DEPARTMENT OF NAVAL ARCHITECTURE, OCEAN & MARINE ENGINEERING

University of Strathclyde Engineering

Financial Risk Management in Shipping Investment

A Machine Learning Approach

By

Mark Clintworth, Luxembourg 2020

Presented for the Degree of Doctor of Philosophy

Supervised by:

Dr. Evangelos Boulougouris

AUTHOR STATEMENT

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Mark Clintworth

Date: May 2020

Abstract

There has been a plethora of research into company credit risk and financial default prediction from both academics and financial professionals alike. However, only a limited volume of the literature has focused on international shipping company financial distress prediction, with previous research concentrating largely on classic linear based modelling techniques. The gaps, identified in this research, demonstrate the need for increased effort to address the inherent non-linear nature of shipping operations, as well as the noisy and incomplete composition of shipping company financial statement data.

Furthermore, the gaps illustrate the need for a workable definition of financial distress, which to date has too often been classed only by the ultimate state of bankruptcy/insolvency. This definition prohibits the practical application of methodologies which should be aimed at the timely identification of financial distress, thereby allowing for remedial measures to be implemented to avoid ultimate financial collapse.

This research contributes to the field by addressing these gaps through i) the creation of a machine learning based financial distress forecasting methodology and ii) utilising this as the foundation for the development of a software toolkit for financial distress prediction. This toolkit enables the practical application of the financial risk principles, embedded within the methodology, to be readily integrated into an enterprise/corporate risk management system. The methodology and software were tested through the application of a bulk shipping company case study utilising 5000 bulk shipping company-year accounting observations for the period 2000-2018, in combination with market and macroeconomic data.

The results demonstrate that the methodology improves the capture of distress correlations, that traditional financial distress models have struggled to achieve. The methodology’s capacity to adequately treat the problem of missing data in company financial statements was also validated.

Finally, the results also highlight the successful application of the software toolkit for the development of a multi-model, real time system which can enhance the financial monitoring of shipping companies by acting as a practical “early warning system” for financial distress.

Acknowledgements

Firstly, I would like to express my gratitude and appreciation to Dr. Evangelos Boulougouris for his encouragement, support and guidance throughout the long road to completion of this PhD. I extend this gratitude also to my management and academic supervisor at the European Investment Bank. I could never have achieved this goal without your unwavering support during my time at the EIB.

I would also like to add a special thanks to the support staff of the Naval Architecture & Marine Engineering dept., and the University library services for their invaluable support. The professionalism and kindness in delivering your service is a credit to Strathclyde University.

To Professor L.M. Delves [†] (University of Liverpool) whose encouragement and support as an undergraduate was invaluable. Mike you remain a great inspiration.

To my friends and EIB colleagues, thanks for being there, looking on patiently and keeping your scepticism and doubts to yourselves. The prime (culpable) suspects, in no order: Christian (Berlin), Chris (Spurs), Dr. Darren O'Neill, Dr. Dimitris Konovessis, Inge Vermeersch, José-Luis Alfaro, Cliona Hurley, Don Shiel, John Lynch, Marjolein Sand, Ma's, BitStop, Sotiris Basiakos, Keith Wallis, Tim Ward & Lloyd Jones. A big "Cheers!" to you all.

Very many thanks also my family; your support never diminishes. Thanks especially to my son Edward and brother Damian, you are simply "always there".

Last, but closest to my heart, to my wife Alex, your love, help, companionship and unfathomable patience kept me going and provided me with the strength to get over the finish line.

To Josh, we love you "Freddy"!

Contents

Table of Figures

List of Tables

Abbreviations

AdaBoost	Adaptive boosting
ANN	Artificial neural networks
AUC	Area under the curve
BDI	Baltic dry index
BEPS	Base erosion and profit shifting
BF	Basis function
BIS	Bank of international settlement
BvD	Bureau van Dijk
CART	Classification and regression trees
CI	Confidence intervals
Clarkson´s	Clarkson´s Shipping Intelligence Network
COMEX	New York commodities exchange
CSR	Corporate, social and responsibility
DD	Distance to default
DJUSST	Dow Jones US Iron and Steel
DWT	Dead weight tonnage
EBITDA	Earnings before interest, taxes, depreciation and amortisation
EBIT	Earnings before interest, taxes
ESG	Environmental, social and governance
FD	Financial distress
FOC	Flag of convenience
GAM	Generalised additive model
GBM	Gradient boosting method
GLM	Generalised linear model
IASB	International accounting standards board
IFRS	International financial reporting standards
IMF	International monetary fund
k-NN	K nearest neighbors algorithm
LDA	Linear discriminant analysis
LR	Logistic regression
MAR	Missing at random
MCMC	Markov Chain Monte Carlo
MARS	Multivariate adaptive regression splines
MI	Multivariate imputation
ML	Machine learning
OECD	Organisation for economic co-operation and development
OLS	Ordinary least squares
P & L accounts	Profit and loss accounts
PMM	Predictive mean matching
RBF	Radial basis function
RF	Random forest
ROA	Return on assets
ROC	Receiver operating characteristics
ROE	Return on equity
SFDP	Shipping financial distress prediction
SMOTE	Synthetic minority over-sampling
SVM	Support vector machine
SEC	US securities and exchange commission
XGB	Extreme gradient boosting

1 Introduction

The principle aim of this chapter is to describe the background, motivation, and objectives of this thesis. This chapter is divided into four sections. The first section provides a background to the issues to be addressed whilst the second describes the research rationale. The third section outlines the research objectives and section four concludes this chapter by presenting the structure of this thesis.

1.1 Background

This research was conceived from the need to address the funding challenges facing the global shipping industry. This was brought about primarily by a combination of ever increasing maritime environmental regulation, an increasing risk-averse commercial banking sector and an aging fleet.

The international shipping industry is responsible for a clear majority of transportation of world trade[1]. In 2018 the global trade continued to increase with exports rising by 9.7% from 2017 reaching a record high of US$19.5 trillion according (UNCTAD, 2019). Simultaneously, growth in global seaborne trade (11 billion tons, 2018) fell to 2,7% from 2017 to 2018 (1970 - 2017 average growth rate of 3%) in comparison to 4,1% the previous year (UNCTAD, 2019).

The financial support necessary to sustain such figures has historically been achieved through debt financing with banks providing much of the capital funding necessary to support the global fleet. However, the banking system failure, which was integral to the financial crisis of 2008, had serious negative impacts on the shipping sector and indeed led to the extraction of the commercial banking sector from all major risk intensive private capital investments. One important regulatory consequence of this was the imposition of increasingly stringent capital adequacy rules by the Basel framework[2] (BIS, 2019) which forced banks to offload the more risky assets from their balance sheets whilst also placing further restrictions on new investments.

The "overheating" in the global economy pre-crisis was reflected very strongly in the shipping sector with an significantly inflated freight rate environment coupled with a peak of newbuilding orders at the beginning of 2008 (Grammenos, 2010). Freight rates collapsed, overcapacity became critical and the inevitable market corrections began with devastating effect.

---

[1] *The International Chamber of Shipping (ICS) estimate 90% but this figure varies somewhat, and accurate figures are difficult to verify.*

[2] *This led directly to the adoption of the Basel Internal Rating System for risk management.*

Since the crisis, commercial bank lending has experienced a severe downturn. Throughout the past 11 years the Petrofin Global Index[3] has almost continuously declined (see Figure 1) with lending from the top 40 banks to shipping standing at $300.7bn (Petropoulos, 2019), its lowest level since the crisis. In contrast, the world merchant fleet has increased during this same period with an estimated 60% increase on 2008 with approximately 1,403m GT with a value of 230,9bn USD on order (Clarksons, 2019). This has resulted in a significant finance gap and a substantial structural shift in the ship financing sector.

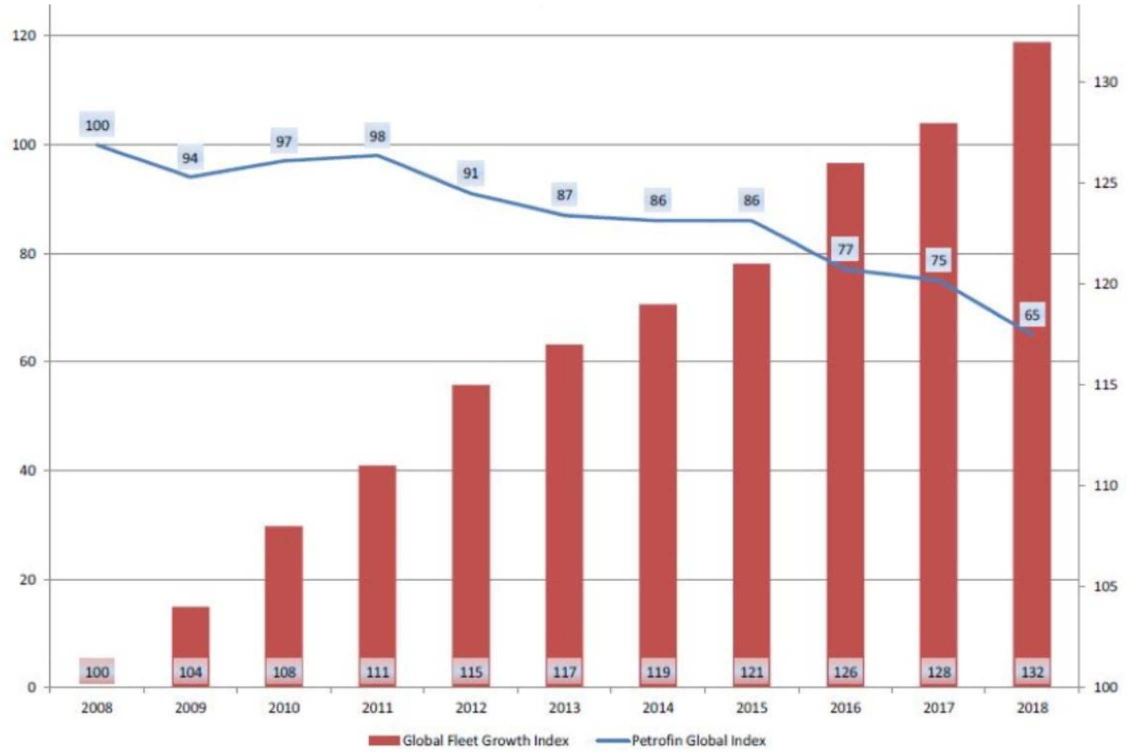


Figure 1: The Petrofin Global Index compared to global fleet growth (Petropoulos, 2019)

Shipping finance has witnessed a switch from traditional European banks to non-banking sources (Chinese leasing companies in particular) and Asian banks. According to Petrofin, Western banks have reduced their shipping portfolio by

[3] *The Petrofin Index for Global Ship Finance was introduced in 2008 as a measurement of bank funding incorporating the top 40 international shipping banks.*

USD 160bn during the past 11 years and their share of the global ship finance market has shrunk from 83% to 58,7%, whilst Asian banks have seen a surge in lending of up to 140% since 2010 (Figure 2). In 2018 Chinese leasing to shipping alone amounted USD 51,3bn compared to USD 47bn in 2017. It is believed that the intervention into the market by Chinese financiers will continue.

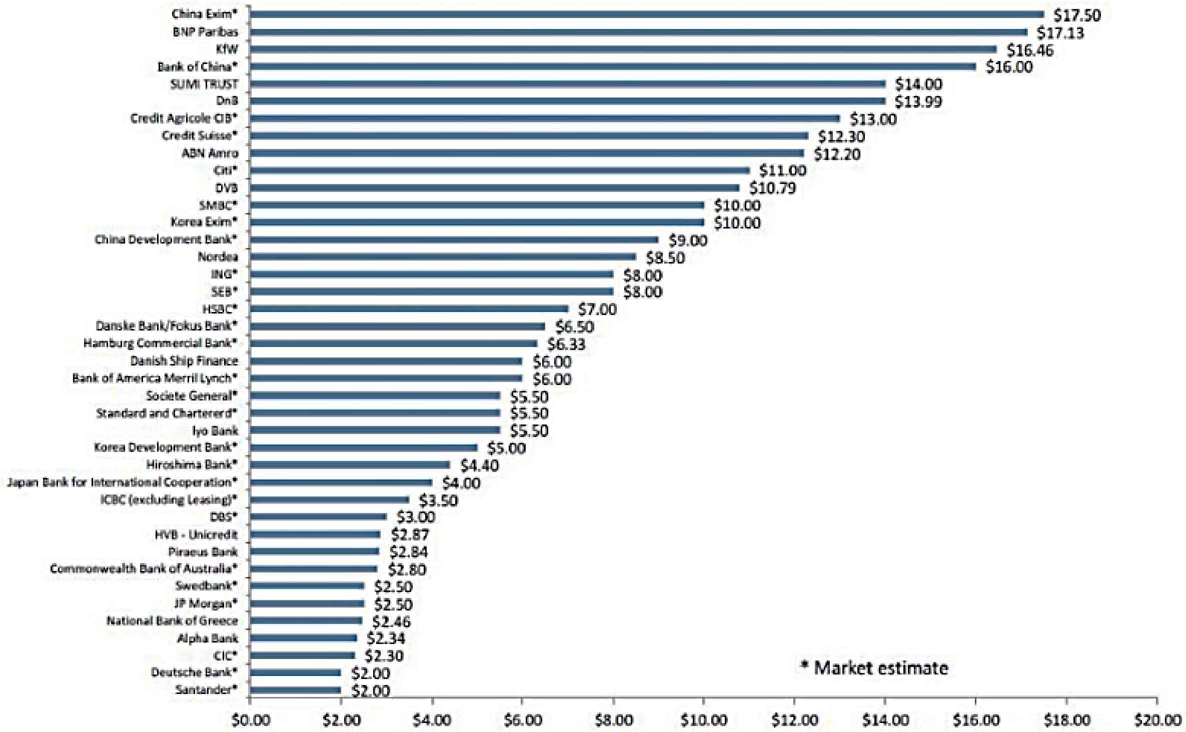


Figure 2: Global bank shipping portfolios as of end 2018 (Petropoulos, 2019)

However, as mentioned above, attempts to bridge the gap were initially taken up in part by Asian banks until relatively recently. The slowdown in the Chinese economy has resulted in a toughening of credit conditions leading to a reduction in bank lending.

The withdrawal of commercial bank lending has seen an increase in sale and leaseback transactions. Leasing provides long term finance option but often with higher pricing. However, in the current low US$ interest rate environment, it remains a viable option. As such, it has increased in popularity in the financing of vessels, especially amongst small and medium sized enterprises (SME) owners which it represents the only available and affordable source of finance. Investment fund finance and lending for family owned concerns has also increased but this has come with an increase in risk pricing.

As regards the equity and bond markets, these have been relatively quiet, with few IPOs (Petropoulos, 2019). Nevertheless, there has been a steady development of the shipping bond market for medium to large companies.

In summary, ship financing, be it from bank lending, leasing or otherwise, remains in increasing demand, especially given the regulatory pressure on the industry to improve its environmental impact and meet climate change goals. As mentioned above, the financing options are becoming more complex with a growing number of funding options through stakeholders with differing demands. However, these stakeholders have one overriding factor in common, the need to protect their investments. The damage to the sector following the crisis, coupled with the sell-off of commercial bank shipping loans and the newly regulated banking industry, has highlighted the need for a more thorough understanding of the risk profile unique to shipping investments. This demands the development of effective probability of default models geared to the specific characteristics of shipping companies and which take due account of the macroeconomic and market environment in which they operate.

1.2 Research rationale

The management and absorption of financial risk is required to meet the needs of both the shipping sector and society's goal of a cleaner and more efficient maritime sector by reducing barriers to investment. The main barrier to investment being the financial risk profile of the sector. As outlined in the previous section, it is becoming increasingly necessary to model the probability of distress of shipping companies more accurately than has been achieved in the past. Therefore, the aim of this research is to examine the extent to which machine learning tools can help detect early signs of shipping company distress and therefore help investors and shipowners alike to manage risk and ultimately reduce the barriers to investment.

This research explores the use of machine learning (ML) models in gauging their ability to capture the correlation in financial distress (FD) which may eliminate the need for unobservable temporal effects. Several established ML models are evaluated together with some more recently established models such as random forest (RF) and extreme gradient boosting (XGB) alongside some more established generalised linear models. The objective is not simply to compare model performance but also to assess their individual capacity to generalise on out of sample data.

Secondly, due to the global nature of the shipping industry, diverse national accounting practices and laws render the identification and collection of complete and consistent financial statements, one of the major challenges in studies such as this. Therefore, the problem of missing accounting values and how to treat them is a major focus of this research.

Finally, the fact that none of the models can be expected to fully capture correlation in FD, solely through the application of accounting data, suggests that there are unobserved macro effects that create correlation in distress. Shipping, being a high-risk sector, will always be highly sensitive to global macroeconomic shifts and stochastic market events. As such a clearer understanding of those features, which accurately represent the risk profile of shipping companies, is essential. This research proposes a distress prediction model that employs not only company level data, but also macroeconomic and market feature definitions aimed at detecting early stages of distress.

1.3 Objectives

This research has two primary objectives. The first is to design and test a forecasting model which attempts to capture correlations between company financial information and other macro events in shipping company financial distress. This model will address:

- issues surrounding the "noisy" nature of company financial data typically associated with shipping companies, such as information skewness, data imbalance and missing accounting information
- the identification of a set of predictor features which represent a predictor set capable of capturing correlation in financial distress prediction for this sector
- the non-company specific event affecting shipping company performance through the inclusion of macroeconomic and market data in the forecasting of distress

The second research objective is to develop the foundation for a machine learning software system which:

- provides a general-purpose tool that is of real practical value
- results in predictions that are as accurate as possible
- takes full advantage of computing architecture, multi-processing, multi-core technologies and cloud computing, in order to both maximise efficiency and reduce execution times to a practical working level.
- is both scalable and modular in form, supporting the addition of more complex deep learning algorithms which require enhanced computing power, including support for larger and more complex datasets
- can be readily applied to a broad class of shipping sub-sector FD learning problems
- must be transparent and open to scrutiny by all stakeholders, investors and particularly regulatory bodies if they are to be accepted as practical tools

Both the model and system software validity and performance will be analysed by applying a suitably representative test case comprising detailed financial statements covering the period 2000-2018 of dry bulk carrier owners/operator companies, worldwide, both listed and non-listed. The case study dataset will also include macroeconomic and bulk commodity data covering the same period.

In conclusion, the aim of this research is to develop a methodology and software toolkit which enables the early detection of FD which not only provides investors and other stakeholders with the means of avoiding some of the costs associated with a bankruptcy filing and subsequent recovery, but also assists shipowners in monitoring their own financial performance.

1.4 Thesis structure

The main body of this thesis is structured as follows:

Chapter 2 - Literature review: This chapter begins with a review of literature on general corporate financial distress and in the context of shipping entities. The second section contains a review of the literature of statistical/machine learning tools for the prediction of financial distress. A third section reviews research into the specific methodologies which form the constituent modules of the model proposed in this research. Following this critical review, a discussion on the gaps identified in the literature is presented, which leads on to the formulation of the research questions and hypothesis. The chapter concludes with description of the proposed contribution of this research and its novelty value within the field.

Chapter 3 - Methodology: This chapter describes the theory behind the development of the model and software architecture. It begins with a description and review of both data analysis and data pre-processing theory. The chapter then progresses to describe and review the classification algorithms, both traditional and complex, which form part of the model. The chapter concludes with discussion and summary.

Chapter 4 - System software and architecture: This chapter describes the high-level technical specification of the proposed "Shipping Financial Distress Predictor" (SFDP) architecture and modules.

Chapter 5 - Case study – The dry bulk shipping sector: This chapter presents the application of the model and SFDP system for a case study in the dry bulk shipping sector. It presents and discusses independent feature selection specifically relevant to bulk shipping companies.

Chapter 6 – Results and discussion: This chapter presents and analyses the results of the case study. It assesses the performance of data pre-processing modelling and individual classifier predictions and examines the efficiency of the SFDP system. It concludes with a discussion of the results.

Chapter 7 - Conclusions: This chapter begins with a summary of the research outcomes and compares these of the research objectives outline in section 1.3 of this chapter. It continues by describing the novelty and contribution of this research whilst discussing the gaps and the recommended future related research. It concludes with final remarks.

1.5 Summary

This chapter has outlined the problem for which this research topic was conceived and provides a high-level set of objectives designed to address the issues concerned. As such, the aim is to contribute to the research field by developing a practical, efficient and transparent methodology and software toolkit which accounts for both the idiosyncratic nature of shipping finance but also harnesses proven machine learning algorithms, both traditional and complex, for the first time in the domain of shipping company financial distress.

2 Literature review

2.1 Overview

This literature review is guided by the research rationale and objectives outlined in chapter 1. The first section reviews research efforts into shipping company financial distress prediction and the literature surrounding the application of machine learning tools for corporate FD prediction. The second section concentrates on the specifics of the SFDP architecture: the predictor or classification algorithms selected; the handling of incomplete financial accounting data; and the selection of core independent variables used in modelling distress prediction. The chapter concludes with a summary and discussion of gaps in the literature.

2.2 Corporate financial distress prediction

2.2.1 Financial distress in shipping companies

The financing of the shipping industry has been traditionally based on bank loans. An important step in bank's credit granting process is the application for the loan and its consequent evaluation by the bank. Numerous bankers, consultants and academics have addressed this matter, presenting financial requirements and attributing values indicating the soundness and the financial strength of the applicant. A critical priority for bank credit risk departments relates to providing an optimal framework for assessing the credit rating of borrowers' as well as loan quality and defining specific quantitative and qualitative criteria that mirror the borrowers' ability to comply with the loan contract terms. The conceptual framework for credit decisions in banking practice has gradually evolved over time. At a simplified level, this has been founded on the three core Cs of credit, that is, borrower's 'character', 'capacity' and 'capital'; frequently, two additional Cs have also been included, that is, 'collateral' and 'conditions' with (Antoniou, A., A. Thanopoulos, 1998; Grammenos, 2010) applying this to shipping credit scenarios.

Credit risk assessment work has often been performed following the construction of 'standardized' models, as noted by (Dimitras, Petropoulos and Constantinidou, 2003). The authors contend, however, that these models, which combine criteria and provide relative weighting to assist the decision-making process of the bank's credit committee, are limited. Their paper presents work on the application of the monotone regression method, UTADIS, and is aimed at the analysis of both credit allocations and the evaluation of the criteria used for the selection of loan applications in shipping industry.

Innovative creditors, nevertheless, have long sought straightforward, timely and solid methods to evaluate credit decisions. (Gavalas and Syriopoulos, 2016) proposed an integrated credit rating model based on a series of critical qualitative and quantitative criteria for bank loan portfolios. The model is applied to and tested on bank financing decisions in the shipping sector as a case study. Again, the authors used a UTADIS based approach in order to assess the relative impact of the selected risk factors on efficient credit rating scoring and loan quality assessment. Based on that, a credit rating and loan quality scoring model was subsequently developed, considering the prioritized risk exposures and evaluating their contribution to an integrated credit analysis framework. Furthermore, this approach was implemented in order to assess the relative impact of the selected risk factors on efficient credit rating scoring and loan quality assessment.

Research into shipping finance distress prediction has been relatively limited to date. These works have tended to focus largely on financial performance predictor/feature selection. They rely on more conventional methods such as binary logistic regression techniques (Antoniou, A., A. Thanopoulos, 1998; Grammenos, Nomikos and Papapostolou, 2008; Kavussanos and Tsouknidis, 2016; Mitroussi *et al.*, 2016; Lozinskaia *et al.*, 2017) and focus on either shipping bond markets or bank shipping debt.

Finally, all these studies demonstrate limited access to longitudinal corporate financial data which would allow for a more thorough assessment of predictive capabilities of the tools available. Moreover, their reliance on linear methodologies limit presented models in their capacity to accurately predict FD in out of sample data.

2.2.2 Machine learning in financial distress prediction modelling

The second strand of this literature review focuses on recent efforts on the application of ML models on FD prediction. Since Altman and Ohlson's work, research in the modelling of corporate financial distress and bankruptcy has been extensive e.g. see (Altman 1977; Shumway 2001; Duffie and Singleton 2003; Hensher and Jones 2007). However, until relatively recently much of this work relied heavily on more traditional classifiers such as logit, probit or linear discriminant analysis which are commonly referred to as generalised linear models (GLM). The financial crisis demonstrated that increased effort was required to develop models with enhanced predictive accuracy, not only for predicting ultimate failure events, but models which also generate indications of the early stages of financial distress. Post-crisis, research has highlighted failures in conventional corporate financial distress prediction models e.g. see (Duffie *et al.*, 2009; Barboza, Kimura and Altman, 2017; Christoffersen, Matin and Mølgaard, 2018). The academic consensus is that conventional statistical techniques have certain restrictive assumptions including linearity, normal distribution, multi-collinearity, auto-correlation, sensitivity to outliers and homoscedasticity which do not sufficiently capture the complex relationships between covariates and FD. These limitations coupled with the need to account for frailty and unobserved heterogeneity have resulted in a switch of focus by industry and academics alike to the application of more complex methods e.g. see (Lessmann *et al.*, 2015; Zhang *et al.*, 2017).

ML methods applied to FD prediction are now well established in the literature, notably (Jones, Johnstone and Wilson, 2015a; Zięba, Tomczak and Tomczak, 2016; Barboza, Kimura and Altman, 2017; Xia *et al.*, 2017). The general conclusion is that 'new age' classifiers outperform traditional (GLMs) models in out of sample generalisation. However, despite research demonstrating the enhanced generalisation performance of ML classifiers, care must be taken with their use and the literature makes repeated reference that the following caveats should be kept in mind when evaluating their exploitation; i) industry standard GLM models based on Altman and Ohlson are still widely used due to their simplicity and relatively good prediction potential and as such their use should be a serious consideration; ii) data quality is paramount and complex ML tools will not compensate for poor data quality; iii) transparency is essential in finance. Investors and regulators demand it and, as much of the literature observe, ML models demonstrate real issues with transparency[4].

Much of the published research has focused on benchmarking (Barboza, Kimura and Altman, 2017) against GLM models as opposed to reviewing the capacity of ML models to adequately predict FD. This research adds to the literature by evaluating the capacity of ML models to capture annual fluctuations in FD of shipping companies.

A rapidly increasing focus in the literature is the application of machine learning (ML) modelling (complex models exhibiting non-linear dependency structures between the covariates and the resulting outcome) in corporate failure prediction see e.g. (Hernandez Tinoco and Wilson, 2013; Jones, Johnstone and Wilson, 2015b; Christoffersen, Matin and Mølgaard, 2018). Much of the previous work has benchmarked the performance of ML models on generalised linear models such as logistic regression (LR). However, it is now widely accepted that generalised linear models result in significantly narrow[5] confidence intervals (CI) of aggregated FD predictions owing to their underlying assumption of conditionally independent observations.

2.3 SFDP model specification

This section consists of a review of the literature surrounding the architectural composition of the SFDP model. It begins by addressing the literature concerning the definition of FD, the *core concept* in developing systems whose prime goal is the prediction of such events. It continues by reviewing research into ML classification methodologies, both traditional and complex, used in this research and includes an examination of imputation methodologies for the treatment of missing data. The imputation methodology review is required to address the problem of missing accounting information in

[4] *The literature particularly singles out that neural networks and "deep learning", algorithms as lacking transparency.*

[5] *Too narrow confidence intervals indicate the existence of a downward bias risk estimation and that the assumption of conditional independence in the covariates is not satisfied.*

shipping company financial statements, which is a frequently encountered problem within the sector. The section concludes with a review of feature (independent variable) selection in FD prediction.

2.3.1 Definition of financial distress

Much of the finance literature define the event as being centred upon the final legal consequence of either the organisation's liquidation or bankruptcy. These are clear legal events, which have a definitive date and can be represented by a dependent variable in a binary classification model variable (Balcaen and Ooghe, 2006). These legal definitions, however, only represent the worse-case scenario of FD and as such present challenges for FD prediction. The process of insolvency is, in many cases, significantly lagged (Hernandez Tinoco and Wilson, 2013). The literature estimates a time gap of up to three years or more between the point at which a company experiences financial distress and the date of a legal declaration of insolvency (Theodossiou, 1993; Hernandez Tinoco and Wilson, 2013). Furthermore, legislation such as the U.S. chapter 11 has brought about changes in the way organisations can be provided time for reorganisation of a company's business, assets and debts in the event of impending insolvency. There are a number of stages a company can encounter before closure: (Wruck, 1990) cites FD, insolvency, filing of bankruptcy and administrative receivership. All of which add to the lag in the final legal declaration.

Within the context of, and prior to, the triggering of the terminal states addressed above, the literature generally follows two approaches regarding the issue of the definition of FD. The first is an accounting features approach, utilising cross sectional annual data, and is widely covered in the default prediction literature e.g. see (Altman, 1968; Ohlson, 1980). This approach utilises historical financial statements which are benchmarked against historical default rates and generally modelled to produce a probability of bankruptcy outcome. The second, is a mixed accounting/market based approach which estimates a company's probability of default based on its distance to default (DD) (Black and Scholes, 1973; Merton, 1974). The DD model utilises both the expected return on assets and the volatility of those returns in order to assess the probability of asset values declining below the value of the company's debt (as a factor of the time to maturity of a company's outstanding debt). Based upon this widely accepted foundation[6], DD is included as a feature in the modelling.

However, recent literature has highlighted the failure of such traditional approaches to encapsulate spatial (annual) fluctuations in FD. Recent publications (Duffie *et al*. 2009; Nickerson and Griffin 2017; Kwon and Lee 2018; Azizpour, Giesecke, and Schwenkler 2018) suggest that simply modelling relationships between observable covariates and FD does

[6] *Moody's Analytics for example.*

not adequately account for latency (unobserved variables) and as such the authors advocate approaches which include frailty[7] or the inclusion of time-varying effects.

2.3.2 Missing accounting values

2.3.2.1 Overview

The problem of missing data is predominant in financial modelling (Kofman, 2003; Burger, Silverman and Vuuren, 2018) and is a particular problem with shipping company accounts e.g. see (Sharife, 2010). This is also true of the raw panel dataset data compiled for the study case using in this research. Missing data leads both to bias as well as loss of information. This literature identifies three accepted methods of treating the missing data issue. The first method is referred to as the "complete case" (Nguyen, Carlin and Lee, 2017) or list-wise deletion approach which discards individual observations (company accounting years) containing missing data to provide only a dataset with complete, observed data. A complete case analysis of the raw data used as the test case for this study involves the removal of approximately 18% of the sample space. It should be noted that the results of complete case analysis potentially reduce the available raw dataset to the point where distortions or bias are introduced.

The second method is referred to as the "omitted variable approach which involves simply removing those covariates with missing values from the dataset (Servaes, 1996; King *et al.*, 2001). However, this approach has a problem when the covariates concerned are particularly correlated with the dependent variable.

The third method is data imputation and is part of a growing field of research to address the challenge of missing values in data. Two primary schools of thought exist for a generalised approach to data imputation. The first methodology, introduced by (Rubin, 1987), is a model-based approach founded upon the concept of multiple imputation (MI). The MI procedure replaces each missing value with a set of potential values that account for the uncertainty around the correct value to impute. The generated multiple imputed data sets are then analysed using standard procedures for complete data and combining the results from these analyses. The second methodology is based on ideas built around the formulation of the expectation-maximization (EM) model, made popular by (Dempster, Laird and Rubin, 2011). The EM approach is basically an iterative method to find maximum likelihood or maximum a posteriori estimates of the missing values. The concept is to handle the missing values as random variables to be removed by integration from the log-likelihood function, as if they were not sampled. A significant disadvantage of EM is the requirement to explicitly model joint multivariate distributions which limits its application to variables deemed to be of substantive relevance (Graham, Cumsille and Elek-

---

[7] *Frailty can be considered a random effect model implemented for" time to event" data. The aim is to account for heterogeneity induced by unobserved features.*

Fisk, 2003) i.e. the effects need to be large enough to be significant. Furthermore, this approach requires the correct specification of usually high-dimensional distributions, even of aspects which have never been the focus of empirical research and for which justification is difficult to realise in practice. For this reason, the MI approach is the focus of this research.

MI has become one of the most widely established methods for handling missing data and is receiving increasing attention in the finance research (Dicesare, 2006; Amel-zadeh *et al.*, 2020). Imputation is fundamentally a further layer of modelling whereby missing values are estimated from other predictor variables in the dataset. Prior to model training or the prediction of new samples, missing values are estimated using an imputation methodology. Such is the evolution in the statistical analysis of missing value problems it has now developed an accepted taxonomy, particularly surrounding the vital issue of understanding the reasons why values are missing. Central to this is the concept of the missingness mechanism (Rubin, 1987) which is a system that compartmentalises the missing data problem into three distinct categories, namely: missing completely at random (MCAR); missing at random (MAR); and missing not at random (MNAR). Missing data is said to be MCAR if the probability of missingness does not depend on any of the other variables relevant to the analysis of scientific interest, observed or missing. The missing mechanism is MAR if the probability of missingness is independent of the relevant unobserved values. The third category MNAR, describes data where the missingness is dependent on both observed and unobserved values.

The commonly adopted theoretical approach, in practice, is MAR. This approach assumes that the reasons for missing data in any sample can be explained by the observed data i.e. information present in the training set is used to estimate the values of other predictors. In company financial statements, the incomplete data is likely to be MAR, with missingness associated with values of relatively complete variables such as firm size, leverage ratios, location, etc. In these instances, multiple imputation methods offer the potential of significantly improved estimates with less bias and greater efficiency (Kofman, 2003).

The approach of "informative missingness" or Missing Not at Random MNAR[8] (Little and Rubin, 2019) suggests that significant bias can be introduced through informative missingness. A prime example in the case of shipping company data is that of "off-shoring" registration in jurisdictions that do not oblige entities to publish complete annual accounts. For various reasons, some well-known, companies wish their financial situation to remain "confidential". In this case there is a clear relationship between the probability of missing values and the related outcome of company financial performance. There is reason however, to hypothesise that treatment of such data as being MAR, and that shipping

[8] *Rubin et al. acknowledge that this should be more clearly described as Missing Not at Random*.

companies with similar recorded feature values can adequately simulate missing observations in such entities. However, care must be taken in treating the imputed values to reflect any departures from the MNAR assumption.

It is important to note the missing values should not be confused with censored data, where entries are missing but something is known about the individual data. For example, it is not uncommon for companies not to publish accounts in the period just before failure. For example (T. Shumway, 2001) deem a company as in FD if the firm delists the following year and “files for any type of bankruptcy within 5 years of delisting”. It is common practice for entities in terminal decline to avoid publishing new accounts.

Note that imputation adds an extra level of uncertainty. It is the application of a predictive model within another predictive model. Furthermore, if resampling is utilised to select tuning parameter values or to estimate performance, the imputation should be incorporated within the resampling. Before performing the final stage of modelling a comprehensive validation of the imputed data should be performed.

2.3.2.2 Shipping and financial secrecy

The different perceptions concerning the status of certain states considered “flags of convenience” (FOC) are due to the fact that their secrecy laws focus on the shipping industry (Sharife, 2010). Shipping companies register their ships under such flags for fiscal and/or regulatory reasons. States providing the statutory environment supporting secrecy to conceal ship ownership are referred to as FOC havens.

The FOC is often one of a shipping company’s international tax planning strategies. Shipping companies often exploit variations in domestic tax law and international taxation standards (Kim and Kim, 2018). This provides them with opportunities to eliminate or significantly reduce taxation and therefore, many multinational corporations use base erosion and profit shifting (BEPS) (OECD, 2013) to erode the corporate tax base.

Global shipping companies flag ships in foreign countries for many reasons, including avoidance of corporate tax, national labour regulations and environmental laws. Furthermore, it makes it easier for companies to lower wage costs through the recruitment of crew from low-wage states. The FOC modus operandi is to require an initial registration fee and annual renewal subscription, and then require minimum or no financial burden on the operating profits from the company’s operations. Moreover, FOC states normally do not impose personal income tax on shareholders or crew residing outside their jurisdiction, nor withholding tax on dividends paid to non-residents. This enables shipping companies to respond flexibly to the fluctuations in the shipping market by securing a smooth cash flow.

2.3.2.3 International accounting standards

The period 2000-2019 saw the gradual global uptake of International Financial Reporting Standards (IFRS) for both public and SME companies. This gradual uptake and multiple changes to the IFRS by the International Accounting Standards Board (IASB) has contributed to inconsistencies which have resulted in certain accounting information either being incomplete or simply not reported. One prime example of this is the reporting of leased assets on company balance sheets prior to the coming into effect of IFRS16 (IFRS Foundation, 2016) in 2019. This was a result of a finding in 2005, by the US Securities and Exchange Commission (SEC), which alleged that US public companies had approximately US$1.25 trillion of off-balance sheet leases. Thus, the IASB deemed that a customer (lessee) leasing assets should recognise and report assets and liabilities arising from those leases. The significance of the missing information problem, prior to IFRS entry into force, varied by industry and region and between companies. However, with shipping companies in particular, the effect on reported assets and financial leverage was substantial. The absence of information regarding vessel leases on the balance sheet meant that investors and analysts were not able to properly compare companies that borrow to buy assets with those that lease assets, without making adjustments.

From 2019 all vessel leases are recognised on the balance sheet for a lessee, with a right of-use asset and a lease liability which will result in changes in profit or loss throughout the life of that lease and which also has an impact on key accounting metrics. According to (Tahtah and Roelofsen, 2016) a result of IFRS16 is that there would be a median debt increase of 24% and a 20% median increase in EBITDA for the transport and infrastructure industry.

2.3.2.4 Summary – Missing accounting values

The extent of the problem of missing values in shipping financial statements *necessitates* its proper inclusion in the development of a viable model for FD prediction in shipping companies. This issue has, to date, not been addressed in the literature.

Finally, the primary objective of this research is the accuracy of the predictions rather than making valid subject related or sector informed inferences. Meaning that the goal is not the regeneration of missing values but to maintain the characteristics of the data distribution and the relationships between features and thereby maintaining the model's overall ability to generalise on out of sample data.

2.3.3 Feature selection for financial distress modelling

The emphasis in this section is on observable covariate (feature) selection. This study is particularly concerned with selecting features, quantitative and qualitative, drawn not purely from company accounts but also from data within the environment within which shipping companies exist and which are considered essential indicators of financial health. Much of the previous work on FD prediction has relied solely on publicly available historical accounting data or on securities market information. However, more recent research has recognised that accounting data alone are not enough to explain the relationship between the covariates and FD prediction. (Balcaen and Ooghe, 2006) argue that if too much emphasis is placed on financial ratios for failure prediction then it is implicitly assumed that all financial distress indicators are contained within financial statements. In order to address this issue, there are many examples in the literature which examine combined approaches using accounting, macroeconomic/market, including qualitative data, in order to provide an enhanced model of financial distress prediction e.g. see (Das *et al.*, 2007; Duffie *et al.*, 2009; Koopman, Lucas and Schwaab, 2011). (Bonfim, 2009) postulate that when macroeconomic features are considered then this leads to an improvement in model results. The consensus within the literature is that macroeconomic dynamics represent an independent contribution in financial distress prediction. As regards shipping this is an issue recognised by (Lyridis, Manos and Zacharioudakis, 2014) for example. However, care must be taken weighing the effects of macroeconomic variables. For example, the results (Ali and Daly, 2010) indicate that the same set of macroeconomic variables display different default rates across the geo-political spectrum. They concluded that GDP, short-term interest rates and total debt explained default risk for two differing economies, the US and Australia, and concluded that US economy is much more susceptible to adverse macroeconomic shocks.

Enhanced corporate, social and responsibility (CSR) procedures are more likely to lead to public disclosure of CSR activities, helping companies develop into more transparent and accountable entities. Improved transparency reduces informational asymmetries between the company and investors and therefore reduces moral hazard, hence improving a company's risk profile. It is also argued that market frictions such as informational asymmetries and agency costs are the prime drivers of "upward sloping supply curves in capital markets", argue (Cheng, Ioannou, 2014). They conclude that a solid CSR environment reduces the capital supply curve slope. For example, (Weber, Scholz and Michalik, 2010) augment financial data with CSR features to investigate their effect in predicting company financial performance and hence improve their credit rating accuracy. They conclude that there is a clear link between company's CSR adoption and its credit rating. Elsewhere, (Reverte, 2012) finds a significant negative relationship between CSR disclosure ratings and the cost of equity capital.

To date, no specific studies investigating agency problems in the context of shipping financial distress have been discovered. However, there are studies that have examined the general agency problem in a wider economic perspective. For example (Bergantino and Veenstra, 2002)) investigate principle agent issues relating to charter party contracts whilst (Rehmatulla and Smith, 2015) examine the issue from a shipping energy efficiency angle whilst also addressing the split incentives problems. The issue of asymmetric information in shipping cost management is addressed by (Shuyong *et al.*, 2009). All of the issues addressed in these papers are relevant to ship finance in that the principle agent components of information asymmetry problems, adverse selection and moral hazard are present in the financier/shipowner relationship.

2.3.4 Classification algorithms

This section reviews the classification algorithms incorporated into the SFPS system modelling of shipping company FD. These algorithms include traditional linear based statistical tools as well as the more commonly used complex models.

The use of logistic regression has frequently been selected for benchmarking in many papers examining the use of ML models for company distress prediction e.g. see (Hernandez Tinoco and Wilson, 2013; Jones, Johnstone and Wilson, 2015b). The consensus is that there is arguably limited value for including it as a modelling tool for the purposes of benchmarking against the array of complex ML tools available today. Nevertheless, it remains the basis of other linear models which have been developed to relax the linear regression assumptions of the linear relationship the covariates and probability (logit of) financial distress. Furthermore, for purposes of modelling transparency the SFDP system incorporates linear based algorithms, namely: dynamic hazard, linear mixed effects, multivariate adaptive regression splines (MARS) and a generalised additive models (GAM).

2.3.4.1 Dynamic hazard model

The data distress events are experienced in continuous time, however, only the interval "accounting period" in which the event occurs is recorded. This use of interval-censoring means that the data can be classed as discrete-time data. The event may take place at any time within a period but is unknown until the accounts become available. Hence, a discrete hazard model is employed using time varying panel data see e.g. (Shumway, 2001; Duffie *et al.*, 2009; Christoffersen, Matin and Mølgaard, 2018; Gupta, Gregoriou and Ebrahimi, 2018). The discrete hazard model is estimated with random effects $\alpha(i)$ and control for unobserved heterogeneity/shared frailty.

2.3.4.2 Linear mixed effects model

Generalised linear models are a group of mixed effect regression models used for regression and classification. However, these models assume that all observations are independent of each other and are hence not appropriate for analysis of several types of correlated data structures, such as panel data. In the data, companies are observed nested within accounting periods. For analysis of this multi-level data, random effects should be added into the regression model to account for the correlation of the data. Random effect models for company distress have been covered extensively in the literature e.g. see (Duffie *et al.*, 2009; Koopman, Lucas and Schwaab, 2011; Chaudhuri, 2013; Christoffersen, Matin and Mølgaard, 2018; Kalak and Hudson, 2018).

2.3.4.3 Multivariate adaptive regression splines (MARS)

MARS (Friedman, 1991) is a one of the family of linear regression models which was developed to deal with non-linearity between variables. It utilises the ordinary least squares (OLS) method to estimate the coefficient of each covariate. However, instead of a variable, each term in a MARS model is a basis function (BF) derived from the original variable. BFs describe the relationship between the predictor variable and the response. MARS then partitions the predictor values into groups, using recursive splitting, and a separate linear regression line is modelled for each group. The connections between the separate regression lines are called knots. The knot is the point at which the model extensions minimize a squared error. Each knot has two spline BFs.

2.3.4.4 Generalised additive models (GAM)

One of the main assumptions of linear regression models is that they require the covariates to be linearly related to the probability of FD (or logit thereof). However, GAMs (Hastie and Tibshirani, 1987) relax this assumption by accounting for the fact that some of the predictors exhibit a continuous, non-linear relationship with FD. Furthermore, non-linear relationships are observed both below and above specific thresholds with respect to shipping company's adjusted financial ratios and as such it is necessary to take account of these non-linear relationships.

Compared with GLMs, GAMs demonstrate superior regularisation capacity thus enabling them to more adequately address problems of overfitting. They also have an advantage over more complex models of being more interpretable and as such, GAMs represent an acceptable solution between the interpretable, yet biased, GLMs, and more complex, "black box" learning algorithms.

Our implementation of company FD prediction utilising GAMs follows along the lines of those documented by e.g. (Berg, 2007; Lohmann and Ohliger, 2017; Christoffersen, Matin and Mølgaard, 2018; Valencia *et al.*, 2019).

2.3.4.5 Artificial neural network (ANN)

ANNs have been extensively covered in the FD prediction business and finance literature. (Tkáč and Verner, 2015) count 412 articles over 20 years and write that much of this research covers FD issues.

ANNs can be described as a non-linear discriminant model. The model is arranged in layers, which for binary classification consist of at least one input and two output class layers and one hidden layer. Each layer consists of one or more nodes, and there are weights to connect the nodes in different layers. ANN has several variations in terms of possible algorithms. The most commonly and widely used back-propagation network is utilised in this study.

Finally, and this is important in the world of finance, a major limitation of ANNs is transparency and therefore they are frequently referred to as 'black box' algorithms.

2.3.4.6 Support vector machines (SVM)

SVMs (Vapnik, 1999) are a class of modelling techniques which were originally developed in the context of classification models. They have been extensively examined in the context of FD prediction (Min and Lee, 2005; Shin, Lee and Kim, 2005; Sun and Li, 2012; Zhang, Hu and Zhang, 2015; Kim, Mun and Bae, 2018).

Unlike the more traditional classifiers e.g. LDA, logit and probit, SVM is relatively robust to observations with the greatest displacement from the hyper-plane i.e. less sensitive to outliers. However, the disadvantage of SVM is that they are susceptible to many of the same limitations as ANNs, particularly in terms of computational scalability, lack of interpretability and ability to handle irrelevant inputs and data of mixed type (Tian, Shi and Liu, 2012).

2.3.4.7 Random forest (RF)

RF (Breiman, 2001) is based on decision tree models or generalised classification and regression trees (CART). It has shown to be relatively robust and is particularly adept at handling outliers and noise in the training set. An RF identifies the importance of each variable in the classification outcome. Therefore, it provides not only the classification of observations, but also information about the determinants of separation among groups. The RF technique repeatedly generates classification functions based on subsets. However, RFs randomly select a subset of characteristics from each node of the tree, avoiding correlation in the bootstrapped sets. The forest is built for several sub-sets that generate the same number of classification trees. The preferred class is defined by a majority of votes, thus providing more precise forecasts and, most importantly, avoiding data overfitting. Our RF fitting model follows (Jones, Johnstone and Wilson, 2015b).

2.3.4.8 Boosting I – Stochastic gradient boosting (GBM)

Boosting algorithms (Schapire, 1990) basically involves the combination of a number of weak classifiers to create an ensemble classifier with an augmented generalised misclassification error rate. This is commonly referred to as boosting. However, it was not until (Freund and Schapire, 1995) developed the adaptive boosting (AdaBoost) algorithm that boosting became an established modelling tool within the machine learning community. Boosting developed further with the introduction by (Friedman, 2002) of stochastic gradient boosting. The basic principles of which are: given a loss function e.g. squared error for regression and a weak learner e.g. regression trees, the algorithm seeks out an additive model that minimizes the loss function.

In the finance literature, e.g. see (Florez-Lopez and Ramon-Jeronimo, 2015; Zhao *et al.*, 2016; Jones, 2017; Krauss, Do and Huck, 2017) attempts are made to demonstrate that gradient boosting enhances the performance of conventional linear models. Moreover, (Jones, 2017) notes that the can function both an exploratory/diagnostic tool and as a "bias eliminating framework" to rank predictors as well as identifying including important non-linear relationships and interaction effects. He concludes that a logit model can enhance the analysis and improve predictive and explanatory performance.

2.3.4.9 Boosting II - Extreme gradient boosting (XGB)

XGB (Chen and Guestrin, 2016) is an enhancement of Friedman's stochastic gradient boosting model. However, it builds on the qualities of gradient boosting by providing a highly scalable model which incorporates regularisation in order to limit overfitting and is developed to uniformly handle sparse data. Furthermore, it is generally optimized for parallel processing i.e. was designed for both speed and performance. The model has drawn attention by being behind many recent winning entries in large scale big-data competitions such as Kaggle[9].

To date, published literature on the application of XGB in FD research is limited (Zięba, Tomczak and Tomczak, 2016; Chang, Chang and Wu, 2018; Carmona, Climent and Momparler, 2019). This is most likely due to its relatively recent uptake in the social science research domain.

2.4 Summary, literature gaps and novelty.

This review has presented research into corporate and specifically shipping company FD prediction. The findings have demonstrated that significant attention has been devoted to the development of corporate distress classification. The literature demonstrates the clear evolution from linear methods, such as logistic regression such as hazard and linear mixed methodologies, to the more complex methodologies of GAM and MARS in the domain of general corporate distress prediction. The review also highlighted the increasing focus on research into the application of more complex machine learning tools in this domain. However, the review has highlighted several gaps in the literature on the issue of FD prediction specific to shipping companies. The outcome of this review is that research to date has made insufficient reference to, or account for, frameworks or theories with respect to:

---

[9] *Kaggle is a global community of data scientists which frequently holds competitions to solve real-world machine learning problems.*

- the application of machine learning models to shipping FD prediction
- the use of sufficiently large, longitudinal, company year financial statement datasets, which are a necessary foundation for an adequately representative analysis
- the reliance on the application of linear modelling in an environment which is proven to be non-linear in nature
- the issue of missing values in shipping financial statements
- the inclusion of a set of statistical evaluation metrics necessary to adequately evaluate model performance
- the inclusion of macro and market features (and externalities)
- the development of a scalable, modular single system which provides stake holders with a real time source of shipping company financial performance metrics

In conclusion, the gaps identified in the literature support the formulation of the following research questions:

1) Does the inclusion of market and macroeconomic data improve the predictive accuracy of shipping company financial distress?
2) Can multiple imputation models improve the FD prediction in the presence of significant amount missing financial statement values?
3) Can complex modelling, using modern classification algorithms, capture latent, unobserved variables that more completely account for correlation in financial distress than previously studies which relied solely on linear modelling techniques?

In summary, the aim of this research is to develop practical financial forecasting methodology which can: i) capture correlations in shipping company financial distress from company level, market and macroeconomic predictors data using modern machine learning techniques; ii) perform effectively in the presence of significant levels of missing values in company financial statements; and iii) form the foundation for a machine learning financial distress, early detection system which could improve the financial monitoring of shipping companies by acting as a practical "early warning system" for financial distress.

Finally, the novelty contributions of this research to the field of shipping company financial distress are as follows:

i. The company level information gathered for this research is the most extensive shipping company dataset utilised to date. The 20 year, longitudinal, data collected for bulk shipping company financial statement data increases the potential for valid inference that, to date, has not been achievable due to the restricted nature of previous studies which have relied primarily on cross-sectional data.
ii. This is an unprecedented application of a formal methodology addressing the problem of missing values in shipping company accounts.

iii. For the first time, a set of both linear based and modern machine learning algorithms have been tested simultaneously on a large set of longitudinal shipping company financial statement, macro-economic and market data.

iv. The development of a unique software tool kit for the development of shipping company FD prediction and general risk management applications.

3 Theoretical framework

3.1 Overview

This chapter begins with a description of the research strategy and design which were formulated to answer the research questions. The main body of the chapter continues with a description of the analytical framework encompassing the statistical and mathematical principles applied in modelling of the SFDP model. Figure 3 depicts how the chapter sections are incorporated into this model. The core of the chapter is divided into four main sections representing principle stages of statistical modelling. Section 3.3 and 3.4 examine issues surrounding missing values and data pre-processing and examines the theory behind missing value handling, statistical data transformation, outliers and class imbalance. The base theories of the classificational algorithms, both traditional and complex, used in this research are presented in section 3.5. Evaluation criteria for classification models is the subject of section 3.6. A summary of the framework is provided in the last section.

3.2 Research strategy and design

Research strategies, according to (Bryman, 2016) refer to the choice between quantitative, qualitative or a combination of the two. Quantitative research emphasises quantification in data collection and analysis, whereas qualitative research emphasises words or observations. This research implements a predominately quantitative strategy but also includes some qualitative elements. However, the qualitative elements are quantified, through categorisation (factors), for example, in order to represent principle-agent and corporate control structure issues, see Table 1.

The research design adopted for this study can be described as a combined fixed/flexible, cohort study which is tested through the application of a suitable case study. The design follows the logical flow and construction of the SFDP model depicted in Figure 3: The SFDP process. The fixed component comes in the form of non-experimental correlation/cohort[10] study, involving an extensive set of global shipping company accounts, observed over time (20 years). The flexible component comes in the form of a case study comprising the bulk shipping sub-sector.

[10] A cohort study is a type of longitudinal study that samples a group and performs cross-sections at intervals through time. In the case of this research a panel study where the companies in the panel share a common characteristic i.e. bulk shipping.

This research design adopted in this study is both *confirmatory* and *exploratory* in nature. For research question 1) the approach is to generate *a posteriori* hypotheses, first through an analysis of the data and then through the identification of potential relations between dependent and independent variables. The reason for selecting this strategy is that previous research indicates the existence of some form of relationship between variables selected in this study, but that there is a deficiency of understanding of both the direction and depth of this relationship. However, as no specific hypotheses exists, regarding the predictive ability of our feature set prior to this research, the study is exploratory with respect to the covariates selected. A further reason for selecting exploratory research is to avoid missing potentially interesting relationships, aiming to minimize the probability of rejecting a real effect or relation; often referred to as β probability or the probability of a type II error.

For research questions 2) and 3) *confirmatory* research design is deemed more appropriate. Previous research into financial distress prediction has produced *a priori* hypotheses stating that:

i. methodologies such as multivariate data imputation techniques can be statistically effective in representing missing data patterns, whilst maintaining the overall integrity of the observed dataset
ii. modern machine learning methodologies can significantly improve the predictive performance of financial distress in corporations over the more traditional linear based modelling techniques

Since neither of these hypotheses has been applied, to date, on a large longitudinal set of shipping company level, market and macroeconomic data, a confirmatory research design is therefore adopted for the research questions 2) and 3).

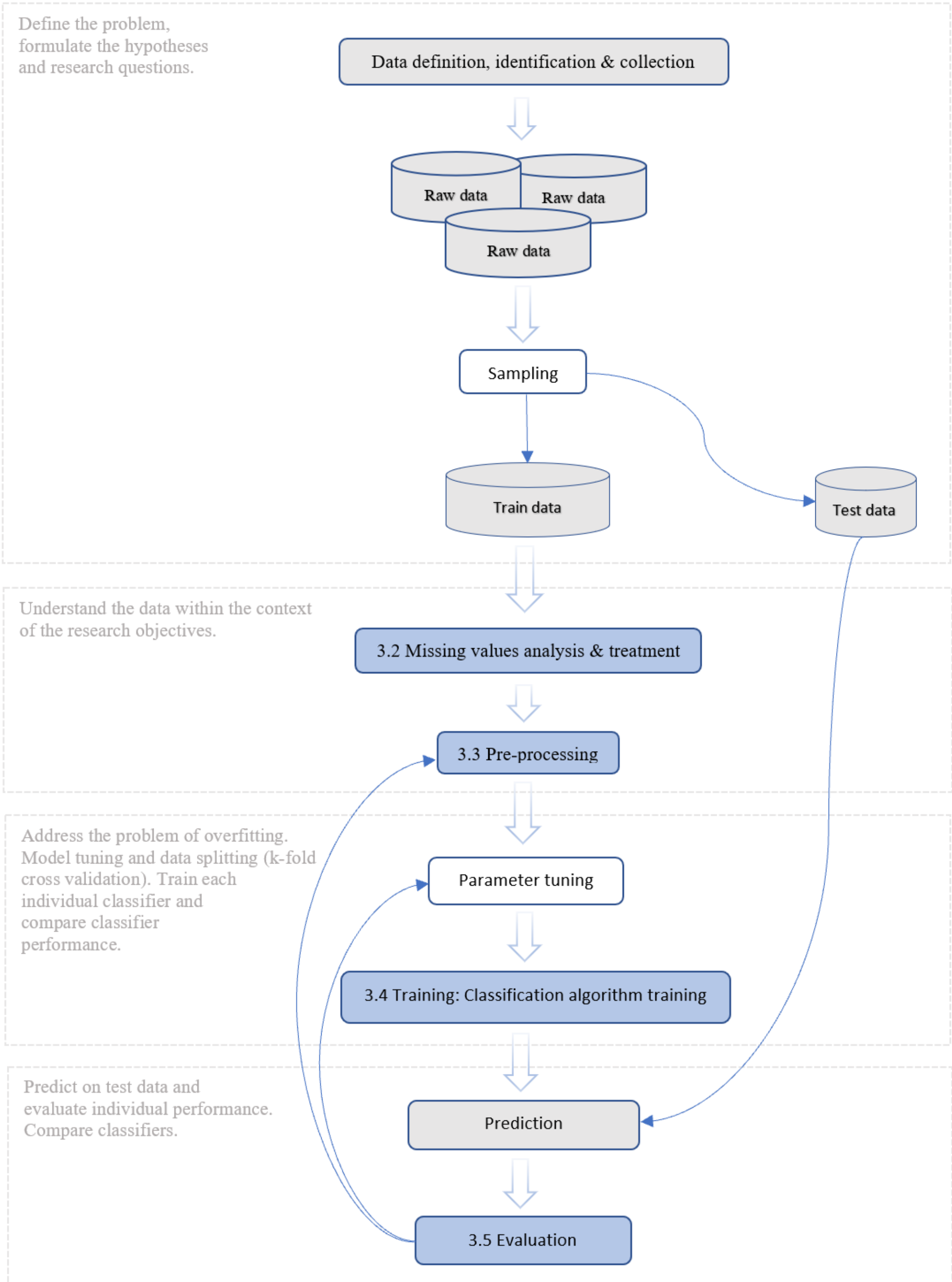


Figure 3: The SFDP process

3.3 Missing value imputation

Following on from the review of missing value literature in 2.3.2, this section describes both the theoretical background behind MI and the practical techniques used for its implementation in this research.

3.3.1 Missingness mechanism

The foundations of MI are based upon the assumptions of the missing data mechanism utilised when estimating the model parameters. The performance of missing data methodologies strongly depends on this mechanism for the generation of the missing values. What follows is a summary of this mechanism.

Let y denote an $n \times p$ dataset, $Y = (y_1, y_2, \dots, y_n)^T$, where $y_i = \left(y_{i1}, \dots, y_{ip}\right)^T$ is a random sample from a p-dimensional multivariate probability distribution $P(Y|\theta)$ regulated by parameters θ. The rows of Y observations are denoted by $Y_i(i = 1,2, \dots, n)$ with columns of $Y_j(\mathrm{j} = 1,2, \dots, n)$ variables. The n x p missingness indicator matrix is $R = \left(r_{jj}\right)$, where

$$r_{i_j} = \begin{cases} 1 & \text{if } y_{ij} \text{ is missing} \\ 0 & \text{if } y_{jj} \text{ is observed} \end{cases}$$

Defining $\Pr\left\{r_{ij} = 0|y_{jj}\right\} = \Pr\{y_{ij}\ observed|y_{ij}\} = p_{ij}$, then R has a probability distribution $P(R|\xi, Y)$ which is regulated by parameters ξ . Given this, the joint probability of the response variables and the missingness indicator variables can be expressed as

$$P(\mathrm{Y}, R|\theta, \xi) = P(Y|\theta)P(R|\xi, Y) \tag{1}$$

Where $P(P|\theta)$ is the marginal distribution of the response variables and $P(R|\xi, Y)$ represents the conditional distribution of missingness. With incomplete data, following, Y_{obs} and Y_{mis} (Little and Rubin, 1987) represent the observed portion and the missing portion of *Y*, where $Y_{obs} = \left\{y_{ij}|r_{ij} = 0\right\}$ and $Y_{mis} = \left\{y_{ij}|r_{ij} = 1\right\}$.

Equation (1) contains two sets of parameters, the parameter of interest, θ and the nuisance[11] parameters, ξ. Inferences on θ are conducted based on the joint probability model (Eq. 1), which depend on how the probability model for the

[11] *Nuisance parameters arise when the complexity of reality and data is such that models with multiple parameters are required. However, inferential interest is confined to a reduced set of parameters.*

missingness is defined i.e. the dependency of missingness on Y. Based on the conditional distribution, (Little and Rubin, 1987) define the missingness mechanism as one of the following:

(1) If $P\big(R|\xi,(Y_{obs},Y_{mis})\big) = P(R|\xi)$, the probability that a data point is missing does not depend on any variables, either observed or unobserved then the missingness mechanism is defined as Missing Completely at Random (MCAR)

(2) If $P(R|\xi,(Y_{obs},Y_{mis}) = P(R|\xi,Y_{obs})$, the probability of missingness depends on only on observed values in the data set then the missingness mechanism is defined as Missing At Random (MAR).

(3) If $P(R|\xi,(Y_{obs},Y_{mis}) \neq P(R|\xi,Y_{obs})$, the missingness depends on both observed and missing responses and is termed Missing Not At Random (MNAR).

Missing completely at random (MCAR) is the most restrictive as the missing values do not depend on either observed values nor missing values. With MCAR, the missing values for a variable are essentially a random sample of the values for that variable. Hence, the distributions of observed and missing values are the same. In contrast to MCAR, MAR is a less restrictive assumption because the missing values can depend only on the observed response variables. In this case, the missing values for a variable are a random sample of the data for that variable but within a sub-group of the observed values. Again, the distribution of missing values is the same as the distribution of observed values with that sub-group. MAR is the most widely used in practice.

Following the conclusions for Missing accounting values in section 2.3.2, including the justification therein, this research follows the MAR missing mechanism approach.

3.3.2 Multiple imputation

Multiple imputation (Rubin, 1987) was developed to solve a problem of survey non-response. Rubin argued that the problem of missing data should be handled in a principled systematic manner rather than ad hoc. His system has three fundamental steps:

1) Create m complete datasets, where $m > 1$ by assigning a value to each missing value m time by extracting m samples from *n* appropriate imputation model given observed values. The imputation model must respect the true distributional relationship between the missing and observed values.

2) The *m* imputed datasets are analysed by treating each imputed dataset as a real complete dataset. Standard complete dataset procedures and software can be used directly.

3) The results from m imputed complete datasets are combined in a simple appropriate way to obtain the so called repeated imputation inference (Rubin, 1987). Combined estimates variances consist of both within imputation and the between imputation variances, such that uncertainties in the imputed data are integrated into the final inference. This approach bypasses the restriction of single imputation, which underestimates the standard errors of the estimates.

The first step, constructing m imputation models to sample from, is the most fundamental part of MI. If a MI model satisfies certain frequentist properties it is termed *proper* (Rubin, 1987). Inferences based on relationships from imputed complete data risk being biased if the imputation model does not preserve the distributional relationships between the missing values and the observed values. For example, if the model does not include variables from the imputed complete data for the inference, then correlations between omitted and imputed variables will be biased towards zero. Furthermore, the between imputation variance typically will be underestimated if the multiple imputations are not based on conditionally independent samples from the imputation model, given Y_{obs}.

A Bayesian model provides the theory for making a repeated imputation inference (Rubin, 1987). When multiple imputations are *proper,* justification is also provided from a frequentist perspective. Let Q be a value to be estimated, e.g. an odds ratio or regression coefficient. The observed data posterior distribution of Q is:

$$E(Q|Y_{obs}) = \int P(Q|Y_{obs}, Y_{mis})P(Y_{mis}|\, Y_{obs})dY_{mis} \quad (2)$$

or, the observed data-posterior distribution of Q is the completed-data posterior distribution of Q averaged over the posterior predictive distribution of Y_{mis}.

If $\hat{Q} = \hat{Q}(Y)$ is used to estimate Q, given complete data, and the squared standard error is given by $U = U(Y)$ then the moment summaries can be obtained from the observed-data posterior distribution:

$$E(Q|Y_{obs}) = E[E(Q|Y_{obs}, Y_{mis})|\mathrm{y}_{obs}] \approx Avg(\hat{Q}), \quad (3)$$

$$V(Q|Y_{obs}) = E[V(Q|Y_{obs}, Y_{mis})|\mathrm{y}_{obs}] + V[E(Q|Y_{obs}, Y_{mis})|\mathrm{y}_{obs}] \approx Avg(\widehat{U}) + (1 + m^{-1})V(\hat{Q}) \quad (4)$$

$\hat{Q}$ and $\widehat{U}$ are produced from the imputed complete data, and $Avg(\hat{Q})$, $Avg(\widehat{U})$ and $V(\hat{Q})$ are the averages and variance over repeated imputations. As a finite number of imputations are used to calculate $Avg(\hat{Q})$, the inflation factor $(1 + m^{-1})$ is used to account for the additional variance (Rubin and Schenker, 1986).

Following the imputation of the missing data, Y_{mis}, through m sets of conditionally independent samples, $Y_{mis}^{(\mathrm{t})}$, from the posterior predictive distribution $P(Y_{mis}|Y_{obs})$, the repeated imputation inference is obtained as follows:

Calculate the repeated estimates $\hat{Q}^{(t)} = \hat{Q}(Y_{obs}, Y_{mis}^{(t)})$ together with the estimated square standard errors,

$\hat{U}^{(t)} = \hat{U}(Y_{obs}, Y_{mis}^{(t)})$, from the imputed completed datasets $\left\{y_{obs}, y_{mis}^{(t)}\right\}(t = 1,2, \dots, m)$. An estimate of Q is then the average of the repeated estimates,

$$\overline{Q} = \frac{1}{m}\int_{t=1}^{m} \hat{Q}^{(t)} \tag{5}$$

The standard error of $\overline{Q}$ is given by

$$T = \{(1 + m^{-1})B + \overline{U}\}^{1/2} \tag{6}$$

Where $B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{Q}^{(t)} - \overline{Q})^2$ is the between imputation variance and $\overline{U} = \frac{1}{m}\sum_{t=1}^{m}\hat{U}^{t}$ is the within imputation variance.

Finally, when the data are complete, it is assumed that the hypothesis test and the confidence interval are based on the standard normal distribution:

$$(\hat{Q} - Q)/\sqrt{\overline{U}} \sim N(0,1) \tag{7}$$

3.3.3 Multivariate missing values

As discussed in section 2.3.2, there are two main approaches for selection and specification of imputation methods joint methods (JM) and full conditional specification (FCS). The JM approach is considered less flexible than FCS and RP when treating complex data sets. Most JM implementations assume that the data originate from a multivariate normal distribution e.g. see (Little, 1988; Honaker and King, 2010; Templ, Kowarik and Filzmoser, 2011). The assumption of normality is inappropriate in the presence of outliers, skewed data, kurtosis and multimodal distributions as it potentially leads to flawed results.

FCS necessitates the specification of an imputation model for all incompletely observed variables. It then imputes values iteratively for each variable. Here, the FCS methodology is applied through the multivariate imputation by chained equations (MICE) algorithm (Groothuis-oudshoorn, 2011).

For each missing value, a density, $f_j\ (Y_j|Y_j-, \Theta_j\ )$, conditional on all other variables is specified, where Θ_j are the imputation model parameters. MICE, which is essentially a MCMC methodology, sequentially reviews each variable with missing values and draws alternately the imputation parameters and the imputed values. This method is summarized as follows.

Algorithm 1: MICE (FCS)

1: Fill in missing data Y^{mis} bootstrapping the observed data Y^{obs}
2: For $j = 1, \dots , p$
i. Sample $\Theta_j^{\star}$ ' from the posterior distribution of the imputation parameters.
ii. Impute $Y_j^{\star}$ from the conditional model $f_j(Y_j|Y_{j^-}, \Theta_j^{\star})$
3: Repeat step 2 K times to allow the Markov chain to reach its stationary distribution.

The methodology splits high-dimensional imputation models into multiple, single dimensional problems. The choice of imputation models in this setting can be varied, e.g. parametric, non-parametric or tree based.

3.3.4 Imputation methods

Although the results of the MI have been the subject of much scrutiny, relatively little is known about the comparative merits of various imputation methodologies that have been produced in recent years. Recent studies comparing available methodologies suggest the use of more flexible imputation methods, where available. Empirical evidence (Akande, Li and Reiter, 2017; Murray, 2018) also suggests that simple default methods such as log-linear models or a default FCS imputation e.g. predictive mean matching (PMM) with linear mean, are not necessarily suitable for practical application.

Flexible non- and semiparametric approaches, such as Bayesian and sequential tree-based methods, have demonstrated their ability to capture certain unanticipated features of the data (Shah *et al.*, 2014; Karim *et al.*, 2018). Empirically these methods can outperform existing simple parametric models or PMM using linear models in simulations.

As such the SFDP model incorporates; i) a flexible non-parametric model with PMM and cubic splines; ii) a CART based RF implementation as imputation methodologies.

3.3.4.1 Additive regression splines

This methodology is based on an alternative to fully parametric methods, called "hot deck" imputation (Chen and Shao, 2000; Harrell, 2015) which consists of replacing the missing value with the response of a "similar" observed variable. One common implementation of a hot deck method is the k-NN technique. The method is simple, it avoids strong parametric assumptions, only eligible and observed values are imputed, and it can easily be applied to differing variable types. The concept is the identification of k completely observed neighbours, for each missing value Y_{ij}, that are *close* with respect to Y_{ij^-}. From an identified set of eligible neighbours, one donor is randomly selected and its value $Y^{\star}_{ij}$ is taken as an imputation for Y_{ij}. *Closeness* can be expressed as a distance measure based on the estimated conditional mean of $Y_j|Y_{j^-}$,

$$d_{i,i'} = |\hat{E}\left(Y^{\mathrm{mis}}_{ij}|Y_{ij^-}\right) - \hat{E}\left(Y^{\mathrm{obs}}_{ij}|Y_{i'j^-}\right)|\,, \tag{8}$$

where Y^{mis}_{ij} denotes instance i of variable Y_j whose value was unobserved, and $Y^{obs}_{i'j}$ denotes instance i' of variable Y_j whose value has been observed $\left(i, i' = 1, \ldots, n\right)$.

Unfortunately, this can be too restrictive as a distance measure based on linear regression models ignores non-linear effects of Y_{j^-} on Y_j. In order to address this restriction (Harrell, 2015) developed an algorithm to account for non-linearity. Bootstrap resamples used for each imputation by fitting a flexible additive model on a sample with replacement from the original data. The model is used to predict all of the original missing and non-missing values for the target variable for the current imputation. The methodology uses flexible parametric additive regression spline models in order to predict target variables. The distance function is formulated as:

$$d^{areg}_{i,i'} = \sum_{l=1}^{L} |\left(f_l\left(Y_{ij^-}\right) - f_l\left(Y_{i'j^-}\right)\right)\beta^{\star}_l|\,, \tag{9}$$

where $f_l(\cdot)$, $l = 1, \ldots, L$ is a cubic spline basis which lead to optimal prediction, according to the coefficient of determination R^2, of a linear transformation of Y_j in the following additive model:

$$c + Y_j d = \alpha + \Sigma l = 1 L f_l\left(Y_j\text{-}\right)\beta_l + \nu \tag{10}$$

A non-parametric bootstrap is used to obtain the values of $\beta^{\star}_l$.

3.3.4.2 Random forest

An alternative approach to the conditional models discussed above define a new class of non-parametric multiple imputation methods based on classification and regression trees (CART) or random forests (RF) algorithms (Doove, Van Buuren and Dusseldorp, 2014). These methods form part of the concept of *recursive partitioning*, which provides for the modelling of internal interactions in the data by sequentially partitioning the data set into homogeneous subsets. The methodology involves growing a RF (Breiman, 2001) of size *k* by bootstrapping the complete cases and (optionally) subsampling the variables. An imputed value is generated by sampling from the k trees and then following the RF procedure to generate a CART imputation.

Assume $X = (X_1, X2, ..., X_p)$ to be a $n \times p$ - dimensional data matrix.

The prediction of missing values using an RF which is trained on the observed parts of the dataset. For an arbitrary variable X_s including missing values at entries $i_{mis}^{(s)} \subseteq \{1, \dots, n\}$ the data can be separated into four parts:

1. The observed values of variable X_s, denoted by $y_{obs}^{(s)}$;
2. the missing values of variable X_s, denoted by $y_{mis}^{(s)}$;
3. the variables other than X_s with observations $i_{obs}^{(s)} = \{1, \dots, n\} \setminus i_{mis}^{(s)}$ denoted by $x_{obs}^{(s)}$; and
4. the variables other than X_s with observations $i_{mis}^{(s)}$ denoted by $x_{mis}^{(s)}$.

Note that $x_{obs}^{(s)}$ is typically not completely observed since the index $i_{obs}^{(s)}$ corresponds to the observed values of the variable X_s. Likewise, $x_{mis}^{(s)}$ is typically not completely missing.

To begin, make an initial guess for the missing values in X using mean imputation or another imputation method. Then, sort the variables $X_s, s = 1, p$ according to the amount of missing values starting with the lowest amount. For each variable X_s, the missing values are imputed by first fitting an RF with response $y_{obs}^{(s)}$ and predictors $x_{obs}^{(s)}$; then, predicting the missing values $y_{mis}^{(s)}$ by applying the trained RF to $x_{mis}^{(s)}$ The imputation procedure is repeated until a stopping criterion is met. Algorithm 2 outlines the steps:

Algorithm 2: Impute missing values with random forest.

Require: X an $n \times p$ matrix, stopping criterion γ

1. make an initial estimate for missing values
2. k ←vector of sorted indices of columns in X w.r.t. increasing amount of missing values
3. while not γ do
4. $\quad X_{old}^{imp}$ ←store previously imputed matrix
5. $\quad$ for s in k do
6. $\qquad$ Fit a random forest: $y_{obs}^{(s)} \sim x_{obs}^{(s)}$;

7. Predict $y_{mis}^{(s)}$ using $x_{mis}^{(s)}$;
8. X_{new}^{imp} ←update imputed matrix, using predicted $\mathrm{y}_{\mathrm{mis}}^{(\mathrm{s})}$;
9. end for
10. update γ.
11. end while
12. return the imputed matrix X^{imp}

The stopping criterion γ is met as soon as the difference between the newly imputed data matrix and the previous one increases for the first time with respect to both variable types if present. Here, the difference for the set of continuous variables N is defined as

$$\triangle_N = \frac{\sum_{j\in N}\left(X_{new}^{imp} - X_{old}^{imp}\right)^2}{\sum_{j\in N}\left(X_{new}^{imp}\right)^2}, \tag{11}$$

and for the set of categorical variables F as

$$\triangle_F = \frac{\sum_{j\in F}\sum_{i=1}^{n} I_{X_{new}^{imp} \neq X_{old}^{imp}}}{\#NA}, \tag{12}$$

where $\#NA$ is the number of missing values in the categorical variables.

After imputing the missing values, the performance is assessed using the normalized root mean squared error for the continuous variables which is defined by

$$\mathrm{NRMSE} = \sqrt{\frac{mean\left(X^{true} - X^{imp}\right)^2}{var\left(X^{true}\right)}}, \tag{13}$$

where x^{true} is the complete data matrix and x^{imp} the imputed data matrix. *Mean* and *var* are used as a short notation for empirical mean and variance computed over the continuous missing values only.

3.3.5 Multiple imputation evaluation criteria

Multivariate imputation prime objective is to provide statistically valid inferences from incomplete data. The quality of an imputation model should be evaluated against this objective. There are a range of measures designed to evaluate the statistical validity of a model. These are:

i. Raw bias (RB) and percent bias (PB) . The raw bias of the estimate $\overline{Q}$ is defined as the difference between the expected value of the estimate and truth:

$$RB = E(\overline{Q}) - Q. RB \text{ (this should tend to zero)} \tag{14}$$

This can be expressed in percent terms:

$$PB = 100 \times |(E(\overline{Q}) - Q)/Q|. \tag{15}$$

For satisfactory performance an upper limit for *PB* of 5% is taken (Demirtas, Freels and Yucel, 2008).

ii. Coverage rate (CR) . This is defined as the proportion of confidence intervals (CI) that contain the true value. The actual rate should at least equal to the nominal rate otherwise the method is too optimistic and leads to false positives. A *CR* below 90%, for a nominal 95% interval, signifies poor quality. A high *CR* (e.g., 0.99) may indicate a too wide confidence interval inefficiency in the method which and indicates over conservative inferences. However, over conservative inferences are generally regarded a lesser evil than too optimistic.

iii. Average width (AW) . An indicator of statistical efficiency CI is the AW of the confidence interval. This should be as small as possible but not such that the *CR* falls below the nominal level.

iv. Root mean squared error (*RMSE*). This metric is a compromise between bias and variance. It evaluates $\overline{Q}$ on accuracy and precision:

$$RMSE = \sqrt{(E(\overline{Q}) - Q)^2} \tag{16}$$

Ideally, *RB* should tend to zero and the coverage should be close to 0.95. Methods termed *randomization - valid* have zero bias and *proper* coverage (Rubin, 1987). If two methods are both *randomization - valid*, then the method with the narrower confidence intervals is deemed more efficient.

A note of caution: although RMSE is widely employed, it is not a suitable metric for the evaluation of multiple imputation methods. The evaluation of methods based solely based on their ability to recreate the true data is not the aim. On the contrary, selecting such methods may be harmful as these might increase the rate of false positives. “Imputation is not prediction” (Van Buuren, 2018).

3.4 Data pre-processing

It has long been accepted that corporate financial statement data clearly demonstrate issues such as skewness, kurtosis and outliers e.g. (Barnes, 1987). Whilst this may be less of a problem for some advanced ML algorithms such as tree-based models, other models such as linear regression, k-NNs and principle component analysis (PCA) are particularly sensitive to such issues. The usual aims of variable transformation in regression are to make the distribution of a covariate, a response variable, or the residual, less skew, more homoscedastic, and closer to normal.

In this section the theoretical background into the methodologies used for the pre-processing of company financial data is presented.

3.4.1 Centering and scaling

Real world datasets contain features whose values range significantly in scale and range. This disparity can reduce the performance of some ML algorithms, particularly those that distinguish observations/feature effectiveness based on a distance measurement e.g. k-NN’s or those that use numerical gradient information in their solution e.g. ANN and SVM models. Models that depend on a measure of the variance such as PCA and those that penalize variables based on the size of their corresponding parameters, like penalized regression, are also affected. However, some algorithms are scale invariant; for example, tree-based models, which bin inputs before independently splitting on feature values, in order to avoid these issues.

In order to mitigate the detrimental effect of such variations in data, one of the most common data transformations can be utilised through the centering (standardizing) and scaling (normalising) of the independent variables. Scaling the data coerce the predictor values to hold a common standard deviation of one. Simple scaling is achieved through the division of each value of the predictor variable by its standard deviation:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{17}$$

To center a predictor variable, it’s mean is subtracted from all the values. The goal is to center the data around 0 and to scale with respect to the standard deviation.

$$x_{centered} = \frac{x-\mu}{\sigma} \quad (18)$$

These operations are aimed at improving the numerical stability of some models.

3.4.2 Data skewness

Many financial models, which attempt to predict the future performance of an asset, assume a normal distribution in which measures of central tendency are equal. If the data are skewed, this kind of model will always underestimate skewness risk in its predictions. The more skewed the data, the less accurate this financial model will be and therefore the greater the need to correct this through transformation. A successfully established form of transformation, which has been applied in financial modelling, is that based on the model developed by (Box and Cox, 1964). This model is considered to be a more effective alternative to the logarithmic approach e.g. see (Jones, Johnstone and Wilson, 2015b).

The Box-Cox approach involves a family of transformations; using maximum likelihood estimation in order to determine the values for the transformation parameter λ and thus minimise variance.

$$x = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & if\ \lambda \neq 0 \\ \log(x) & if\ \lambda = 0 \end{cases} \quad (19)$$

However, in the Box-Cox methodology, x represents strictly positive covariate values. In order to overcome this limitation, (Yeo and Johnson, 2000) proposed a family of distributions that can be used without restrictions on x , whilst retaining the properties of Box-Cox. These transformations are defined by:

$$\psi(\lambda, x) = \begin{cases} \{(x+1\rangle^{\lambda} - 1\}/\lambda & (x \geq \lambda \neq 0\rangle, \\ \log(x+1\rangle & (x \geq \lambda = 0\rangle, \\ -\{(-x+1\rangle^{2-\lambda} - 1\}/(2-\lambda\rangle & (x < \lambda \neq 2\rangle, \\ -\log(-x+1\rangle & (x < \lambda = 2\rangle. \end{cases} \quad (20)$$

If x is strictly positive, then the Yeo-Johnson transformation is the same as the Box-Cox power transformation of $x + 1$. If x is strictly negative, then the Yeo-Johnson transformation is the Box-Cox power transformation of $(-x + 1)$, but with power $2 - \lambda$. With both negative and positive values, the transformation is a mixture of these two, so different powers are used for positive and negative values.

3.4.3 Outliers

Financial ratios have long been prevalent explanatory variables used in FD prediction models. However, these features can exhibit heavily skewed distributions due to of the presence of outliers. This is indeed an issue with a relatively large set of globally diverse shipping company financial statements. Furthermore, as some models[12] are particularly sensitive to outliers it is therefore necessary to address the issue of outliers as a fundamental part of the data pre-processing stage (prior to modelling). (Tsai and Cheng, 2012; Nyitrai and Virág, 2019) document the most common approaches to this challenge as either the omission of observations or features containing outlier values or the winsorization[13] of such extreme values. As regards the omission of data this can result in diminishing the overall signalling effect of the remaining data thus introducing bias, especially if the data sample size is small. With small sample sizes, apparent outliers might be a result of a skewed distribution where there is insufficient data for the skewness to be apparent. Also, the outlying data may be an indication of an idiosyncratic section of the subject field that has only recently been subject of sampling. Finally, winsorization is considered unsuitable as it potentially suffers from the same drawbacks as simple omission.

Nevertheless, initial tests on the data financial data from the test case dataset indicated that spatial sign (Serneels, De Nolf and Van Espen, 2006) transformation, used in conjunction with centring and scaling of the sample data produced improved results over both omission and winsorization approaches. The spatial signed procedure projects the independent variables onto a multidimensional sphere which, in effect, places the variables at the same distance from the centre of the sphere. Every sample is divided by its squared norm and is depicted as:

$$x_{ij}^{*} = \frac{x_{ij}}{\sum_{=1}^{P} x_{ij}^{2}}. \quad (21)$$

This is a multi-variate operation, transforming all predictors as a group.

Centering and scaling the data prior to applying this transformation is necessary as the denominator is intended to measure the squared distance to the center of the predictor's distribution.

3.4.4 Class imbalance

This section discusses the approach to model tuning in order to increase the sensitivity of the minority class. Class or data imbalance is a frequent issue in classification modelling and is a condition where a significant majority of the training observations belong to one class. This is a challenge as classification algorithms are generally trained under the

[12] *Linear models and to a certain extent ANNs and SVMs*
[13] *Replace outliers with the largest value from those not considered outliers. Usually percentage based.*

assumption that class ratios in the training data are balanced. However, in real-world datasets this assumption is frequently violated. This is particularly the case with company default prediction where the training dataset contains thousands of company/year financial statements and where the majority class is heavily skewed towards the non-distressed entities.

The problem has received considerable attention in the literature (Kang and Cho, 2006; Wang and Japkowicz, 2008; Kim, Kang and Kim, 2015) as it is a major cause of degradation in the performance of classification algorithms. There are two main reasons behind this. The first is associated with the objective function of classification models which is the arithmetic accuracy. This is a ratio of the number of correctly classified observances over the number of total observances. However, in the presence of severe data imbalance, the data will misrepresent the arithmetic accuracy as the “accuracy” reported is highly dependent on the classification accuracy of majority class observations. In short, with very imbalanced samples, most standard classifiers will tend to learn how to predict the majority class. This is particularly the case in company financial performance assessment as bankruptcy is a relatively rare event[14] i.e. the arithmetic accuracy of the generated classifier will tend to be deceptively high due to the elevated accuracy for the non-bankrupt majority class. The second cause of degradation is the distortion of decision boundaries caused by imbalanced class distribution. With significant imbalance the classification boundary of the majority class tends to encroach on the decision boundary of the minority class. This results in a distortion of the class boundary in favour of majority class leading to a decrease in the accuracy for minority class.

There are two generally accepted approaches to the resolution of the imbalance issue. Firstly, if there exists a priori knowledge of a class imbalance, then the approach is to select a training set sample containing equal event rates at the initial data collection stage. In this case, as opposed to requiring the model deal with the imbalance, the sample frequency is simply balanced before modelling. However, in these circumstances, the test sample needs to be consistent with the real-world state and should reflect its natural imbalance so that a true estimate of the out of sample performance is achieved. Secondly, if an a priori sampling approach is not possible or considered unsuitable, then there are post hoc, in model training, sampling approaches to address imbalance. These include under-sampling, over-sampling, synthetic minority over-sampling (SMOTE) and more advanced sampling methods such as cost sensitive and boosting algorithms. Testing is limited to first three ad hoc sampling methods discussed briefly above. Up-sampling simulates or imputes additional data points to improve class balance, whilst down-sampling reduces the number of samples to improve class balance. SMOTE (Chawla *et al.*, 2002), is a data sampling methodology combining both up and down sampling utilising observations from the sample classes.

---

[14] *World Bank and OEDC figures report global long-term average commercial company default rates of around 4 to 6 percent.*

Finally, care should be taken when modifying training samples as resampled estimates can introduce bias. Despite this, as the results reflect, resampling methods can still be effective at tuning the models.

3.5 Classification algorithms

This section reviews the classification algorithms incorporated into the modelling of shipping company FD in the SFDP system. These algorithms include traditional linear based statistical tools as well as the more commonly used complex models. Artificial neural networks (ANN), support vector machines (SVM), random forest (RF), gradient boosting method (GBM) and extreme gradient boosting (XGB).

The results are then benchmarked against more traditional but nevertheless accepted GLM models used in FD prediction: a generalised[15] hazard model and a mixed effects model; GAM and MARS. The inclusion of these generalised linear models is to provide a balanced comparison with complex models. Their inclusion is performed in the name of model transparency as, following the much-quoted Ockham[16]'s razor principle, that if two models demonstrate similar predictive power then the model which is more transparent is preferable.

In this study, ML mechanisms are utilised to distinguish between financially distressed and non-financially distressed companies based on characteristics such as profitability, liquidity, leverage, size, and growth measures. This section briefly reviews each of these mechanisms, considering each one's specific goals, mathematical modelling, and learning algorithms. The research question surrounds the accurate identification of the category, financially distressed or non-financially distressed to which each observation belongs. This research tests the following algorithms on both the base model and full model:

3.5.1 Generalised linear models

The use of logistic regression has frequently been selected for benchmarking in many papers examining the use of ML models for company distress prediction e.g. see (Hernandez Tinoco and Wilson, 2013; Jones, Johnstone and Wilson, 2015b). The consensus is that there is arguably limited value for including it as a modelling tool for the purposes of benchmarking against the array of complex ML tools available today. Nevertheless, it remains the basis of other linear models which have been developed to relax the linear regression assumptions of the linear relationship the covariates and probability (logit of) financial distress. This research includes the use of some linear based models which have been demonstrated to perform well against the more complex ML tools. One issue frequently observed in the literature is the

[15] *We refer to both as generalised linear effects models (GLM)*
[16] *William of Ockham (c. 1287–1347); scholastic philosopher, and theologian.*

use of cross-sectional company, market and macroeconomic data when forecasting for financial distress modelling. However, this static approach does not adequately account for corporate distress fluctuation over time. As such it compromises the generalisation capability of linear models thereby rendering them poor competitors in any benchmarking exercise. The linear modelling described below also examines the capacity of models which do account for time dependent covariates.

Dynamic hazard model

The discrete hazard model is estimated with random effects $\alpha(i)$ and control for unobserved heterogeneity/shared frailty.

The logit link is specified as:

$$P(Y_{i,t} = 1|Y_{i,t-1} = 0, X_{i,t}) = \frac{e^{\alpha(t)+\alpha(t)_i^{'}\beta}}{1+e^{\alpha(t)+\alpha(t)_i^{'}\beta}} \quad (22)$$

Where $P(Y_{i,t} = 1)$ denotes the probability of an event for company i at time t and $\alpha(t)$ the baseline hazard rate. Regression parameters (β) are estimated through the partial likelihood function. As regards the baseline hazard specification, following (T. Shumway, 2001) the natural logarithm of a company's annual age is utilised.

Mixed effects model

The model used here is a mixed model including the usual mixed effects for the regressors plus the random effects.

Extending the generalised linear mixed effects model conditional mean, μ_{it} from

$$E[Y_{ij}|x_{ij}]$$

to

$$E[Y_{ij}|v_i, x_{ij}]$$

the form of the generalised linear mixed model predictor is obtained:

$$\eta_{it} = x_{it}^{'}\beta + z_{it}^{'}v_j \quad (23)$$

Where x is the vector of i predictors over t periods while β represents the mixed-effects regression coefficients. The variables exhibiting random effects are represented by vector z with v_j random effects.

The link function $g(.)$ is used to convert μ_{it} to η_{it}.

$$g(\mu_{it}) = logit(\mu_{it}) = \log\left[\frac{\mu_{it}}{1-\mu_{it}}\right] = \eta_{it}. \quad (24)$$

Hence the conditional probability of a response which includes random effects is

$$P(Y_{it} = 1 | v_t, x_{it}, z_{it}) = g^{-1}(\eta_{it}) = \Psi(\eta_{it}) \quad (25)$$

where the inverse link function $\Psi(\eta_{it})$ is the cumulative distribution function $\Psi(\eta_{ij}) = [1 + \exp(-\eta_{ij})]^{-1}$.

Multivariate adaptive regression splines (MARS)

As described earlier, MARS utilises the ordinary least squares (OLS) method to estimate the coefficient of each covariate. However, instead of a variable, each term in a MARS model is a basis function (BF) derived from the original variable. BFs describe the relationship between the predictor variable and the response. MARS then partitions the predictor values into groups, using recursive splitting, and a separate linear regression line is modelled for each group. The connections between the separate regression lines are called knots. The knot is the point at which the model extensions minimize a squared error. Each knot has two spline BFs. These are denoted as:

$$h^{+}(x;t) = [+(x-t)]_{+}$$

$$h^{-}(x;t) = [-(x-t)]_{+}$$

where $h(x)$ is the basis function and t is a univariate knot.

The model is subsequently constructed from N basis functions $\{h_n(x)_{n=1}^{N}\}$:

$$Y = \beta_0 + \sum_{n=1}^{N} \beta_n h_n(x) + \varepsilon \quad (26)$$

whereβ_n is the coefficient for the n^{th} BF.

Generalised additive models (GAM)

One of the main assumptions of linear regression models is that they require the covariates to be linearly related to the probability of FD (or logit thereof). However, GAMs (Hastie and Tibshirani, 1987) relax this assumption by accounting for the fact that some of the predictors exhibit a continuous, non-linear relationship with FD. Furthermore, non-linear relationships are observed both below and above specific thresholds with respect to shipping company's adjusted financial ratios and as such it is necessary to take account of these non-linear relationships.

Compared with GLMs, GAMs demonstrate superior regularisation capacity thus enabling them to more adequately address problems of overfitting. They also have an advantage over more complex models of being more interpretable and as such, GAMs represent an acceptable solution between the interpretable, yet biased, GLMs, and more complex, "black box" learning algorithms.

Our implementation of company FD prediction utilising GAMs follows along the lines of those documented by e.g. (Berg, 2007; Lohmann and Ohliger, 2017; Christoffersen, Matin and Mølgaard, 2018; Valencia *et al.*, 2019).

In short, GAM is an additive modelling technique where the predictive variable effects are acquired through smoothing functions which can be both linear and non-linear. The basic structure is:

$$g(E(Y)) = \alpha + s_1(x_1) + \cdots + s_p(x_p) \tag{27}$$

where Y is the dependent variable, $E(Y)$ is the expected value, and $g(Y)$ the link function between the expected value and the predictor variables $x_1, \ldots, x_p$.

The terms $s_1(x1), \ldots, s_p(x_p)$ represent non-parametric[17] "smoother", functions.

As the *smoother* function basis, (thin plate) regression splines are selected, which can be expressed as a linear combination of a finite set of basis functions which are not dependent upon dependent variable Y. A regression spline of order q is defined as:

$$s(x) = \sum_{l=1}^{K} B_{l,q}(x)\beta_l = B^0\beta \tag{28}$$

[17] *In this instance the term non-parametric means that the predictor function shapes are determined by the data as opposed to parametric functions that are defined by a set of parameters.*

where B is the model matrix of basis function, $B_{p,1}(x)$, . . . , $B_{p,K}(x)$ depict the basis functions and $\beta = [\beta_1 : \beta_2 : \ldots : \beta_p]$ are the coefficients. The number of basis functions is dependent upon the number of $inner\ knots$; a set of ordered, values of x_j and the spline order. So, if m denotes the number of inner knots, the number of basis functions $K = p + 1 + m$.

3.5.2 Complex models

Artificial neural network (ANN)

Artificial Neural Network have been extensively covered in the FD prediction business and finance literature[18].

They are described as being a non-linear discriminant model. The model is arranged in layers, which for binary classification consist of at least one input and two output class layers and one hidden layer. Each layer consists of one or more nodes, and there are weights to connect the nodes in different layers. The method has several variations in terms of possible algorithms. The most common and widely used back-propagation network is utilised.

For a typical single hidden layer binary neural network classifier, there are inputs (X), one hidden layer (Z) and two output classes (Y). Derived features Z_m are created from linear combinations of the inputs, and then the target Y_k is modelled as function of the linear combinations of the Z_m,

$$Z_m = \alpha(\alpha_{om} + \alpha_m^T X), m = 1, \ldots, M.$$

$$T_k = \beta_{ok} + \beta_k^T Z, k = 1, \ldots, K.$$

$$f_k(X) = g_k(T), k = 1, \ldots, K.$$

Where Z = $(Z_1, Z_2, Z_3, \ldots, Z_M)$, and T = $(T_1, T_2, T_3, \ldots, T_K)$.

The activation function, $(v)$is basically represented as $\frac{1}{1+e^{-v}}$.

The output function $g_k(T)$ allows a final transformation of the vector of outputs T.

For K-class classification the identity function $g_k(T)$ is estimated using the softmax[19] function (final ANN layer):

---

[18] (Tkáč and Verner, 2015) count 412 articles over 20 years and write that much of this research covers FD issues.
[19] *The softmax function is often used in the final layer of an ANN classifier. It represents a non-linear variant of multinomial logistic regression.*

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{k} e^{T_l}} \tag{29}$$

Whilst ANNs are considered capable of treating dynamic non-linear relationships they generally have less capacity in handling large numbers of irrelevant inputs, data of mixed type, outliers, or missing data. They also have a significant potential for over-fitting. However, the use of weight decay attenuates the size of the parameter estimates, resulting in smoother classification boundaries. Model averaging also help limit over-fitting. This is discussed further in the results discussion section.

Finally, and this is important in the world of finance, a major limitation of ANNs is transparency and they are frequently referred to as ‘black box’ algorithms.

Support vector machines (SVM)

Support vector achines (Vapnik, 1999) are a class of modelling techniques which were originally developed in the context of classification models. They have been extensively examined in the context of FD prediction (Min and Lee, 2005; Shin, Lee and Kim, 2005; Sun and Li, 2012; Zhang, Hu and Zhang, 2015; Kim, Mun and Bae, 2018).

In order to process non-linear decision boundaries, SVMs expand the feature space. This is achieved using various types of kernels such as linear, non-linear, polynomial, Gaussian kernel, Radial basis function (RBF) and sigmoid etc. This study utilises the RBF kernel. A hyper-plane divides p-dimensional space into two halves; where a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data point of any class.

Classification is based on the sign of the test observation. The ideal is completely separable observation sets, as this would allow SVM to build a model with 100% accuracy. In finance, this is practically impossible as economic variables are influenced by noise in empirical data and are often biased. For classification problems involving partially separable groups, the SVM method allows the inclusion of a margin of error. In general, the number of variables is not a constraint on the optimisation problem. The algorithm associated with the quantitative model establishes a classification mechanism, calibrating parameters using a training set (i.e. the algorithm learns from the training data). The resulting classification scheme can then be applied to predict the grouping or classification of new observations. The validation set is usually evaluated by comparing the classification given by the model with the actual group to which the observation belongs. The validation and training sets are independent: no observations are common between them.

The optimisation problem can be summarised as:

$$\text{Minimise } \tfrac{1}{2} w^{\mathrm{T}} w = C \sum_{i=1}^{M} \xi_i \tag{30}$$

Subject to $y_i \, [w^T \, \phi(x_i) + b] \; \geq 1 - \xi_i$

where i = 1 , 2 , . . . , M, $\xi_i \geqslant 0$ are the margins of error related to classification cost C, y_i are the classifications in the training set, and φ(x) transforms space $\mathbb{R}^M$. One advantage of this technique is that φ(x) does not need to be known, since a kernel function (K(x) = K(x_i , x_j)) is applied so that K(x) = $\varphi(x_i)^{T} . \varphi(x_j)$. The kernel function is pre-determined in the algorithm and a solution to the optimisation problem (Eqs. (1) and (2)). The traditional kernel functions are:

$$K(x_i, x_j) = < x_i, x_j > \tag{31}$$

and

$$K(x_i, x_j) = \; e^{(-\gamma \|x_i - x_j\|^2)} \tag{32}$$

where γ is a positive constant. Eq. (31) is called the 'linear kernel' and eq. (32) is the 'radial basis function' (RBF). The linear kernel does not provide strong predictability in non-separable datasets, due to the complexity of the empirical analysis, but the results are easily interpreted by users. Meanwhile, although the RBF kernel is difficult to analyse, or even discuss, it provides superior predictions in non-separable cases.

Unlike the more traditional classifiers e.g. LDA, logit and probit, SVM is relatively robust to observations with the greatest displacement from the hyper-plane i.e. less sensitive to outliers. However, the disadvantage of SVM is that they are susceptible to many of the same limitations as ANNs, particularly in terms of computational scalability, lack of interpretability and ability to handle irrelevant inputs and data of mixed type (Tian, Shi and Liu, 2012).

Random forest (RF)

Random forest (Breiman, 2001) is based on decision tree models or generalised classification and regression trees (CART). It has shown to be relatively robust and is particularly adept at handling outliers and noise in the training set. The importance of each variable in the classification outcome is identified by a RF. Therefore, it provides not only the classification of observations, but also information about the determinants of separation among groups. The RF technique repeatedly generates classification functions based on subsets. However, RFs randomly select a subset of characteristics from each node of the tree, avoiding correlation in the bootstrapped sets. The forest is built for several sub-sets that generate the same number of classification trees. The preferred class is defined by a majority of votes, thus providing more precise forecasts and, most importantly, avoiding data overfitting. Our RF fitting model follows (Jones, Johnstone and Wilson, 2015b):

1. Training data divided into B sets. For b=1 to B:
   i. A bootstrap sample of Z^*, size N is drawn from the training data.
   ii. An RF tree, T_b is 'grown' from the bootstrapped data by process of recursion for each terminal node of the tree until the min. node size n_{min} is attained:
   iii.
      - Randomly select m variables from the full p set of predictors.
      - Select the best variable split point from m.
      - Split the node into two child nodes.
2. Produce the ensemble of trees $\{T_b\}_1^B$

For a discrete outcome variable, let $\hat{C}_b(x)$ be the class prediction of the b^{th} RF tree. Then $\hat{C}_{rf}^B(x)$= majority vote $\{\hat{C}_b(x)\}_1^B$

Boosting I - Stochastic gradient boosting (GBM)

The basic principle GBM is: given a loss function e.g. squared error for regression and a weak learner e.g. regression trees, the algorithm seeks out an additive model that minimizes the loss function. The algorithm is typically initialized with the best guess of the response e.g. the mean of the response in regression. The gradient or residual is calculated, and a model fitted to the residuals to minimize the loss function. This model is used with the previous model and the process is repeated for a number of user-specified iterations.

The classification GB modelling takes the following form:

Initialized all predictions to the sample log-odds: $f_i^0 = \log \frac{-\wedge}{1-\hat{p}}$. (33)

for $j = 1 \ldots M$

Compute the residual (i.e. gradient) $z_i = y_i - \hat{p}$

Randomly sample the training data.

Train a tree model on the random subset using the residuals as the outcome.

Compute the terminal node estimates of the Pearson residuals:

$$r_i = \frac{\frac{1}{n}(y_i - \hat{p})}{\frac{1}{n}\hat{p}(1-\hat{p})} \tag{34}$$

Update the current model using $f_i = f_i + \lambda f_i^{(j)}$

When trees are used as the base learner, basic gradient boosting has two tuning parameters: tree depth and number of iterations. One formulation of stochastic gradient boosting models an event probability by

$$\hat{p}_i = \frac{1}{1+exp[-f(x)]}, \tag{35}$$

where $f(x)$ is a model prediction in the range of [- ∞, ∞]. For example, an initial estimate of the model could be the sample log odds,

$$f_i^0 = \log \frac{\hat{p}}{1-\hat{p}}, \tag{36}$$

where p is the sample proportion of one class from the training set.

The algorithm can be tailored by applying a suitable loss function and gradient. This algorithm can be transformed into a stochastic gradient boosting context by including random sampling at the start of the inner loop. Shrinkage can also be implemented as a closing step.

In summary, the finance literature, e.g. see (Florez-Lopez and Ramon-Jeronimo, 2015; Zhao *et al.*, 2016; Jones, 2017; Krauss, Do and Huck, 2017) attempts to demonstrate that that gradient boosting enhances the performance of conventional linear models. Moreover, (Jones, 2017) notes that the can function both an exploratory/diagnostic tool and as a “bias eliminating framework” to rank predictors as well as identifying including important non-linear relationships and interaction effects. He concludes that a logit model can enhance the analysis and improve predictive and explanatory performance.

Boosting II - Extreme gradient boosting (XGB)

Extreme Gradient Boosting (Chen and Guestrin, 2016) is an enhancement of Friedman’s stochastic gradient boosting model. However, it builds on the qualities of gradient boosting by providing a highly scalable model which incorporates regularisation in order to limit overfitting and is developed to uniformly handle sparse data. Furthermore, it is generally optimized for parallel processing i.e. was designed for both speed and performance. The model has drawn attention by being behind many recent winning entries in large scale big-data competitions such as Kaggle.

To date, published literature on the application of XGB in FD research is limited (Zięba, Tomczak and Tomczak, 2016; Chang, Chang and Wu, 2018; Carmona, Climent and Momparler, 2019). This most likely due to its relatively recent uptake in the social science research domain.

3.6 Classification model evaluation

The classification performance of each model/classification combination is performed using their respective area under the curve (AUC) of the Receiver Operating Characteristics (ROC). The ROC originated in the 1940's for use in radar signal analysis and one of its first recorded uses in ML was (Spackman, 1989).

However, ROC/AUC method has its limitations and as such H measures (Hand, 2009) are also employed in the evaluation of models. The H measure is a robustness check on the AUC results. This metric addresses the main problem associated with the AUC, that of the handling of misclassification costs across different classifiers. The AUC does not apply the same misclassification cost distributions to individual classifiers i.e. it utilises different metrics when evaluating different classification rules. And as such its use should be limited to the broad comparison of individual classifiers as an AUC may rank the individual models adequately but perform inadequately in terms of the level of the predicted probabilities.

The log loss function is also used to compare the calibrated probabilities. The log loss function measures the accuracy of a classification model by penalising false classifications. The basic premise is in minimising the log loss in order to maximise the accuracy of the classifier. In order to calculate log loss the classifier assigns a probability to each class in place of assigning the most likely class.

Mathematically log loss is defined as:

$$H_j = -\frac{1}{n}\sum_{i \in R}(y_i \log(\hat{p}) + (1 - y_i)\log(1 - \hat{p})) \qquad (37)$$

where H_{jt} is the model's log score (loss) of model j in year t, y_i is a dummy equal to 1 if company i financially distressed, $\hat{p}$ is the predicted probability of distress of firm i by model j, R is the sample of active companies and n is the number of companies in R. A perfect score is zero.

The Log Loss metric considers the probabilities underlying model models, and not only the final output of the classification. The stronger the probabilities correspond to a lower Log Loss. As log loss is a measure of entropy or uncertainty, a low log loss means a low entropy. The measure is similar to the Accuracy value derived from the confusion matrix, but it will favour models that more strongly the distinguish classes and useful for comparing not only model output but on their individual probabilistic outcome.

3.7 Summary

This chapter provided an overview of the theoretical framework which forms the foundation for the statistical models and techniques employed in this thesis. The chapter is split into four main sections reflecting the main stages of the SFDP process flows. The process commences with the data identification, collection, review and preparation stage. In statistical modelling/machine learning processing, this is probably the most important and resource consuming of all the steps in the process. The quality and usefulness of the information derived from this stage directly affect the ability of models to learn. The discussion of data preparation in this chapter was separated into section 3.3 Missing value imputation and section 3.4 Data pre-processing. The missing value issue would ordinarily be treated as an integral part of the pre-processing stage. However, the complexity of missing data handling methodologies, as well as the relatively large, longitudinal dataset and the high percentage of missing values, warrants a distinct discussion in order to highlight the magnitude of the task faced in correctly managing the problem. Furthermore, pre-processing of data prior to imputation risks the “lock-in” of any bias which is resident within the original dataset.

The chapter closed with section 3.5 Classification models, explaining the theory behind the FD prediction algorithms which form at the core of the SFDP, and section 3.6 Classification model which describes the evaluation techniques and metrics necessary to validate the model.

In conclusion, the chapter described the theoretical background which provides the foundation for the SFDP system architecture described in chapter 4.

4 System software and architecture

4.1 Overview

The Shipping Financial Distress Prediction (SFDP) system architecture and component modules[20] are described here. The chapter is split into four main sections, respecting the flow and structure of the theoretical framework described in chapter 3, namely: multiple imputation; data pre-processing; model tuning and training; and model testing and evaluation. The SFDP system specification overview is presented in Figure 4 outlining the four modules and their respective functional components.

4.2 Architectural foundation

The SFDP system is based on the principles of the "Reactive Architecture" (Bonér *et al.*, 2014). Also referred to as reactive systems, the aim is to develop responsive, resilient, elastic, and message driven systems (see Figure 5):

- Responsive systems enable rapid and consistent response times. This is essential when training multiple models, which may involve a significant time lapse (many hours) when multiple classifier models are involved.
- Resilient systems are responsive following failure of individual components. System failures should be contained within each component and so isolating components from each other ensure that elements of the system can fail and recover without compromising the whole system.
- Elastic systems stay responsive under varying workload. They react to variations in the input rate by increasing or decreasing the resources allocated to service these inputs. They achieve elasticity in a cost-effective way on commodity hardware and software platforms.
- Message driven systems rely on asynchronous messaging between components providing loose coupling, isolation and location transparency.

Whilst the message driven aspects are not yet part of the SFDP system, it can be readily included in later releases.

[20] *The term, "module", in this instance, refers to a standalone block of code that provides a set of specific and tightly coupled functionality, which defines and enforces logical boundaries within the said module.*

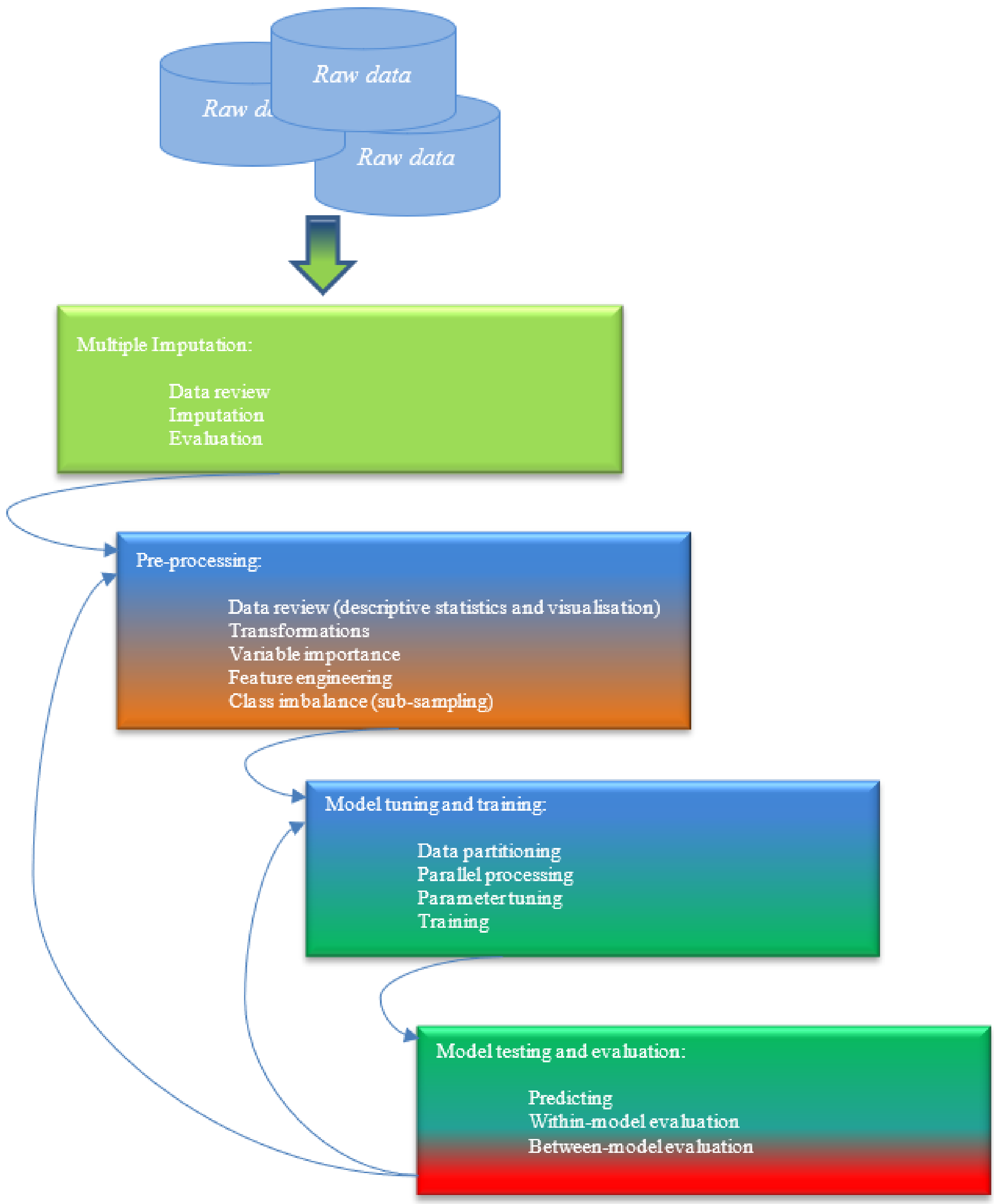


Figure 4: SFDP system specification overview

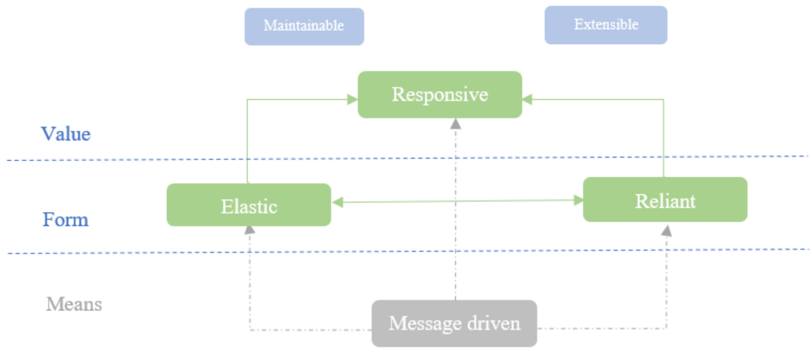


Figure 5: The "Reactive Architecture"

The SFDP system is written in the R(R Core Team, 2000) programming language. R is a widely used and well-established computer language and environment used for statistical computing and graphics. R provides a range of statistical linear and non-linear modelling, tests, time-series analysis and graphical techniques. It is also highly extensible and runs on a wide variety of operating systems such as UNIX, FreeBSD and Linux as well as Windows and MacOS.

The SFDP functionality also takes advantage of many of the statistical, modelling and peripheral support, *packages* (Wickham, 2015). These *packages* are collections of functions and data sets written in R by third party statisticians and machine learning experts. For a view of how the SFDP system fits into a corporate software architecture see Figure 6.

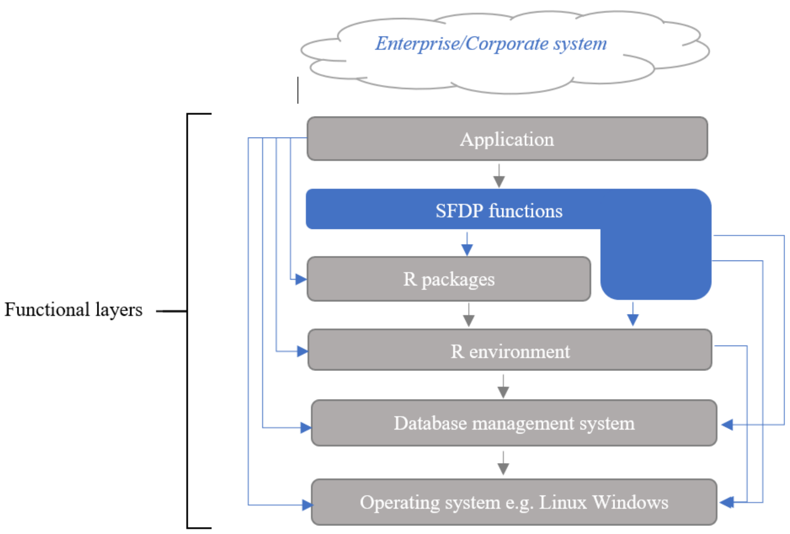


Figure 6: Corporate functional software architecture

4.3 SFDP functional composition

The SFDP function suite is composed of five main conceptual modules which implement a set of data import, missing data handling, pre-processing, classification and evaluation functions. The composition of the system modules and function is depicted in Figure 7.

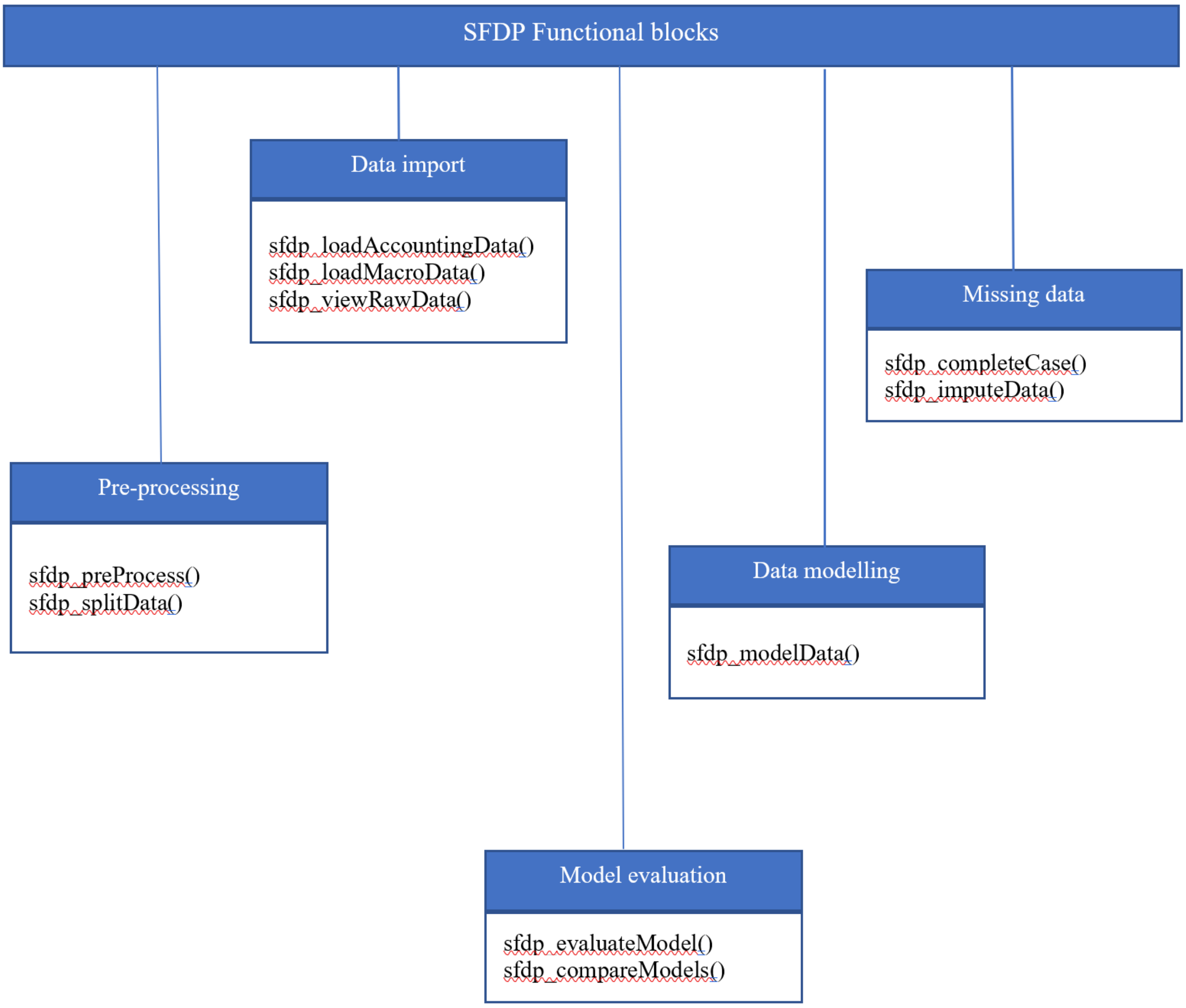


Figure 7: SFDP function block composition

4.4 Documentation

SFDP's documentation generated using the Roxygen2[21] package for documentation in R, along with custom extensions. The documentation pages include embedded code extracts explaining application. The code is dynamically extracted from test cases in the source code, ensuring that the extracts are automatically updated. The print output and plots from the extracts are also dynamically generated through custom Roxygen extensions and embedded in the documentation page. Similarly, additional customisation embeds a table of options, default values, valid types, values, and descriptions in the model documentation.

4.5 Summary

The SFDP toolkit is a set of functions that risk assessment teams can use to build financial risk assessment software applications for shipping company distress prediction, with limited experience of the underlying technology. The toolkit, constructed using the methodologies described in chapter 3, is extensible and can be used on multiple software platforms. The main novelty at the core of the system is in its focus on missing data management, which to the authors knowledge, is the first of its kind.

In order to test and evaluate both the methodologies and the toolkit a test case derived from the bulk shipping market is presented in chapter 5. The results of this test case are presented in chapter 6.

[21] For more information ref: https://cran.r-project.org/web/packages/roxygen2/index.html

5 Case study: The bulk shipping sector

5.1 Overview

The bulk carrier fleet is an essential element of the global economy. No other single mode of transport has the capacity to deliver the raw material, neither the nature of, nor quantities, that are required to sustain global economic development. In 2018, the world fleet totalled approximately 116,857 ships (Equasis, 2019), with bulk carrier vessel accounting for approximately 12,000 vessels or 10% of the global fleet. The dry bulk market is very diversified and volatile. It comprises three major sectors: iron ore and steel, coal and food and over 30 other commodities. In the same year, the bulk fleet took the delivery of 26,7% of the total gross tonnage, more than any other vessel type, followed by oil tankers (25%), container ships (23,5%) and gas carriers (13%). The three major dry bulk commodities represented more than 40% of total dry cargo shipments in 2018 (UNCTAD, 2019). Containerized trade contributed with 24%, minor bulks with 25,8% and the remaining volumes consisted of dry cargo including break bulks.

The size and diversity of bulk carrier owner/operating companies, coupled with the bulk fleet's equally diverse vessel types and cargos, identifies them a valid representative subset of global shipping commercial entities. This, therefore, qualifies the subsector as an appropriate case study for the SFDP model.

5.2 Empirical context

The case study develops three main ex-ante models for FD prediction comprising three sets of predictive features: company level predictors, including financial statement ratios; macroeconomic indicators; and bulk shipping market features. In model 1, the independent variable selection is taken solely from company level predictors. Model 2 adds macroeconomic indicators to model 1. Model 3 combines bulk carrier market specific indicators with model 1. Finally, model 4 includes all three sets of features.

The next step is to complete an analysis of missing company level values in the raw dataset. This dataset is then subject to both case wise deletion and data imputation techniques (see sections 2.3.2 and 3.2) in order to examine corresponding model performance. This is performed prior to any other pre-processing operations so as to avoid locking in any bias within the original raw data.

Following the missing value stage, all the data are subject to pre-processing by applying variance stabilizing and skewness transformation. For the testing of linear based models, however, further data pre-processing is required to account for the assumptions previously addressed in the literature review (2.2.2). This is essential as it provides for a more balanced, level playing field when comparing linear based models with more complex ML algorithms. Hence allowing for a more thorough evaluation of the full potential of linear based modelling.

5.3 Data sources

The raw dataset used for company level financials is sourced from unconsolidated statements sources from the Orbis company database (Bureau-Van-Dijk, 2019) and consists of over 5000 global dry bulk shipping company year financial statements for the period 2000-2018. The shipping specifics are primarily generated from Clarkson's, Shipping Intelligence Network (Clarksons, 2019) whilst macroeconomic is data drawn from various data sources: OECD database (OECD, 2019); the World Bank (World Bank, 2019); IMF (International Monetary Fund, 2018).

Filters are applied at company level in the raw data to exclude financial companies; such entities differ from other corporates particularly as regards their asset base, accounting standards and regulatory status. Furthermore, in order to avoid modelling distortions, holding companies are also filtered where they do not demonstrate that their holding entities prime business drivers are bulk shipping,

There is no filtering on company size as there is a requirement to account for interactions between size and other variables in the models, thereby allowing for the modelling of companies of different sizes.

5.4 Dependent variable – Outcome and hypotheses

The dependent variable is a binary variable, FD, representing the state (distressed or not distressed) of the company in any discrete accounting period. The definition of FD in companies follows on from (Pindado, Rodrigues and de la Torre, 2008) and outlines the following primary conditions to be fulfilled in predicting company financials distress. The hypothesis follows that a company is distressed when any one of the following events occur i) the company's EBITDA to expenses ratio is short of its expenses for two consecutive years; ii) the company suffers from negative growth for two consecutive years; iii) when a formal default event has been triggered; iv) failed to publish accounts for the following year (Christoffersen, Matin and Mølgaard, 2018). This definition also implies that companies experiencing FD in a single period can recover; therefore, recurrent events are implicitly modelled.

5.5 Independent feature selection

The independent variable selection in this study was primarily driven by the specific nature of the dry bulk shipping sub-sector. The information was quantified through the inclusion of company level features (financial and non-financial) as well as market and macroeconomic characteristics. The sector's risk framework can be largely described through financial features relating to the capital intense and cash flow dependent nature of the industry and through market and macroeconomic features which reflect a highly cyclical sub-sector with a high sensitivity to; global and regional economic growth; fuel prices; the balance of supply and demand amongst others. This framework is developed using a feature set common to corporate risk assessment, which is based on the aforementioned empirical studies, but also includes a set of predictors through which it attempts to capture dry bulk company idiosyncrasies, including but not restricted to the following:

- With capital intensive companies such as shipping, close attention should be payed to the level and structure of debt. This includes the increasingly prevalent use of sale and leaseback as a finance option in the acquisition of vessels.
- Price competition is intense and is driven by the commodity nature of the bulk sector. This can lead to a periodic oversupply of tonnage. Furthermore, peak-to-trough price declines can be significant.
- Technological risk in the form of air and waterborne pollutants is becoming a major issue in the face of increasing environmental regulation.
- Flag state risk factors play a significant role in determining company FD risk. Therefore, the more an individual flag state regulation reflects the globally accepted environmental, safety and employment and financial regulatory standards, the lower the risk.
- The overall marketability of a company fleet is enhanced by a modern, technically advanced vessel base. This is an issue with ever increasing environmental regulations which in turn result in an expanding environmentally sensitive customer base.
- The breadth of the route network affects a shipping line's market position make it more attractive to global customers as operators with a global route network have a competitive advantage over regional players. The International shipping market is highly fragmented, with the largest operators having a relatively modest share of the market. However, companies that operate purely domestically may be protected by cabotage laws aimed at restricting competition. Having a route network with broad geographic coverage can serve as a natural hedge against weak demand and help an operator ride out cyclical downturns.
- The diversity of a company’s fleet, in terms of vessels size and type can improve end market and customer demand. Operators with multiple classes of vessels (tankers, containerships, and bulk commodity ships) and of various sizes, or those that participate in commercial pools, can carry a broad range of commodities and attract a more diverse customer base. A diverse customer base of reputable charterers limits counterparty risk and adds stability to revenues.
- Companies with a high degree of operating efficiency should generate relatively better profit margins during all market conditions. Cost structure, as a measures of asset utilization and efficiency (revenue or cost per unit of capacity), and operating profit margins are important efficiency indicators. Emphasis should be placed on

operating cost structure as it tends to be a more consistent differentiating factor than revenue generation, which varies with market conditions.

Based on the above, a set of independent variables were selected and tested. Finally, with the context of the above discussion the predictor variables are separated into three distinct groups:

i. Company level predictors, including financial and non-financial indicators.
ii. Dry bulk market/macroeconomic indicators
iii. Global macroeconomic indicators

The full list of explanatory variables can be found in Table 2.

5.5.1 Company level predictors

5.5.1.1 Financial predictors

The financial ratio mix selected in this study follows recent standard practice in generic company default prediction (Hernandez Tinoco and Wilson, 2013; Jones, Johnstone and Wilson, 2016; Barboza, Kimura and Altman, 2017). The selected indicators are those most widely used by banks and ratings agencies (Standard & Poor's, 2019). Furthermore, when selecting this feature set, careful consideration was taken to account for the idiosyncratic nature of financing issues within the bulk carrier shipping sector e.g. see (Grammenos, Nomikos and Papapostolou, 2008; Lyridis, Manos and Zacharioudakis, 2014; Kavussanos and Tsouknidis, 2016; Wang, Woo and Mileski, 2016; Lozinskaia *et al.*, 2017; Standard & Poors, 2019). This resulted in the selection of the following financial statement ratios: i) liquidity and solvency; ii) earnings and profitability; iii) cash flow; iv) growth/change indicators (ROE); v) leverage and capital structure ratios; vi) activity ratios and vii) investment in capital expenditure to total assets.

Table 1: Corporate ownership concentration (based on Bureau van Dijk Independence indicators)

Indicator	**Sub-level**	**Notes**
Low ownership concentration Independent companies - those with known recorded shareholders, each of them having less than 25% of direct or total ownership of the company	**1**	Companies with six or more shareholders and/or companies in whose case the sum of direct ownership is above 75%
	2	Companies with 4 or 5 shareholders and/or companies that are the ultimate owners of another company (given that the information is included in a source), even when its shareholders are not mentioned.
	3	Companies with 1 to 3 shareholders
Medium-low ownership concentration Companies with known recorded shareholders with ownerships below 50%, but with one or more shareholders with ownership percentages above 25%	**4**	Companies with six or more shareholders and/or companies in whose case the sum of direct ownership is above 75%
	5	Companies with 4 or 5 shareholders and/or companies that are the ultimate owners of another company (given that the information is included in a source), even when its shareholders are not mentioned.
	6	Companies with 1 to 3 shareholders
Medium- high ownership concentration Companies with known recorded shareholders that have a total or calculated ownership above 50%	**7**	Companies with a sum of direct percentage of ownership is 50.01% or higher
	9	Also assigned to companies in whose case an ultimate owner is mentioned in a source, although its ownership percentage is known
High ownership concentration Companies with a recorded shareholder that has a direct ownership above 50%	**9**	
Unknown concentration Companies with an unknown degree of ownership concentration	**10**	

Table 2: Bulk carrier case study: Predictive feature set

Category	Sub-category	Feature	Description
Company			
	Financials	ROE	Return on equity (ROE)
		ROA	Return on assets (ROA)
		ProfitM	Profit margin
		GrossM	Gross margin
		EBITDAM	EBITDA margin
		EBITM	EBITM
		NetAssetT	Net asset turnover
		Current R	Current ratio
		LiquidityR	Liquidity ratio
		SolvencyR	Solvency ratio
		Gearing	Gearing (debt/equity)
	Non-financials	Age	Age
		Employees	Registered full time employees
		BvDII	Bureau van Dijk Independence indicator
Dry Bulk market		1YTC	1-year TC
		3YTC	3-year TC
		OrderBook	Order book / Total fleet
		NBPdex	New build price index
		SHPdex	Second-hand price index
		LaidUp	Inactive tonnage / Total Fleet
		Scrap	Scrapping rate
		HFOSpot	HFO (spot)
		MDOSpot	MDO (spot)
		WSTOre [22]	WSTOre (iron)
		WSTCc	WSTCc (coking coal)
		WSTSc	WSTSc (steaming coal)
		WSTGr	WSTGrain
		WSTMinor	WSTMinor (minor bulk)
		BDI	BDI (price)
Macroeconomic (core)		GDP	Real GDP / real GDP Growth
		LTI	Interest rate (short term)
		STI	Interest rate (long term)
		Inflation	Inflation
		debtToGDP	Public Debt / GDP
		Unemployment	Unemployment
		Insolvency	Company bankruptcy rates
		Copper	Copper (COMEX)
		SteelDex	DJUSST (Dow Jones historical iron & steel price)

[22] *Features DBM08-11 are World Seaborne Trade figures – Clarksons 2019*

5.5.1.2 Company level non-financial predictors

Non-financial specific company characteristics are described in this section. Company age, employee numbers, charter policy, corporate governance predictors are included.

Company age

The period taken to create and establish a shipping company is a critical component in assessing its likelihood of success. It is hypothesised that a company's age plays a role in the prediction of FD and that younger[23], less experienced companies, typically demonstrate more of a tendency to FD than do their more established counterparts (Pompe and Bilderbeek, 2005; Kavussanos and Tsouknidis, 2016). Furthermore, younger companies have an obvious disadvantage of being less able to generate a sufficient balance of retained earnings to protect the organisation from financial difficulties.

Employment

This company employee count is a control variable common in company FD models in the literature. It is used as a proxy for firm size.

Corporate governance

Much research has been conducted in the field of corporate governance structures e.g. see (Daily and Dalton 1994; Andreou, Louca, and Panayides 2014; Giannakopoulou, Thalassinos, and Theodoros 2016; Lozinskaia et al. 2017;) with the general consensus being that corporate structures have a concrete part to play in company FD.

It is widely accepted that strong governance structures enhance company operating and financial performance as well as aid in mitigating agency risk. Furthermore, previous studies also point to a larger board size being more optimal in maximising efficiency in shipping companies.

[23] *However, attention must be taken in assigning an appropriate "age" to shipping companies which can be identified as "spinoffs" of established organisations.*

Corporate ownership concentration indicators (see Table 1) are used as a company level corporate governance proxy in order to attempt to capture and quantify corporate governance effects. The values reflecting ownership concentration and range from 1 (low) to 8 (high). Previous studies have also established a link between effective corporate governance and ownership concentration. Based on this system (Horobet *et al.*, 2019) have reported that companies demonstrating a high to medium level of ownership concentration display lower performance volatility, compared with companies having lower degrees of concentration.

However, on a note of caution. In the case of shipping companies, a lack of transparency brought about by an elaborate system of ship registration; a nebulous beneficial ownership structure; and the frequent use of financial "off-shoring" renders a complete assessment of individual shipping company corporate governance particularly challenging. Therefore, only a reasonably approximation can be achieved on this control variable but in a manner which at least "underestimates" rather than exaggerate its potential predictive effect.

Charter policy

A company policy of incorporating a preference for time-charters for its vessels as opposed to operating them in the spot markets is inherently more stable and provides for more predictable cash flows (Kavussanos and Tsouknidis, 2016). A dummy variable ranging from 1 (only time charter) to 0 (only spot market) is employed. This variable is intended to capture company risk aversion towards the freight market. It also reflects their expectations towards the shipping market conditions.

5.5.2 Macroeconomic (core)

Previous literature has demonstrated that macroeconomic features do contribute to company financial performance (Figlewski, Frydman and Liang, 2012; Jones, Johnstone and Wilson, 2015b). Here the hypothesis is that eight broad macroeconomic indicators relating to the global economic status of the economy and financial market conditions are related to FD. These features[24] are:

- Real GDP/Real GDP growth
- Interest rates (short and long term)
- Inflation rates

[24] *Sources: Figures are acquired global basis. They are cross referenced through the World Bank & OECD. USD based statistics have been selected as a basis, taking into consideration that the majority of shipping finance is contracted or fixed to this currency.*

- Public debt to GDP
- Unemployment rates
- Insolvency (company bankruptcy rates)
- Copper prices
- Steel prices

Real GDP is an inflation adjusted measure of the value of goods and services and is a widely quoted indicator overall economic growth. The real GDP growth rate is expressed as a percentage that indicates the rate of change in the real GDP year on year. A strong and growing economy is positively correlated with lower default risks.

Interest rates affect the economy through consumer and business spending, inflation and growth. They also influence equity and bond prices. High interest rates negatively affect consumer and business spending and make it difficult for businesses to borrow funds for growth. In turn, high rates have a particularly negative affect on highly leveraged, capital intense sectors such as shipping which increase the pressure on debt servicing and cash flows. The interest rates are lagged on a 12-month period to reflect the period taken for the full effects of rate changes to be reflected in the economy.

Inflation is the rate at which the level of prices for goods and services is rising and consumer purchasing power falling. It is a heavily cited economic indicator. High or exceptionally low (deflation) inflation is viewed as negative for economic outlook. High inflation is viewed as symptomatic of weaker economic outlook, leading to increased risk of default.

The debt-to-GDP ratio compares a country's public debt to its GDP. It indicates a country's ability to pay back its debts by comparing what a country owes with what it produces. A high ratio of public debt to GDP is a generally considered sign of broad economic fragility. This indicator is therefore expected to be positively associated with increased default rates.

Unemployment rates are a strong barometric indicator of economic health with higher rates obviously associated with business failure rates.

Commercial bankruptcy rates are viewed as a firm indicator of general economic conditions. Peaks in failure rates are associated with recession, a severe reduction in commercial lending or market crashes. Hence higher failures rates are associated with poor economic conditions and lead to a higher probability of default.

Finally, copper and steel prices have been included as covariates. Steel is one of the world economy's fundamental materials used in construction, transport and a wide array of industrial and domestic appliances. Copper is considered a reliable leading indicator of economic strength. It has widespread applications in most sectors of the economy both domestically and industrially and is used extensively in electronics and power generation and transmission.

5.5.3 Dry bulk market

The following is a short description of the features selected for the market prevailing over the period of interest.

Some of these market indicators can be considered macroeconomic but they are included in the "dry bulk market" grouping as they are specific to bulk trade/business/commodity interests. Hence, an attempt is made to introduce a clearer differentiation in the modelling between these features and "core" macroeconomic indicators.

Dry bulk "stocks" (Time charter 1 & 3 year and BDI)

Consistent with (Grammenos, Nomikos and Papapostolou, 2008; Kavussanos and Tsouknidis, 2016) the bulk time-charter rates are utilised in order to capture the bulk shipping market conditions prevailing during each accounting period captured. Freight rates directly affect cash-flows necessary to cover operational costs service debt. They also affect asset value which is an integral financial covenant (loan to value) in shipping finance contracts.

The Baltic Dry Index (BDI) is an index generated through 20 key dry bulk routes, measured on a time charter and voyage basis. The BDI that is published by the Baltic Exchange and covers Handymax, Panamax, and Capesize dry bulk carriers transporting a range of commodities including coal, iron ore and grain. The BDI is a widely accepted barometer for the bulk shipping market.

Fleet level covariates

The amount of shipping tonnage (dwt) available, on order, inactive (laid-up), and market for demolition all have an effect on the bulk market (Kavussanos and Tsouknidis, 2016). Here, the following covariates are introduced; i) the new build shipping order book, captured through the order book/fleet total ratio, which has an effect on future freight rates and therefore cash flows; ii) the ratios laid up tonnage/fleet total and vessel scrapping/fleet total. The hypothesis here is that both inactive tonnage and vessel demolition rates indicate diminished freight rates which adversely affect cashflows.

A Newbuilding Price Index (Clarksons, 2019) is also employed. This is formulated by averaging the US$ per dwt values of new build prices for bulk carriers and used as a percentage year on year change.

Commodity and fuel prices

The market for bulk carrier transportation is very much derived and dependent upon the demand for core commodities such as iron ore, coke and agricultural basics. Demand for core commodities can be volatile and sensitive to regional economic, geopolitical, environmental and climate fluctuation. This volatility has a direct influence on freight rates and hence the cash flows of shipping companies. Our commodity covariates include the annual pricing of the following:

- Iron ore
- Coking coal
- Steaming coal
- Grain
- Minor bulk

Depending on ship types and types of operations, marine fuel costs (IFO and MGO) can represent upward of 50% of the total vessel operating costs (Kavussanos and Visvikis, 2016). Therefore, any study on shipping company financial health should include marine fuel pricing.

5.6 Summary

This chapter opened with an economic overview of the bulk shipping sector and highlighted the importance of the bulk sector to the global economy and overall supply chain. The company ownership/operating and financial structure, as well as the diverse bulk fleet and cargo composition, renders the sub-sector an accurate representation of the commercial shipping sector. The structure and character of the sub-sector was then addressed in order to establish an appropriate identification and definition of both the dependent variable and the independent predictor variables, required for bulk shipping FD prediction.

This chapter concludes by describing the composition of a commercial shipping entity dataset which can act as a reliable testbench for the SFDP model and toolkit.

6 Results and discussion

6.1 Overview

This chapter presents and discusses the results from the SFDP bulk shipping case study. The discussion begins with section 6.2 which briefly outlines the dataset structures used as input into the individual and complete data model tests. Section 6.3 addresses missing value assessment and management. The section starts by examining the raw data and then presents and discusses the performance of imputation methodologies, using the missing shipping company financial data. Section 6.4 examined the effect of each of the predictor variables on the FD dependent variable. The results of the application of individual classifiers, on each of the three covariate data sets (company, macroeconomic and market, as described in section 5.5) are presented in section 6.5. This section also covers the pre-processing and parameterisation (training and cross validation) stages of each classifier prior to presenting the FD prediction results. The results of classification tests using the final, complete data model are presented in section 6.6. A summary of the chapter is provided in section 6.7.

6.2 Test data structure

The test plan consists of four high-level data model structures comprising the three covariate datasets plus a complete model which combines all three datasets. Data model 1 represents company level data (see section 5.5.1); model 2 include company level feature set plus macroeconomic data see section (see section 5.5.2); model 3 includes model 1 plus bulk market data (see section 5.5.3) and finally model 4 represents a full model including all these datasets.

Model 1, the company level dataset, is used to perform the analysis of missing values and to test the performance of missing value management options including complete case treatment and data imputation techniques. It is the only dataset with missing values.

All four data models are then used to test

i. the performance of the individual classification modules
ii. the contribution of each of the covariate datasets to the prediction of FD

6.3 Missing value - analysis and management

As discussed extensively in previous chapters, missing value analysis and management is a core component of this research. This section starts by examining the raw company level dataset in order to review the extent of the missing value problem (section 6.3.1). This will inform the decision-making process concerning the missing data management.

Following the initial data analysis, a simulated dataset is generated from the bulk carrier dataset with missing values replaced by values generated through the observed values in the dataset. The outcome of this phase is a view on the performance of a set of various imputation algorithms. This is described in section 6.3.2. Finally, a subset of the tested algorithms is selected for integration into the final SFDP system for the case study evaluation.

6.3.1 Data analysis

The structured analysis of any dataset with missing values is necessary to avoid seriously compromising inferences which may be derived from the observed information contained within that dataset. Potential bias due to missing data depends on both the missingness mechanism, which was behind the prior loss of data, and the analytical methods applied to treat the missingness.

The first objective was to analyse the financial statement raw data in order to ascertain the extent of the missingness i.e. which individual features were affected and to what degree. The analysis summary in Table 3 shows that of 5,368 company financial statements collected, only 1,483 were fully complete, meaning that approximately 72% of financial statements are only partially complete. However, at the individual financial ratio level, the missingness level is 17.6% with 10,405 out of 59,048 accounting ratio values not recorded in the dataset. A breakdown of the missing values on an individual ratio level can be found in Table 4.

The results demonstrate that there is a relatively high level of observed accounting values, 82.4%. This indicates that, if the MAR assumption is applied, there is sufficient information present in the observed values for multi-variate imputation to yield beneficial results (in terms of reduced bias and efficiency) when compared with complete case treatment. Moreover, a complete case treatment option would result in only 27% of the financial statement observation being available for analysis. This would result in a loss of 32,300 observed financial ratio values which are present within the 3,885 incomplete financial statements, which contain significant levels of potentially exploitable information.

Table 3: Missing financial statement value analysis

Company financial statements (annual)	5,368			
Financial ratios	11			
Company accounts				
Complete financial statements	1,483	containing	16,313	observed financial ratio values
Incomplete financial statements	3,885	containing	32,330	observed financial ratio values
% complete observations	27.6%			
% incomplete observations	72.4%			
Financial ratio values				
Total ratio values	59,048			
Missing ratio values	10,405			
Observed ratio values	48,643			
Fraction missingness	17.6%			

Table 4: Missing value level per accounting ratio

Accounting ratio	**Missing values**	**% missing**
ProfitM	2839	53%
GrossM	1617	30%
EBITM	1586	30%
Gearing	1241	23%
EBITDAM	862	16%
ROA	806	15%
NetAssetT	676	13%
ROE	481	9%
SolvencyR	128	2%
LiquidityR	104	2%
CurrentR	65	1%

Cut-off thresholds for the levels of missing data where not utilised in this analysis. The literature (Jakobsen *et al.*, 2017; Madley-Dowd *et al.*, 2019) suggests that such thresholds have a limited evidence base to support them. The limited number of studies which have investigated bias and efficiency, in datasets containing increasing levels of missingness, produced inconclusive results.

A matrix plot of the missing data distribution is represented in Figure 8, with grey denoting missing data. This provides a visual description of the missing data pattern and what proportion of data is missing.

A plot of the pairwise correlation point-biserial correlation coefficients between covariate pairs is shown in Figure 9. Variables are assigned TRUE or FALSE based on their missing data status and these Boolean vectors are correlated to the native variables. The resulting correlation matrix has rows with variable names + _is.na - the rows represent the converted variables, while the columns are the original variables.

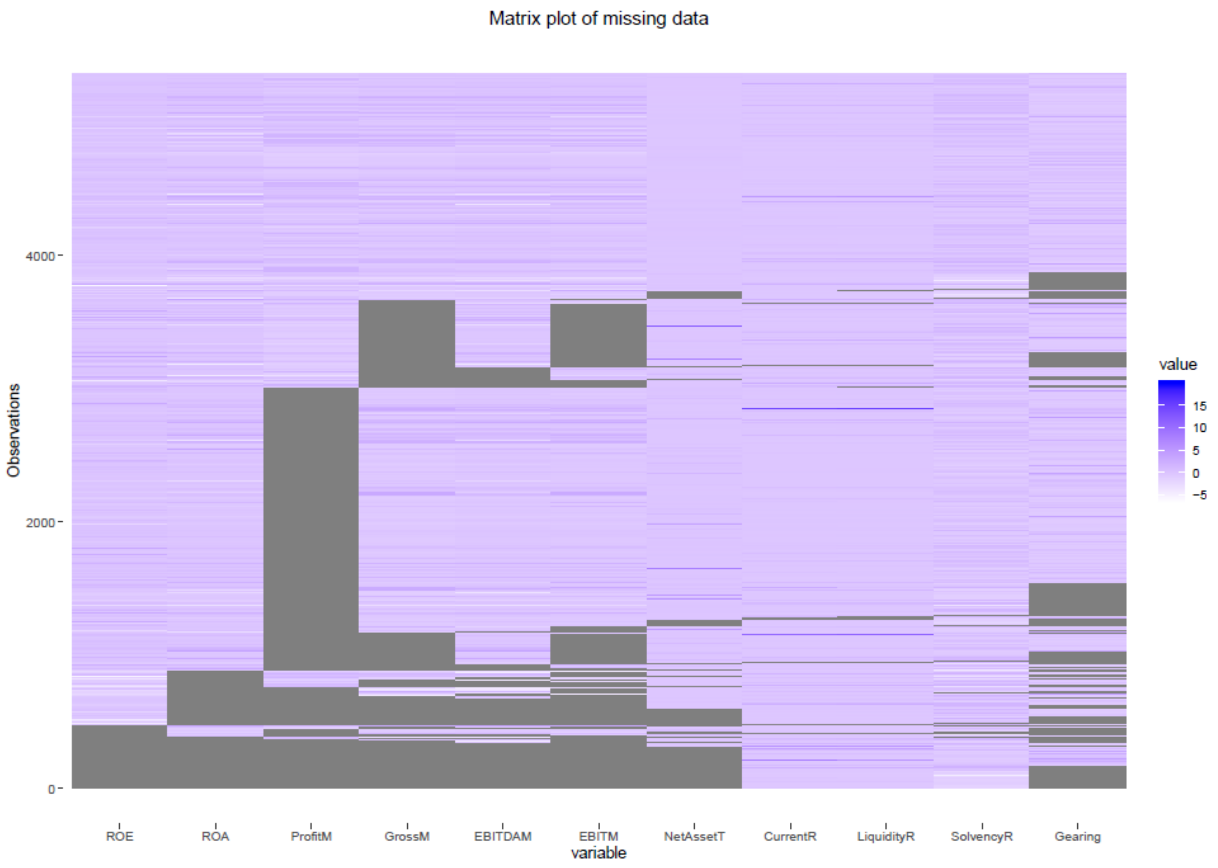


Figure 8: Matrix plot of missing accounting data

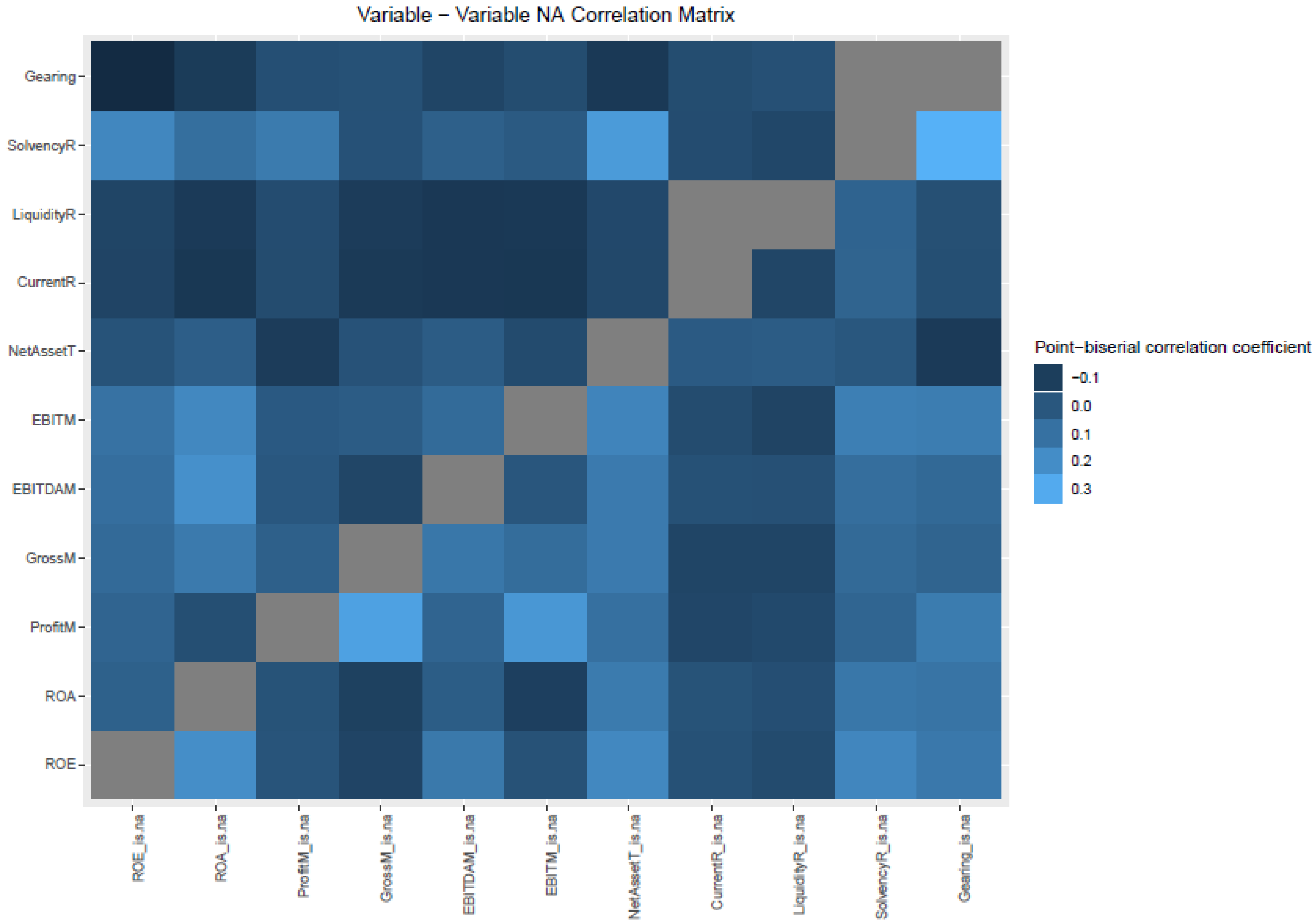


Figure 9: Raw accounting data - observed v missing (NA) correlation coefficients

6.3.2 Imputation algorithms – Simulated data evaluation

Having analysed the missing data, the next step is to develop a platform for testing the performance of various missing data methodologies on simulated data. This is achieved through the creation of a dataset which has no missing values, and which is based on the original data frame. The dimensions of this simulated data, including correlations between variables, correspond to the original data set.

Missing values are then replaced in the simulated data using the MCAR, MAR, MNAR and MAP patterns described earlier. With MCAR NAs are replaced randomly. MNAR, NAs are replaced based on the same covariate values while with MAR, the NAs are replaced based on all other covariate values. MAP represents a combination of any of the previous three missing data patterns.

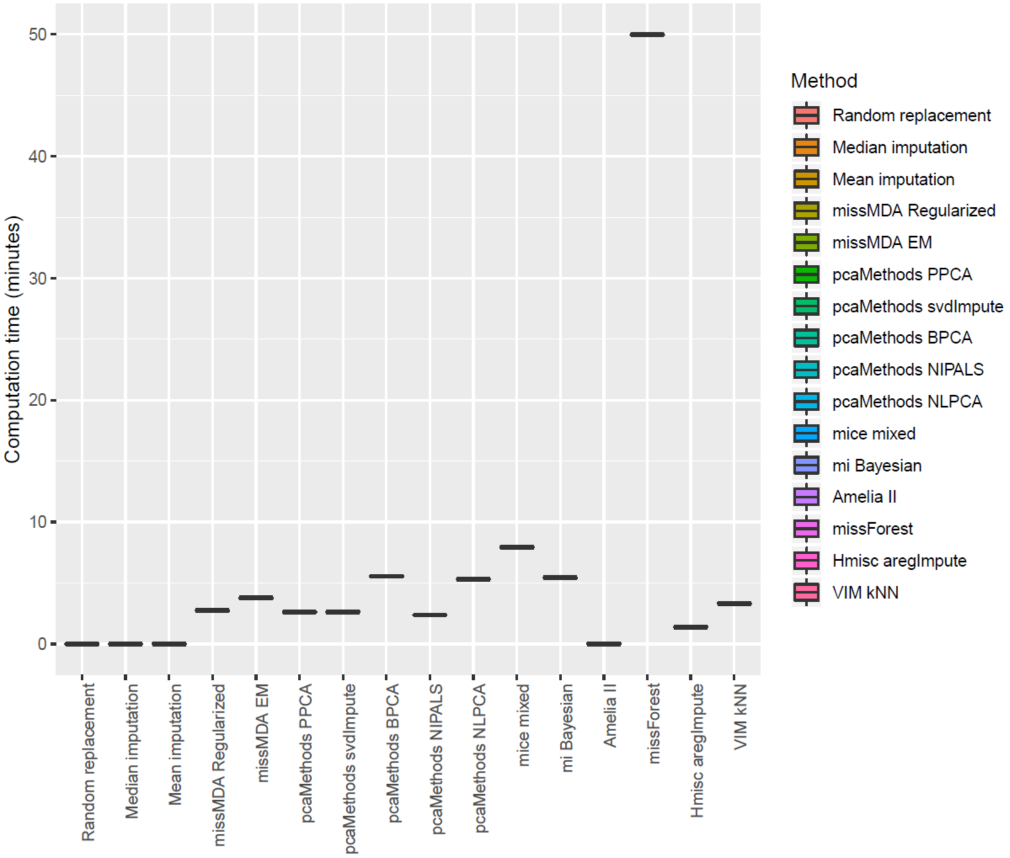


Figure 10: Missing data algorithm computational time comparison

The computational times of each of the algorithms on the simulated dataset are depicted in Figure 10. As can be seen, most computational times are comparable on the simulated dataset apart from the random forest algorithm, which take up to 5 times as long to process the dataset. This is indeed born out in the tests on our real dataset, discussed in section 6.3.3.

The following Figures (11, 12 & 13) show the performance results of each classifier based on the simulated missing data patterns. As discussed in section 3.3 the data in this case study is generally considered to be MAR (but with elements which can be considered MCAR or MNAR).

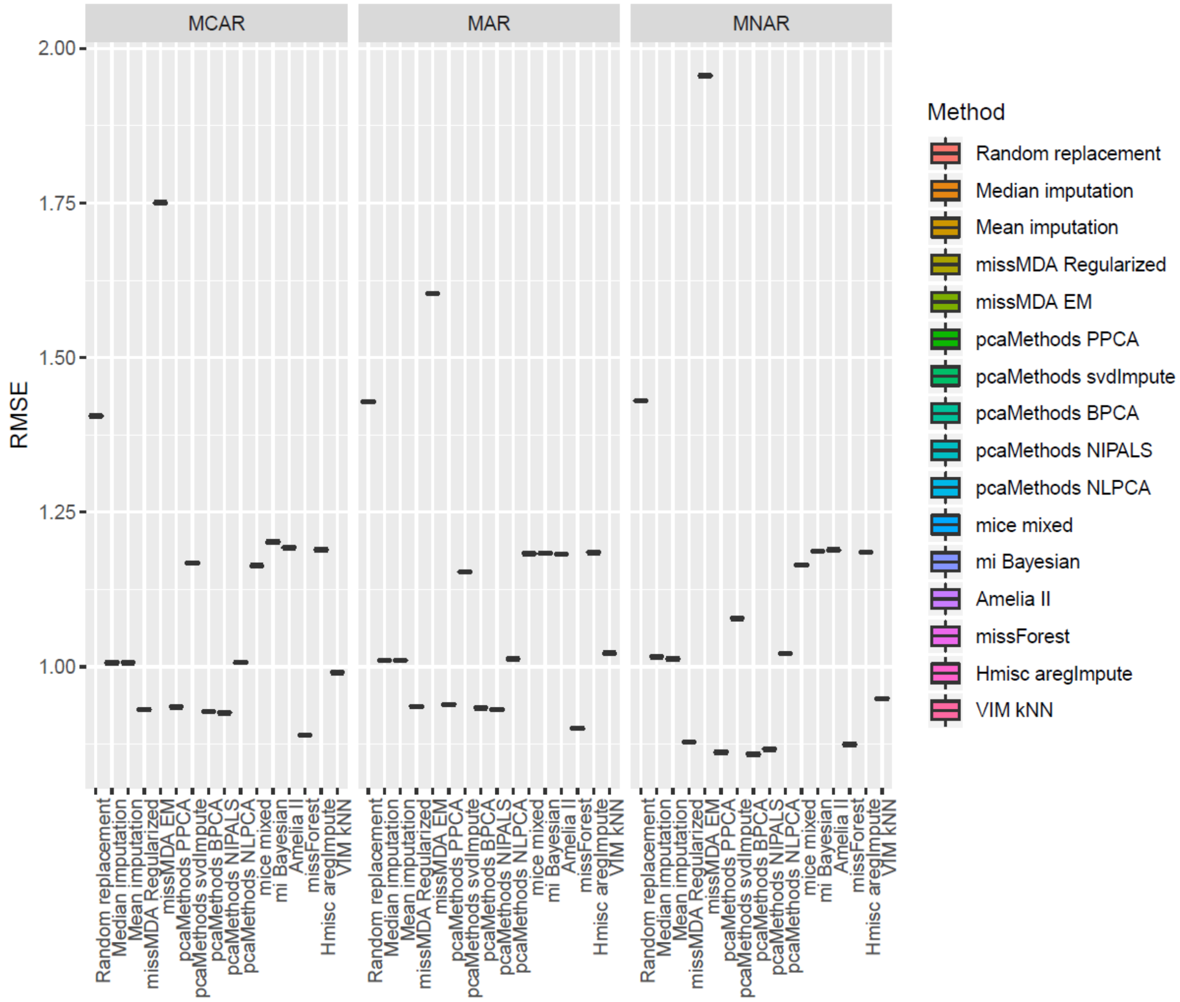


Figure 11: Imputation algorithms root mean square error (RMSE) performance comparisons – Simulated data set

The results based on the RMSE evaluation are shown in Figure 12. It is noted, however, that the RMSE is insufficient for the evaluation of imputation methods (Van Buuren, 2018). The claim is that it focuses to heavily on assessing the discrepancy between real and imputed data. In other words, one cannot base imputation method evaluations solely on their ability to recreate the true data and that such methods may increase the rate of false positives. As Van Burren points out, “imputation is not prediction”.

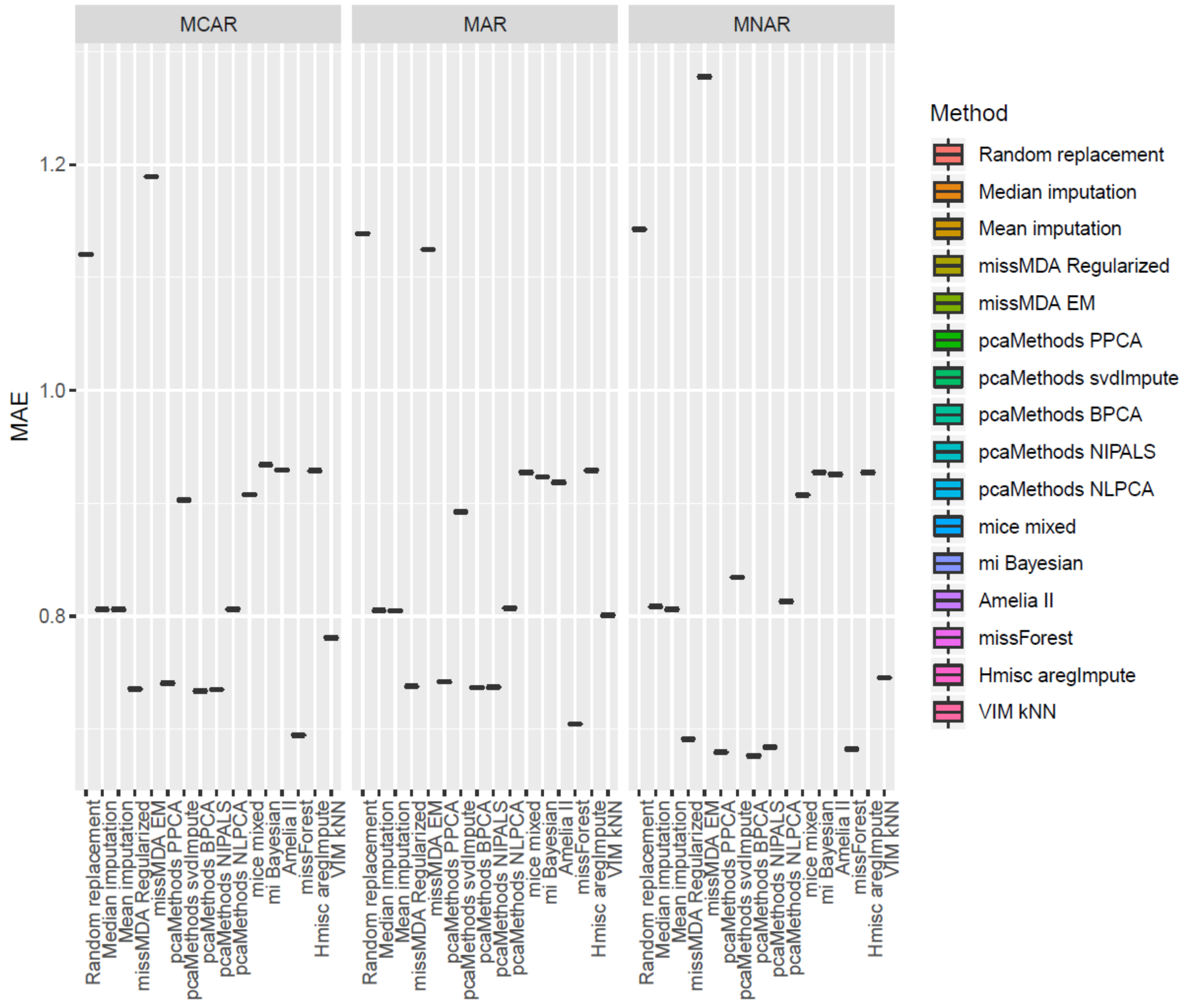


Figure 12: Imputation algorithms mean absolute error (MAE) performance comparisons – Simulated data set

The results from the Kolmogorov-Smirnov (KS) tests can be found in Figure 13. The KS is a non-parametric method for testing whether two samples are from the same population. The KS value corresponds to the maximum vertical distance between the empirical distribution functions of the samples. For missing value method evaluation, it is used as a diagnostic tool to compare the distributions of observed and imputed data. For any variable exhibiting missing data there is an empirical distribution of observed values. Following imputation there is then a distribution of imputed values for those variables exhibiting missing values.

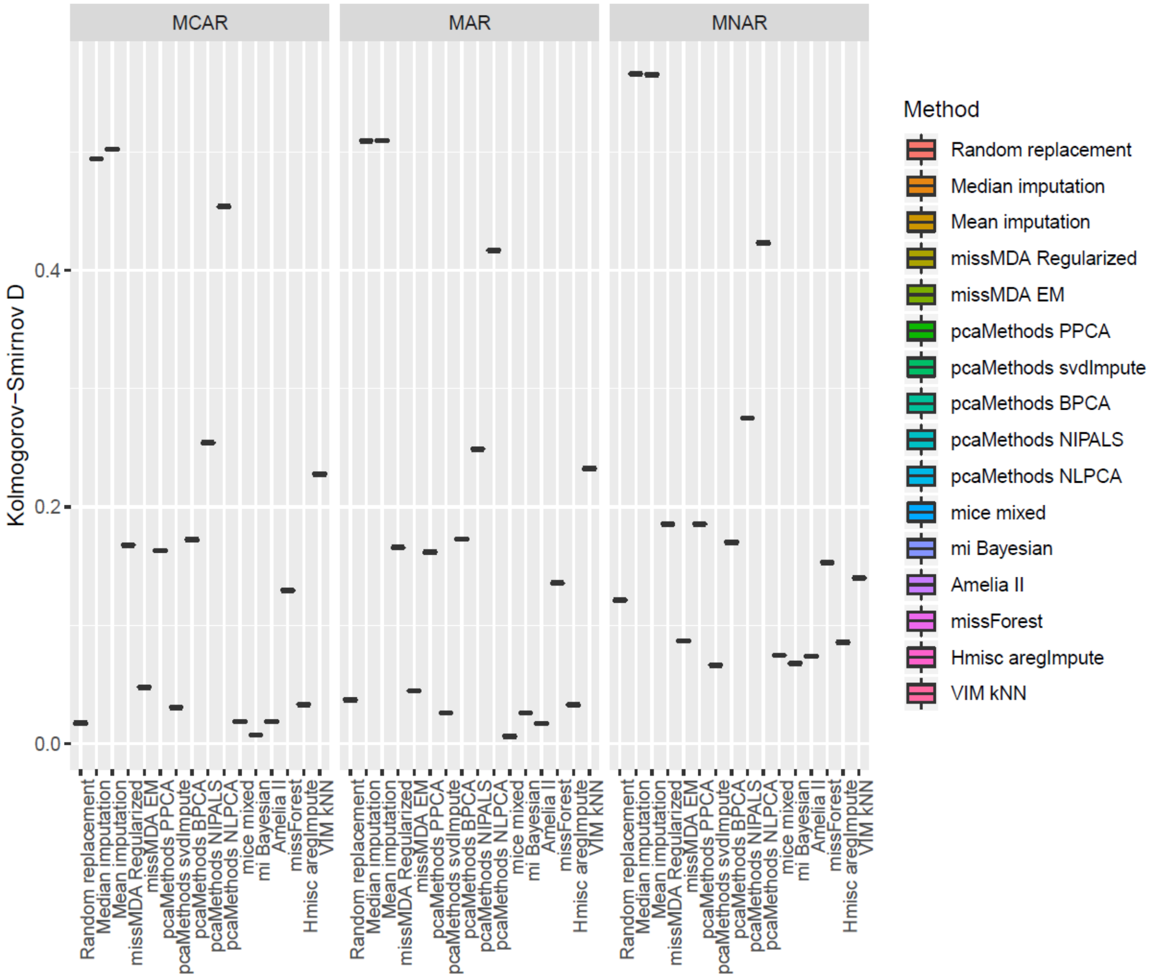


Figure 13: Imputation algorithms Kolmogorov-Smirnov (KS) performance comparisons – Simulated data set

In summary, performance results from the test on simulated data appear relatively inconclusive for MCAR, MAR and MNAR patterns in the performance metrics. However, given that the Hmisc (an implementation of additive regression splines imputation (Harrell, 2015)) results for MAR on a KS scale resulted in a value below .05 qualifies its inclusion as a candidate for further testing on the real dataset. Random forest imputation performed well on average over all three tests. Given these results combined with the findings from previous research, Hmisc and random forest imputation where selected together with k-NN (for calibration) for use as evaluation of imputation management techniques in this case study.

6.3.3 Imputation results

Further to the conclusions presented in the previous section, three imputation algorithms: Hmisc (additive regression splines (Harrell, 2015)), random forest (RF) and k-NN were selected in order to evaluate their performance in a real-world, practical setting using the bulk carrier company case study data, with the aim of identifying a suitable imputation methodology which will be used to in the testing of FD prediction using each of the four data model sets.

Each of the three algorithms were evaluated on their ability to obtain statistically valid inferences from the incomplete financial data set. The effect of imputation on each individual financial covariate is first examined.

The basic testing assumption is the null hypothesis, which states that, the two distributions, observed and imputed are drawn from the same data sample. As such, the performance metrics utilised are the Welch t-test P and the KS D values. For example, a 95% significance level indicates that where imputed covariates show a $p < .05$, the null hypothesis should be rejected. Also, KS D values should be as close to 0 as possible. These results are presented in the following section and are visualised using histograms and boxplots depicting the distributions of original and imputed covariate values. Although the goal is to have the two distributions as similar as possible, differences do not necessarily signal problematic imputation. Distributions of missing and observed data can differ, with missingness still being attributed to missing at random. The MAR assumption, by definition, is that differences in distributions are necessarily explainable by other variables in the data set. However, dramatic differences between the imputed and observed data do indicate problems and, in a context with many imputed variables, it is helpful to have some screening devices to identify these potential problems. The empirical density plots used here act as flags for potential problems with the imputed estimates.

Figure shows the distribution histogram overlay of imputed and observed values with rows representing each financial covariate and columns for imputation algorithms. Similarly, an alternative view on the two distributions is provided in **Error! Reference source not found.** which provide a boxplot representation of the dataset covariates. At this stage, no d ata pre-processing was performed on the pre-imputation dataset as any bias will be “locked into” the data prior to training and validation.

A visualisation of the hierarchical clustering of the missing data is provided in Figure 16**.** The bifurcations approaching a length of 0 (to the left of the plots) represent closer relationships in terms of missing data - i.e. those variables in one group are more likely to be missing together compared to the rest. An important aspect of this is that the patterns should not alter too much pre- and post-imputation.

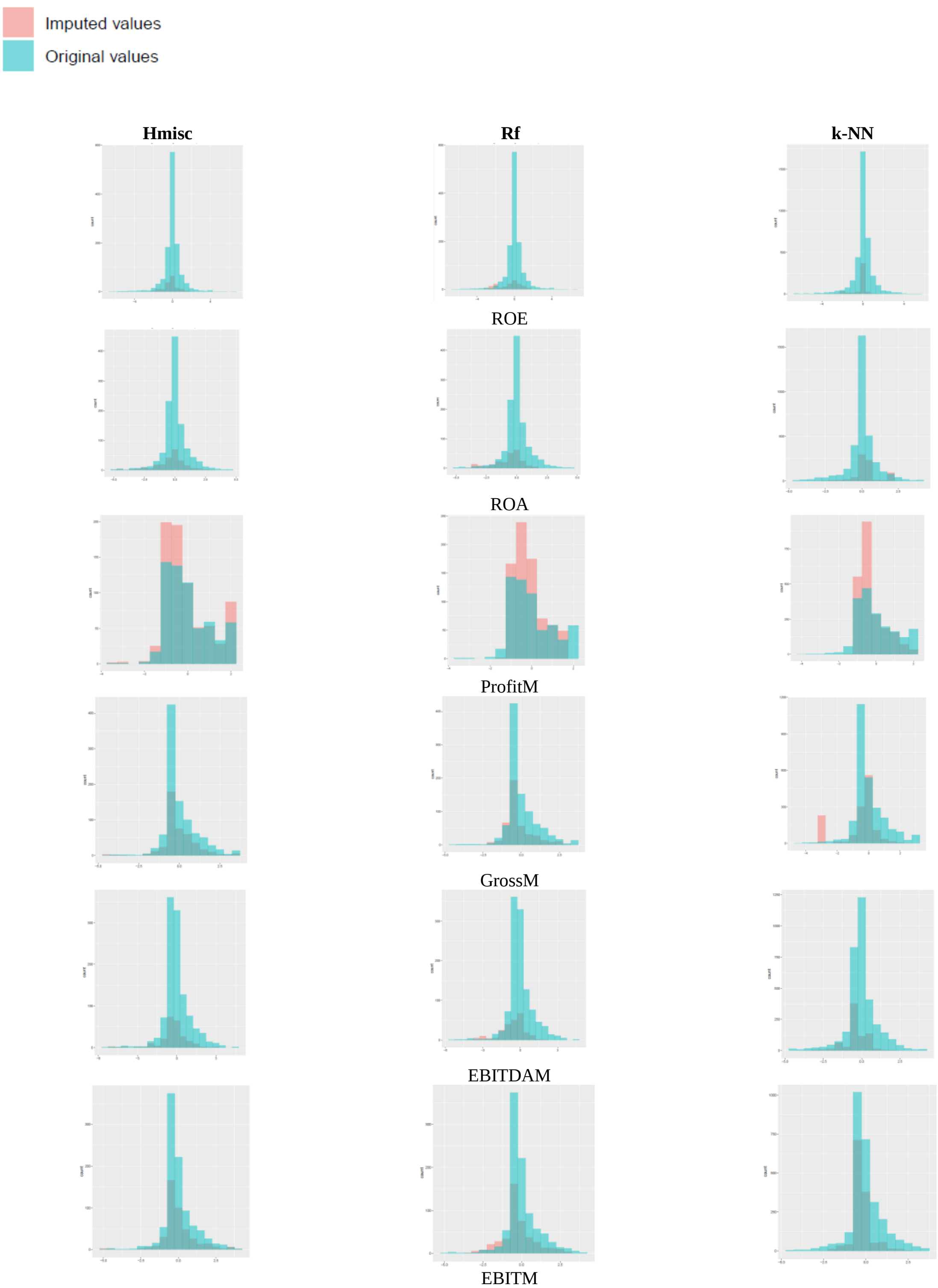


Figure 14: a) Overlaid histograms of imputed and original values

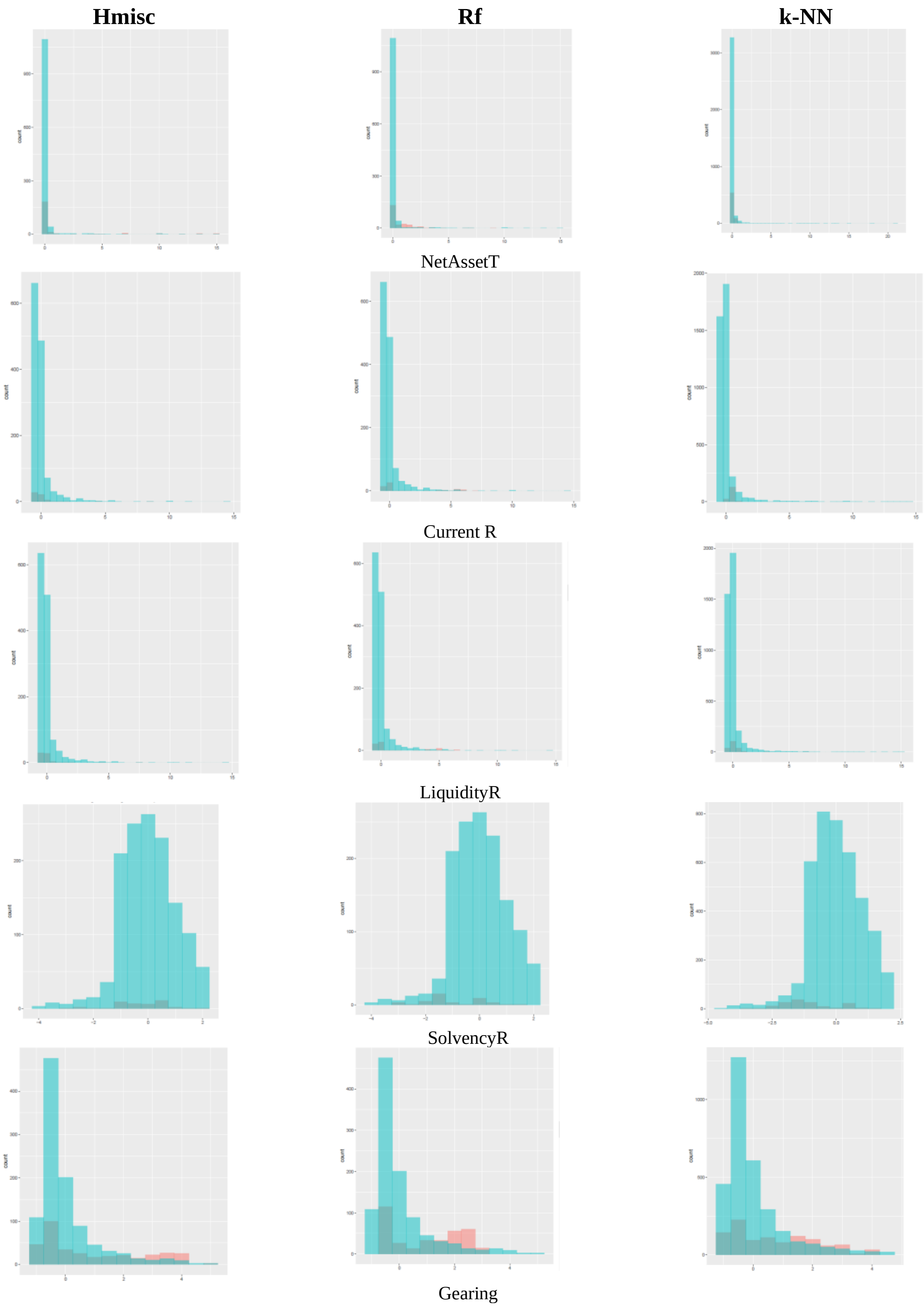


Figure 14: b) Overlaid histograms of imputed and original values

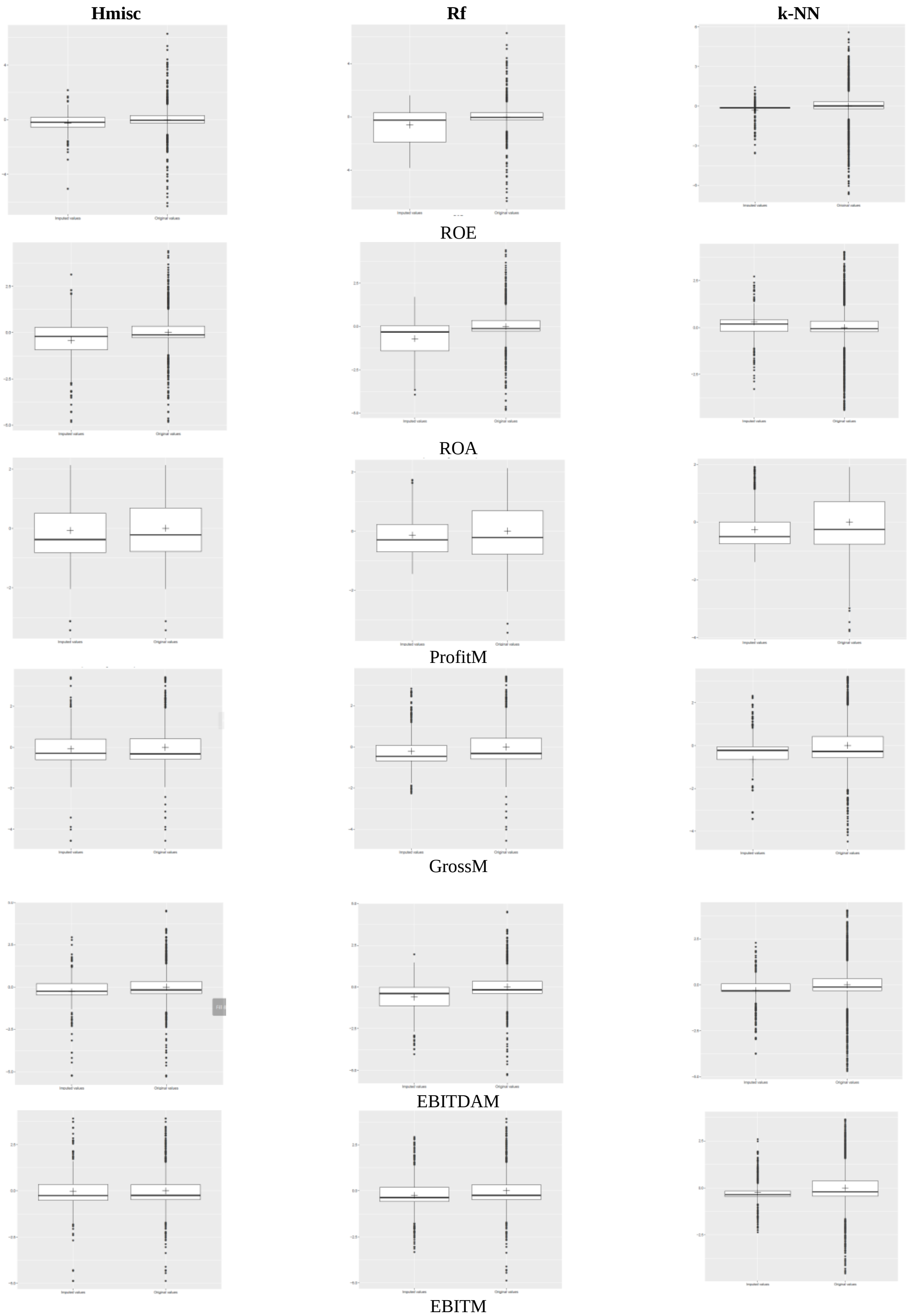


Figure 15: a) Boxplots of original and imputed values

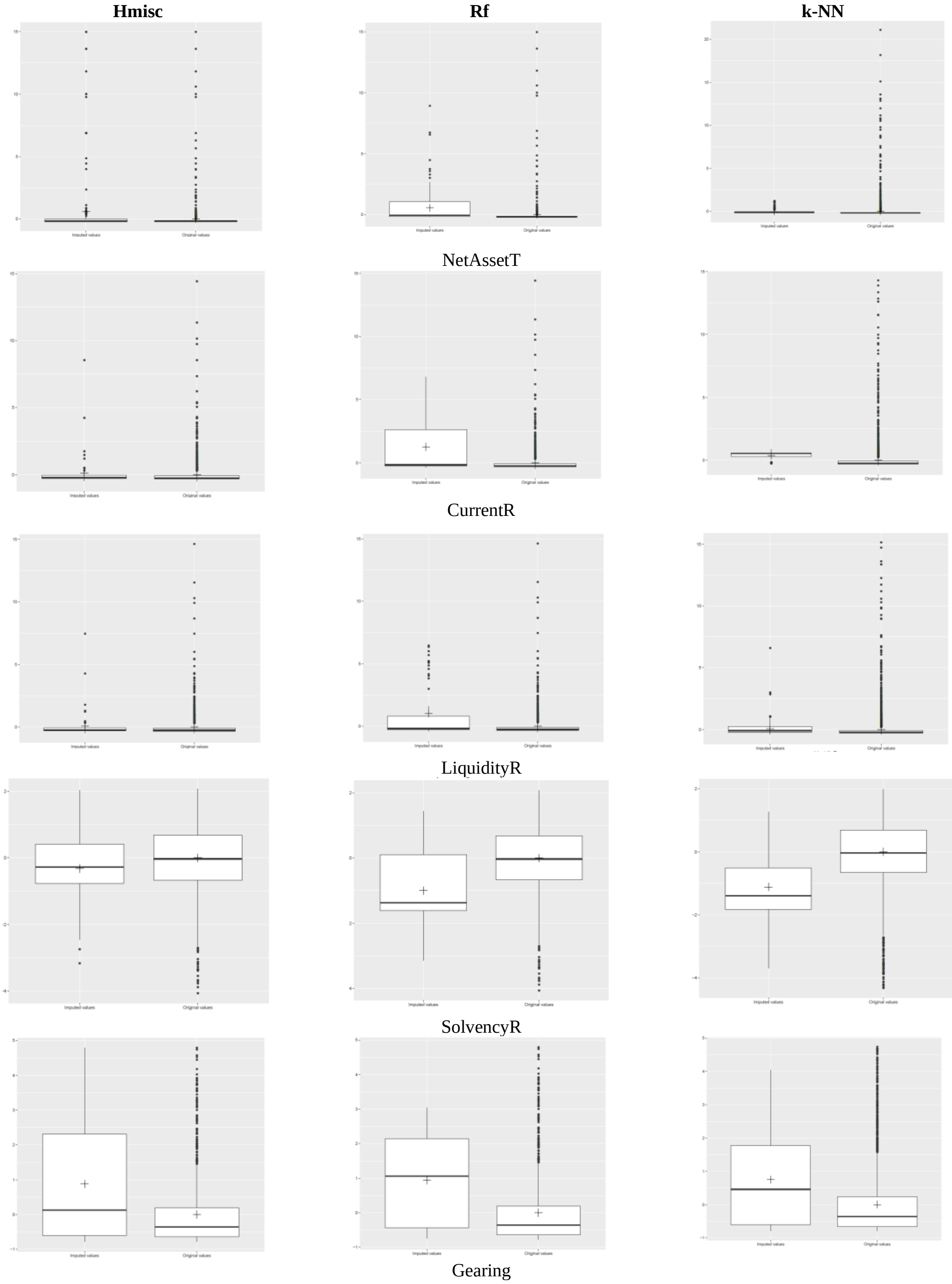


Figure 15: b) Boxplots of original and imputed values

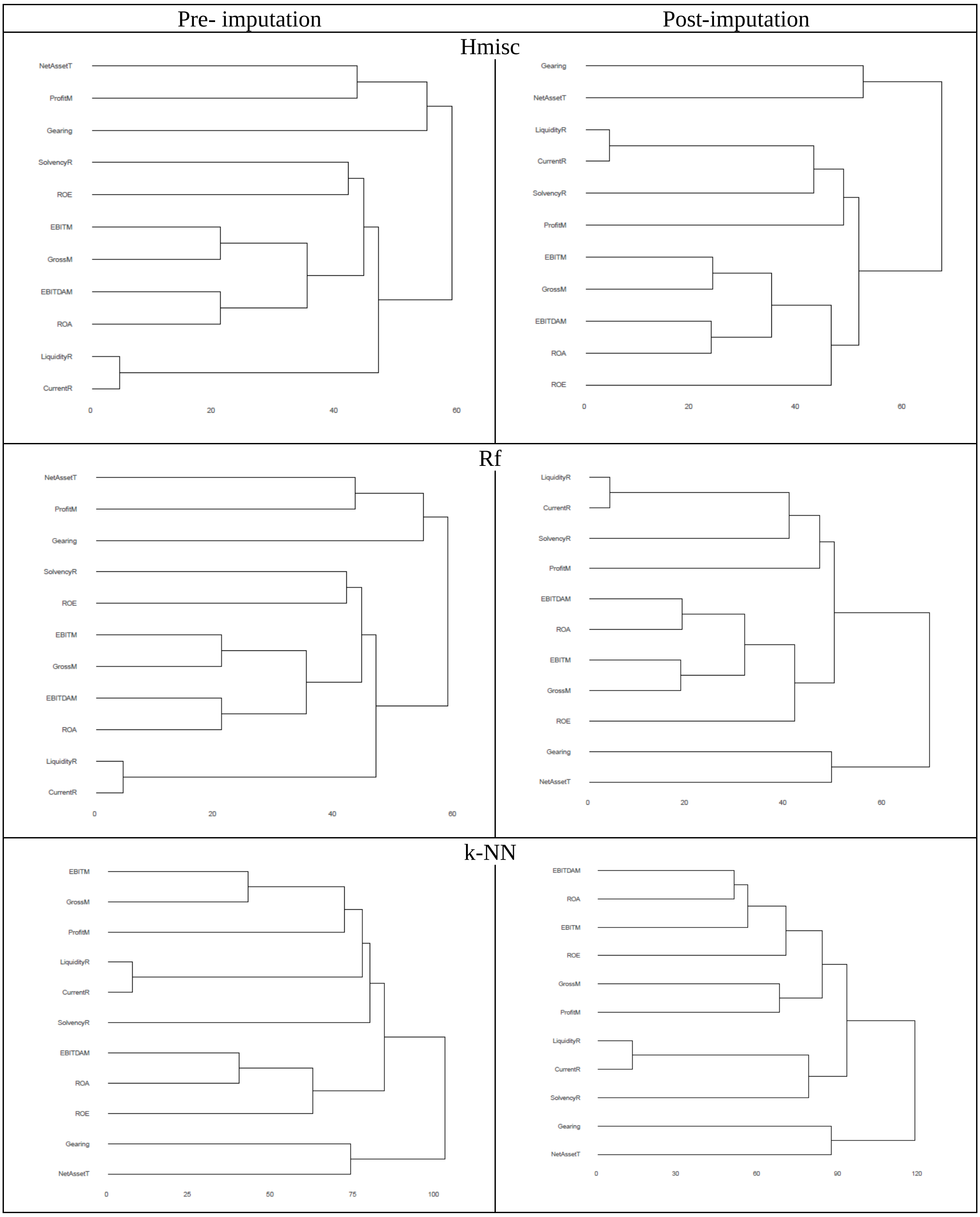
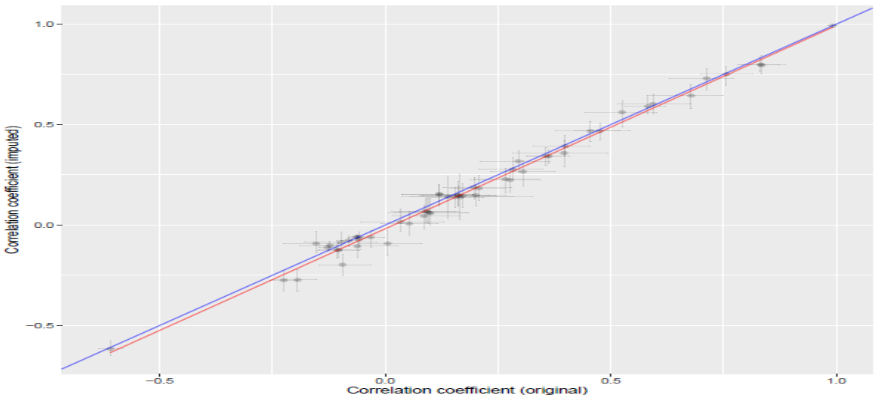
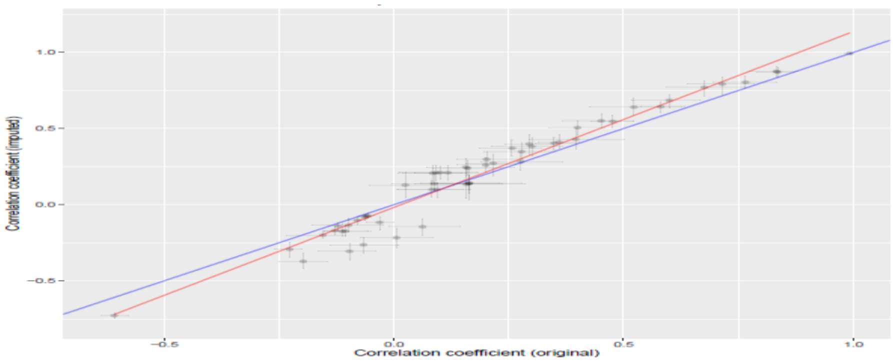


Figure 16: Original and imputed variable clusters - pre and post imputation

Hmisc



Rf



k-NN

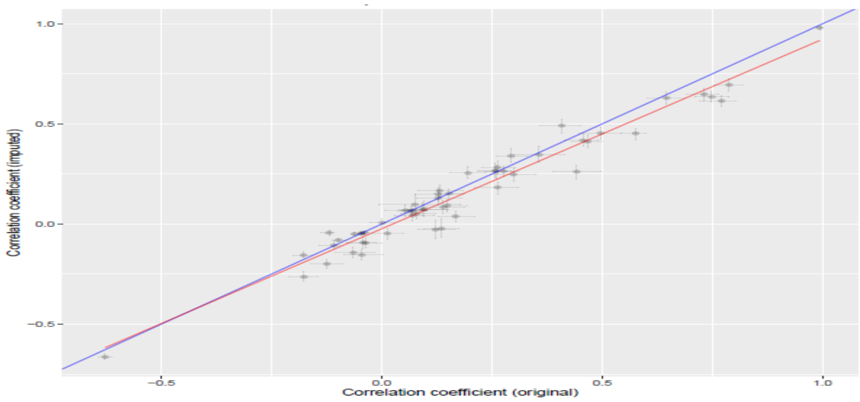


Figure 17: Correlation coefficient scatter plots

Finally, Figure 17 shows a plot of the bootstrapped correlation coefficients from the original and imputed data sets. This was generated through applying 20 iterations in the diagnostics function to obtain bootstrapped correlation coefficients with 95% confidence intervals. The correlation coefficients are represented by the dots and the red line. The blue line (intercept 0, slope 1) and the red correlation line should be aligned. With k-NN imputation the line is turned clockwise and has a lower slope than 1 (regression to the mean).

Table 5: Imputation evaluation metrics

	HM			**RF**			**k-NN**	
	Welch t-test P	**KS test D**		**Welch t-test P**	**KS test D**		**Welch t-test P**	**KS test D**
ROE	1.18E-03	1.92E-01		1.79E-08	2.81E-01		2.38E-17	4.07E-01
ROA	6.09E-06	1.73E-01		9.22E-16	2.96E-01		4.91E-16	2.31E-01
ProfitM	1.75E-01	8.16E-02		2.88E-03	1.15E-01		2.11E-21	1.87E-01
GrossM	1.46E-01	4.57E-02		7.80E-05	1.56E-01		9.11E-61	1.85E-01
EBITDAM	5.16E-04	8.62E-02		6.55E-16	2.84E-01		3.09E-22	3.56E-01
EBITM	5.97E-01	5.38E-02		8.22E-05	1.55E-01		3.47E-25	2.27E-01
NetAssetT	9.31E-04	1.14E-01		1.28E-09	4.29E-01		2.60E-01	3.56E-01
CurrentR	4.50E-01	1.66E-01		2.65E-04	2.72E-01		1.25E-36	7.46E-01
LiquidityR	5.96E-01	1.30E-01		3.95E-04	2.36E-01		3.05E-01	2.70E-01
SolvencyR	5.32E-02	1.76E-01		9.52E-07	5.14E-01		8.33E-25	5.43E-01
Gearing	3.47E-19	2.80E-01		4.75E-32	4.18E-01		3.88E-55	3.22E-01

6.3.4 Imputation summary

The Welch's t-test and KS-D test were used as statistical measurements to evaluate if the imputed data was statistically close enough to the observed data. The p-values of the Welch's t-test showed that only Hmisc and RF consistently imputed values across the covariates such that the null hypothesis of equal means could not be rejected, e.g. for a p-value for a covariate above 0.05 indicating that the null hypothesis at the 5% level cannot be rejected. For example, the p-value of the liquidity ratio, "LiquidR", for imputation method Hmisc is 0.596, which indicates that the null hypothesis should not be rejected. However, the low p-value of the variables for ROA and gearing, indicate that further investigation is warranted.

The KS D values for all three imputed data sets indicate a close approximation to the validation data than dissimilarity, since all values in Table 5 are closer to 0 than 1. This can be interpreted as an indicator of limited loss of information from the imputed data from the models. However, Hmisc consistently demonstrates slightly superior performance of the three.

6.4 Feature set analysis

The section begins by examining the contribution of each of the three covariate sets which form the foundation of the four models. Figure 18 shows the ranked importance of the four feature sets (models) using the LASSO (Tishbirani, 1996) method. This method places a constraint on the sum of the absolute values of the model parameters by applying a shrinking or regularization process to penalize the regression variable coefficients, shrinking some of them to zero. During features selection process the variables that still have a non-zero coefficient after the shrinking process are selected to be part of the model. The goal of this process is to minimize the prediction error.

All the accounting ratios represent a significant predictive power in Model 1 with the liquidity ratio seemingly being the least ranked predictor. The inclusion of the macroeconomic data in Model 2 maintains the significance of the accounting covariates but signals short term interest rates, unemployment and debt to GDP as additional significant macro indicators. Model 3 introduces bulk market indicators to the accounting data with new build prices index, shipyard orderbook and laid up fleet indicators having significant effect on the distress prediction. Finally, Model 4 containing all covariates from the previous three models appears consistent with the previous three models in ranking predictors.

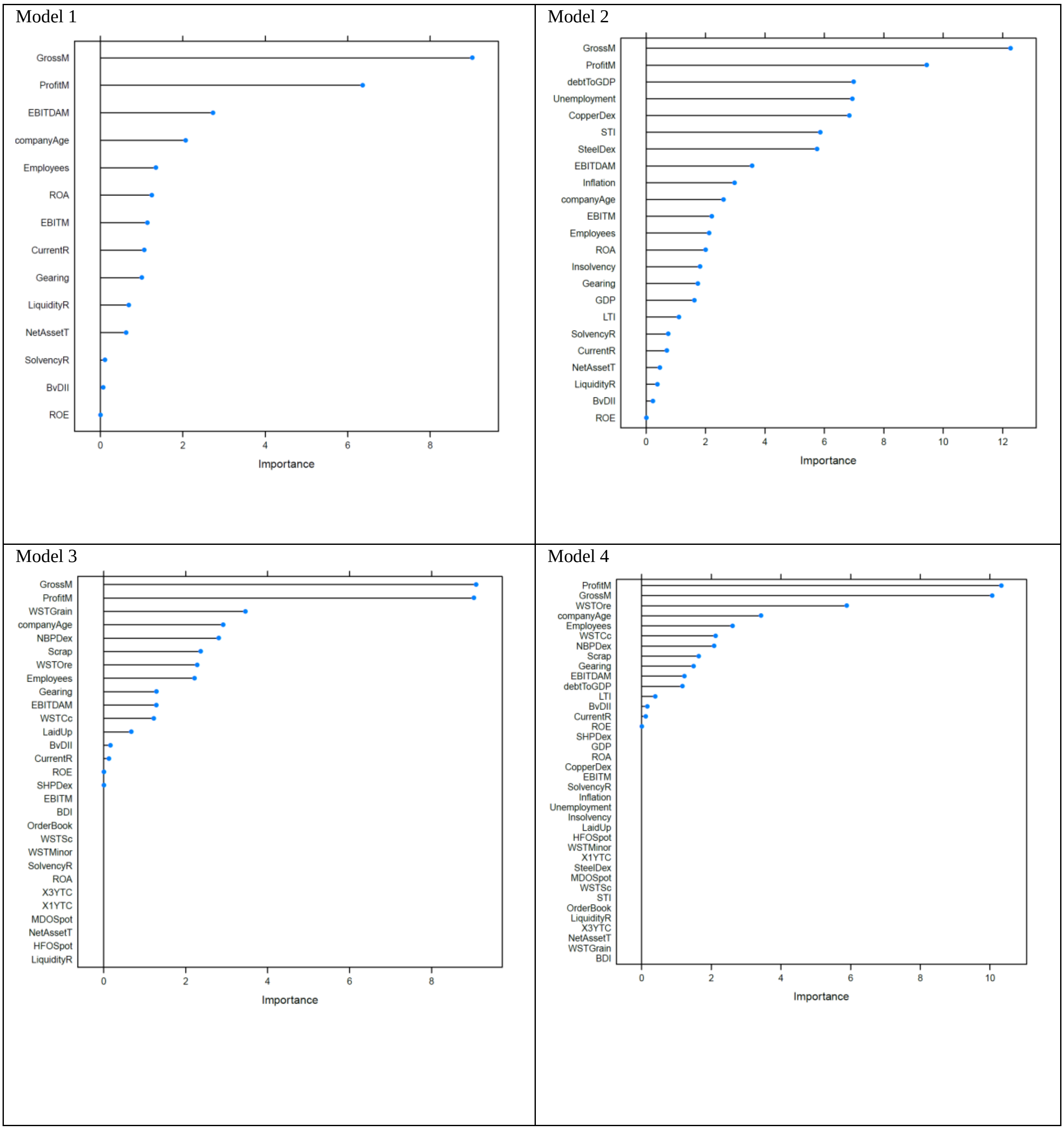


Figure 18: Variable importance - linear based classification (model 4 predictor set)

The results from the random forest classifier are plotted in Figure 19 for all four data models. Variable importance is the mean decrease of accuracy over all cross validated predictions, when a given variable is permuted after training, but before prediction. Covariate importance is the mean decrease of accuracy by each individual cross validated prediction. It basically represents the number or proportion of observations that are incorrectly classified by removing the covariate in question from the model. For example, removing profit margin (ProfitM) from the model would result in an additional misclassification of approximately 80 observations. Variable importance is used mainly to rank the usefulness of covariates. A clear interpretation of the absolute values of variable importance is hard to do well.

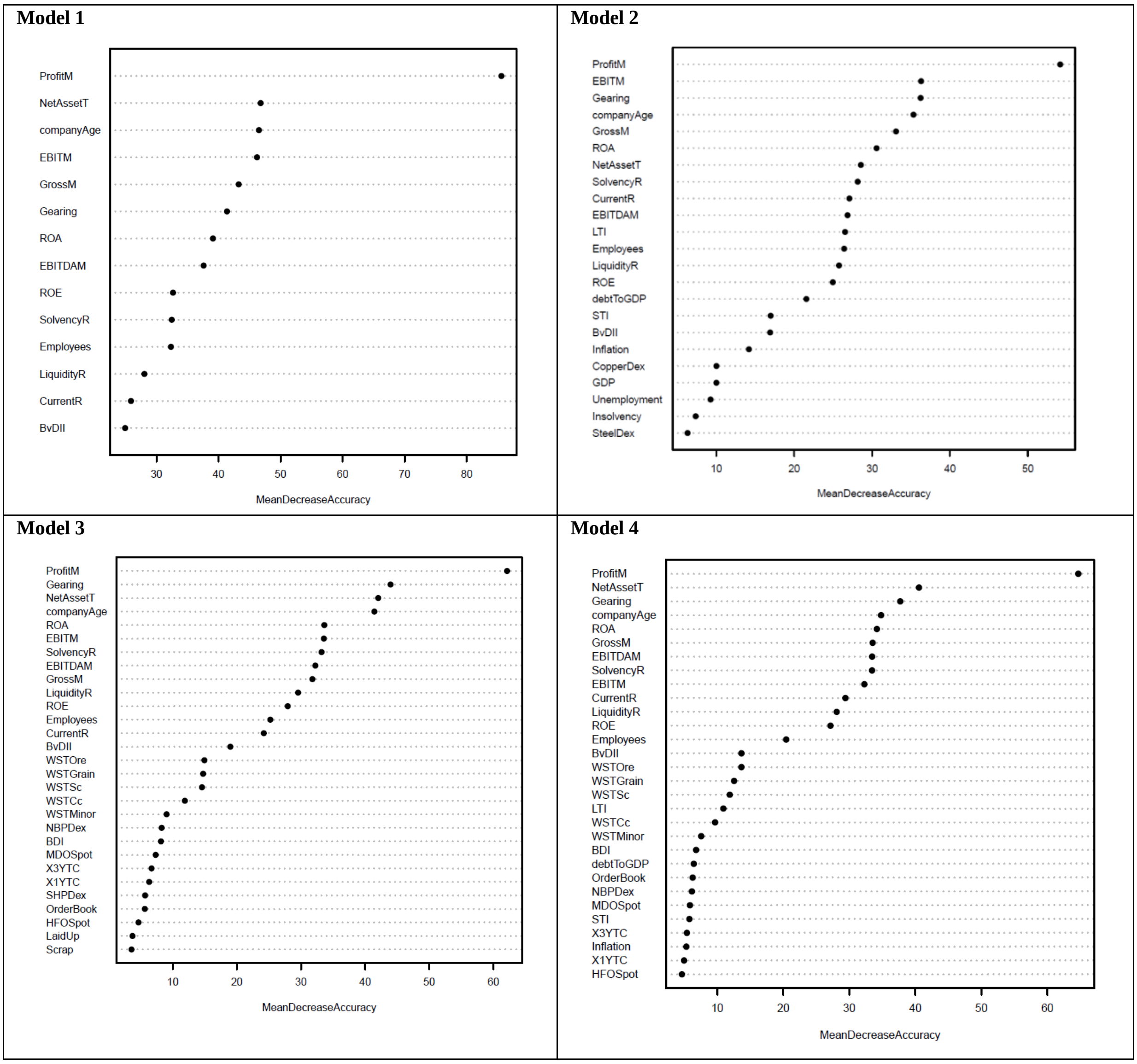


Figure 19: Results from tree (RF) model classification on the full data model showing variable importance

Figure 20: Feature importance and correlations – Model 1

Finally, all linear models where tested on data sets with and without i) highly correlated predictors and ii) covariates marked as statistically insignificant. No change in performance was registered and all results presented are those derived from the complete datasets.

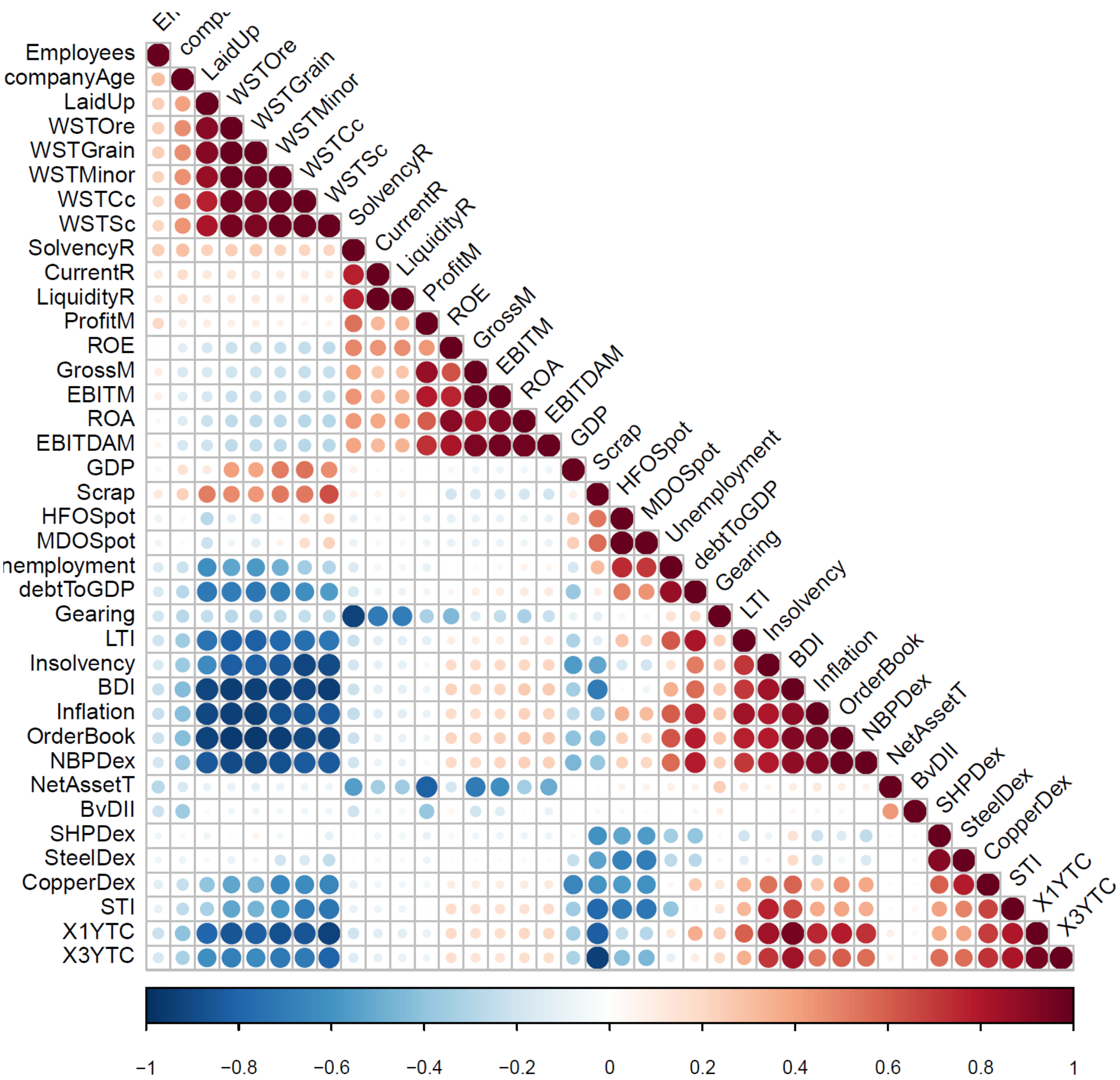


Figure 21: Correlation matrix – Model 4

6.5 Feature set modelling – results

6.5.1 Performance metrics

As previously discussed, the finance literature extensively covers the subject of comparisons between linear based algorithms and complex modelling in the domain of company distress prediction (classification). However, the aim in this research is to not only compare classification algorithms based on predictor set capacity but also to analyse their individual FD prediction capacity and the effects of augmenting company level financial data with both market and macroeconomic data.

The assessment of the classification performance of each model/classification combination is performed using their respective area under the curve (AUC) of the receiver operating characteristics (ROC). For ROC clarification, companies marked as exhibiting the pre-defined status of "distressed" were assigned a positive value (1) and a zero (0) to companies not considered so. As this exercise is being approached from an investor/banking perspective heavier costs are placed on the false negative prediction counts. That is those companies predicted as "normal" by the models when in fact the actual state is "distressed". This is reflected in the Type II error metric whereby the lower the value the less prone classifier is at yielding false negative responses. The sensitivity metric records the levels of correctly predicted "distressed" cases (true positives).

Finally, the results must reflect the nature of both data imbalance (Japkowicz and Stephen, 2002) and the costs of misclassification. This imbalance is common in financial statement datasets, especially with bankruptcy as it is a relatively rare event. Account must be taken for this imbalance otherwise the risk of bias in favour of the predominant class will be significant. If the dependent variable is heavily biased towards then any predictions on that data will tend toward that class. This imbalance is present in the case study dataset. In order to address this imbalance the H measures (Hand, 2009) metric is utilised. This is a robustness check on the AUC results. This metric addresses the main problem associated with the AUC, that of the handling of misclassification costs across different classifiers. The AUC does not apply the same misclassification cost distributions to individual classifiers i.e. it utilises different metrics when evaluating different classification rules. And as such its use should be limited to the broad comparison of individual classifiers as an AUC may rank the individual models adequately but perform inadequately in terms of the level of the predicted probabilities.

The log loss function is also used to compare the calibrated probabilities. The log loss function measures the accuracy of a classification model by penalising false classifications. The basic premise is in minimising the log loss to maximise the accuracy of the classifier. In calculating the log-loss, a classifier assigns a probability to each class in place of assigning the most likely class.

Mathematically log loss is defined as:

$$H_j = -\ \frac{1}{n}\sum_{i \in R}(y_i \log{(\hat{p})} + (1 - y_i) \log{(1 - \hat{p})}) \tag{38}$$

where H_{jt} is the model's log score (loss) of model j in year t, y_i is a dummy equal to 1 if company i financially distressed, $\hat{p}$ is the predicted probability of distress of firm i by model j, R is the sample of active companies and n is the number of companies in R. A perfect score is zero.

The log loss metric considers the probabilities of the underlying model prediction and not only the final output of the classification. The stronger the probabilities correspond to a lower log loss. As log loss is a measure of entropy or uncertainty, a low log loss means a low entropy. The measure is similar to the Accuracy value derived from the confusion matrix, but it will favour models that more strongly the distinguish classes and useful for comparing not only model output but on their individual probabilistic outcome.

6.5.2 Model performances

The results discussed below are those produced from classification modelling performed with the random forest imputed dataset. As will be discussed in section 6.6, the RF imputed dataset consistently produced the highest sensitivity and, therefore, the lowest type II error rate of all of the imputation techniques. Although the Hmisc technique produced the highest overall AUC results of the imputation techniques, they were only marginally higher than RF. However, the prime goal of FD classification is not only high overall performance but also the avoidance of false negative results.

The four models and nine classifiers are compared using the ROC data produced following predictions made on the hold-out datasets. In the diagrams below, each classifier test produces a ROC curve (bottom right) and classification density distributions (top left). The "Density by tag" shows the "Active" company prediction density (pink or 0) compared "Distressed" company predictions (blue or 1). Ideally, these distributions should be as separated as possible indicating the classifiers ability to make clear predictions. All model results based on two years to (t-2) distress prediction with a macro/market data lag of one year to distress and a 95% confidence level.

Before moving onto the data model discussions, it is important to look first at the results from the mixed effect model. The motivation behind its inclusion in the study is that the model allows for unobservable macro effects which create correlation in distresses, notwithstanding those related to the observed covariates. The results of running the 2000–2018 (rf imputed), model 1 data set through a mixed effects algorithm, produced fixed effect (observed feature set) estimates indicating an overall strong effect on FD prediction. However, the various covariate mixes of company, macroeconomic and market predictors produced estimated standard deviation estimation of the random intercept of between 2.4 and 0.639, e.g., for $\hat{\sigma} = 0.639$, a change of one standard deviation in the random intercept implies a 1.89 ($e^{0.639}$) times higher odds of a company being distressed. Therefore, it is assumed that there is a non-negligible random effect, which accounts for an unobserved correlation in distress. This notion of an unobserved temporal effect is one of the hypotheses driving this study.

6.5.2.1 Model 1 - Company level covariate set

This section presents the results for each of the nine classification model applications on the company level covariate set (model 1). The results for the traditional classifiers are shown in Figure 22 and those for complex classifiers found in Figure 23.

As can be seen from the "Density by Tag" plot, the hazard classifier performs quite poorly on the separation of outcomes. Furthermore, the ROC curve for this classifier also reports an AUC of 0.61, confirming the distribution result. This indicates that the hazard classifier is barely able to separate the classes (a value of 0.50 would mean that the algorithm could not separate the classes at all) on this dataset. The linear mixed effect classifier does demonstrate a reasonable ability to separate the "Distress" and "Active" outcomes and demonstrates improved predictive power over the hazard model with and AUC of approximately 0.83.

The results produced by the MARS and GAM classifiers, which have been developed to better accommodate the non-linear aspects of the data, demonstrate the clear ability of these algorithms to separate the data and perform relatively well, with AUC values of approximately 0.89 and 0.87 respectively.

The complex classifier results demonstrate an overall improvement on the four linear based classifiers. However, compared with MARS and GAM, the improvements are not as pronounced as previous reported in the literature. The tree-based algorithms (RF, GBM and XGB) outperform the ANN and SVM but only marginally. However, the tuning of the tree-based models was almost set at default and improved results would, more than likely, be achieved through reviewing the parameterisation. Overall, GBM exhibits the highest AUC with 0.93.

These results should, however, be used with caution. Overall AUC levels are only one indicator of the predictive power of classification algorithms and do not indicate the over predictive capacity of a classifier, as will be discussed further in the chapter.

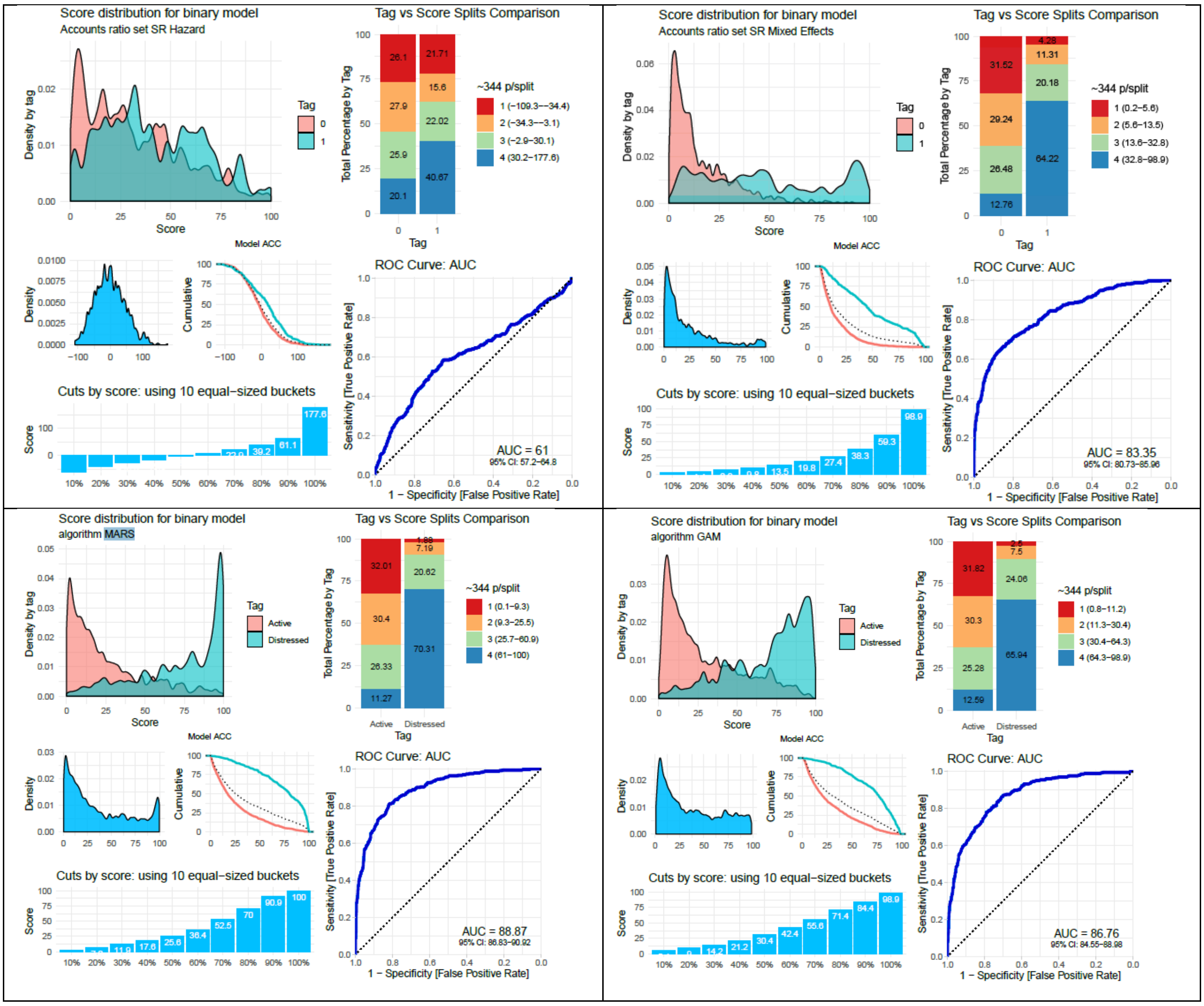


Figure 22: Model 1 - ROC results for traditional classifiers

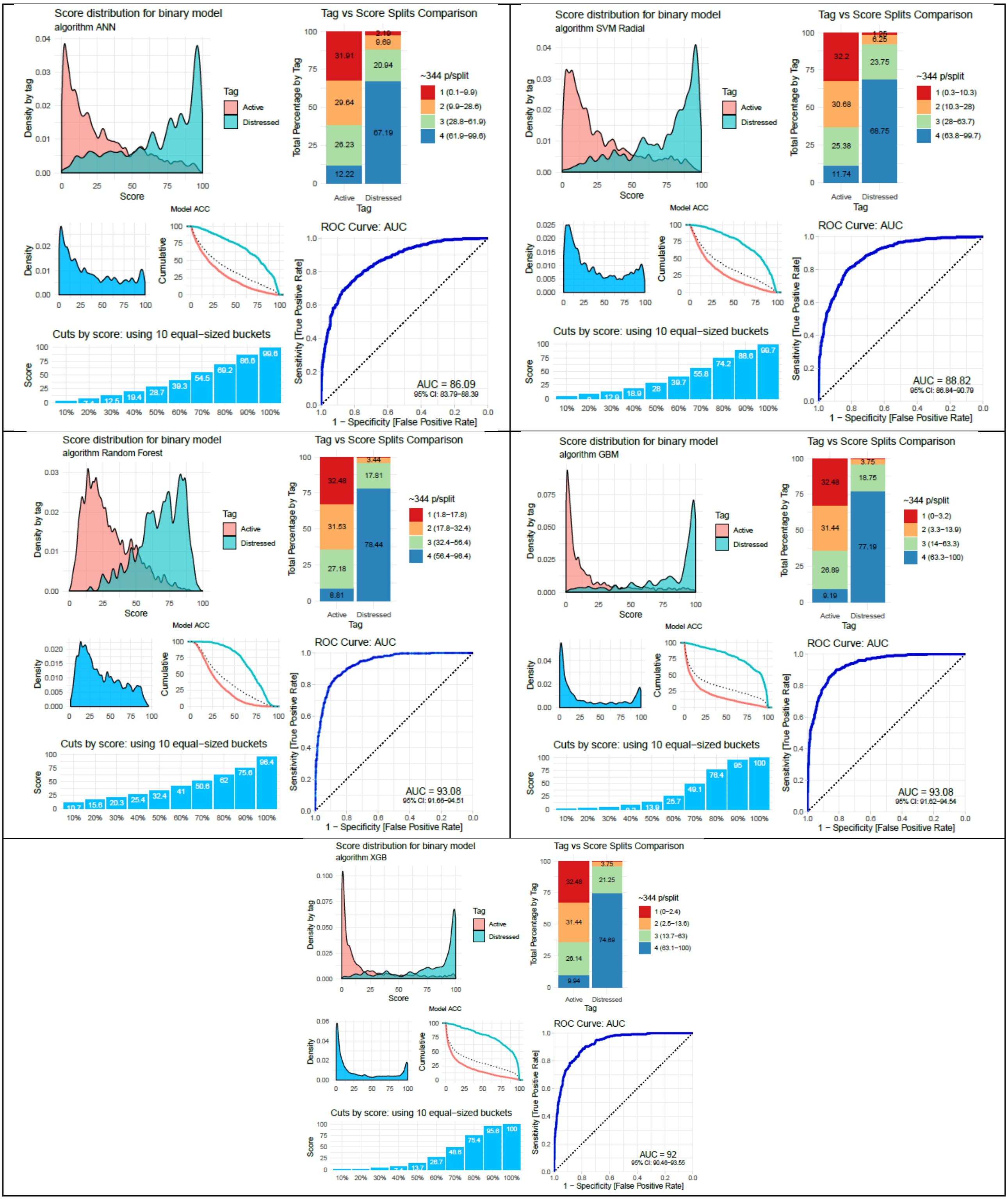


Figure 23: Model 1 - ROC results for complex classifiers

6.5.2.2 Model 2 - Macroeconomic covariate combined with model 1

This section presents the results for each of the nine classification model applications on the macroeconomic covariate set combined with the company level covariates (model 2). The results for the traditional classifiers are shown in Figure 24 and Figure 25, for the complex classifiers. The results show little change in the AUC figures compared with those for model 1. This indicates that the introduction of macroeconomic predictors has little overall effect on FD predictive power of the classifiers. This is consistent with the results from the feature set analysis (section 6.4) which indicates that only short term interest rates, inflation and debt to GDP have relatively strong predictive power. However, the combined effect of these predictors with the company levels predictors is imperceptible in the combined predictive power of model 2.

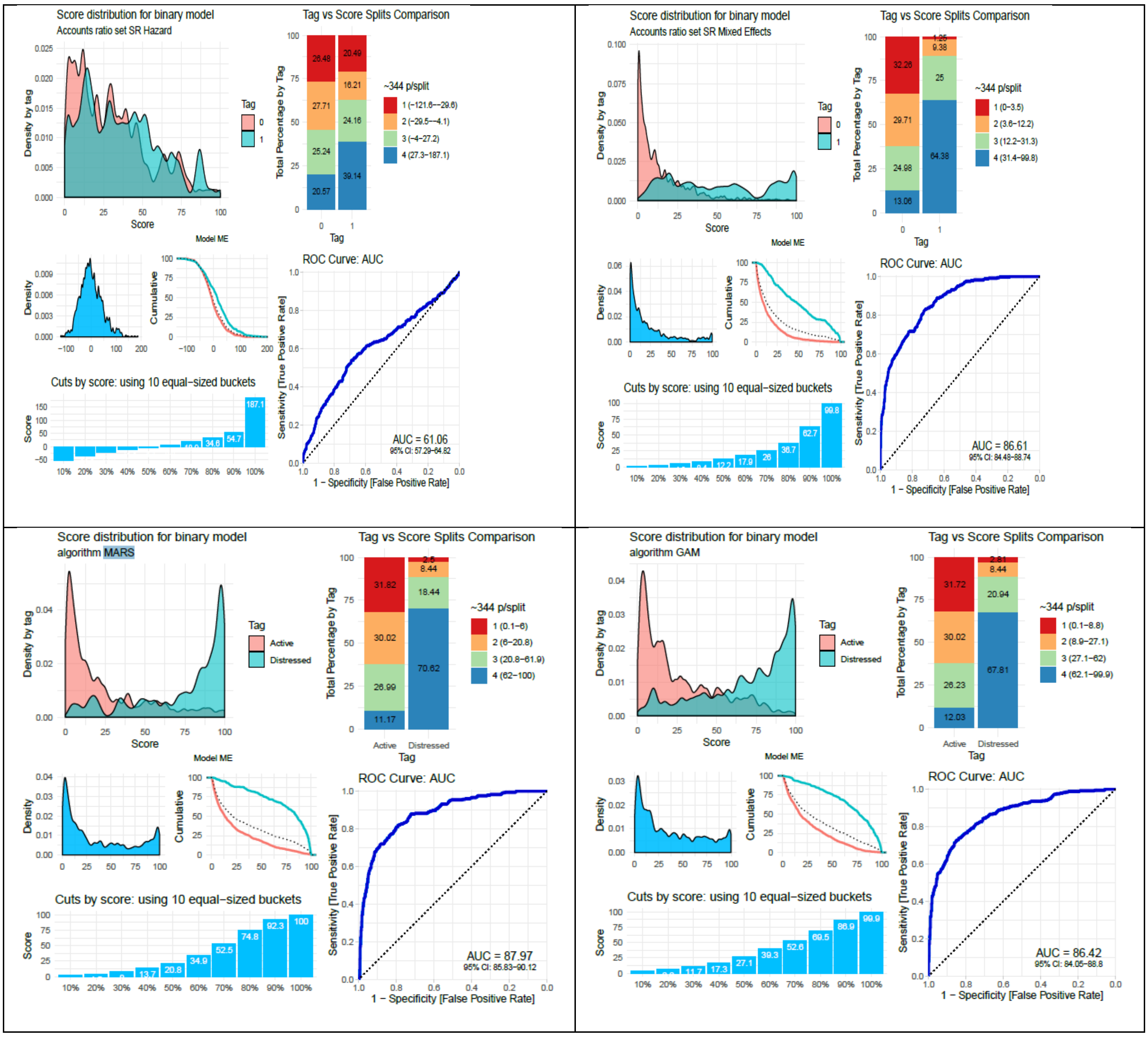


Figure 24: Model 2 - ROC results for traditional classifiers

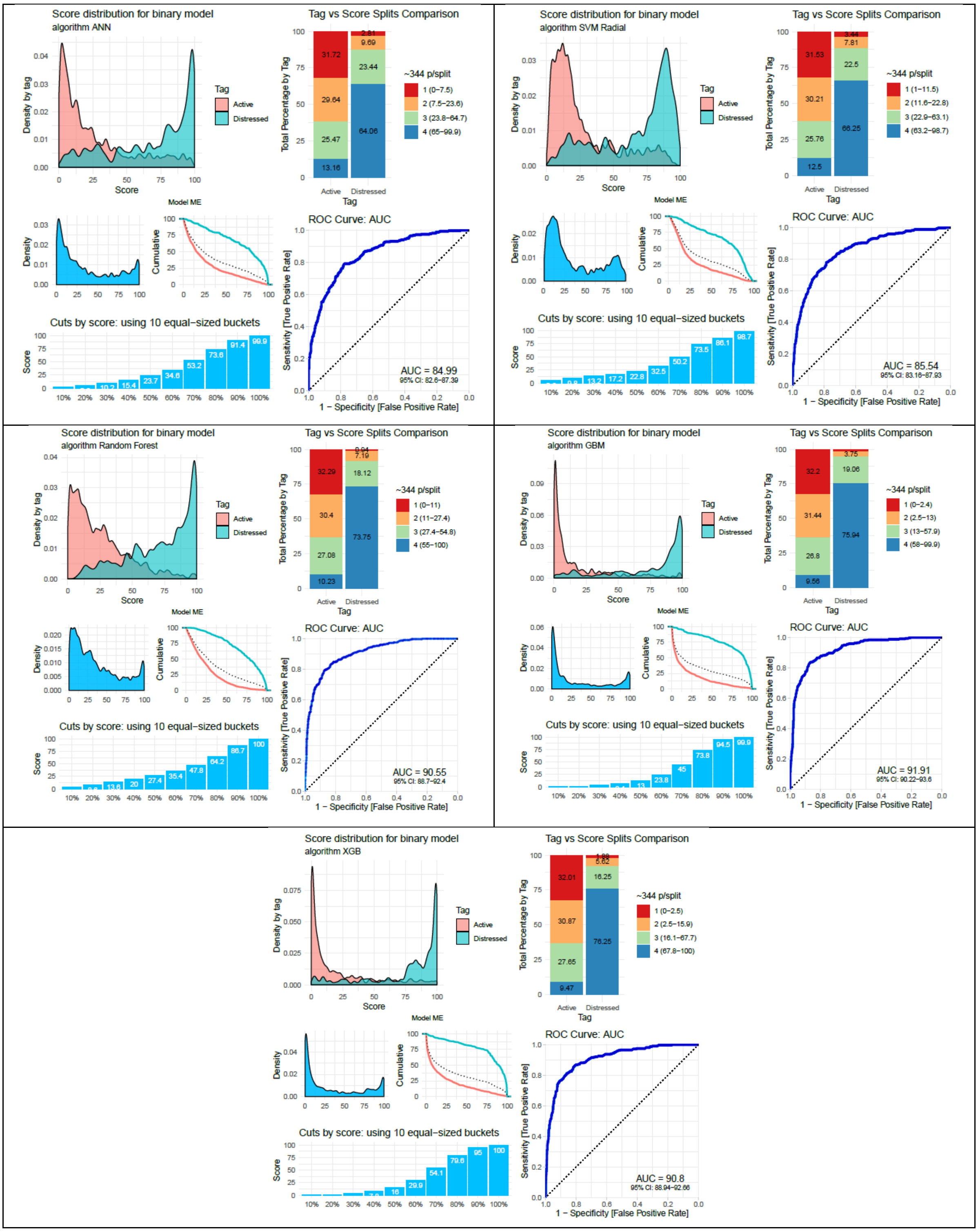


Figure 25: Model 2 – ROC results for complex classifiers

6.5.2.3 Model 3 - Bulk market covariate set with model 1

This section presents the results for each of the nine classification model applications on the bulk market covariate set combined with the company level covariates (model 3). The results for the traditional classifiers are shown in Figure 26 and Figure 27, for the complex classifiers. Again, as with model 2, the introduction of the bulk market predictors has little demonstrative effect on the ROC results for any of the classifiers. However, there is a positive effect on the sensitivity and type II error rates for each of the classifiers. As indicated in section 6.4, the market indicators for freight rates of iron ore and grain , scrap prices and the new build price index do appears to have a positive effect on the ability of model 3 to improve the false negative reporting rate when market predictors are combined with company level information.

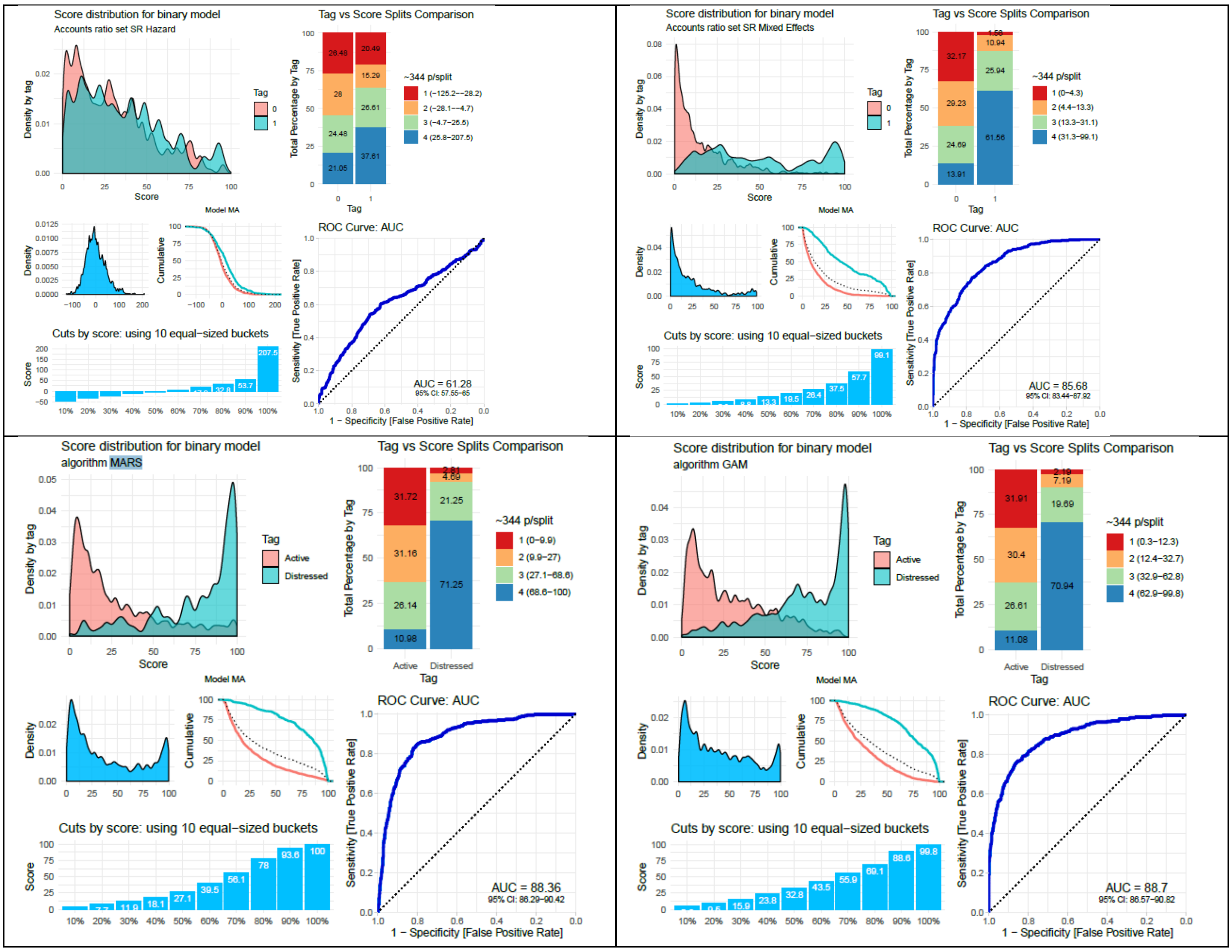


Figure 26: Model 3 – ROC results for traditional classifiers

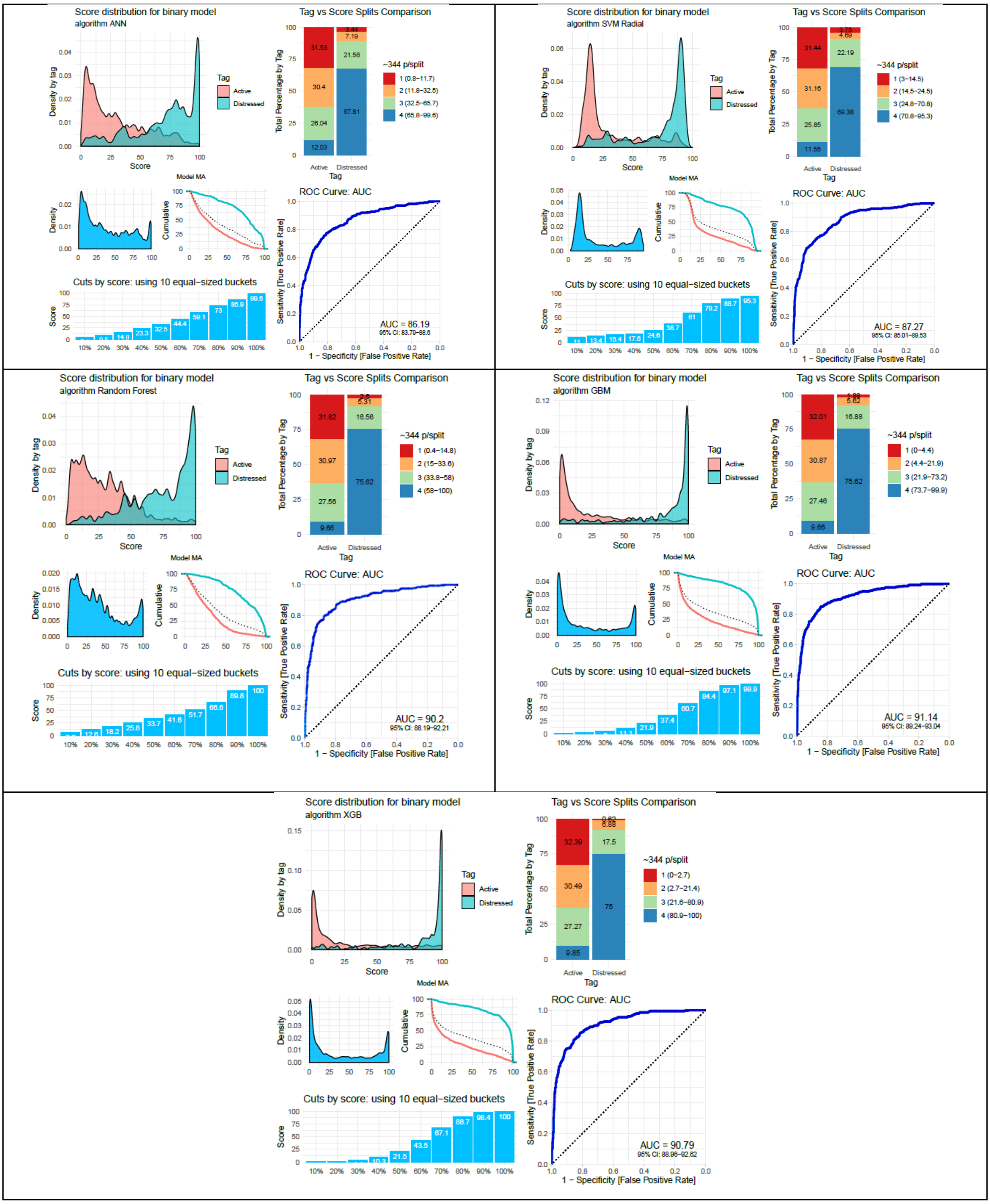


Figure 27: Model 3 - ROC results for complex classifiers

6.5.2.4 Model 4 – Complete covariate set

This section presents the results for each of the nine classification model applications on the combined covariate sets (model 4). The results for the traditional classifiers are shown in Figure 28 and those for the complex classifiers in Figure 29. The results from this model show little difference from that of model 3 which again indicates the strength of the company levels data combined with some market features in predicting FD in bulk shipping companies.

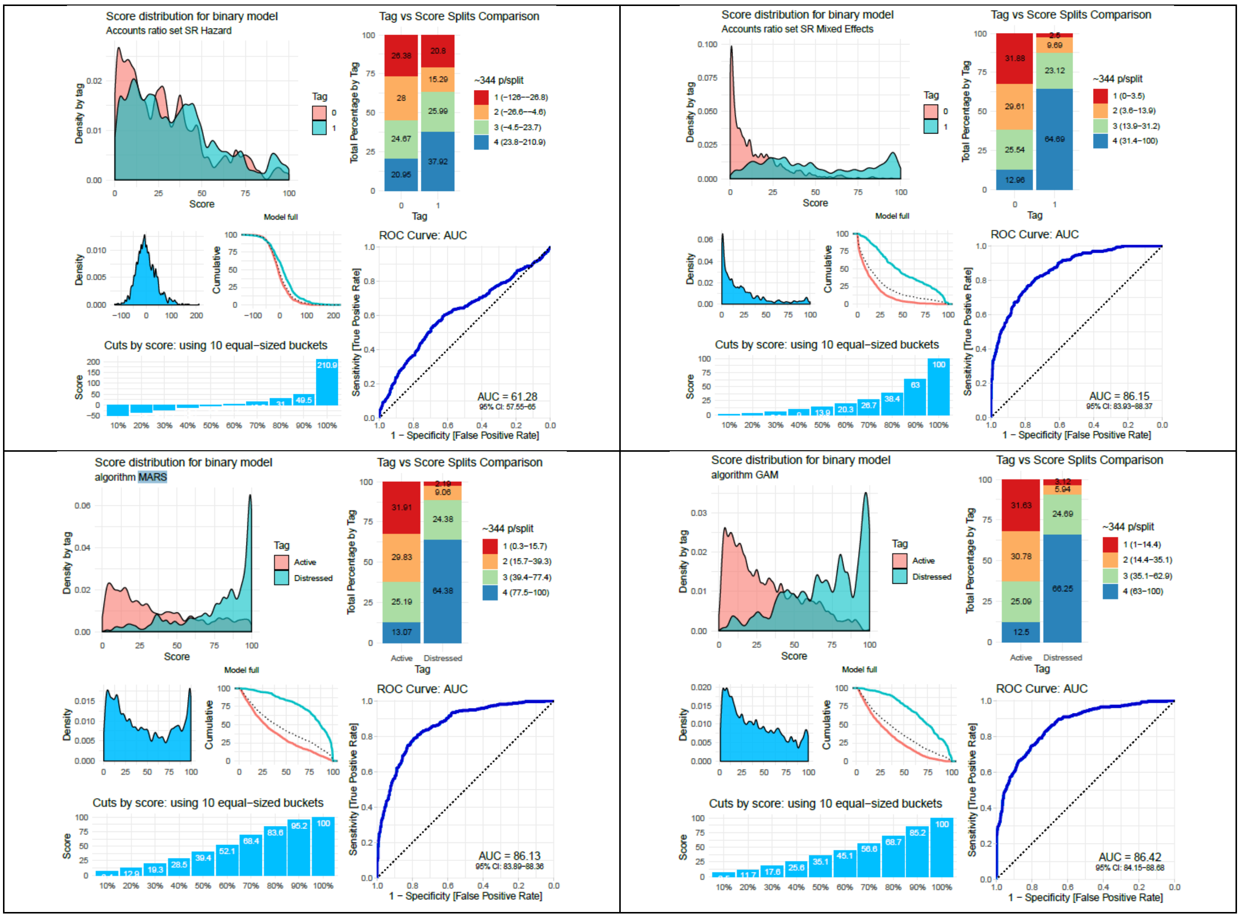


Figure 28: Model 4 – ROC results for traditional classifiers

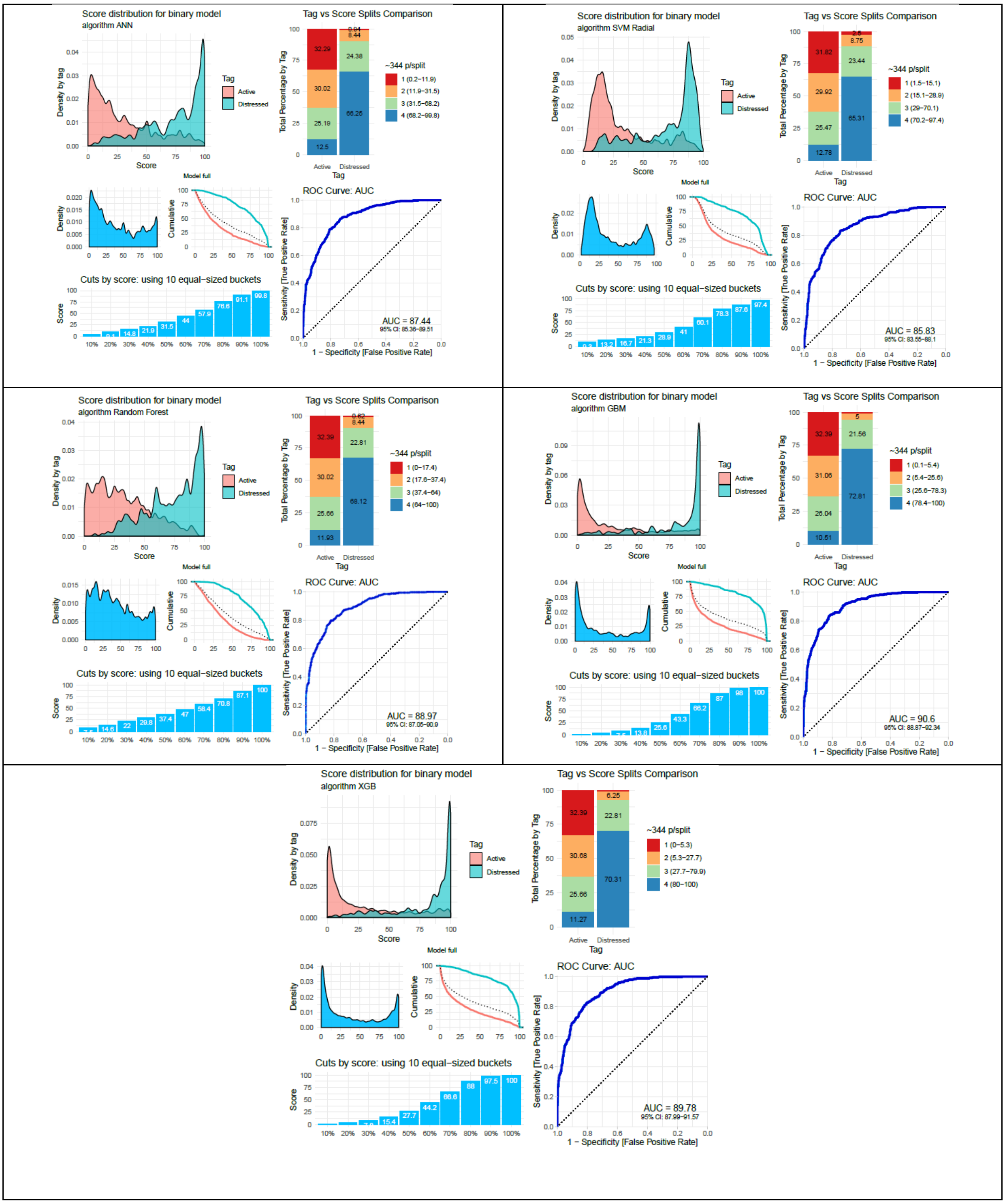


Figure 29: Model 4 – ROC results for complex classifiers

An interesting point to note is that the log loss figures for model 4 are largely unaffected by the introduction of both macroeconomic and market covariates despite their unobserved effectiveness in adding to the overall FD predictive power. In other words, where the macro and market predictor effectiveness is difficult to observe (or masked by stronger company predictors), their inclusion is not to the detriment of overall predictive power of the models tested.

In summarising, the tree-based classifiers demonstrate the most effectiveness in shipping company FD prediction. However, MARS and GAM have demonstrated strong enough predictive capacity to be included as part of any system for FD prediction due their transparency of operation as well as their predictive capacity. As regards the four model performances, the information contained in the company level data combined with some of the market predictors mention above, appears to represent a more parsimonious but effective predictor set, regardless the classifier selection. However, as indicated in section 6.4, macroeconomic and market indicators do play some demonstrable role in prediction even though this is largely hidden by the predictive power of the company level data. Care should be taken not to dismiss their effectiveness; however marginal the results may appear. This should be a subject of further investigation.

6.6 Consolidated results

This section focuses on the effect of applying each of the different missing value treatments (imputation techniques and a complete case scenario) on the classification results. The results are partitioned into four sets, the first three represent the results on an imputation method basis whilst the fourth set contains the "complete case" missing data management results. In this section, the focus is placed only on seven of the best performing classifiers with hazard and mixed effects extracted from the classifiers sets which are examined here. The results summarised here represent the performance of the classifiers using the metrics discussed in section 6.5.1.

The results are presented in five tables based primarily of the type of missing data technique used. The multivariate imputation methodologies are presented in Table 6 for k-NN, Table 7 for Hmisc and Table 8 for RF imputation. The results for the "complete case" treatment can be found in Table 9. A summary of the aggregated performances of each of the missing data techniques are found in Table 10. These results are presented for each of the four data models.

With respect to the individual classifiers, their overall performance on ROC metrics and log loss indicate better overall performance of the more complex models (ANN, SVM, RF, GBM and XBG) compared with the traditional classifiers (MARS and GAM). The performance difference between the two sets is not as pronounced as described in previous research findings, however, the performance of the complex models can be improved by further hyperparameter optimisation (particularly in the case of XGB). Furthermore, a closer examination of the ROC metrics, particularly the type II error and H measure results, indicate higher performances from each of the three tree-based models of RF, GBM and XGB over the linear models, ANN and SVM. This is also true of their corresponding tree-based log loss results.

As regards overall classifier performance, Figure 30 and Figure 31 indicate the superior performance of the tree-based classifiers (GBM, XGB and RF), which is consistent with results, published to date, of similar research comparing classification algorithms in finance. This performance is somewhat reflected in Figure 32 where the distress rate prediction capacity of each classifier is plotted. However, the interesting point depicted here, is that none of the classifiers failed to track the events leading up to or immediately following the 2008 crash. It is apparent from these annual plots that the algorithms found difficulty tracking evolution of FD rates between 2007 and 2011. This indicates that the SFDP performance seriously degrades in the presence of a sudden "seismic" macroeconomic event.

Figure 30 is also a visual demonstration the apparent absence of evolution of each of the algorithm's FD performance as the macroeconomic and then market features are introduced.

Taking the averaged results for each data model across each of the four missing data management methods (see Table 10) provides a sense of the evolution of the covariate modelling across these methods. The overall AUC show that the Hmisc method has the highest response of the three MI methods across the four data models. However, this is only marginally greater overall than the RF MI results. Of the other main ROC indicators, RF MI demonstrates superior performance on sensitivity, type II error and H measures. Both RF and Hmisc MI show similar performance levels on log loss results. Of all of the four techniques, however, the complete case treatment consistently outperforms each of the MI techniques, except for log loss results, where the RF MI scores lowest (therefore ranks highest). Therefore, given the bias or distortions introduced by the removal of large numbers of incomplete observations (annual accounts), through the allocation of a "complete case" approach, the RF multi-variate imputed data set may represent a more reliable approach. However, further investigation would be the first step before deciding against the "complete case" management.

The results of this study suggest that the tree-based classifiers may require minimal data intervention as their predictive performance appears largely immune to the shape and structure of input variables. Our findings confirm previous research that these classifiers seem to be robust to outliers and missing values; and appear largely insensitive to the monotone transformation of input variables. The strong out-of-sample predictive performance of these classifiers, particularly on the more challenging longitudinal test sample, also appears to confirm previous findings that these models are resilient to model over-fitting, which typically manifests in weaker out-of-sample predictive accuracy relative to the training sample (Hastie et al., 2009; Schapire and Freund, 2012). However, as reported by (Christoffersen, Matin and Mølgaard, 2018) the differences between the results achieved through complex models compared with the linear based models is not as pronounced as previous studies have suggested e.g. (Jones, Johnstone and Wilson, 2015b).

Finally, using linear classifiers which are flexible enough to model non-linear input-output relationships reveal better overall predictive performance than generalise linear models. For example, the GAM and MARS classifiers consistently outperformed mixed effects model and compared relatively favourably with the more complex models.

Table 6: Classifier performance – All models (K-NN imputation)

Model 1

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.874	0.716	0.849	28.44%	0.386	0.549
GAM	0.858	0.731	0.815	26.88%	0.369	0.416
ANN	0.864	0.741	0.830	25.94%	0.429	0.476
SVM radial	0.882	0.747	0.829	25.31%	0.402	0.400
RF	0.920	0.744	0.890	25.63%	0.497	0.360
GBM	0.911	0.734	0.897	26.56%	0.508	0.363
XGB	0.902	0.706	0.885	29.38%	0.459	0.414

Model 2

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.808	0.876	0.599	12.41%	0.348	0.547
GAM	0.880	0.828	0.580	17.23%	0.450	0.419
ANN	0.870	0.813	0.554	18.75%	0.406	0.466
SVM radial	0.865	0.813	0.552	18.75%	0.402	0.422
RF	0.897	0.880	0.655	12.03%	0.452	0.365
GBM	0.898	0.853	0.613	14.68%	0.466	0.374
XGB	0.896	0.865	0.626	13.54%	0.475	0.418

Model 3

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.862	0.741	0.808	25.94%	0.365	0.551
GAM	0.866	0.734	0.825	26.56%	0.401	0.411
ANN	0.869	0.744	0.831	25.63%	0.382	0.450
SVM radial	0.862	0.744	0.841	25.63%	0.396	0.424
RF	0.889	0.719	0.858	28.13%	0.418	0.361
GBM	0.900	0.744	0.869	25.63%	0.471	0.357
XGB	0.891	0.734	0.857	26.56%	0.435	0.405

Model 4

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.880	0.803	0.798	19.69%	0.434	0.570
GAM	0.742	0.800	0.684	20.00%	0.158	0.990
ANN	0.855	0.763	0.780	23.75%	0.389	0.455
SVM radial	0.872	0.800	0.799	20.00%	0.356	0.443
RF	0.881	0.800	0.798	20.00%	0.399	0.367
GBM	0.893	0.766	0.825	23.44%	0.447	0.364
XGB	0.886	0.766	0.808	23.44%	0.399	0.415

Table 7: Classifier performance – All models (Hmisc imputation)

Model 1

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.902	0.794	0.844	20.63%	0.318	0.549
GAM	0.924	0.841	0.824	15.94%	0.477	0.366
ANN	0.916	0.813	0.845	18.75%	0.443	0.414
SVM radial	0.928	0.844	0.864	15.63%	0.440	0.359
RF	0.93	0.713	0.944	28.75%	0.587	0.214
GBM	0.935	0.675	0.942	32.50%	0.436	0.229
XGB	0.927	0.647	0.950	35.31%	0.453	0.335

Model 2

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.919	0.809	0.874	19.06%	0.267	0.544
GAM	0.917	0.819	0.844	18.13%	0.308	0.34
ANN	0.913	0.791	0.856	20.94%	0.302	0.403
SVM radial	0.915	0.838	0.850	16.25%	0.396	0.359
RF	0.947	0.750	0.931	25.00%	0.385	0.212
GBM	0.942	0.784	0.934	21.56%	0.442	0.215
XGB	0.941	0.763	0.935	23.75%	0.426	0.327

Model 3

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.892	0.744	0.891	25.63%	0.178	0.546
GAM	0.897	0.806	0.820	19.38%	0.085	0.95
ANN	0.902	0.809	0.842	19.06%	0.306	0.394
SVM radial	0.886	0.784	0.830	21.56%	0.236	0.366
RF	0.907	0.700	0.908	30.00%	0.341	0.234
GBM	0.917	0.738	0.904	26.25%	0.307	0.237
XGB	0.9	0.659	0.936	34.06%	0.382	0.346

Model 4

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.88	0.672	0.884	32.81%	0.289	0.574
GAM	0.772	0.678	0.866	32.19%	0.317	0.99
ANN	0.885	0.747	0.839	25.31%	0.271	0.409
SVM radial	0.875	0.744	0.837	25.63%	0.366	0.38
RF	0.882	0.606	0.922	39.38%	0.443	0.243
GBM	0.912	0.684	0.920	31.56%	0.449	0.231
XGB	0.903	0.625	0.940	37.50%	0.427	0.337

Table 8: Classifier performance – All models (Rf imputation)

Model 1

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.889	0.781	0.829	21.88%	0.443	0.549
GAM	0.868	0.772	0.792	22.81%	0.382	0.449
ANN	0.861	0.753	0.797	24.69%	0.422	0.482
SVM radial	0.888	0.809	0.811	19.06%	0.381	0.431
RF	0.931	0.856	0.855	14.38%	0.538	0.398
GBM	0.931	0.838	0.868	16.25%	0.544	0.398
XGB	0.920	0.800	0.859	20.00%	0.499	0.438

Model 2

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.880	0.772	0.821	22.81%	0.428	0.545
GAM	0.864	0.759	0.811	24.06%	0.413	0.432
ANN	0.850	0.731	0.816	26.88%	0.358	0.470
SVM radial	0.855	0.728	0.828	27.19%	0.353	0.447
RF	0.905	0.794	0.869	20.63%	0.493	0.406
GBM	0.919	0.806	0.884	19.38%	0.538	0.397
XGB	0.908	0.819	0.845	18.13%	0.505	0.433

Model 3

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.884	0.828	0.817	17.19%	0.411	0.556
GAM	0.887	0.809	0.799	19.06%	0.448	0.446
ANN	0.862	0.797	0.767	20.31%	0.421	0.471
SVM radial	0.873	0.778	0.781	22.19%	0.400	0.474
RF	0.902	0.819	0.846	18.13%	0.500	0.420
GBM	0.911	0.863	0.812	13.75%	0.528	0.411
XGB	0.908	0.872	0.783	12.81%	0.502	0.446

Model 4

Classifier	AUC	Sensitivity	Specificity	Type II Err	H Measure	Log loss
MARS	0.861	0.838	0.712	16.25%	0.356	0.572
GAM	0.864	0.769	0.775	23.13%	0.388	0.950
ANN	0.874	0.806	0.776	19.38%	0.363	0.475
SVM radial	0.858	0.778	0.784	22.19%	0.389	0.476
RF	0.890	0.831	0.775	16.88%	0.431	0.397
GBM	0.906	0.850	0.786	15.00%	0.470	0.407
XGB	0.898	0.841	0.772	15.94%	0.439	0.438

Table 9: Classifier performance – All models (complete case treatment)

Model 1

Classifier	AUC	Sensitivity	Specificity	Type II err	H Measure	Log loss
MARS	0.833	0.659	0.846	34.09%	0.443	0.493
GAM	0.932	0.864	0.853	13.64%	0.382	0.482
ANN	0.933	0.909	0.846	9.09%	0.422	0.487
SVM radial	0.934	0.932	0.788	6.82%	0.381	0.431
RF	0.949	0.909	0.848	9.09%	0.538	0.358
GBM	0.943	0.909	0.838	9.09%	0.544	0.461
XGB	0.949	0.932	0.864	6.82%	0.499	0.410

Model 2

Classifier	AUC	Sensitivity	Specificity	Type II err	H Measure	Log loss
MARS	0.905	0.841	0.798	15.91%	0.428	0.501
GAM	0.881	0.795	0.791	20.45%	0.413	0.514
ANN	0.895	0.795	0.822	20.45%	0.358	0.497
SVM radial	0.901	0.795	0.859	20.45%	0.353	0.426
RF	0.942	0.955	0.798	4.55%	0.493	0.392
GBM	0.951	0.955	0.843	4.55%	0.538	0.446
XGB	0.934	0.886	0.822	11.36%	0.505	0.428

Model 3

Classifier	AUC	Sensitivity	Specificity	Type II err	H Measure	Log loss
MARS	0.810	0.659	0.801	34.09%	0.411	0.544
GAM	0.771	0.773	0.770	22.73%	0.448	0.990
ANN	0.844	0.750	0.785	25.00%	0.421	0.493
SVM radial	0.879	0.841	0.767	15.91%	0.400	0.423
RF	0.901	0.773	0.859	22.73%	0.500	0.397
GBM	0.894	0.795	0.851	20.45%	0.528	0.447
XGB	0.914	0.864	0.822	13.64%	0.502	0.409

Model 4

Classifier	AUC	Sensitivity	Specificity	Type II err	H Measure	Log loss
MARS	0.900	0.886	0.809	11.36%	0.356	0.513
GAM	0.887	0.841	0.830	15.91%	0.388	0.990
ANN	0.883	0.727	0.832	27.27%	0.363	0.522
SVM radial	0.878	0.818	0.725	18.18%	0.389	0.449
RF	0.925	0.864	0.814	13.64%	0.431	0.419
GBM	0.938	0.955	0.743	4.55%	0.470	0.451
XGB	0.926	0.841	0.843	15.91%	0.439	0.424

Table 10: Model performance metrics – averaged metrics

Model 1						
	AUC	**Sensitivity**	**Specificity**	**Type II Err**	**H Measure**	**Log loss**
k-NN	0.887	0.731	0.856	26.88%	0.436	0.425
Hmsic	0.923	0.761	0.887	23.9%	0.450	0.352
Rf	0.898	0.801	0.830	19.87%	0.458	0.449
Complete case	0.92	0.87	0.84	12.66%	0.46	0.45
Model 2						
k-NN	0.873	0.847	0.597	15.34%	0.429	0.430
Hmisc	0.928	0.793	0.889	20.7%	0.361	0.343
Rf	0.883	0.773	0.839	22.72%	0.441	0.447
Complete case	0.92	0.86	0.82	13.96%	0.44	0.46
Model 3						
k-NN	0.877	0.737	0.841	26.29%	0.410	0.423
Hmisc	0.900	0.749	0.876	25.1%	0.262	0.439
Rf	0.890	0.824	0.801	17.63%	0.458	0.461
Complete case	0.86	0.78	0.81	22.08%	0.46	0.53
Model 4						
k-NN	0.858	0.785	0.785	21.47%	0.369	0.515
Hmisc	0.873	0.679	0.887	32.1%	0.366	0.542
Rf	0.879	0.816	0.768	18.39%	0.405	0.513
Complete case	0.91	0.85	0.80	15.26%	0.41	0.54

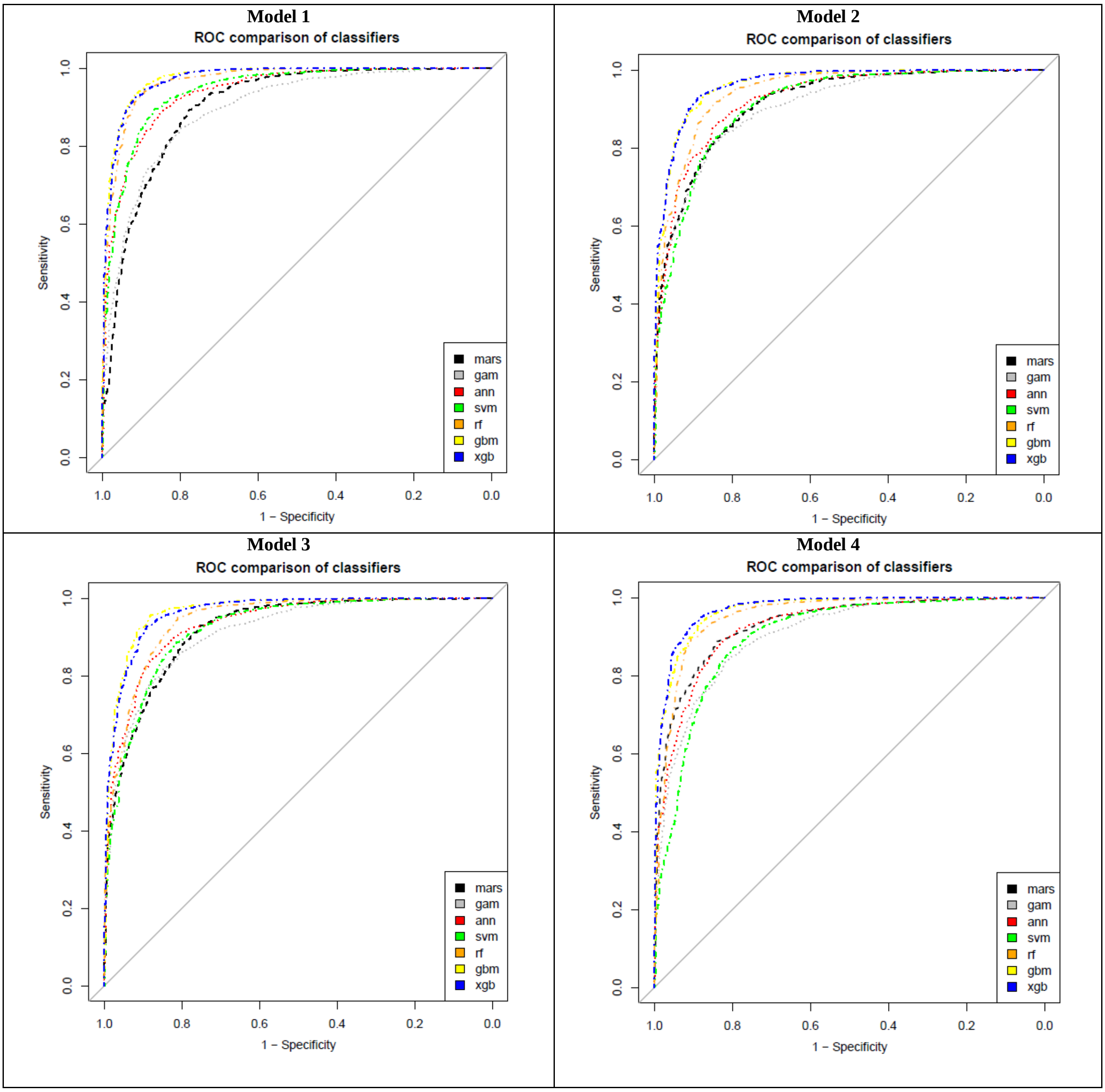


Figure 30: Individual classifier comparison ROC curves

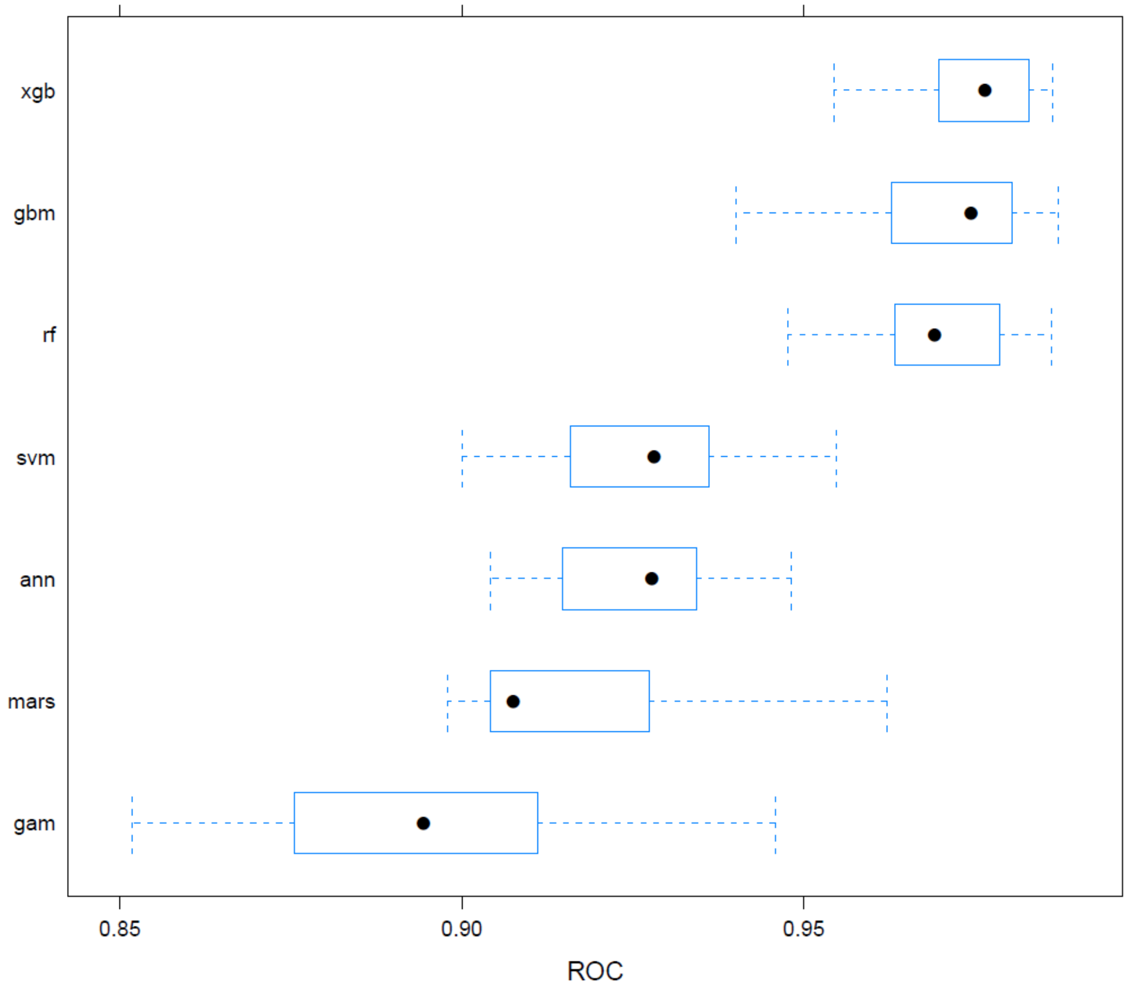


Figure 31: Box plots – Comparison of classifiers

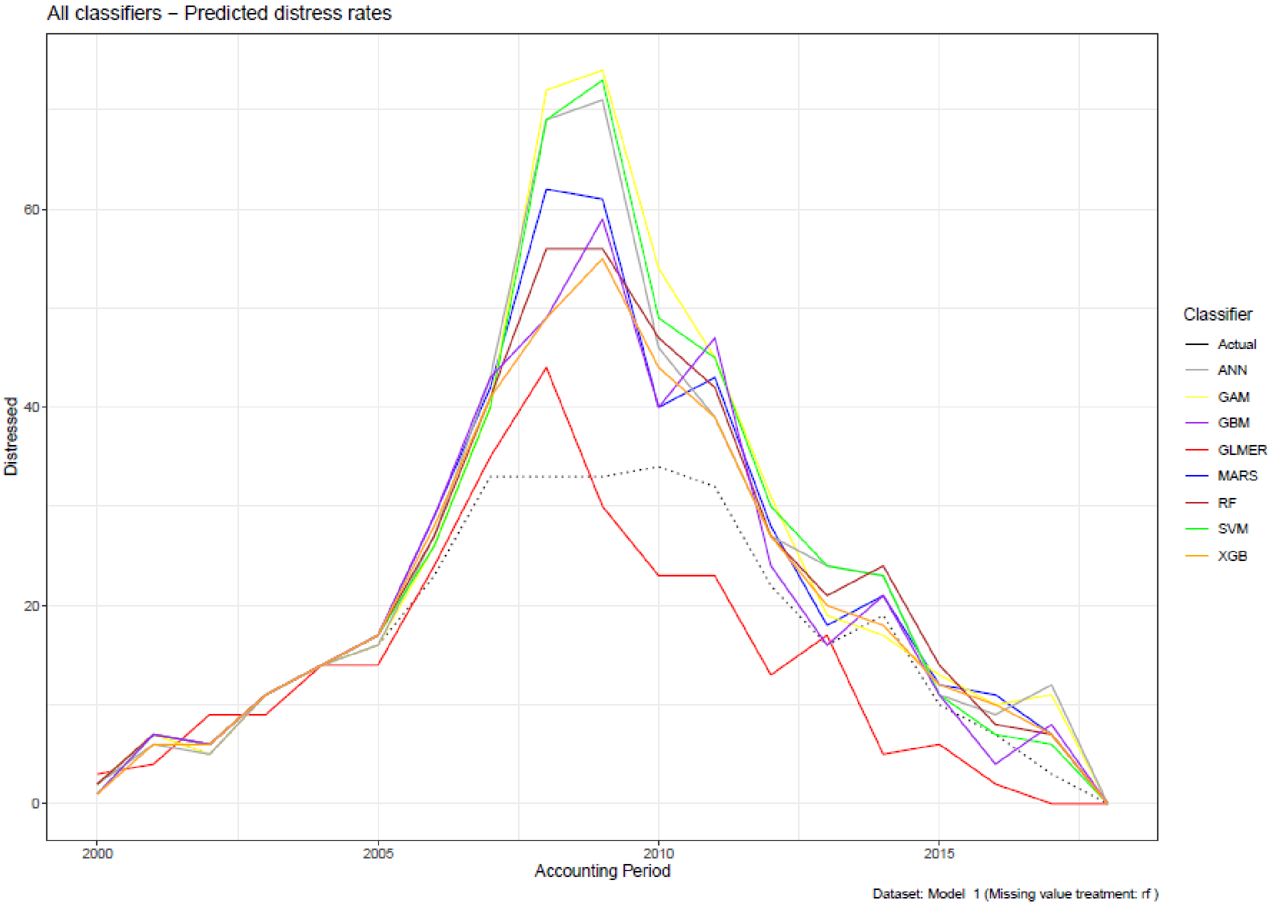


Figure 32: Distress rate predictors - All classifiers

6.7 Summary

A presentation and discussion of the case study test results was presented in this chapter. The chapter began with the results from the missing value analysis and management. The missing value analysis demonstrated that approximately 18% of the case study dataset contained missing values. It also demonstrated little autocorrelation and multi-collinearity, indicating the importance of each of the selected covariates in the dataset. This suggested that case wise deletion or the removal of covariates would introduce bias however, based on the metric selected for the evaluations of the modelling in this study, the complete case treatment of missing values produced the best overall ROC results compared with the MI techniques tested. The results also indicate the value of data imputation, with the random forest methodology demonstrating the best performance amongst the imputation techniques and producing ROC results close to those of complete case treatment; and even improving on the log loss performance of complete case.

The contribution of the macroeconomic and market based predictors in the enhancement of FD prediction was demonstrated as being present but masked by the stronger company level financial covariates. This is most evident with the tree-based classifiers where the ability to extract information from “parsimonious” predictor sets is strongest.

Finally, the results confirm the general consensus in the literature that the more complex classifiers perform strongest on corporate FD prediction than linear based classifiers. This is especially evident from the results for the tree-based classifiers RF, GBM and XGB. However, an import point of interest is the performance of linear based classifiers, MARS and GAM, which demonstrate performance levels close enough to those of the more complex classifiers to warrant their inclusion in FD modelling, given their transparent nature.

7 Conclusions

This final thesis chapter discusses the conclusions resulting from this research. The chapter begins with a review of the research objectives and discusses how and to what degree these objectives were achieved. The research results have demonstrated a practical novelty in its contribution to the field of shipping FD prediction and this is described in section 7.2. The research has also precipitated reflections as to how this work can be developed through further research effort and these are described in section 7.3. The chapter closes with some final remarks in section 7.4.

7.1 Achievements against the objectives

This research had two primary objectives. The first was to design and test a forecasting methodology, the Shipping Financial Distress Prediction (SFDP) model. The goal was to capture correlations between company financial information and other macro events in shipping company financial distress. The second objective was to take the SFDP model structure and construct a machine learning tool kit. This tool kit aims at providing software developers a foundation for the construction of a financial risk software application which can be readily integrated into a corporate/enterprise wide risk assessment system.

7.1.1 SFDP model objectives

This first set of research objectives addressed:

- issues surrounding the “noisy” nature of company financial data typically associated with shipping companies, such as information skewness, data imbalance and missing accounting information
- the identification of a set of predictor features which represent a predictor set capable of capturing correlation in financial distress prediction for the sector
- the non-company specific event affecting shipping company performance through the inclusion macroeconomic and market data in the forecasting of distress.

The first and arguably most important confirmation recorded in this section is that, as commonly understood in the literature, obtaining complete and accurate financial statements from shipping companies on a global basis is somewhat challenging. Shipping companies by their very nature are founded upon mobile assets which operate on an international basis and which can change ownership frequently during their economic life. Companies and their vessels are not necessarily registered in the states of their owners. Some states have different accounting laws and practices with some positively promoting financial secrecy. Balance sheets, even when available, do not necessarily reflect the true state of asset valuation and P&L accounts do not necessarily reflect the true costs or revenue attained in any single accounting period. As an example, even on a global basis, only in 2018 was international financial regulation changed to force companies to properly and transparently account for the sale and leaseback of vessels. For any study, whose goal is to construct a system which can predict financial distress, with a generalisation out of sample rate accurate enough to be used in practice, must be executed in this context. It must collect a dataset large enough to not only to include as many shipping entities as possible and over a significantly wide timeframe in order to allow the algorithms to deal with problems inherent with missing information.

This study assessed the predictive performance of nine binary classifiers on a large sample of international bulk carrier shipping company financial statement data. This longitudinal data set was gathered for accounting periods from 2000 to 2018 and contains over 5000 company/year observations. Prior to modelling the data were examined in order to assess how to deal with the problem of missing values. It was discovered that around 72% of the company records contain some missing covariate values, with a total of circa 18% missing values in the dataset. Modelling was then formulated using four, reconstructed, versions of the data, three were developed using established imputation techniques and one dataset formulated by including only those records/observations which contained complete information ("complete case").

Four models where then formulated and tested on three data models. The first model, involving only company financial level data, formed a foundation for the construction of the other three models. It is tested whether the market and macroeconomic variables add information that is not contained in financial statements in order to complement this data and improve the performance of default prediction models. The results presented in this study indicate that whilst they contribute to the predictive capacity of financial distress models, the direct impact, of individual macro and market predictors, is less directly observable than their indirect impact which is captured though company level financial data…

Prior research into the use of company accounts, market and macroeconomic variables used to predict shipping company FD have been largely restricted to linear regression models and the results obtained from these models have been contentious, given the inherent limitations of linear models. Furthermore, most of these studies have focused on modelling using only relatively small, cross-sectional data sets which limits their predictive capacity and have not sufficiently accounted for frailty. A large, longitudinal data set allows for both improved accuracy of the estimates and provides for the inclusion of spatial effects in capturing dynamic relations. The models confirm previous findings in the finance literature of the importance of accounting for the existence of a time-varying frailty variable persisting in the presence of unobserved heterogeneity. Failure to allow for such unobserved factors leads to downward bias in risk estimates. The existence of this frailty in bulk shipping company FD risk is demonstrated using a linear mixed model applied to the panel data. Testing is then executed on the ability of a set of more complex non-linear models on their ability to capture the effects of frailty in the data. The results indicate that the complex classifiers such as extreme gradient boosting, gradient boosting and random forests outperformed other classifiers on the panel data samples. However, the performance of two of the linear based models, MARS and GAM where sufficiently performant to warrant further investigation into their inclusion in a system designed for distress prediction. Particularly due to their more transparent methodologies compared with the more complex models.

7.1.2 SFDP tool kit

The second research objective is to develop the foundation for a machine learning software system which:

- provides a general-purpose tool that is of real practical value
- results in predictions that are as accurate as possible
- takes full advantage of computing architecture, multi-processing, multi-core technologies and cloud computing, in order to both maximise efficiency and reduce execution times to a practical working level
- is both scalable and modular in form, supporting the addition of more complex deep learning algorithms which require enhanced computing power, including support for larger and more complex datasets
- can be readily applied to a broad class of shipping sub-sector FD learning problems
- must be transparent and open to scrutiny by all stakeholders, investors and particularly regulatory bodies if they are to be accepted as practical tools

Both the model and system software were tested on a case study using detailed financial statements covering the period 2000-2018 of dry bulk carrier owners/operator companies, worldwide, both listed and non-listed.

Section 4 described the SFDP system architecture and functionality. The application of this, through the case study in section 5, demonstrated the practical values of the system. A real dataset was established which composed of shipping company financial data, over a 20-year period, as well as macro and market data for the same period (lagged). This dataset formed input to a rudimentary financial risk application which was developed specifically for the demonstration of the SFDP toolkit functionality. Basing this system on a multi-platform, parallel processing and scalable software suite, R ensures the SFDP software meets the objective set out.

Furthermore, using the bulk shipping sub-sector as a case study, allowed for a completely representative sub-set of global shipping and demonstrated that the system can be applied to other shipping sub-sectors with little or no modification to the base functionality.

Finally, the inclusion of linear mixed models, and more complex linear based algorithms such as MARS and GAM, provides for transparency required by the potential user community and regulators.

7.2 Contribution to the field

The aims and intended contributions, outlined in section 2.4, defined the goals of the methodology and system software which were developed as the prime objectives of this research. The application of the case study is primarily aimed at identifying the degree to which these research objectives have been achieved. In this context, and with reference to the results discussed in section 6, the contributions to the field which have been realised through this research are as follows:

i. The successful development of a methodology incorporating a) advanced linear modelling, which have the capacity to account for non-linear effects and b) advanced tree-based classification machine leaning algorithms, demonstrated a clear performance advantage over generalised linear models in the domain of FD prediction. The predictive capacity of the methodology was successfully tested against a large longitudinal, noisy shipping dataset. Furthermore, it was demonstrated that the inclusion of macro-economic and market data, augmenting company level data, contribute to the FD predictive capacity of both linear and complex classification algorithms. However, it must be noted that the contribution of these features was demonstrated as having limited *direct* effect. Nevertheless, it is reasonable to believe that macro and market information is already absorbed within the reported financial information. This was successfully demonstrated through the utilisation of a large longitudinal company financial dataset which was collected for the case study.

ii. The effective application of a formal a multivariate imputation (MI) methodology which addresses the problem of missing values in shipping company accounts. This produced results (see section 6.3) which improved, or at least matched, the predictive capacity of the classifiers incorporated into the methodology, when compared with case or covariate wide (complete case) deletion. This is an area which has previously received little or no attention in the research of shipping company accounting data.

iii. The successful development, application and testing of a scalable and flexible software toolkit for shipping company FD modelling has been demonstrated by the application of the system to an industry specific case study based on the global bulk shipping sector. The 20-year longitudinal dataset of over 5000 bulk shipping company year accounts provided for an adequately representable test bed for this research. This is, to the best knowledge of the author, the first practical realisation of such as system which has been applied to such large company/year longitudinal dataset. The results have demonstrated the successful implementation of this toolkit/methodology and how such a system can be readily incorporated into a corporate risk management system. Investors and shipowners alike can track performance and save the costs associated with severe financial distress through the implementation of such a system when deployed in a real-time environment, i.e. the system can act as a financial *early warning system* for all stakeholders. This is unprecedented in shipping company risk management software application development.

7.3 Further research

This research has demonstrated that shipping company FD prediction can be improved through a structured, formal application of available statistical tools, comprising both traditional and modern complex ML algorithms. The research has however, also identified the following areas which would benefit from further research:

i. Extending the company level predictor set to widen the unobserved predictors not yet captured by currently identified independent variable set. For example, sustainability and fleet risk profile data
ii. Further extend to classification capacity of the toolkit through the inclusion of deep learning techniques

7.4 Final remarks

In summary, the goal of this research was to develop a practical methodology and software toolkit which enables the early detection of FD in shipping companies. The aim was to provide investors and other stakeholders with the means of avoiding some of the costs associated with a bankruptcy filing and subsequent recovery but could also assist shipowners in monitoring their own financial performance. This has clearly been demonstrated through the application of the bulk shipping test case. Unfortunately, the research also demonstrated that, given the nature of the available data, even complex models do not achieve an accuracy level, at present, to be relied upon for use as anything other than an early 'warning system' for financial distress. However, this research has concluded that, with a sufficiently large test set coupled with heavy and methodologically sound and transparent data pre-processing, viable and practical shipping company FD warning system is possible.

8 References

Akande, O., Li, F. and Reiter, J. (2017) 'An Empirical Comparison of Multiple Imputation Methods for Categorical Data', *American Statistician*, 71(2), pp. 162–170. doi: 10.1080/00031305.2016.1277158.

Ali, A. and Daly, K. (2010) 'Macroeconomic determinants of credit risk: Recent evidence from a cross country study', *International Review of Financial Analysis*. Elsevier Inc., 19(3), pp. 165–171. doi: 10.1016/j.irfa.2010.03.001.

Altman, E. I. (1968) 'Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy', *The Journal Of Finance*, 23(4), pp. 589–609.

Altman, E. I. (1977) 'ZETA ANALYSIS A new model to identify bankruptcy risk of corporations', *Journal of Banking a*, 1, pp. 29–54.

Amel-zadeh, A. *et al.* (2020) 'Machine Learning-Based Financial Statement Analysis ∗'.

Andreou, P. C., Louca, C. and Panayides, P. M. (2014) 'Corporate governance, financial management decisions and firm performance: Evidence from the maritime industry', *Transportation Research Part E: Logistics and Transportation Review*. Elsevier Ltd, 63, pp. 59–78. doi: 10.1016/j.tre.2014.01.005.

Antoniou, A., A. Thanopoulos, and C. G. (1998) 'An Attempt to Quantify Individual Factors of the Five C's of Credit Risk Analysis in Bank Shipping Finance.', *Department of Shipping, Trade and Finance, City University Business School, London.*

Balcaen, S. and Ooghe, H. (2006) '35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems', *British Accounting Review*, 38(1), pp. 63–93. doi: 10.1016/j.bar.2005.09.001.

Barboza, F., Kimura, H. and Altman, E. (2017) 'Machine learning models and bankruptcy prediction', *Expert Systems with Applications*. Elsevier Ltd, 83, pp. 405–417. doi: 10.1016/j.eswa.2017.04.006.

Barnes, P. (1987) 'The Analysis and Use of Financial Ratios: A Review Article', *Journal of Business Finance & Accounting*, 14(4), pp. 449–461. doi: 10.1111/j.1468-5957.1987.tb00106.x.

Berg, D. (2007) 'Bankruptcy prediction by generalized additive models', *John Wiley & Sons, Ltd.* doi: 10.1002/asmb.

Bergantino, A. and Veenstra, A. (2002) 'Principal agent problems in shipping: An investigation into the formation of charter contracts', *IAME 2002 Conference Proceedings*.

BIS (2019) *Minimum capital requirements for market risk.*

Black, F. and Scholes, M. (1973) 'The pricing of options and corporate liabilities', *Journal of Political Economy*, 81(3), pp. 637–657. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85015692260&partnerID=40&md5=18e58714f8857b5a350e8ca2d6eae1c6.

Bonér, J. *et al.* (2014) 'The reactive manifesto.(2014)', *URL: http://www. reactivemanifesto. org*, 8.

Bonfim, D. (2009) 'Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics', *Journal of Banking and Finance*. Elsevier B.V., 33(2), pp. 281–299. doi: 10.1016/j.jbankfin.2008.08.006.

Box, G. and Cox, D. (1964) 'The Analysis of Transformations', *Journal of the Royal Statistical Society*, (Series B (Methodological)), pp. 211–252.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Bryman, A. (2016) *Social research methods*. Oxford university press.

Bureau-Van-Dijk (2019) *Orbis*, *Orbis*. Available at: https://orbis.bvdinfo.com/version-201989/Home.serv?product=orbisneo.

Burger, S., Silverman, S. and Vuuren, G. van (2018) 'Deriving Correlation Matrices for Missing Financial Time-Series Data', *International Journal of Economics and Finance*, 10(10), p. 105. doi: 10.5539/ijef.v10n10p105.

Van Buuren, S. (2018) *Flexible imputation of missing data*. Chapman and Hall/CRC.

Carmona, P., Climent, F. and Momparler, A. (2019) 'Predicting failure in the U.S. banking sector: An extreme gradient boosting approach', *International Review of Economics and Finance*. Elsevier Ltd, 61(March), pp. 304–323. doi: 10.1016/j.iref.2018.03.008.

Chang, Y. C., Chang, K. H. and Wu, G. J. (2018) 'Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions', *Applied Soft Computing Journal*. Elsevier B.V., 73, pp. 914–920. doi: 10.1016/j.asoc.2018.09.029.

Chaudhuri, A. (2013) 'Bankruptcy Prediction Using Bayesian, Hazard, Mixed Logit and Rough Bayesian Models: A Comparative Analysis', *Computer and Information Science*, 6(2). doi: 10.5539/cis.v6n2p103.

Chawla, N. V. *et al.* (2002) 'SMOTE: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357. doi: 10.1613/jair.953.

Chen, J. and Shao, J. (2000) 'Nearest Neighbor Imputation for Survey Data', *Journal of official statistics*, 16(2), pp. 113–132.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. doi: 10.1145/2939672.2939785.

Cheng, Ioannou, S. (2014) 'Corporate Social Responsibility and access to finance', *Academy of Management Journal*, 51(2), pp. 315–334. doi: 10.1002/smj.

Christoffersen, B., Matin, R. and Mølgaard, P. (2018) 'Can Machine Learning Models Capture Correlations in Corporate Distresses?', *SSRN Electronic Journal*, pp. 1–34. doi: 10.2139/ssrn.3273985.

Clarksons (2019) *Shipping Intelligence Network, Shipping Intelligence Network*. Available at: https://sin.clarksons.net/.

Daily, C. M. and Dalton, D. R. (1994) 'Corporate governance and the bankrupt firm: An empirical assessment', *Strategic Management Journal*, 15(8), pp. 643–654. doi: 10.1002/smj.4250150806.

Das, S. R. *et al.* (2007) 'Common Failings - How Corporate defaluts are correlated', LXII(1), pp. 93–117.

Demirtas, H., Freels, S. A. and Yucel, R. M. (2008) 'Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment', *Journal of Statistical Computation and Simulation*. Taylor & Francis, 78(1), pp. 69–84.

Dempster, Laird and Rubin (2011) *MAXIMUM LIKELIHOOD ESTIMATION FROM INCOMPLETE DATA VIA EM-TYPE ALGORITHMS BT - Advanced Medical Statistics*, *Advanced Medical Statistics*. Available at: http://www.worldscientific.com/doi/abs/10.1142/9789812388759_0028%5Cnpapers3://publication/doi/10.1142/9789812388759_0028.

Dicesare, G. (2006) 'Imputation, Estimation and Missing Data in Finance'.

Dimitras, A. I., Petropoulos, T. and Constantinidou, I. (2003) 'Multi-criteria Evaluation of Loan Applications in Shipping', 246(September 2002), pp. 237–246. doi: 10.1002/mcda.332.

Doove, L. L., Van Buuren, S. and Dusseldorp, E. (2014) 'Recursive partitioning for missing data imputation in the presence of interaction effects', *Computational Statistics and Data Analysis*. Elsevier B.V., 72, pp. 92–104. doi: 10.1016/j.csda.2013.10.025.

Duffie, D. *et al.* (2009) 'Frailty Correlated Default', *Journal of Finance*, 64(5), pp. 2089–2123. Available at: http://econpapers.repec.org/article/blajfinan/v_3a64_3ay_3a2009_3ai_3a5_3ap_3a2089-2123.htm.

Duffie, D. and Singleton, K. (2003) 'Credit Risk: Pricing, Measurement and Management', *Princeton University Press*, (October), pp. 1–6. doi: 10.2139/ssrn.600545.

Equasis (2019) *The world Merchant Fleet 2018*, *Equasis statistics service*. Available at: http://www.equasis.org/Fichiers/Statistique/MOA/Documents availables on statistics of Equasis/Equasis Statistics - The world fleet 2018.pdf.

Fielding, N. G., Lee, R. M. and Blank, G. (2008) *The SAGE handbook of online research methods*. Sage.

Figlewski, S., Frydman, H. and Liang, W. (2012) 'Modeling the effect of macroeconomic factors on corporate default and credit rating transitions', *International Review of Economics and Finance*. Elsevier Inc., 21(1), pp. 87–105. doi: 10.1016/j.iref.2011.05.004.

Florez-Lopez, R. and Ramon-Jeronimo, J. M. (2015) 'Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal', *Expert Systems with Applications*, 42(13), pp. 5737–5753. doi: 10.1016/j.eswa.2015.02.042.

Freund, Y. and Schapire, R. (1995) 'A decision-theoretic generalization of on-line learning and an application to boosting', *Computational learning theory*, 55(1), pp. 119–139. doi: 10.1006/jcss.1997.1504.

Friedman, H. (1991) 'Multivariate Adaptive Regression Splines', *The annals of Statistics*, 19(1), pp. 1–141.

Friedman, J. H. (2002) 'Stochastic gradient boosting', *Computational Statistics and Data Analysis*, 38(4), pp. 367–378. doi: 10.1016/S0167-9473(01)00065-2.

Gavalas, D. and Syriopoulos, T. (2016) 'An integrated credit rating and loan quality model : application to bank shipping finance', 8839(August). doi: 10.1080/03088839.2014.904948.

Giannakopoulou, E. N., Thalassinos, E. I. and Theodoros, V. (2016) 'Corporate governance in shipping : an overview', 8839(February). doi: 10.1080/03088839.2015.1009185.

Graham, J. W., Cumsille, P. E. and Elek-Fisk, E. (2003) 'Methods for handling missing data, Research Methods in Psychology, Vol. 2 of Handbook of Psychology'. New York, John Wiley & Sons.

Grammenos, C. (2010) *The handbook of maritime economics and business.* Taylor & Francis.

Grammenos, C. T., Nomikos, N. K. and Papapostolou, N. C. (2008) 'Estimating the probability of default for shipping high yield bond issues', *Transportation Research Part E: Logistics and Transportation Review*, 44(6), pp. 1123–1138. doi: 10.1016/j.tre.2007.10.005.

Groothuis-oudshoorn, K. (2011) 'MICE : Multivariate Imputation by Chained Equations in R', *Journal Of Statistical Software*, VV(Ii), pp. 1–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22289957.

Gupta, J., Gregoriou, A. and Ebrahimi, T. (2018) 'Empirical comparison of hazard models in predicting SMEs failure', *Quantitative Finance*. Routledge, 18(3), pp. 437–466. doi: 10.1080/14697688.2017.1307514.

Hand, D. J. (2009) 'Measuring classifier performance : a coherent alternative to the area under the ROC curve', *Machine Learning*, pp. 103–123. doi: 10.1007/s10994-009-5119-5.

Harrell, F. E. (2015) *Springer Series in Statistics Regression Modeling Strategies*.

Hastie, T. and Tibshirani, R. (1987) 'Generalized additive models: some applications', *Journal of the American Statistical Association*. Taylor & Francis, 82(398), pp. 371–386.

Hensher, D. A. and Jones, S. (2007) 'Forecasting corporate bankruptcy: Optimizing the performance of the mixed logit model', *Abacus*, 43(3), pp. 241–364. doi: 10.1111/j.1467-6281.2007.00228.x.

Hernandez Tinoco, M. and Wilson, N. (2013) 'Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables', *International Review of Financial Analysis*. Elsevier Inc., 30, pp. 394–419. doi: 10.1016/j.irfa.2013.02.013.

Honaker, J. and King, G. (2010) 'What to do about missing values in time-series cross-section data', *American Journal* *of Political Science*, 54(2), pp. 561–581. doi: 10.1111/j.1540-5907.2010.00447.x.

Horobet, A. *et al.* (2019) 'Ownership Concentration and Performance Recovery Patterns in the European Union', *Sustainability (Switzerland)*, 11(4). doi: 10.3390/su11040953.

IFRS Foundation (2016) 'IFRS 16 Leases - Effects Analysis', *International Financial Reporting Standard*, (January), p. 104. Available at: https://cdn.ifrs.org/-/media/project/leases/ifrs/published-documents/ifrs16-effects-analysis.pdf.

International Monetary Fund (2018) *IMF Data, IMF data*. Available at: https://www.imf.org/en/Data.

Jakobsen, J. C. *et al.* (2017) 'When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts', *BMC Medical Research Methodology*. BMC Medical Research Methodology, 17(1), pp. 1–10. doi: 10.1186/s12874-017-0442-1.

Japkowicz, N. and Stephen, S. (2002) 'The class imbalance problem: A systematic study', *Intelligent Data Analysis*, 6(5), pp. 429–449. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-33845536164&partnerID=40&md5=c3e993a2f034141157fdf3a66165f088.

Jones, S. (2017) 'Corporate bankruptcy prediction: a high dimensional analysis', *Review of Accounting Studies*, 22(3), pp. 1366–1422. doi: 10.1007/s11142-017-9407-1.

Jones, S., Johnstone, D. and Wilson, R. (2015a) 'An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes', *Journal of Banking and Finance*, pp. 72–85. doi: 10.1016/j.jbankfin.2015.02.006.

Jones, S., Johnstone, D. and Wilson, R. (2015b) 'An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes', *Journal of Banking and Finance*. Elsevier B.V., 56, pp. 72–85. doi: 10.1016/j.jbankfin.2015.02.006.

Jones, S., Johnstone, D. and Wilson, R. (2016) 'Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks', *Journal of Business Finance and Accounting*, pp. 1–32. doi: 10.1111/jbfa.12218.

Kalak, I. El and Hudson, R. (2018) 'The effect of size on the failure probabilities of SMEs: An empirical study on the US market using discrete hazard model', *Kent Academic Repository*, pp. 294–306.

Kang, P. and Cho, S. (2006) 'EUS SVMs: Ensemble of Under-Sampled SVMs for Data Imbalance Problems BT - Neural Information Processing', in King, I. et al. (eds). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 837–846.

Karim, S. A. A. *et al.* (2018) 'Rational cubic spline interpolation for missing solar data imputation', *Journal of Engineering and Applied Sciences*. Medwell Journals, 13(9), pp. 2587–2592.

Kavussanos, M. G. and Tsouknidis, D. A. (2016) 'Default risk drivers in shipping bank loans', *Transportation Research Part E*. Elsevier Ltd, 94, pp. 71–94. doi: 10.1016/j.tre.2016.07.008.

Kavussanos, M. G. and Visvikis, I. D. (2016) *The International Handbook of Shipping Finance*, *The International Handbook of Shipping Finance*. doi: 10.1057/978-1-137-46546-7.

Kim, M. J., Kang, D. K. and Kim, H. B. (2015) 'Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction', *Expert Systems with Applications*. Elsevier Ltd, 42(3), pp. 1074–1082. doi: 10.1016/j.eswa.2014.08.025.

Kim, S. M. and Kim, J. (2018) 'the OECD BEPS Package', 49(2), pp. 221–238.

Kim, S., Mun, B. M. and Bae, S. J. (2018) 'Data depth based support vector machines for predicting corporate bankruptcy', *Applied Intelligence*, 48(3), pp. 791–804. doi: 10.1007/s10489-017-1011-3.

King, G. *et al.* (2001) 'Analyzing incomplete political science data: An alternative algorithm for multiple imputation', *American political science review*. Cambridge University Press, 95(1), pp. 49–69.

Kofman, P. (2003) 'Using Multiple Imputation in the Analysis of Incomplete Observations in Finance', *Journal of Financial Econometrics*, 1(2), pp. 216–249. doi: 10.1093/jjfinec/nbg013.

Koopman, S. J., Lucas, A. and Schwaab, B. (2011) 'Modeling frailty-correlated defaults using many macroeconomic covariates', *Journal of Econometrics*, 162(2), pp. 312–325. doi: 10.1016/j.jeconom.2011.02.003.

Krauss, C., Do, X. A. and Huck, N. (2017) 'Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500', *European Journal of Operational Research*, 259(2), pp. 689–702. doi: 10.1016/j.ejor.2016.10.031.

Lessmann, S. *et al.* (2015) 'Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research*, 247(1), pp. 124–136. doi: 10.1016/j.ejor.2015.05.030.

Little, R. J. A. (1988) 'Missing-data adjustments in large surveys', *Journal of Business & Economic Statistics*. Taylor & Francis Group, 6(3), pp. 287–296.

Little, R. J. A. and Rubin, D. B. (2019) *Statistical analysis with missing data*. 3rd edn. John Wiley & Sons.

Little and Rubin (1987) 'Statistical analysis with missing data', *Hoboken, NJ: Wiley*.

Lohmann, C. and Ohliger, T. (2017) 'Nonlinear Relationships and Their Effect on the Bankruptcy Prediction', *Schmalenbach Business Review*, 18(3), pp. 261–287. doi: 10.1007/s41464-017-0034-y.

Lozinskaia, A. *et al.* (2017) 'Determinants of the probability of default: the case of the internationally listed shipping corporations', *Maritime Policy and Management*. Routledge, 44(7), pp. 837–858. doi: 10.1080/03088839.2017.1345018.

Lyridis, D. V., Manos, N. D. and Zacharioudakis, P. G. (2014) 'Modeling the dry bulk shipping market using macroeconomic factors in addition to shipping market parameters via artificial neural networks', *International Journal of Transport Economics*, 41(2), pp. 231–253.

Madley-Dowd, P. *et al.* (2019) 'The proportion of missing data should not be used to guide decisions on multiple imputation', *Journal of Clinical Epidemiology*. Elsevier Inc, 110, pp. 63–73. doi: 10.1016/j.jclinepi.2019.02.016.

Merton, R. C. (1974) 'ON THE PRICING OF CORPORATE DEBT: THE RISK STRUCTURE OF INTEREST RATES*', *The Journal of Finance*. John Wiley & Sons, Ltd (10.1111), 29(2), pp. 449–470. doi: 10.1111/j.1540-6261.1974.tb03058.x.

Min, J. H. and Lee, Y.-C. (2005) 'Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters', *Expert Systems with Applications*, 28(4), pp. 603–614. doi: 10.1016/j.eswa.2004.12.008.

Mitroussi, K. *et al.* (2016) 'Performance drivers of shipping loans : An empirical investigation', *Intern. Journal of Production Economics*. Elsevier, 171, pp. 438–452. doi: 10.1016/j.ijpe.2015.09.041.

Murray, J. S. (2018) 'Multiple imputation: A review of practical and theoretical findings', *Statistical Science*, 33(2), pp. 142–159. doi: 10.1214/18-STS644.

Nguyen, C. D., Carlin, J. B. and Lee, K. J. (2017) 'Model checking in multiple imputation: An overview and case study', *Emerging Themes in Epidemiology*. BioMed Central, 14(1), pp. 1–12. doi: 10.1186/s12982-017-0062-6.

Nyitrai, T. and Virág, M. (2019) 'The effects of handling outliers on the performance of bankruptcy prediction models', *Socio-Economic Planning Sciences*. Elsevier, 67(August 2018), pp. 34–42. doi: 10.1016/j.seps.2018.08.004.

OECD (2013) *Addressing base erosion and profit shifting*, *Addressing Base Erosion and Profit Shifting*. doi: 10.1787/9789264192744-en.

OECD (2019) *OECD Data*, *OECD Data*. Available at: https://data.oecd.org/.

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

Petropoulos, T. (2019) 'Key Developments and Growth in Global Ship Finance', *Petrofin Bank Research*, (October 2019), pp. 1–10.

Pindado, J., Rodrigues, L. and de la Torre, C. (2008) 'Estimating financial distress likelihood', *Journal of Business Research*, 61(9), pp. 995–1003. doi: 10.1016/j.jbusres.2007.10.006.

Pompe, P. P. M. and Bilderbeek, J. (2005) 'The prediction of bankruptcy of small- and medium-sized industrial firms', *Journal of Business Venturing*, 20(6), pp. 847–868. doi: 10.1016/j.jbusvent.2004.07.003.

R Core Team (2000) 'R language definition', *Vienna, Austria: R foundation for statistical computing*.

Rehmatulla, N. and Smith, T. (2015) 'Barriers to energy efficiency in shipping: A triangulated approach to investigate the principal agent problem', *Energy Policy*, 84. doi: 10.1016/j.enpol.2015.04.019.

Reverte, C. (2012) 'The impact of better corporate social responsibility disclosure on the cost of equity capital', *Corporate Social Responsibility and Environmental Management*, 19(5), pp. 253–272. doi: 10.1002/csr.273.

Rubin, D. B. (1987) 'Multiple Imputation for Nonresponse in Surveys. en.'

Rubin, D. B. and Schenker, N. (1986) 'Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse', *Journal of the American Statistical Association*. Taylor & Francis, 81(394), pp. 366–374. doi: 10.1080/01621459.1986.10478280.

Schapire, R. E. (1990) 'The Strength of Weak Learnability', *Machine Learning*, 5(2), pp. 197–227. doi: 10.1023/A:1022648800760.

Serneels, S., De Nolf, E. and Van Espen, P. J. (2006) 'Spatial sign preprocessing: A simple way to impart moderate robustness to multivariate estimators', *Journal of Chemical Information and Modeling*, 46(3), pp. 1402–1409. doi:

10.1021/ci050498u.

Servaes, H. (1996) ‘The value of diversification during the conglomerate merger wave’, *The Journal of Finance*. Wiley Online Library, 51(4), pp. 1201–1225.

Shah, A. D. *et al.* (2014) ‘Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE : A CALIBER Study’, 179(7), pp. 764–774. doi: 10.1093/aje/kwt312.

Sharife, K. (2010) ‘Flying a questionable flag: Liberia’s lucrative shipping industry’, *World Policy Journal*, 27(4), pp. 111–118. doi: 10.1162/wopj.2011.27.4.111.

Shin, K.-S., Lee, T. S. and Kim, H.-J. (2005) ‘An application of support vector machines in bankruptcy prediction model’, *Expert Systems with Applications*, 28(1), pp. 127–135. doi: 10.1016/j.eswa.2004.08.009.

Shumway (2001) ‘Forecasting Bankruptcy More Accurately: A Simple Hazard Model’, *The Journal of Business*, 74(1), pp. 101–124. doi: 10.1086/209665.

Shumway, T. (2001) ‘Forecasting bankruptcy more accuratemy - A simple Hazard model’. Journal of business, pp. 101–124.

Shuyong, L. *et al.* (2009) ‘Cost control strategy of shipping enterprises under asymmetric information’, in *2009 International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2009*, pp. 785–788. doi: 10.1109/ICMTMA.2009.302.

Spackman, K. A. (1989) ‘Signal detection theory: Valuable tools for evaluating inductive learning’, in *Proceedings of the sixth international workshop on Machine learning*. Elsevier, pp. 160–163.

Standard & Poor’s (2019) ‘Corporate methodology: Ratios and adjustments’, *Standard & Poor’s Rating Service: RatingsDirect*, (1), pp. 1–39. Available at: WWW.STANDARDANDPOORS.COM/RATINGSDIRECT.

Standard & Poors (2019) ‘Key Credit Factors For The Transportation Infrastructure Industry’, (44), pp. 1–15.

Sun, J. and Li, H. (2012) ‘Financial distress prediction using support vector machines: Ensemble vs. individual’, *Applied Soft Computing Journal*. Elsevier B.V., 12(8), pp. 2254–2265. doi: 10.1016/j.asoc.2012.03.028.

Tahtah, J. and Roelofsen, E. (2016) *A study on the impact of lease capitalisation IFRS 16 : The new leases standard*, *PwC* *report*. Available at: www.pwc.com.

Templ, M., Kowarik, A. and Filzmoser, P. (2011) 'Iterative stepwise regression imputation using standard and robust methods', *Computational Statistics and Data Analysis*, 55(10), pp. 2793–2806. doi: 10.1016/j.csda.2011.04.012.

Theodossiou, P. T. (1993) 'Predicting shifts in the mean of a multivariate time series process: An application in predicting business failures', *Journal of the American Statistical Association*, 88(422), pp. 441–449. doi: 10.1080/01621459.1993.10476294.

Tian, Y., Shi, Y. and Liu, X. (2012) 'Recent advances on support vector machines research', *Technological and Economic Development of Economy*, 18(1), pp. 5–33. doi: 10.3846/20294913.2012.661205.

Tishbirani, R. (1996) 'Regression shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288. Available at: https://statweb.stanford.edu/~tibs/lasso/lasso.pdf.

Tkáč, M. and Verner, R. (2015) 'Artificial neural networks in business: Two decades of research', *Applied Soft Computing*, 38, pp. 788–804. doi: 10.1016/j.asoc.2015.09.040.

Tsai, C. F. and Cheng, K. C. (2012) 'Simple instance selection for bankruptcy prediction', *Knowledge-Based Systems*. Elsevier B.V., 27, pp. 333–342. doi: 10.1016/j.knosys.2011.09.017.

UNCTAD (2019) *Review of Maritime Transport*.

Valencia, C. *et al.* (2019) 'Generalized additive model with embedded variable selection for bankruptcy prediction: Prediction versus interpretation', *Cogent Economics and Finance*. Cogent, 7(1), pp. 1–14. doi: 10.1080/23322039.2019.1597956.

Vapnik, V. N. (1999) 'An overview of statistical learning theory', *IEEE Transactions on Neural Networks*, 10(5), pp. 988–999. doi: 10.1109/72.788640.

Wang, B. X. and Japkowicz, N. (2008) 'Boosting Support Vector Machines for Imbalanced Data Sets BT - Foundations of Intelligent Systems', in An, A. et al. (eds). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 38–47.

Wang, G. W. Y., Woo, S. and Mileski, J. (2016) 'The relative efficiency and financial risk assessment of shipping companies', 8839(January). doi: 10.1080/03088839.2013.873546.

Weber, O., Scholz, R. and Michalik, G. (2010) 'Incorporating sustainability criteria into credit risk management', *Business strategy and the …*, 50(November 2008), pp. 39–50. Available at: http://onlinelibrary.wiley.com.sci-hub.io/doi/10.1002/bse.636/abstract.

Wickham, H. (2015) *R packages: organize, test, document, and share your code*. ' O'Reilly Media, Inc.'

World Bank (2019) *World Bank Open Data*, *World Bank Open Data*. Available at: https://data.worldbank.org/.

Wruck, K. H. (1990) 'Financial distress, reorganization, and organizational efficiency', *Journal of Financial Economics*, 27(2), pp. 419–444. doi: 10.1016/0304-405X(90)90063-6.

Xia, Y. *et al.* (2017) 'A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring', *Expert Systems with Applications*, 78, pp. 225–241. doi: 10.1016/j.eswa.2017.02.017.

Yeo, I. N. and Johnson, R. . (2000) 'A new family of power transformations to improve normality or symmetry', *Biometrika*, 87(4), pp. 954–959.

Yu, Fan *et al.* (2005) 'Accounting transparency and the term structure of credit spreads ARTICLE IN PRESS', *Journal of Financial Economics*, 75, pp. 53–84. doi: 10.1016/j.jfineco.2004.07.002.

Zhang, C. *et al.* (2017) 'An up-to-date comparison of state-of-the-art classification algorithms', *Expert Systems with Applications*. Elsevier Ltd, 82, pp. 128–150. doi: 10.1016/j.eswa.2017.04.003.

Zhang, L., Hu, H. and Zhang, D. (2015) 'A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance', *Financial Innovation*, 1(1). doi: 10.1186/s40854-015-0014-5.

Zhao, H. *et al.* (2016) 'Portfolio selections in P2P lending: A multi-objective perspective', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2075–2084. doi: 10.1145/2939672.2939861.

Zięba, M., Tomczak, S. K. and Tomczak, J. M. (2016) 'Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction', *Expert Systems with Applications*, 58, pp. 93–101. doi: 10.1016/j.eswa.2016.04.001.

Appendix A: Ethics

There are five generally accepted broad principles of ethics when performing research. Within the general context of "do good" beneficence and "do no harm" (non-malfeasance) a researcher should endeavour to:

i. obtain informed consent from potential research participants;
ii. minimise the risk of harm to participants;
iii. protect their anonymity and confidentiality;
iv. refrain from employing deceptive practices; and
v. ensure that participants have the right to withdraw from one's research (and are aware of their right to do so).

Ethical issues are contained in various stages in research (Bryman 2008) and are closely interrelated with the type of research conducted. Considerations essential to ethics behaviour (Fielding, Lee and Blank, 2008) can be broken down into four main areas:

- Lack of informed consent
- Harm to participants
- Invasion of privacy
- Deception

The areas are addressed by the principles in the UK Data Protection Act (1998) and each of the above have been a provided due consideration throughout this research.

Dealing with sensitive financial and market data is at best a minefield and therefore great care and preparation of security and confidentiality is necessary. This study, however, has relied predominantly on the collection of secondary data from publicly accessible database sources. As such, no primary data collection from individuals has been conducted. Care has also been taken not to divulge the names of individual companies or the involvement of individual financial institutions (other than as sources of publicly accessible data).