##  3rd Qu.:   -0.05297   3rd Qu.:      8.2   3rd Qu.: 0.02255
##  Max.   :    6.93899   Max.   :349750.3   Max.   :51.35897
##
```

```
##      TD.TC          Asset.Turnover     Current.Ratio
##  Min.   : 0.00000   Min.   : 0.00000   Min.   :   0.000
##  1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.:   1.094
##  Median : 0.00000   Median : 0.01319   Median :   3.197
##  Mean   : 0.23639   Mean   : 0.22897   Mean   :  11.199
##  3rd Qu.: 0.06177   3rd Qu.: 0.16807   3rd Qu.:   8.729
##  Max.   :34.60248   Max.   :25.09792   Max.   :1099.079
##
##   Quick.Ratio           CFO.CL            Altman.Z
##  Min.   :  -0.0011   Min.   :-99.2811   Min.   :-6550.182
##  1st Qu.:   0.8475   1st Qu.: -4.1902   1st Qu.:   -0.076
##  Median :   2.9002   Median : -1.4553   Median :    0.000
##  Mean   :  10.6448   Mean   : -2.9311   Mean   :  -10.131
##  3rd Qu.:   8.2271   3rd Qu.: -0.2749   3rd Qu.:    7.733
##  Max.   :1098.9893   Max.   :152.6471   Max.   :  288.969
##
##   ln.Employees         ln.TR              ln.TA            CA.TA
##  Min.   :-1.8891   Min.   :-11.5129   Min.   :-4.699   Min.   :0.
0000
##  1st Qu.: 0.0000   1st Qu.: -1.9384   1st Qu.: 1.224   1st Qu.:0.
1689
##  Median : 0.0000   Median :  0.0000   Median : 2.180   Median :0.
4417
##  Mean   : 0.2141   Mean   : -0.5975   Mean   : 2.138   Mean   :0.
4944
##  3rd Qu.: 0.0000   3rd Qu.:  0.7021   3rd Qu.: 3.059   3rd Qu.:0.
8485
##  Max.   : 6.9078   Max.   :  5.3863   Max.   : 7.759   Max.   :1.
0000
##
##     Cash.TA           Cash.CL              NWC.TA
##  Min.   :0.00000   Min.   :   0.0000   Min.   :-77.42895
##  1st Qu.:0.09101   1st Qu.:   0.5085   1st Qu.: -0.09290
##  Median :0.26765   Median :   2.3096   Median : -0.02111
##  Mean   :0.37214   Mean   :  10.0496   Mean   : -0.28710
##  3rd Qu.:0.63867   3rd Qu.:   7.6186   3rd Qu.:  0.00487
##  Max.   :1.00000   Max.   :1090.9000   Max.   :  0.99875
##
##      NI.TA              TL.TA             EBITDA.TA
##  Min.   :-107.4446   Min.   :  0.00000   Min.   :-107.4037
##  1st Qu.:  -0.7298   1st Qu.:  0.05024   1st Qu.:  -0.4183
##  Median :  -0.2262   Median :  0.15134   Median :  -0.1242
##  Mean   :  -0.9650   Mean   :  1.05712   Mean   :   0.0484
##  3rd Qu.:  -0.0511   3rd Qu.:  0.41213   3rd Qu.:  -0.0107
##  Max.   :  23.5216   Max.   :127.91209   Max.   : 708.7017
##
##      RE.TA
##  Min.   :-8178.618
##  1st Qu.:   -6.639
##  Median :   -2.143
##  Mean   :  -26.101
##  3rd Qu.:   -0.646
##  Max.   :    0.953
##
```

```r
mydata <- mydata[,-1] # remove company name
print(table(mydata$Listed.Delisted))
```

```
##
##    0    1
##   42 1233
```

```r
str(mydata)
```

```
## 'data.frame':    1275 obs. of  25 variables:
##  $ Listed.Delisted: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ROA            : num  -0.151 -0.211 -0.461 -0.131 -0.151 ...
##  $ ROC            : num  -0.159 -0.327 -0.492 -0.134 -0.158 ...
##  $ ROE            : num  -0.254 -0.526 -0.84 -0.241 -0.284 ...
##  $ Gross.Margin   : num  1 1 1 1 1 1 1 1 0 1 ...
##  $ ROCE           : num  -0.254 -0.526 -0.84 0 -0.284 ...
##  $ SGA.Margin     : num  84261 112264 40101 5559 9858 ...
##  $ TD.TE          : num  0 0.0102 0 0 0 ...
##  $ TD.TC          : num  0 0.0101 0 0 0 ...
##  $ Asset.Turnover : num  0 0 0.00001 0 0.00002 0.00043 0.00097 0.
00001 0.00003 0.00001 ...
##  $ Current.Ratio  : num  20.3898 0.0476 13.5801 12.238 7.8093 ...
##  $ Quick.Ratio    : num  20.2753 0.0438 13.3504 12.238 7.7131 ...
##  $ CFO.CL         : num  -3.782 -0.364 -10.94 -3.786 -6.238 ...
##  $ Altman.Z       : num  21.356 -31.175 11.499 -0.357 33.114 ...
##  $ ln.Employees   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ln.TR          : num  -11.5 -11.5 -10.8 -10.8 -10.4 ...
##  $ ln.TA          : num  1.829 0.587 1.383 -0.811 0.445 ...
##  $ CA.TA          : num  0.7613 0.0201 0.2966 0.7576 0.2912 ...
##  $ Cash.TA        : num  0.7399 0.0123 0.2852 0.5373 0.2835 ...
##  $ Cash.CL        : num  19.8183 0.0293 13.0558 8.6773 7.6017 ...
##  $ NWC.TA         : num  -0.016 -0.4081 -0.0104 0.1584 -0.0296 ..
.
##  $ NI.TA          : num  -0.137 -0.317 -0.411 -0.232 -0.24 ...
##  $ TL.TA          : num  0.0374 0.4217 0.0218 0.0619 0.0373 ...
##  $ EBITDA.TA      : num  -0.135 -0.312 0 -0.263 -0.24 ...
##  $ RE.TA          : num  -0.422 -23.66 -15.237 -0.232 -6.038 ...
```

```r
print(prop.table(table(mydata$Listed.Delisted)))
```

```
##
##          0          1
## 0.03294118 0.96705882
```

```
set.seed(1234)
splitIndex <- createDataPartition(mydata$Listed.Delisted, p = .50, l
ist = FALSE, times = 1)
trainSplit <- mydata[ splitIndex,]
testSplit <- mydata[-splitIndex,]
ctrl <- trainControl(method = "cv", number = 10)
write.csv(trainSplit, "C:/Users/Khaled/Downloads/SMEsTrain.csv",row.
names = FALSE)
write.csv(testSplit, "C:/Users/Khaled/Downloads/SMEsTest.csv",row.na
mes = FALSE)
print(prop.table(table(trainSplit$Listed.Delisted)))

## 
##          0          1 
## 0.03918495 0.96081505

print(prop.table(table(testSplit$Listed.Delisted)))

## 
##         0         1 
## 0.0266876 0.9733124

train <- trainSplit

tbmodel1 <- train(factor(Listed.Delisted) ~ ., data = trainSplit, me
thod = "treebag", trControl = ctrl)

## Loading required package: ipred

## Loading required package: plyr

## 
## Attaching package: 'plyr'

## The following object is masked from 'package:DMwR':
## 
##     join

## Loading required package: e1071

predictors <- names(trainSplit)[names(trainSplit) != 'Listed.Deliste
d']
pred1 <- predict(tbmodel1$finalModel, testSplit[,predictors])

mean(testSplit$Listed.Delisted == as.numeric(pred1)-1)

## [1] 0.978022

auc <- roc(testSplit$Listed.Delisted, as.numeric(pred1)-1)
print(auc)

## 
## Call:
```

```
## roc.default(response = testSplit$Listed.Delisted, predictor = as.
numeric(pred1) -     1)
##
## Data: as.numeric(pred1) - 1 in 17 controls (testSplit$Listed.Deli
sted 0) < 620 cases (testSplit$Listed.Delisted 1).
## Area under the curve: 0.75

plot.roc(testSplit$Listed.Delisted,as.numeric(pred1)-1)
```

```
trainsplit1 <- train
trainsplit1$Listed.Delisted <- as.factor(trainsplit1$Listed.Delisted
)
smotedata <- SMOTE(Listed.Delisted ~ ., trainsplit1, perc.over = 100
, perc.under=200)
smotedata$Listed.Delisted <- as.numeric(smotedata$Listed.Delisted) -
1
print(prop.table(table(smotedata$Listed.Delisted)))

##
##   0   1
## 0.5 0.5

tbmodel2 <- train(factor(Listed.Delisted) ~ ., data = smotedata, met
hod = "treebag", trControl = ctrl)

pred2 <- predict(tbmodel2$finalModel, testSplit[,predictors])

auc <- roc(testSplit$Listed.Delisted, as.numeric(pred2)-1)

print(auc)

##
## Call:
## roc.default(response = testSplit$Listed.Delisted, predictor = as.
numeric(pred2) -     1)
##
## Data: as.numeric(pred2) - 1 in 17 controls (testSplit$Listed.Deli
sted 0) < 620 cases (testSplit$Listed.Delisted 1).
## Area under the curve: 0.82

mean(testSplit$Listed.Delisted == as.numeric(pred2)-1)

## [1] 0.7095761

plot.roc(testSplit$Listed.Delisted,as.numeric(pred2)-1)
```

Appendix 3: Chapter 8's LCs' R Report

```
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(pROC)

## Type 'citation("pROC")' for a citation.

## 
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
## 
##     cov, smooth, var

library(DMwR)

## Loading required package: grid

setwd("C:/Users/Khaled/Downloads")
mydata <- read.csv("Large3.csv",header=TRUE)
summary(mydata)

##     Excel.Company.ID Listed.Delisted       ROA           
##  IQ126981528:  2     Min.   :0.0000   Min.   :-0.663018  
##  IQ7652776  :  2     1st Qu.:1.0000   1st Qu.: 0.005697  
##  IQ875280   :  2     Median :1.0000   Median : 0.040017  
##  IQ100656194:  1     Mean   :0.8904   Mean   : 0.038488  
##  IQ10525123 :  1     3rd Qu.:1.0000   3rd Qu.: 0.073088  
##  IQ105597   :  1     Max.   :1.0000   Max.   : 0.478332  
##  (Other)    :283                                         
##       ROC                ROE             Gross.Margin   
##  Min.   :-0.91922   Min.   :-19.43217   Min.   :-0.3622  
##  1st Qu.: 0.00000   1st Qu.:  0.00000   1st Qu.: 0.1692  
##  Median : 0.05227   Median :  0.09272   Median : 0.3424  
##  Mean   : 0.05281   Mean   :  0.45530   Mean   : 0.3765  
##  3rd Qu.: 0.10272   3rd Qu.:  0.15909   3rd Qu.: 0.5617  
##  Max.   : 0.60028   Max.   :136.06926   Max.   : 1.0000  
##                                                          
##       ROCE             SGA.Margin           TD.TE         
##  Min.   :-19.39802   Min.   :0.00000   Min.   : 0.00000  
##  1st Qu.:  0.00000   1st Qu.:0.03871   1st Qu.: 0.02089  
##  Median :  0.09186   Median :0.14606   Median : 0.28873  
##  Mean   :  0.97543   Mean   :0.19761   Mean   : 0.61942  
##  3rd Qu.:  0.15610   3rd Qu.:0.30652   3rd Qu.: 0.58236  
##  Max.   :288.00000   Max.   :1.04301   Max.   :27.09083  
## 
```

```
##      TD.TC          Asset.Turnover   Current.Ratio      Quick.Rat
io
##  Min.   : 0.00000   Min.   :0.0000   Min.   : 0.0000   Min.   : 0
.0000
##  1st Qu.: 0.02382   1st Qu.:0.3316   1st Qu.: 0.8871   1st Qu.: 0
.5255
##  Median : 0.22777   Median :0.6712   Median : 1.3855   Median : 0
.9044
##  Mean   : 0.48984   Mean   :0.9003   Mean   : 2.0252   Mean   : 1
.5391
##  3rd Qu.: 0.38695   3rd Qu.:1.1808   3rd Qu.: 2.0544   3rd Qu.: 1
.4378
##  Max.   :27.20878   Max.   :4.7202   Max.   :82.7306   Max.   :70
.6010
##
##      CFO.CL            Altman.Z         ln.Employees        ln.TR
##  Min.   :-1.64592   Min.   :-20.5550   Min.   : 0.000   Min.   :-
1.027
##  1st Qu.: 0.02895   1st Qu.:  0.6132   1st Qu.: 5.407   1st Qu.:
4.868
##  Median : 0.36223   Median :  2.4212   Median : 6.702   Median :
5.949
##  Mean   : 0.52248   Mean   :  3.4673   Mean   : 6.589   Mean   :
6.186
##  3rd Qu.: 0.76652   3rd Qu.:  3.8528   3rd Qu.: 8.213   3rd Qu.:
7.469
##  Max.   : 9.51479   Max.   :101.2057   Max.   :12.315   Max.   :1
6.353
##
##      ln.TA            CA.TA           Cash.TA            Cash.CL
##  Min.   : 2.609   Min.   :0.0000   Min.   :0.00000   Min.   : 0.0
000
##  1st Qu.: 5.111   1st Qu.:0.1765   1st Qu.:0.02766   1st Qu.: 0.1
135
##  Median : 6.330   Median :0.3136   Median :0.06940   Median : 0.2
991
##  Mean   : 6.638   Mean   :0.3663   Mean   :0.12966   Mean   : 0.7
768
##  3rd Qu.: 8.121   3rd Qu.:0.5348   3rd Qu.:0.17753   3rd Qu.: 0.8
103
##  Max.   :17.217   Max.   :0.9920   Max.   :0.90669   Max.   :18.4
638
##
##      NWC.TA             NI.TA              TL.TA          EBITDA.
TA
##  Min.   :-0.58114   Min.   :-2.36858   Min.   :0.0000   Min.   :-
0.80035
##  1st Qu.:-0.03112   1st Qu.: 0.00000   1st Qu.:0.3340   1st Qu.:
0.02567
```

```
##  Median : 0.00000   Median : 0.03625   Median :0.4701   Median :
0.09657
##  Mean   : 0.03827   Mean   : 0.00132   Mean   :0.4997   Mean   :
0.09696
##  3rd Qu.: 0.11001   3rd Qu.: 0.08123   3rd Qu.:0.6139   3rd Qu.:
0.15325
##  Max.   : 0.82794   Max.   : 0.52555   Max.   :3.6174   Max.   :
0.80271
##
##      RE.TA
##  Min.   :-13.40457
##  1st Qu.: -0.09439
##  Median :  0.03753
##  Mean   : -0.16925
##  3rd Qu.:  0.18263
##  Max.   :  0.66780
##
```

```
mydata <- mydata[,-1] # remove company name
print(table(mydata$Listed.Delisted))
```

```
##
##   0   1
##  32 260
```

```
str(mydata)
```

```
## 'data.frame':    292 obs. of  25 variables:
##  $ Listed.Delisted: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ROA            : num  0.0602 0.0711 0.0475 0.0133 0.0598 ...
##  $ ROC            : num  0.1109 0.0851 0.0747 0.0144 0.0683 ...
##  $ ROE            : num  -0.186 0.148 0.187 0.126 0.137 ...
##  $ Gross.Margin   : num  0.471 0.196 0.323 0.576 0.288 ...
##  $ ROCE           : num  -0.182 0.148 0.187 0.119 0.137 ...
##  $ SGA.Margin     : num  0.1801 0.0564 0.0689 0.0581 0.0432 ...
##  $ TD.TE          : num  0.277 0.344 1.266 0.267 0.519 ...
##  $ TD.TC          : num  0.217 0.256 0.559 0.211 0.342 ...
##  $ Asset.Turnover : num  0.576 0.813 0.709 0.197 0.834 ...
##  $ Current.Ratio  : num  0.32 2.318 0.785 4.892 1.801 ...
##  $ Quick.Ratio    : num  0.295 1.426 0.696 3.653 0.744 ...
##  $ CFO.CL         : num  0.401 1.095 0.376 1.817 0.509 ...
##  $ Altman.Z       : num  2.244 4.584 1.128 0.491 2.231 ...
##  $ ln.Employees   : num  5.49 7.34 9.3 6.37 6.11 ...
##  $ ln.TR          : num  3.96 7.35 8.54 5.34 5.31 ...
##  $ ln.TA          : num  4.42 7.61 8.88 7.13 5.52 ...
##  $ CA.TA          : num  0.141 0.236 0.263 0.234 0.307 ...
##  $ Cash.TA        : num  0.0396 0.0286 0.1909 0.1618 0.0138 ...
##  $ Cash.CL        : num  0.0899 0.2812 0.5692 3.3766 0.0811 ...
##  $ NWC.TA         : num  -0.339 0.1057 -0.2189 0.0247 0.1799 ...
##  $ NI.TA          : num  -0.0853 0.0904 0.0533 0.065 0.0742 ...
##  $ TL.TA          : num  0.587 0.38 0.723 0.276 0.423 ...
##  $ EBITDA.TA      : num  0.112 0.1485 0.1407 0.0676 0.1985 ...
##  $ RE.TA          : num  -0.0593 0.2537 -0.0342 -0.6092 0.3025 ..
.
```

```
print(prop.table(table(mydata$Listed.Delisted)))
```

```
## 
##        0        1 
## 0.109589 0.890411
```

```
set.seed(1234)
splitIndex <- createDataPartition(mydata$Listed.Delisted, p = .50, l
ist = FALSE, times = 1)
trainSplit <- mydata[ splitIndex,]
testSplit <- mydata[-splitIndex,]
ctrl <- trainControl(method = "cv", number = 10)
write.csv(trainSplit, "C:/Users/Khaled/Downloads/LargeTrain.csv",row
.names = FALSE)
write.csv(testSplit, "C:/Users/Khaled/Downloads/LargeTest.csv",row.n
ames = FALSE)
print(prop.table(table(trainSplit$Listed.Delisted)))
```

```
## 
##        0        1 
## 0.109589 0.890411
```

```
print(prop.table(table(testSplit$Listed.Delisted)))
```

```
## 
##        0        1 
## 0.109589 0.890411
```

```
train <- trainSplit
```

```
tbmodel1 <- train(factor(Listed.Delisted) ~ ., data = trainSplit, me
thod = "treebag", trControl = ctrl)
```

```
## Loading required package: ipred
```

```
## Loading required package: plyr
```

```
## 
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:DMwR':
## 
##     join
```

```
## Loading required package: e1071
```

```
predictors <- names(trainSplit)[names(trainSplit) != 'Listed.Deliste
d']
```

```
pred1 <- predict(tbmodel1$finalModel, testSplit[,predictors])

mean(testSplit$Listed.Delisted == as.numeric(pred1)-1)

## [1] 0.9178082

auc <- roc(testSplit$Listed.Delisted, as.numeric(pred1)-1)
print(auc)

##
## Call:
## roc.default(response = testSplit$Listed.Delisted, predictor = as.
numeric(pred1) -     1)
##
## Data: as.numeric(pred1) - 1 in 16 controls (testSplit$Listed.Deli
sted 0) < 130 cases (testSplit$Listed.Delisted 1).
## Area under the curve: 0.89

plot.roc(testSplit$Listed.Delisted,as.numeric(pred1)-1)
```

```
trainsplit1 <- train
trainsplit1$Listed.Delisted <- as.factor(trainsplit1$Listed.Delisted
)
smotedata <- SMOTE(Listed.Delisted ~ ., trainsplit1, perc.over = 100
, perc.under=200)
smotedata$Listed.Delisted <- as.numeric(smotedata$Listed.Delisted) -
1
print(prop.table(table(smotedata$Listed.Delisted)))

##
##   0   1
## 0.5 0.5

tbmodel2 <- train(factor(Listed.Delisted) ~ ., data = smotedata, met
hod = "treebag", trControl = ctrl)

pred2 <- predict(tbmodel2$finalModel, testSplit[,predictors])

auc <- roc(testSplit$Listed.Delisted, as.numeric(pred2)-1)

print(auc)

##
## Call:
## roc.default(response = testSplit$Listed.Delisted, predictor = as.
numeric(pred2) -     1)
##
## Data: as.numeric(pred2) - 1 in 16 controls (testSplit$Listed.Deli
sted 0) < 130 cases (testSplit$Listed.Delisted 1).
## Area under the curve: 0.89
```

```
mean(testSplit$Listed.Delisted == as.numeric(pred2)-1)

## [1] 0.9041096

plot.roc(testSplit$Listed.Delisted,as.numeric(pred2)-1)
```

```
cbind(Actual=testSplit$Listed.Delisted,Model1=(as.numeric(pred1)-1),
model2=(as.numeric(pred2)-1))
```