Financial risk assessment in shipping: A holistic machine learning based methodology

Mark Clintworth[1], Dimitrios Lyridis[2] and Evangelos Boulougouris[3]

[1]Department of Naval Architecture and Marine Engineering, University of Strathclyde, Glasgow, Scotland, UK. Email: mbc@maritime-research.org. [2]National Technical University of Athens, Greece. [3]Maritime Safety Research Centre, University of Strathclyde, Glasgow, Scotland, UK.

Abstract Corporate financial distress (FD) prediction models are of great importance to all stakeholders, including regulators and banks, who rely on acceptable estimates of default risk, for both individual borrowers and bank loan portfolios. Whilst this subject has been covered extensively in finance research, its application to international shipping companies has been limited while the focus has mainly been on the application of traditional linear modelling, using sparse, cross-sectional financial statement data. Insufficient attention has been paid to the noisy and incomplete nature of shipping company financial statement information. This study contributes to the literature through the design, development and testing of a novel holistic machine learning methodology which integrates predictor evaluation and missing data analysis into the distress prediction process. The model was validated using a longitudinal dataset of over 5,000 company year-end financial statements combined with macroeconomic and market predictors. We applied this methodology first for individual company level distress prediction before testing the models' ability to provide accurate confidence intervals by backtesting conditional value-at-risk estimations of the distress rates for bank portfolios. We conclude that, by adopting a holistic approach, our methodology can enhance financial monitoring of company loans and bank loan portfolios thereby providing a practical "early warning system" for financial distress.

Keywords: financial distress, machine learning, multi-variate imputation, random forest, extreme gradient boosting, generalised additive modelling, conditional value-at-risk.

Acknowledgements

We wish to thank the Editor-in-Chief and the two anonymous reviewers for their excellent contributions to, and time spent on, this paper. We believe that the comments made have contributed greatly to the revised version's clarity and content.

DRAFT

Introduction

The banking system failure, integral to the financial crisis of 2009, had serious negative impacts on the shipping sector and indeed led to the withdrawal of the commercial banking sector from all major risk intensive private capital investments. One important regulatory consequence was the imposition of increasingly stringent capital adequacy rules by the Basel framework (BIS 2019), which forced banks to offload the more risky assets from their balance sheets whilst also placing further restrictions on new investments. Furthermore, the crisis highlighted the financial system's weakness in financial risk management and distress prediction. This led directly to the adoption of the Basel Internal Rating System for risk management which demanded the development of effective probability of default models geared to the specific characteristics of companies, taking at the same time due account of the macroeconomic environment.

Consequently, it is becoming increasingly important to model the probability of distress of shipping companies more accurately than before. Our research approaches this challenge through the development of a new methodology consisting of three core elements: i) the application of modern machine learning (ML) classification algorithms; ii) supplementing financial statement data with macroeconomic and market predictors and iii) the application of modern multiple imputation techniques for the analysis and treatment of missing accounting information. Each of these elements has been previously addressed in various fields of endeavour including, to a limited degree, the shipping sector. However, to the authors knowledge, none has adopted a holistic methodology combining all three. The rationale behind this approach is to examine if, by simultaneously accounting for the non-linear relationship between default risk and the independent variables; exogenous macro effects; and the effects of incomplete data, our model can improve on the predictive power of traditional corporate distress modelling.

We first evaluate the performance of ML classification models in the prediction of financial distress (FD) which may eliminate the need for unobservable temporal effects. We appraise the relatively recently established models random forests (RF) and extreme gradient boosting (XGB), alongside an extended linear classifier, i.e., the generalised additive model (GAM). The objective is not simply to compare model performance (e.g. accuracy) but assess their individual confidence intervals (CI) thereby measuring a model's true capacity to generalise on out of sample data. A rapidly increasing focus in the literature is the application of ML modelling (complex models exhibiting non-linear dependency structures between the co-variates and the resulting outcome) in corporate failure prediction (Christoffersen, Matin, and Mølgaard 2018; Jones, Johnstone, and Wilson 2015a; Hernandez Tinoco and Wilson 2013). Much of the earlier work has focused on benchmarking 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 confidence intervals (CI)[1] of aggregated FD predictions owing to their underlying assumption of conditional independence.

The fact that none of the models fully capture correlation in FD, solely through the application of accounting data, suggests the existence of unobserved macro effects creating 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 events, which accurately represent the risk profile of shipping companies, is essential. We develop distress prediction models employing not only firm level data but also macroeconomic and market indicators to detect early stages of distress.

A major aspect of our methodology is the tackling of the problem of missing accounting data. The global nature of the shipping industry, and diverse national accounting practices and laws, render the identification and collection of complete and consistent financial statements one of the major challenges. Therefore, the problem of missing accounting values and how to treat them is a major focus of this research. Our methodology is trained and tested by applying a test case comprising raw data compiled from detailed financial statements covering the period 2000-2018 of worldwide dry bulk carrier owners/operator companies, both listed and non-listed. Financial institutions and ultimate owner (parent) companies are not included due to the bias potential introduced through group level accounting practices.

Finally, the early detection of FD provides investors with ways of avoiding some of the costs associated with a bankruptcy filing and recovery. However, models must be transparent[2] and open to scrutiny by all stakeholders, investors and particularly regulatory bodies if they are to be accepted as practical tools.

The rest of the paper is structured as follows. In section 2 we review the relevant literature; section 3 details data issues; section 4 describes the methodologies and models used and the results are presented and discussed in section 5. Section 6 presents the conclusions and recommendation for further research.

Literature Review

This literature review comprises five threads. The definition of FD is first addressed as its clear and accepted definition is core to this research. The second thread reviews research efforts into shipping company FD prediction followed by a review of the literature surrounding the application of machine learning tools in corporate FD prediction. The section also covers a review of the literature relating to incomplete financial accounting data and the selection of core independent variables.

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

[2] A common problem with many ML models is their lack of interpretability and so they are often described as being black box algorithms.

Definition of financial distress

Much of the literature defines FD as being centred upon the final legal consequence of an organisation's liquidation or bankruptcy. This legal event is represented by a dependent variable in a binary classification model variable (Balcaen and Ooghe 2006). This definition, however, only represents the worse-case scenario of FD and therefore presents 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 FD and the date of a legal declaration of insolvency (Theodossiou 1993; Hernandez Tinoco and Wilson 2013). For example, U.S. Chapter 11 legislation has brought about changes in the way organisations can be given time to reorganise their business, assets and debts in the event of impending insolvency. There are a number of stages a company can encounter before closure, for example (Wruck 1990) cites FD, insolvency, filing of bankruptcy and administrative receivership.

Prior to, the triggering of the terminal state, the literature generally follows two approaches. The first is an accounting features approach, utilising cross sectional annual data, and is widely covered in the default prediction literature (Altman 1968; Ohlson 1980). This utilises historical financial statements which are benchmarked against historical default rates and generally modelled to produce a probability of bankruptcy.

The second approach is a mixed accounting/market approach which estimates a company's probability of default founded 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). Accepting this as a foundation[3], we also include DD as a feature in our modelling.

Shipping company financial distress prediction

Research into shipping FD prediction has been relatively limited to date. Earlier works have focused largely on financial performance predictor/feature selection, relying on, the more conventional, binary logistic regression techniques (Antoniou, A., A. Thanopoulos 1998; C. Th Grammenos, Nomikos, and Papapostolou 2008; Kavussanos and Tsouknidis 2016; Mitroussi et al. 2016; Lozinskaia et al. 2017). Moreover, research interest was either on shipping bond markets or bank shipping debt.

The financing of the shipping industry has, traditionally, relied heavily on bank loans. A critical priority for bank credit risk departments concerns the provision of an optimal framework for assessing the credit rating of borrowers' as well as of loan quality. This includes defining specific quantitative and qualitative criteria mirroring the borrowers' ability to comply with the loan contract terms. Traditionally, this has been founded on the five core Cs of credit: the borrower's 'character', 'capacity' and 'capital', 'collateral' and 'conditions' (Antoniou, A., A.

[3] Moody's Analytics for example.

Thanopoulos 1998; C. T. Grammenos 2010) applied 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). These authors contended that models which combine criteria and provide relative weighting to assist the decision-making process of a bank's credit committee, are limited. Their paper presents work on the application of the monotone regression method, UTADIS, and was aimed at the analysis of both credit allocations and the evaluation of the criteria used for the selection of loan applications in the shipping industry. Gavalas and Syriopoulos (2016) proposed an integrated credit rating model founded on a series of critical qualitative and quantitative criteria for bank loan portfolios. The model was applied to, and tested on, bank financing decisions in the shipping sector as a case study. Again, the authors used a UTADIS approach to assess the relative impact of the selected risk factors on efficient credit rating scoring and loan quality assessment.

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

Machine learning application to financial distress prediction

Since the works of Altman (1968) and Ohlson (1980), research relating to the modelling of corporate FD and bankruptcy has been extensive (Altman 1977; Shumway 2001; Duffie and Singleton 2003; Hensher and Jones 2007). However, until recently, much of this work relied heavily on more traditional classifiers e.g. logit, probit or linear discriminant analysis models, which are commonly referred to as generalised linear models (GLM). The financial crisis of 2009, however, demonstrated that more research effort was required to develop models with enhanced predictive accuracy, not only for predicting ultimate failure events, but models which also detect the early stages of FD. Post crisis, research has highlighted failures in conventional corporate FD prediction models (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, multicollinearity, autocorrelation, 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 (Lessmann et al. 2015; Zhang et al. 2017). ML methods applied to FD prediction are now well established in the literature (Jones, Johnstone, and Wilson 2015b; Zięba, Tomczak, and Tomczak 2016; Xia et al. 2017; Barboza, Kimura, and Altman 2017) and the general conclusion is that 'new age' classifiers outperform transitional GLMs in out of sample generalisation.

The application of ML modelling (complex models exhibiting non-linear dependency structures between the covariates and the resulting outcomes) in corporate failure prediction is increasingly prevalent in the literature (Hernandez Tinoco and Wilson 2013; Jones, Johnstone, and Wilson 2015b; de Moraes Barboza, Kimura, and Altman 2015; Christoffersen, Matin, and Mølgaard 2018). Much of the previous work has attempted to benchmark the performance of ML models on GLMs. However, despite research demonstrating the enhanced generalisation performance of ML classifiers compared with GLMs, care must be taken as transparency is paramount in finance

(as is demanded by investors and regulators) and, as the literature notes, ML models involve issues of transparency[4] .

Missing accounting values

The problem of missing data is predominant in financial modelling (Kofman 2003; Burger, Silverman, and Vuuren 2018) and is a common feature of shipping company accounts (Sharife 2010; Merk 2020). This is also true of the raw panel dataset compiled for the study case used in this research. This issue has, to date, not been addressed in the shipping finance literature. There are various reasons for incomplete financial accounts and here we cite two examples which are a common feature in shipping company accounts. Firstly, open registries or "flags of convenience" (FOC) concede favourable tax environments to shipping companies (Merk 2020), and hence have become a part of shipping company tax planning. Shipping companies often exploit variations in domestic tax law and international taxation standards (S. M. Kim and Kim 2018; Merk 2020). This provides them with opportunities to eliminate or significantly reduce taxation and therefore, many multinational corporations use base erosion and profit shifting (BEPS) to reduce the corporate tax base (OECD 2013).

A second reason might be the application of 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, coupled with multiple changes to the IFRS by the International Accounting Standards Board (IASB), have 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). 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 an entity (lessee) which leases vessels should recognise and report assets and liabilities arising from those leases. 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.

There are three accepted approaches to the problem of missing data in statistical analysis. The first method is referred to as the "complete case" (Nguyen, Carlin, and Lee 2017) or list-wise deletion approach which discards incomplete individual observations (company accounting years) and results in a residual dataset containing complete, observed data. The second method is referred to as the "omitted variable" approach which involves simply removing those covariates with missing values from the dataset (Honaker and King 2010). The third method is data imputation and is part of a growing field of research to address the challenge of missing values in data. In this research we focus on the multivariate imputation (MI) methodology (Rubin 1987). MI has become one of the most widely used methods for handling missing data and is receiving increasing attention in financial research (DiCesare 2006; Amel-zadeh et al. 2020).

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

Finally, the primary objective of this research is the accuracy of predictions rather than making valid subject related or sector informed inferences[5].

Predictor selection

Much of the earlier 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. According to Balcaen and Ooghe (2006), if too much emphasis is placed on financial ratios for failure prediction then it is implicitly assumed that all FD indicators are contained within financial statements. There are many examples in the literature which examine combined approaches using accounting, macroeconomic/market, and qualitative data, in order to provide an enhanced model of FD prediction (Das et al. 2007; Duffie et al. 2009; Koopman, Lucas, and Schwaab 2011). Furthermore, Bonfim (2009) postulates that when macroeconomic indicators are considered, this leads to an improvement in model results. The consensus in the literature is that macroeconomic dynamics represent an independent contribution in FD prediction. As regards shipping, this is an issue recognised by (Lyridis, Manos, and Zacharioudakis 2014) for example. Furthermore, recent literature has highlighted the failure of such traditional approaches to encapsulate spatial (annual) fluctuations in FD. Numerous 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 not adequately account for latency (unobserved variables) and so the authors advocate approaches which include frailty[6] or the inclusion of time-varying effects.

Methodological background

This section briefly outlines the analytical framework encompassing the main principles applied in our methodology, namely missing value treatment, data pre-processing, feature selection, classification algorithms and model evaluation metrics (a more detailed discussion can be found in Appendix A).

With respect to the treatment of missing values, we provide both a complete case and a multivariate imputation analysis of the raw data. For the complete case data set, we simply select those records from the raw data which contain non-null values for all independent variables. For data imputation, we begin by assuming that our missing accounting information is missing at random (MAR). This approach assumes that the reasons for missing data in

[5] The goal is not the regeneration of missing values but to maintain the characteristics of the data distribution and the relationships between features, thereby maintaining the model's overall ability to generalise on out of sample data.

[6] 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.

any sample can be explained by the observed data, i.e., the probability of missing values is dependent upon observed data as opposed to the values of missing data.

Once the issue of missing data has been addressed, we then examine and pre-process our resulting dataset for skewness, kurtosis and outliers (Barnes 1987). Pre-processing is performed using a variation of the Box-Cox (Box and Cox 1964) transformation (Yeo and Johnson 2000), and outlier treatment is applied through spatial sign (Serneels, De Nolf, and Van Espen 2006). As company default is a relatively rare event (M. J. Kim, Kang, and Kim 2015), the dataset is imbalanced with the "distressed" class being in the (significant) minority class. In order to account for this, we tested several sampling methods, with down sampling (reducing the instances of the majority class) producing the most effective results (out of sample generalisation) on our test dataset. Once transformed and sampled, the task of feature analysis and selection is undertaken. At this stage, the data are examined for multicollinearity and an assessment of the level of contribution to the dependent variable of independent variables. Random forest modelling (Breiman 2001) was selected for this task (Zhou, Zhou, and Li 2016a; Lakshmipadmaja and Vishnuvardhan 2018).

The final dataset is partitioned into training and test sets on a ratio of 70:30 before applying classification modelling on the training data. As the literature suggests, there are many examples of research into the effectiveness of many machine learning algorithms in the financial and economics domain. The general consensus (Jones, Johnstone, and Wilson 2015a; Son et al. 2019) is that tree-based algorithms consistently outperform models such as artificial neural network (ANNs) or support vector machines (SVMs). Our research corroborates their conclusions, however, in this paper we only report our results from the best performing tree-based classifier, extreme gradient boosting (XGB). As a benchmark, we also report the results generated from the implementation of one of the best performing linear based classifiers, GAM. The inclusion of an extended generalised linear model is to provide a balanced comparison with the complex model. Their inclusion is performed in the name of model transparency: Following the Ockham's Razor principle, if two models demonstrate similar predictive power then the more transparent model which is preferable.

Finally, the results are compared using a variety of metrics necessary for the accurate assessment of classification performance. Specifically, receiver operating characteristics (ROC), H Measure[7] (Hand 2009) and log loss (Bickel 2007) metrics are used with a focus on the ability of the models to accurately predict the minority "distressed" class (Appendix A).

[7] The H measure is a robustness check on the ROC results. This metric addresses the main problem associated with the ROC, that of the handling of misclassification costs across different classifiers.

Data – Bulk shipping case study

The bulk carrier fleet is an essential part of the global economy. An Equasis report (2019) notes that, as of 2018, the world fleet totalled 116,857 ships (1,361,920 GT), with dry bulk carrier vessels totalling 11,929 vessels (457,648 GT), accounting for 33.6% of the global fleet in GT terms. Furthermore, UNCTAD (2019) report that in 2018 the bulk fleet took delivery of 26,7% of the total of newly built GT, more than any other vessel type, followed by oil tankers (25%), containerships (23,5%) and gas carriers (13%). The dry bulk market is very diversified and volatile with bulk shipping comprising three major sectors: iron ore, coal and grain as well as other minor commodities, e.g. steel, forest products and minerals. According to UNCTAD (2019), the major dry bulk commodities represented more than 40% of total dry cargo tons shipped in 2018, with containerized cargo contributing 24%, minor bulks with 25,8% (the remaining volumes consisted of dry cargo including break-bulk).

Dry bulk shipping market is characterised by the large number of small-scale shipowners, few market barriers and transparent transactions (Wu, Yin, and Sheng 2018). Furthermore, dry bulk freight rates are determined predominantly by market dynamics with no individual shipowner or charterer having a significant effect on rates. In short, the dry bulk market can be viewed as a perfectly competitive market (Yin, Wu, and Lu 2019) and, therefore, a viable test case for this study.

Empirical context

The study develops four main ex-ante models for estimating FD likelihood, to test the predictive power of three sets of independent variables (Table 1): financial statement ratios; macroeconomic indicators; and bulk shipping market predictors. In Model 1, the independent variable selection is made solely from financial statement ratios. Model 2 adds macroeconomic indicators to company level financial data. Model 3 comprises financials and market related covariates. Finally, Model 4 comprises all three sets of covariates.. Missing company level financial data are subjected to both case wise deletion and data imputation in order to examine corresponding model performance.

Data sampling

The raw dataset used for company level financials is sourced from unconsolidated statements of the Orbis company database (Bureau-Van-Dijk 2019) and consists of over 5,000 global dry bulk shipping company yearly statements for the period 2000-2018. The shipping specifics are primarily drawn from Clarkson's Shipping Intelligence Network (Clarksons 2019), whilst macroeconomic data is drawn from two data sources, the OECD (2019) and the World Bank (2019). At company level we apply filters to our 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 we see the need to account for interactions between size and other variables in the models, thereby allowing for the modelling of companies of different sizes.

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. Our definition of FD in companies follows Pindado, Rodrigues, and de la Torre (2008) and we outline the following primary conditions to be fulfilled in predicting company financial distress. The hypothesis follows that a company is distressed when any of the following events occurs 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 (Hernandez Tinoco and Wilson 2013); 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; we therefore implicitly model recurrent events.

Independent variable 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 as well as market and macroeconomic indicators (Table 1). The sector's risk framework is largely described by financial features relating to the capital intensity 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; and the balance of supply and demand.

Table 1: Predictive features for the FD model

Results discussion

In this section we present and discuss the results from the application of our methodology to the bulk shipping case study. We first present the results from the missing value analysis and treatments, followed by the results produced through the application of our two classifiers, GAM and XGB, to the four data models.

Missing Values

Missingness analysis

The first objective was to analyse the financial statement data in order to ascertain the extent of the missingness. Table 2 shows that, of the 5,368 company financial statements collected, only 1,483 were complete, with approximately 72% of them being 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 Table 3Table 3.

Table 2: Missing financial statement value analysis

DRAFT

The results demonstrated a relatively high level of observed accounting values, 82.4%. This indicated that, if the MAR assumption is applied, there is sufficient information present in the observed values for multivariate imputation to yield beneficial results (in terms of reduced bias and efficiency), when compared with complete case treatment. 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 in the 3,885 incomplete financial statements; these contain significant levels of potentially exploitable information. A matrix plot of the missing data distribution is represented in

Figure 1, with grey denoting missing data.

A plot of the pairwise correlation point-biserial correlation coefficients, between covariate pairs, is shown in

Figure 2. Variables are assigned TRUE or FALSE depending on their missing data status and these Boolean vectors are correlated to the native variables.

Table 3: Missing value level per accounting ratio

Figure 1: Matrix plot of missing accounting data

Figure 2: Raw accounting data - observed v missing (NA) correlation coefficients

Post imputation evaluation

The RF algorithm was evaluated on its ability to obtain statistically valid inferences from the incomplete financial data set and the results indicate a limited loss of information from the imputed data (see Appendix B).

A distribution histogram overlay of imputed and observed values is found in Figure 3, depicting the distributions of original and imputed accounting ratio values. Although the goal is to have the two similar distributions, differences do not necessarily signal problematic imputation. The empirical density plots act as flags for potential problems with the imputed estimates. At this stage, no data pre-processing was performed on the pre-imputation dataset, as any bias would have been “locked into” the data prior to training and validation.

DRAFT

Figure 4 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.

Figure 3: Overlaid histograms of imputed and original values

Figure 4: Correlation coefficient scatter plot

The final stage of missing value analysis was to examine the out of sample classification results when modelling using both the imputed and complete case datasets. The results of imputation (see Appendix B) indicate that, as discussed previously, the removal of circa 72% of records (containing incomplete data) involved the introduction of bias. Indeed, further examination employing RF feature set analysis (Figure 5), highlights the differences in contribution to the dependent variable, provided by the independent variables (depending upon the constitution of the individual data sets).

Table 4: Missing value treatments – RF classification

For example, in this case we observe a greater contribution by the current, gearing and solvency ratios in the RF MI dataset than in the complete case (CC) data, where profitability plays a greater role in establishing the distressed state. The RF importance analysis shows that the CC data set results in an over-weighted importance given to financial ratios which are not generally accepted as the most suitable in forecasting corporate financial health e.g. see (Son et al. 2019). Furthermore, it is also noted that the removal of so much data, in order to distil the CC dataset, involves a significant diminution of over 50%, of the minority class (distressed), which further increases the risk of bias.

Figure 5: Missing value treatment – Data set feature importance comparison

Prediction model evaluations

We evaluated the prediction power of the four data models utilising the GAM and XGB classification algorithms. We first examined variable correlations and cross reference the data models[8] with an analysis of feature

[8] Following the results of the missing value analysis, the RF MI data set was used as the accounting base for this analysis. However, the methodology was designed with the assumption that this decision will be for a corporate "risk department".

importance, based on the permutation results from an RF feature evaluation of the dataset. The accounting feature distributions shown in Figure 6 confirm the non-normal, skewed and kurtosis nature of the dataset, consistent with corporate company panel data. This suggested that the use of a non-parametric correlation test (Spearman 1904) was most suitable to assess correlations between the features. The correlation matrices produced from the Spearman tests on each of the four datasets can be seen in

Figure 7.

The results show correlations in both accounting (e.g. solvency and gearing ratios) and bulk market (e.g. freight and time charter rates) feature sets, with some significant enough to warrant closer examinations of the data. This was performed using the RF feature importance methodology: given the bias in mean decreased impurity measurement, when predictor variables are highly correlated (Strobl et al. 2008), we used both unconditional and conditional permutational analysis. The results shown in

Figure 8 were then used to identify the best performing variable permutations for our models. This information was used to inform feature selection testing for each of the three feature sets.

Figure 6: Accounting feature distributions

Figure 7: Feature correlation matrices

Figure 8: Unconditional and conditional permutation tables for each feature set

The classification performance results presented in Table 5, show that, in Model 1 (the accounting ratio set), both GAM and XGB classifiers detect contributions from the full feature set; as indicated by sensitivity, type II error, log loss and H Measure results. This is in contrast to the unconditional RF feature importance results (

Figure 8), which show strong contributions from asset turnover, liquidity, current and solvency ratios, as well as gearing ratios, indicating that the remaining ratios have limited predictive contributions. Furthermore, according to the unconditional permutation analysis, only identified profit margin as providing strong input to the distress rate.

For Model 2, the results following the introduction of bulk market indicators to company financials concur with the feature importance analysis in that the addition of freight rate information as the only market predictor produces slightly better results for sensitivity, type II error and log loss for GAM. However, the figures for XGB indicate that this algorithm can perform optimally when the complete market feature set is combined with the accounting information. The feature set analysis indicates that for Model 3, the long-term interest rates and inflation play strong roles in the predictive power of the model. This is borne out in the results for both GAM and XGB for Model 3 with the strongest metrics produced when limiting the macroeconomic indicators to these features. Finally, combining all the feature sets into Model 4 appears to demonstrate that both GAM and XGB perform most optimally when the company financials are combined with freight rate and interest rate information. This is consistent with results produced from Model 2 and 3 tests.

The results shown in Table 5 also indicate that the FD prediction power of XGB is improved over that of the GAM classifier, albeit only marginally. However, as reported in (Christoffersen, Matin, and Mølgaard 2018), the differences between the results achieved through complex model compared with the GAM model is not as pronounced as in previous studies e.g. (Jones, Johnstone, and Wilson 2015a). A comparison of the example overall classifier performance with optimal H Measure cost settings for Model 4 (complete dataset) is shown in

Figure 9. Again, with this methodology, the discussion of the optimal cost setting for the H Measure would be a decision for the credit committee.

Finally, we illustrate the use of the methodology to predict the number of companies entering into distress through Figure 10. This compares the realized percentage of firms in distress to the model predicted values from both classifiers. The models are estimated on an expanding window of data with a 2-year lag (t-2) to the forecasted data set e.g. the forecast for 2010 distress rates are estimated using 2003-2008 data.

Table 5: Classifier/Feature set - performance summary

Figure 9: Classifier performance overview – Model 4

Figure 10: Aggregated distress predictions – Model 4

DRAFT

Portfolio application

In the previous section we illustrated how our methodology can be used to predict individual company distress. Here we expand the use of the methodology by examining its capacity to assist with the assessment of shipping portfolio risks. A comparison is made of the models' 95% Value-At-Risk (VaR) values with the realised portfolio distress rates. This can help banks with their estimated shortfall (ES) assessments.[9] The individual company banking information provided a foundation for the construction of bank portfolio related data for this study. Five banks where selected on the diversification of their bulk fleet exposure over the period 2005-2015 (11 portfolio data sets). The GAM and XGB algorithms were used to estimate the 95% VaR of the FD rates for each portfolio. The results are summarised in Figure 11. The solid vertical lines represent the VaR estimates where the line reaches or exceeds the realised figures (dot) and the broken line represents those that fall below the realised VaR. The results show that GAM had 37 VaR violations and XGB producing 34, with neither performing optimally over the 2008-10 period covering the financial crisis.

Figure 11: VaR – Model estimations v actuals

Conclusions

This study has introduced a novel methodology for the prediction of financial distress in dry bulk shipping focusing on the noisy and incomplete nature of shipping financial statement information. The methodology comprises a unique combination of features starting with a flexible definition of FD for shipping company distress. In addition, our methodology includes tools for the analysis and treatment of missing accounting information and incorporates modern machine learning tools.

The main conclusions of the study are, firstly, that multivariate imputation can help shed light on the nature and structure of missing accounting data as well as providing ML tools for its effective treatment. Secondly, we determined that the ML classification technique, XGB, showed an improvement over the use of GAM modelling for FD prediction. However, our results indicate that the GAM algorithm, has a predictive power comparable to that of more complex ML algorithms. Furthermore, the transparent nature of the GAM algorithm, over more complex "black-box" algorithms, could help facilitate the acceptance of this methodology by regulators.

The bulk shipping market case study also demonstrated that the introduction of non-corporate level, macroeconomic and market predictors do not perceptibly improve the predictive power of modern ML tools. The methodology revealed that whilst macro and market predictors do contribute to FD predictive modelling power, XGB can generalise as effectively by simply utilising a parsimonious accounting feature set. The results indicate

[9] *A limitation of the VaR is that it does not quantify the size of the loss once the loss is greater than the confidence threshold. It only informs on the minimum expected loss. Hence conditional VaR (CVaR) or Expected shortfall are used to estimate the value of the loss when the loss exceeds the VaR threshold.*

that if sufficiently extensive longitudinal accounting data is available, then adequate macroeconomic and market information is captured therein, thus enabling advanced ML algorithms to generalise effectively on out of sample data (without the inclusion of additional non company level features).

The results, however, have indicated some limitations of the methodology which require further investigation. Given the noisy and incomplete nature of the available data, even complex models do not achieve an accuracy level, at present, to be relied upon to be used in anything other than an early 'warning system' for FD. Furthermore, its predictive power was compromised through the financial crisis of 2008, in that the model failed completely to handle systemic shock at both company and portfolio levels. In short, further research is required to examine techniques for addressing the problem of tail end events and to improve upon the treatment of missing accounting information.

In summary, this study demonstrated how the methodology could be used by banking credit risk departments to detect early signs of distress at both individual company and investment portfolio levels, providing for the dynamic monitoring of individual shipping company loans as well as portfolio risk.

DRAFT

References

Altman, Edward I. 1977. "ZETA ANALYSIS A New Model to Identify Bankruptcy Risk of Corporations." *Journal of Banking A* 1: 29–54.

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

Amel-zadeh, Amir, Jan-peter Calliess, Stephen Roberts, and Daniel Kaiser. 2020. "Machine Learning-Based Financial Statement Analysis ∗."

Antoniou, A., A. Thanopoulos, and C. Grammenos. 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.*

Azizpour, S, K Giesecke, and G Schwenkler. 2018. "Exploring the Sources of Default Clustering." *Journal of Financial Economics* 129 (1): 154–83. https://doi.org/10.1016/j.jfineco.2018.04.008.

Balcaen, Sofie, and Hubert Ooghe. 2006. "35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems." *British Accounting Review* 38 (1): 63–93. https://doi.org/10.1016/j.bar.2005.09.001.

Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. "Machine Learning Models and Bankruptcy Prediction." *Expert Systems with Applications* 83: 405–17. https://doi.org/10.1016/j.eswa.2017.04.006.

Barnes, Paul. 1987. "The Analysis and Use of Financial Ratios: A Review Article." *Journal of Business Finance & Accounting* 14 (4): 449–61. https://doi.org/10.1111/j.1468-5957.1987.tb00106.x.

Berg, Daniel. 2007. "Bankruptcy Prediction by Generalized Additive Models." *John Wiley & Sons, Ltd.* https://doi.org/10.1002/asmb.

Bickel, J Eric. 2007. "Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules." *Decision Analysis* 4 (2): 49–65.

BIS. 2019. *Minimum Capital Requirements for Market Risk*.

Black, F, and M Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81 (3): 637–57.

Bonfim, Diana. 2009. "Credit Risk Drivers: Evaluating the Contribution of Firm Level Information and of Macroeconomic Dynamics." *Journal of Banking and Finance* 33 (2): 281–99. https://doi.org/10.1016/j.jbankfin.2008.08.006.

Box, G, and D Cox. 1964. "The Analysis of Transformations." *Journal of the Royal Statistical Society*, no. Series B (Methodological): 211–52.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Bureau-Van-Dijk. 2019. "Orbis." Orbis. 2019. https://orbis.bvdinfo.com/version-201989/Home.serv?product=orbisneo.

Burger, Schalk, Searle Silverman, and Gary van Vuuren. 2018. "Deriving Correlation Matrices for Missing

Financial Time-Series Data." *International Journal of Economics and Finance* 10 (10): 105. https://doi.org/10.5539/ijef.v10n10p105.

Buuren, Stef Van. 2018. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–57. https://doi.org/10.1613/jair.953.

Christoffersen, Benjamin, Rastin Matin, and Pia Mølgaard. 2018. "Can Machine Learning Models Capture Correlations in Corporate Distresses?" *SSRN Electronic Journal*, 1–34. https://doi.org/10.2139/ssrn.3273985.

Clarksons. 2019. "Shipping Intelligence Network." Shipping Intelligence Network. 2019. https://sin.clarksons.net/.

Das, Sanjiv R, Darrell Duffie, Nikunj Kapadia, and Leandro Saita. 2007. "Common Failings - How Corporate Defaluts Are Correlated" LXII (1): 93–117.

DiCesare, Giuseppe. 2006. "Imputation, Estimation and Missing Data in Finance."

Dimitras, Augustinos I, Theodore Petropoulos, and Isabella Constantinidou. 2003. "Multi-Criteria Evaluation of Loan Applications in Shipping" 246 (September 2002): 237–46. https://doi.org/10.1002/mcda.332.

Duffie, Darrell, Leandro Saita, Guillaume Horel, and Andreas Eckner. 2009. "Frailty Correlated Default." *Journal of Finance* 64 (5): 2089–2123.

Duffie, Darrell, and Ken Singleton. 2003. "Credit Risk: Pricing, Measurement and Management." *Princeton University Press*, no. October: 1–6. https://doi.org/10.2139/ssrn.600545.

Equasis. 2019. "The World Merchant Fleet 2018." Equasis Statistics Service. 2019. http://www.equasis.org/Fichiers/Statistique/MOA/Documents availables on statistics of Equasis/Equasis Statistics - The world fleet 2018.pdf.

Gavalas, Dimitris, and Theodore Syriopoulos. 2016. "An Integrated Credit Rating and Loan Quality Model : Application to Bank Shipping Finance" 8839 (August). https://doi.org/10.1080/03088839.2014.904948.

Grammenos, C. T. 2010. *The Handbook of Maritime Economics and Business*. 2nd ed. London: LLP.

Grammenos, C. Th, N. K. Nomikos, and N. C. Papapostolou. 2008. "Estimating the Probability of Default for Shipping High Yield Bond Issues." *Transportation Research Part E: Logistics and Transportation Review* 44 (6): 1123–38. https://doi.org/10.1016/j.tre.2007.10.005.

Hand, David J. 2009. "Measuring Classifier Performance : A Coherent Alternative to the Area under the ROC Curve." *Machine Learning*, 103–23. https://doi.org/10.1007/s10994-009-5119-5.

Hastie, Trevor, and Robert Tibshirani. 1987. "Generalized Additive Models: Some Applications." *Journal of the American Statistical Association* 82 (398): 371–86.

Hensher, David A., and Stewart Jones. 2007. "Forecasting Corporate Bankruptcy: Optimizing the Performance of the Mixed Logit Model." *Abacus* 43 (3): 241–364. https://doi.org/10.1111/j.1467-6281.2007.00228.x.

Hernandez Tinoco, Mario, and Nick Wilson. 2013. "Financial Distress and Bankruptcy Prediction among Listed Companies Using Accounting, Market and Macroeconomic Variables." *International Review of Financial Analysis* 30: 394–419. https://doi.org/10.1016/j.irfa.2013.02.013.

Honaker, James, and Gary King. 2010. “What to Do about Missing Values in Time-Series Cross-Section Data.” *American Journal of Political Science* 54 (2): 561–81. https://doi.org/10.1111/j.1540-5907.2010.00447.x.

IFRS Foundation. 2016. “IFRS 16 Leases - Effects Analysis.” *International Financial Reporting Standard*, no. January: 104.

Jones, Stewart, David Johnstone, and Roy Wilson. 2015a. “An Empirical Evaluation of the Performance of Binary Classifiers in the Prediction of Credit Ratings Changes.” *Journal of Banking and Finance* 56: 72–85. https://doi.org/10.1016/j.jbankfin.2015.02.006.

———. 2015b. “An Empirical Evaluation of the Performance of Binary Classifiers in the Prediction of Credit Ratings Changes.” *Journal of Banking and Finance*. https://doi.org/10.1016/j.jbankfin.2015.02.006.

Kavussanos, Manolis G, and Dimitris A Tsouknidis. 2016. “Default Risk Drivers in Shipping Bank Loans.” *Transportation Research Part E* 94: 71–94. https://doi.org/10.1016/j.tre.2016.07.008.

Kim, Myoung Jong, Dae Ki Kang, and Hong Bae Kim. 2015. “Geometric Mean Based Boosting Algorithm with Over-Sampling to Resolve Data Imbalance Problem for Bankruptcy Prediction.” *Expert Systems with Applications* 42 (3): 1074–82. https://doi.org/10.1016/j.eswa.2014.08.025.

Kim, Sang Man, and Jongho Kim. 2018. “The OECD BEPS Package” 49 (2): 221–38.

Kofman, P. 2003. “Using Multiple Imputation in the Analysis of Incomplete Observations in Finance.” *Journal of Financial Econometrics* 1 (2): 216–49. https://doi.org/10.1093/jjfinec/nbg013.

Koopman, Siem Jan, Andr Lucas, and Bernd Schwaab. 2011. “Modeling Frailty-Correlated Defaults Using Many Macroeconomic Covariates.” *Journal of Econometrics* 162 (2): 312–25. https://doi.org/10.1016/j.jeconom.2011.02.003.

Kwon, T Y, and Y Lee. 2018. “Industry Specific Defaults.” *Journal of Empirical Finance* 45: 45–58. https://doi.org/10.1016/j.jempfin.2017.10.002.

Lakshmipadmaja, D., and B. Vishnuvardhan. 2018. “Classification Performance Improvement Using Random Subset Feature Selection Algorithm for Data Mining.” *Big Data Research* 12: 1–12. https://doi.org/10.1016/j.bdr.2018.02.007.

Lessmann, S, B Baesens, H.-V. Seow, and L C Thomas. 2015. “Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research.” *European Journal of Operational Research* 247 (1): 124–36. https://doi.org/10.1016/j.ejor.2015.05.030.

Lohmann, Christian, and Thorsten Ohliger. 2017. “Nonlinear Relationships and Their Effect on the Bankruptcy Prediction.” *Schmalenbach Business Review* 18 (3): 261–87. https://doi.org/10.1007/s41464-017-0034-y.

Lozinskaia, Agata, Andreas Merikas, Anna Merika, and Henry Penikas. 2017. “Determinants of the Probability of Default: The Case of the Internationally Listed Shipping Corporations.” *Maritime Policy and Management* 44 (7): 837–58. https://doi.org/10.1080/03088839.2017.1345018.

Lyridis, Dimitrios V., Nikolaos D. Manos, and Panayotis G. Zacharioudakis. 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): 231–53.

Merk, Olaf M. 2020. “Quantifying Tax Subsidies to Shipping.” *Maritime Economics & Logistics*. https://doi.org/10.1057/s41278-020-00177-0.

Merton, Robert C. 1974. “ON THE PRICING OF CORPORATE DEBT: THE RISK STRUCTURE OF INTEREST RATES*.” *The Journal of Finance* 29 (2): 449–70. https://doi.org/10.1111/j.1540-

6261.1974.tb03058.x.

Mitroussi, K, W Abouarghoub, J J Haider, S J Pettit, and N Tigka. 2016. "Performance Drivers of Shipping Loans : An Empirical Investigation." *Intern. Journal of Production Economics* 171: 438–52. https://doi.org/10.1016/j.ijpe.2015.09.041.

Moraes Barboza, Flavio Luiz de, Herbert Kimura, and Edward Altman. 2015. "Machine Learning Models and Bankruptcy Prediction." *Expert Systems With Applications*. https://doi.org/10.1016/j.eswa.2017.04.006.

Nguyen, Cattram D., John B. Carlin, and Katherine J. Lee. 2017. "Model Checking in Multiple Imputation: An Overview and Case Study." *Emerging Themes in Epidemiology* 14 (1): 1–12. https://doi.org/10.1186/s12982-017-0062-6.

Nickerson, J, and J M Griffin. 2017. "Debt Correlations in the Wake of the Financial Crisis: What Are Appropriate Default Correlations for Structured Products?" *Journal of Financial Economics* 125 (3): 454–74. https://doi.org/10.1016/j.jfineco.2017.06.011.

OECD. 2013. *Addressing Base Erosion and Profit Shifting*. *Addressing Base Erosion and Profit Shifting*. Vol. 9789264192. https://doi.org/10.1787/9789264192744-en.

———. 2019. "OECD Data." OECD Data. 2019. https://data.oecd.org/.

Ohlson, James A. 1980a. "Financial Ratios and the Probabilistic Prediction of Bankruptcy." *Journal of Accounting Research* 18 (1).

———. 1980b. "Financial Ratios and the Probabilistic Prediction of Bankruptcy." *Journal of Accounting Research* 18 (1): 109. https://doi.org/10.2307/2490395.

Pindado, Julio, Luis Rodrigues, and Chabela de la Torre. 2008. "Estimating Financial Distress Likelihood." *Journal of Business Research* 61 (9): 995–1003. https://doi.org/10.1016/j.jbusres.2007.10.006.

Rubin, D B. 1987. "Multiple Imputation for Nonresponse in Surveys. En."

Serneels, Sven, Evert De Nolf, and Pierre J. Van Espen. 2006. "Spatial Sign Preprocessing: A Simple Way to Impart Moderate Robustness to Multivariate Estimators." *Journal of Chemical Information and Modeling* 46 (3): 1402–9. https://doi.org/10.1021/ci050498u.

Shah, Anoop D, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE : A CALIBER Study" 179 (7): 764–74. https://doi.org/10.1093/aje/kwt312.

Sharife, K. 2010. "Flying a Questionable Flag: Liberia's Lucrative Shipping Industry." *World Policy Journal* 27 (4): 111–18. https://doi.org/10.1162/wopj.2011.27.4.111.

Shumway, Tyle. 2001. "Forecasting Bankruptcy More Accuratemy - A Simple Hazard Model." Journal of business.

Son, H., C. Hyun, D. Phan, and H.J. Hwang. 2019. "Data Analytic Approach for Bankruptcy Prediction." *Expert Systems with Applications* 138: 112816. https://doi.org/10.1016/j.eswa.2019.07.033.

Spackman, Kent A. 1989. "Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning." In *Proceedings of the Sixth International Workshop on Machine Learning*, 160–63. Elsevier.

Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." *The American Journal of Psychology* 15 (1): 72–101. https://doi.org/10.2307/1412159.

Stekhoven, Daniel J, and Peter Bühlmann. 2012. “MissForest — Non-Parametric Missing Value Imputation for Mixed-Type Data” 28 (1): 112–18. https://doi.org/10.1093/bioinformatics/btr597.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. “Conditional Variable Importance for Random Forests.” *BMC Bioinformatics* 9 (1): 307.

Tahtah, Jay, and Erik Roelofsen. 2016. “A Study on the Impact of Lease Capitalisation IFRS 16 : The New Leases Standard.” *PwC Report*.

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): 441–49. https://doi.org/10.1080/01621459.1993.10476294.

UNCTAD. 2019. “Review of Maritime Transport.”

World Bank. 2019. “World Bank Open Data.” World Bank Open Data. 2019. https://data.worldbank.org/.

Wruck, Karen Hopper. 1990. “Financial Distress, Reorganization, and Organizational Efficiency.” *Journal of Financial Economics* 27 (2): 419–44. https://doi.org/10.1016/0304-405X(90)90063-6.

Wu, Yijie, Jingbo Yin, and Pan Sheng. 2018. “The Dynamics of Dry Bulk Shipping Market under the Shipping Cycle Perspective: Market Relationships and Volatility.” *Transportation Research Record* 2672 (11): 1–9. https://doi.org/10.1177/0361198118756622.

Xia, Y, C Liu, Y Li, and N Liu. 2017. “A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring.” *Expert Systems with Applications* 78: 225–41. https://doi.org/10.1016/j.eswa.2017.02.017.

Yeo, I.N., and R.A Johnson. 2000. “A New Family of Power Transformations to Improve Normality or Symmetry.” *Biometrika* 87 (4): 954–59.

Yin, Jingbo, Yijie Wu, and Linjun Lu. 2019. “Assessment of Investment Decision in the Dry Bulk Shipping Market Based on Real Options Thinking and the Shipping Cycle Perspective.” *Maritime Policy and Management* 46 (3): 330–43. https://doi.org/10.1080/03088839.2018.1520400.

Yu, Fan, Nai-Fu Chen, Sanjiv Das, Darrell Duffie, Jan Ericsson, Philippe Jorion, Raymond Kan, Ken Singleton, Ram Willner, and F Yu. 2005. “Accounting Transparency and the Term Structure of Credit Spreads ARTICLE IN PRESS.” *Journal of Financial Economics* 75: 53–84. https://doi.org/10.1016/j.jfineco.2004.07.002.

Zhang, Chongsheng, Changchang Liu, Xiangliang Zhang, and George Almpanidis. 2017. “An Up-to-Date Comparison of State-of-the-Art Classification Algorithms.” *Expert Systems with Applications* 82: 128–50. https://doi.org/10.1016/j.eswa.2017.04.003.

Zhou, Qifeng, Hao Zhou, and Tao Li. 2016a. “Cost-Sensitive Feature Selection Using Random Forest: Selecting Low-Cost Subsets of Informative Features.” *Knowledge-Based Systems* 95: 1–11. https://doi.org/10.1016/j.knosys.2015.11.010.

———. 2016b. “Cost-Sensitive Feature Selection Using Random Forest: Selecting Low-Cost Subsets of Informative Features.” *Knowledge-Based Systems* 95: 1–11.

Zięba, M, S K Tomczak, and J M Tomczak. 2016. “Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction.” *Expert Systems with Applications* 58: 93–101. https://doi.org/10.1016/j.eswa.2016.04.001.

DRAFT

Tables

Table 1: Predictive features for the FD model.

Category	Feature	Description
Company	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)
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 [10]	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)

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

DRAFT

Table 2: 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%			

DRAFT

Table 3: 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%

DRAFT

Table 4: Missing value treatments – RF classification

Missing value Treatment	Accuracy	Kappa	Sensitivity	Specificity	Log loss	AUC
RF Imputation	0.764	0.449	0.776	0.761	0.486	0.849
Complete case	0.810	0.497	0.830	0.806	0.516	0.869

DRAFT

Table 5: Classifier/Feature set - performance summary

Classifier	Accurac	Kappa	Sensitivity	Specificity	Type II err	log loss	H Measu	AUC	Predictor limits
					Model 1				
GAM	0.775	0.465	0.773	0.775	22.66%	0.480	0.361	0.858	None
GAM	0.777	0.467	0.766	0.781	23.44%	0.495	0.352	0.836	Solvency/Liquidity/Current/Gearing/NetAssetT/ProfitM
GAM	0.761	0.440	0.766	0.760	23.44%	0.507	0.349	0.829	Solvency/Liquidity/Current/Gearing/NetAssetT
XGB	0.811	0.526	0.762	0.826	23.83%	0.418	0.441	0.884	None
XGB	0.788	0.474	0.727	0.807	27.34%	0.433	0.432	0.865	Solvency/Liquidity/Current/Gearing/NetAssetT/ProfitM
XGB	0.785	0.476	0.754	0.794	24.61%	0.457	0.429	0.860	Solvency/Liquidity/Current/Gearing/NetAssetT
					Model 2				
GAM	0.775	0.455	0.742	0.785	25.78%	0.469	0.376	0.857	None
GAM	0.774	0.458	0.758	0.779	24.22%	0.462	0.377	0.859	WST Ore/Grain/Cc/Sc
GAM	0.776	0.459	0.750	0.783	25.00%	0.462	0.377	0.860	WST Ore/Grain/Cc/Sc + IYTC
GAM	0.779	0.465	0.750	0.788	25.00%	0.462	0.377	0.860	WST Ore/Grain/Cc/Sc + IYTC+Scrap
GAM	0.776	0.457	0.742	0.786	25.78%	0.463	0.377	0.858	excl. SHPDex/LaidUp/MDO/IFO
XGB	0.818	0.541	0.766	0.834	23.44%	0.424	0.439	0.880	None
XGB	0.807	0.507	0.723	0.833	27.73%	0.423	0.438	0.878	WST Ore/Grain/Cc/Sc
XGB	0.816	0.534	0.758	0.833	24.22%	0.424	0.438	0.878	WST Ore/Grain/Cc/Sc + IYTC
XGB	0.810	0.516	0.734	0.833	26.56%	0.421	0.438	0.877	WST Ore/Grain/Cc/Sc + IYTC+Scrap
XGB	0.806	0.510	0.742	0.825	25.78%	0.422	0.438	0.877	Excl. SHPDex/LaidUp/MDO/IFO
					Model 3				
GAM	0.777	0.462	0.754	0.783	24.61%	0.465	0.383	0.858	None
GAM	0.776	0.469	0.781	0.774	21.88%	0.471	0.384	0.861	LTI
GAM	0.766	0.445	0.758	0.768	24.22%	0.463	0.384	0.860	LTI/STI
GAM	0.765	0.437	0.738	0.773	26.17%	0.463	0.383	0.859	LTI/STI/Inflation
GAM	0.767	0.450	0.766	0.768	23.44%	0.462	0.384	0.860	LTI/STI/Employment
XGB	0.820	0.538	0.742	0.844	25.78%	0.428	0.430	0.879	None
XGB	0.806	0.500	0.711	0.834	28.91%	0.413	0.429	0.878	LTI
XGB	0.809	0.513	0.730	0.833	26.95%	0.418	0.429	0.878	LTI/STI
XGB	0.820	0.538	0.742	0.844	25.78%	0.425	0.430	0.880	LTI/STI/Inflation
XGB	0.818	0.529	0.727	0.846	27.34%	0.422	0.431	0.882	LTI/STI/Employment
					Model 4				
GAM	0.773	0.430	0.680	0.801	32.03%	0.458	0.390	0.829	None
GAM	0.769	0.444	0.738	0.779	26.17%	0.474	0.386	0.821	
GAM	0.772	0.454	0.754	0.778	24.61%	0.462	0.404	0.859	WST Ore/Grain/Cc/Sc + LTI/STI
XGB	0.820	0.538	0.742	0.844	25.78%	0.428	0.440	0.879	None
XGB	0.787	0.470	0.719	0.808	28.13%	0.443	0.430	0.860	
XGB	0.800	0.502	0.750	0.815	25.00%	0.423	0.440	0.879	

Table 6: Imputation evaluation metrics

Ratio	Mean original	SD original	Mean imputed	SD imputed	Welch _ttest P	KS test D
ROE	2.127E+00	1.520E+01	-2.692E+00	2.112E+01	1.607E-06	2.285E-01
ROA	4.876E+00	2.361E+01	-6.922E+00	2.836E+01	1.852E-24	2.752E-01
ProfitM	3.585E+01	3.365E+01	3.423E+01	2.244E+01	7.524E-02	1.164E-01
GrossM	1.701E+01	2.594E+01	1.433E+01	2.238E+01	6.130E-04	8.708E-02
EBITDAM	7.488E+00	2.291E+01	1.223E+00	2.018E+01	5.186E-14	1.931E-01
EBITM	1.172E+01	2.425E+01	9.787E+00	2.311E+01	1.390E-02	6.456E-02
NetAssetT	8.528E+00	4.466E+01	2.377E+01	4.208E+01	4.496E-16	3.454E-01
CurrentR	2.722E+00	6.568E+00	9.535E+00	1.369E+01	4.980E-09	4.873E-01
LiquidityR	2.523E+00	6.221E+00	9.620E+00	1.477E+01	5.515E-10	4.359E-01
SolvencyR	3.698E+01	3.161E+01	-2.905E+00	4.365E+01	2.186E-21	5.279E-01
Gearing	1.414E+02	1.777E+02	3.200E+02	2.443E+02	1.685E-90	4.319E-01

DRAFT

Figures

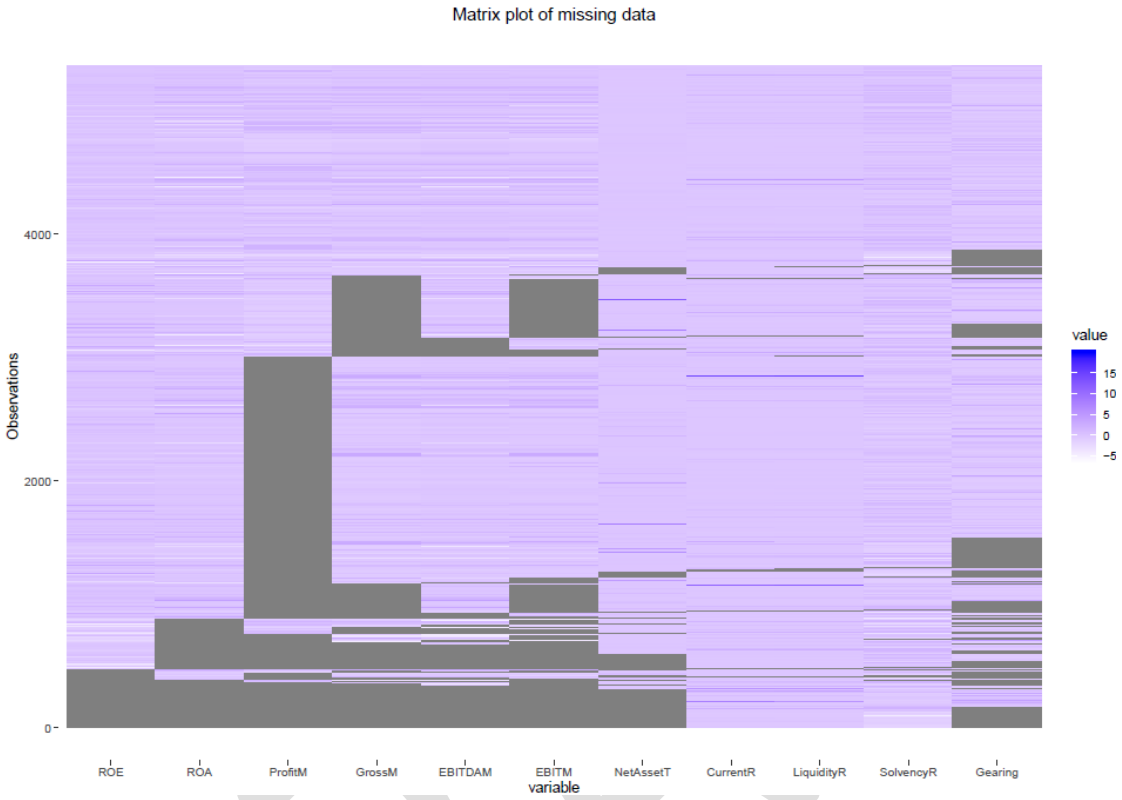


Figure 1: Matrix plot of missing accounting data

DRAFT

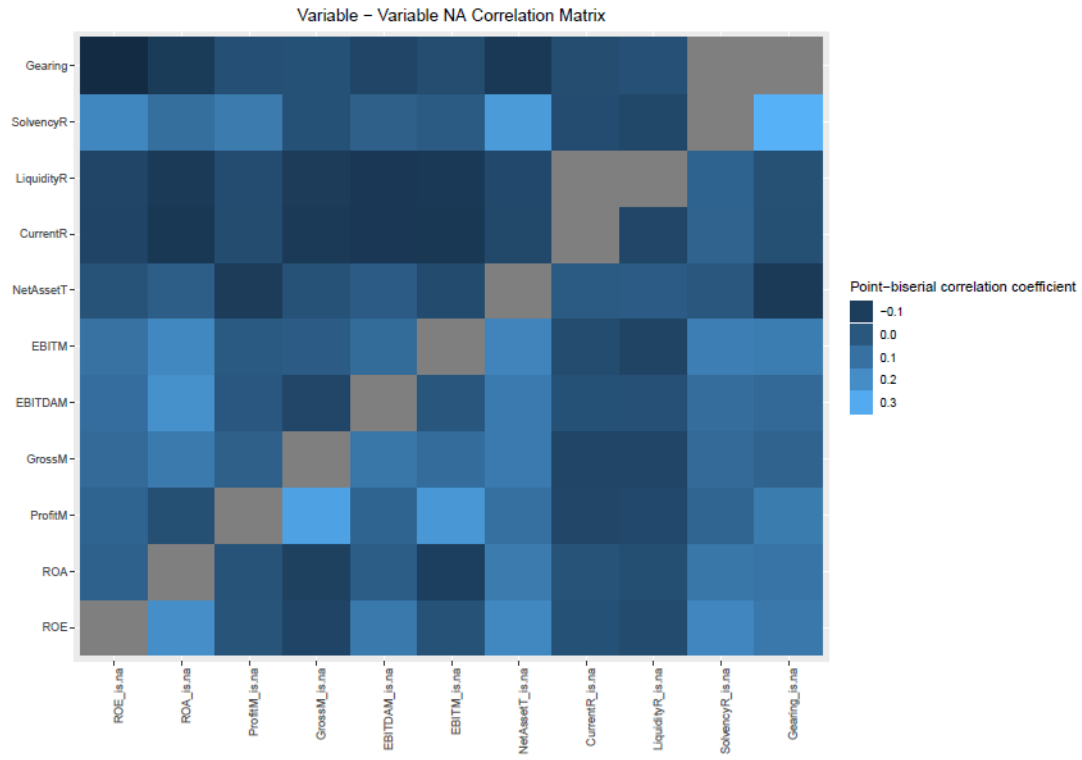


Figure 2: Raw accounting data - observed v missing (NA) correlation coefficients

DRAFT

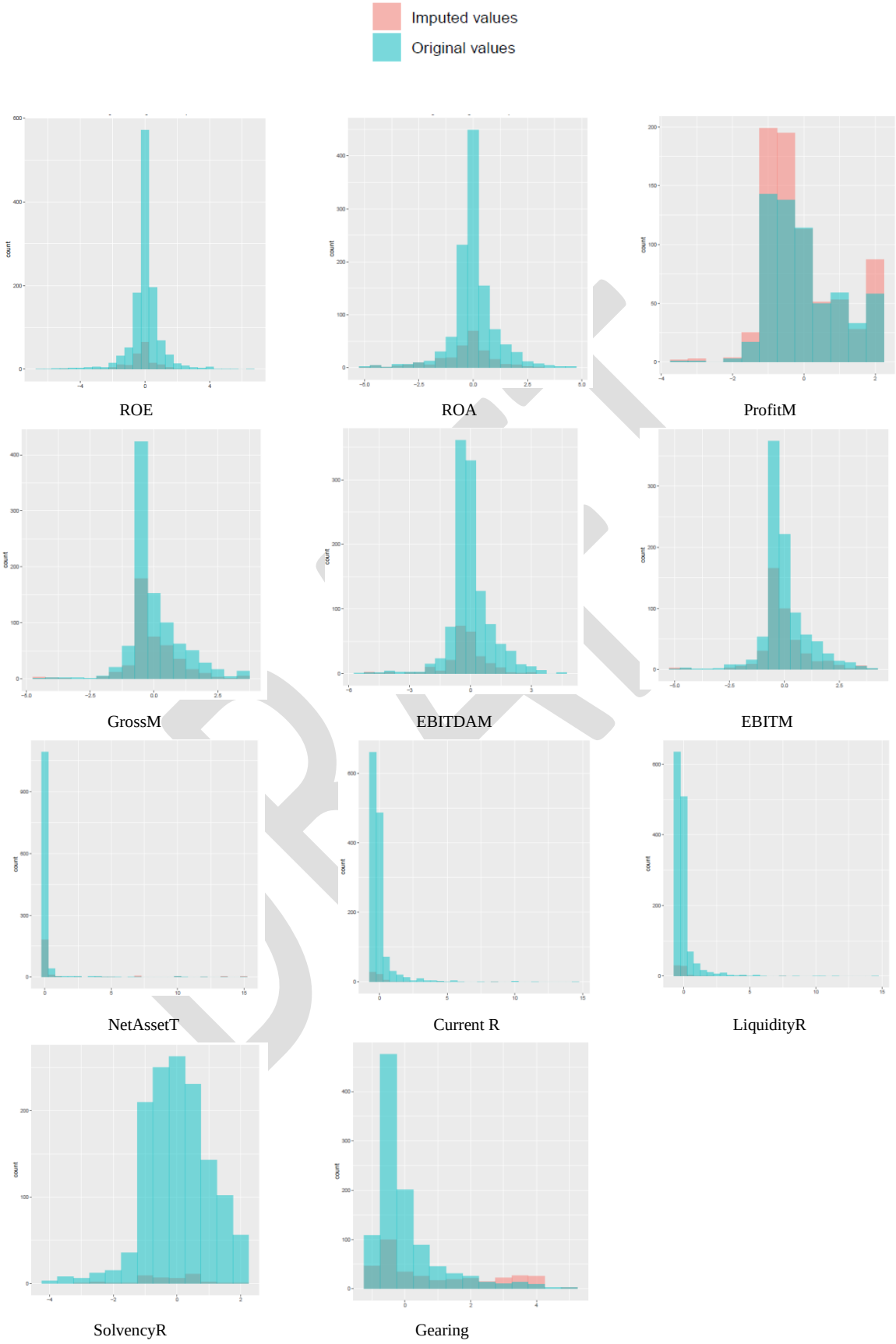


Figure 3: Overlaid histograms of imputed and original values

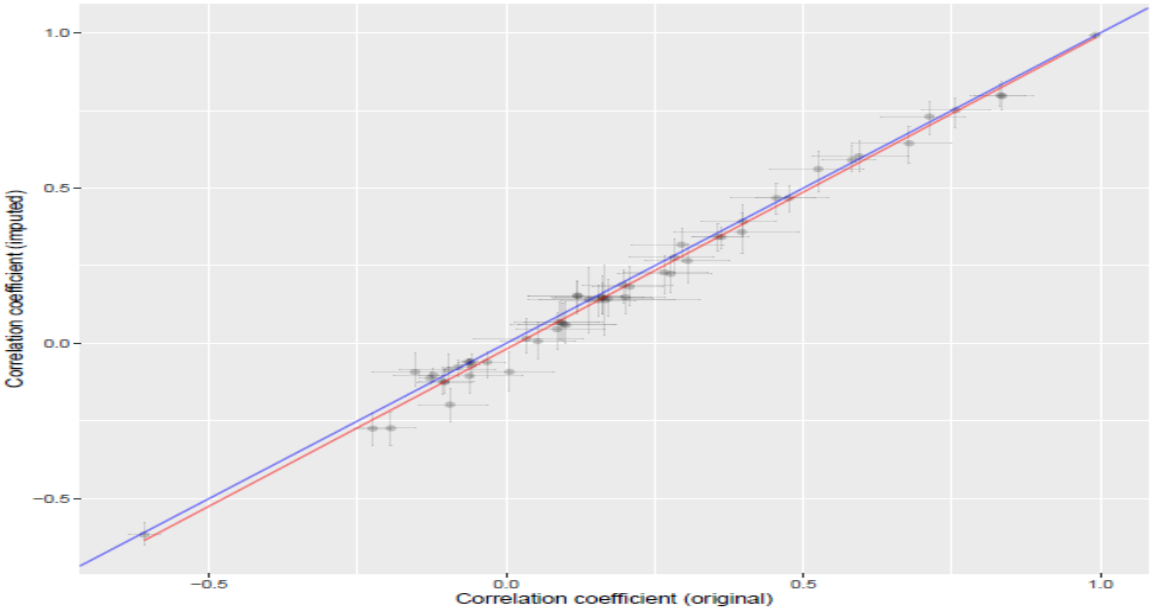
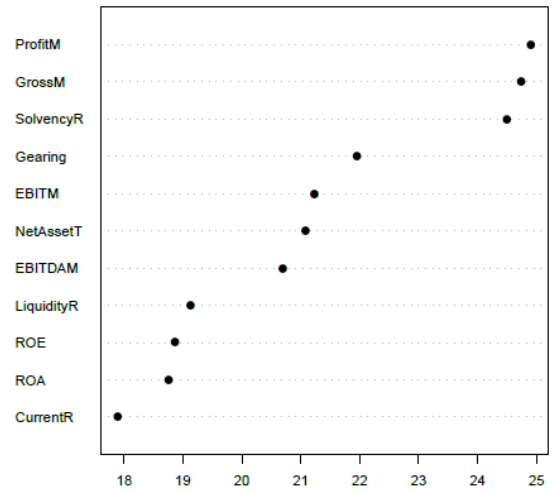


Figure 4: Correlation coefficient scatter plot

DRAFT

Complete case treatment



RF Imputation

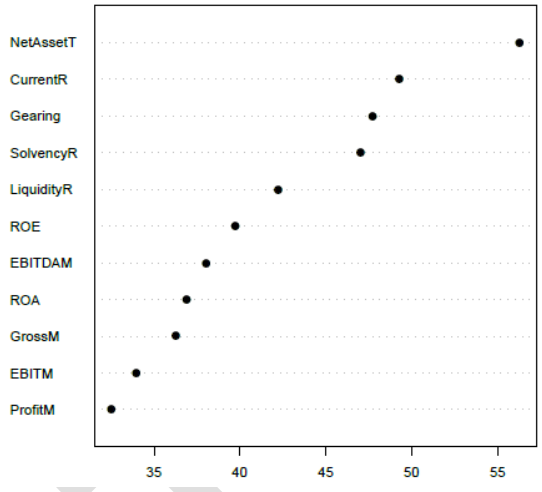


Figure 5: Missing value treatment – Data set feature importance comparison

DRAFT

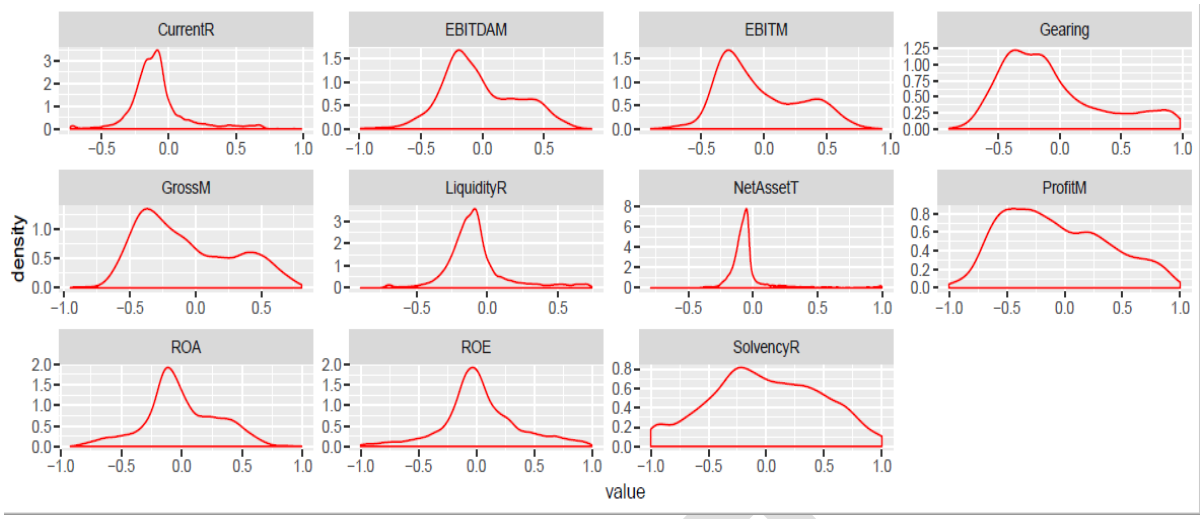
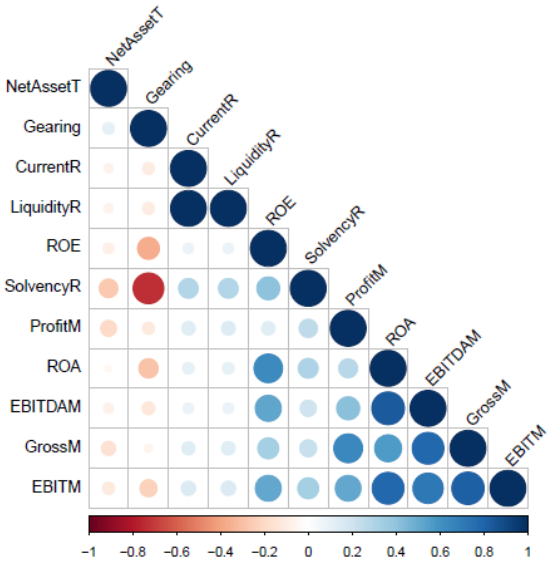


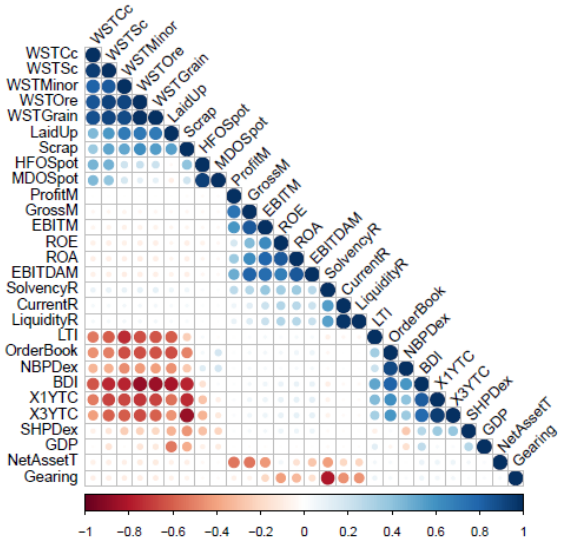
Figure 6: Accounting feature distributions

DRAFT

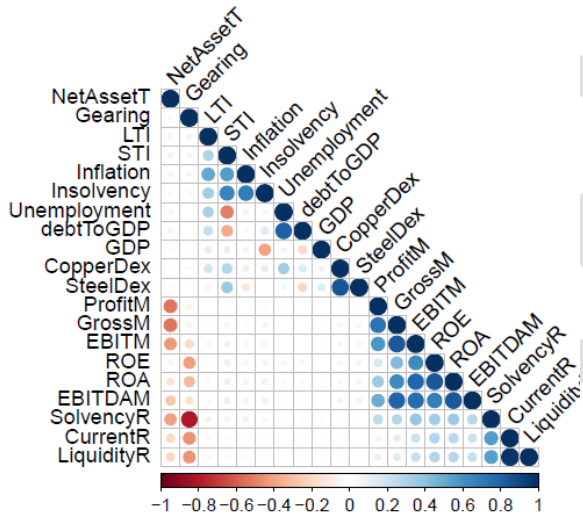
Model 1



Model 2



Model 3



Model 4

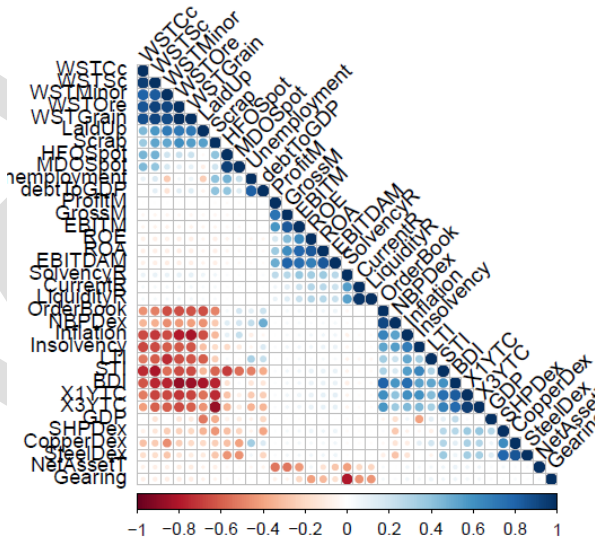


Figure 7: Feature correlation matrices

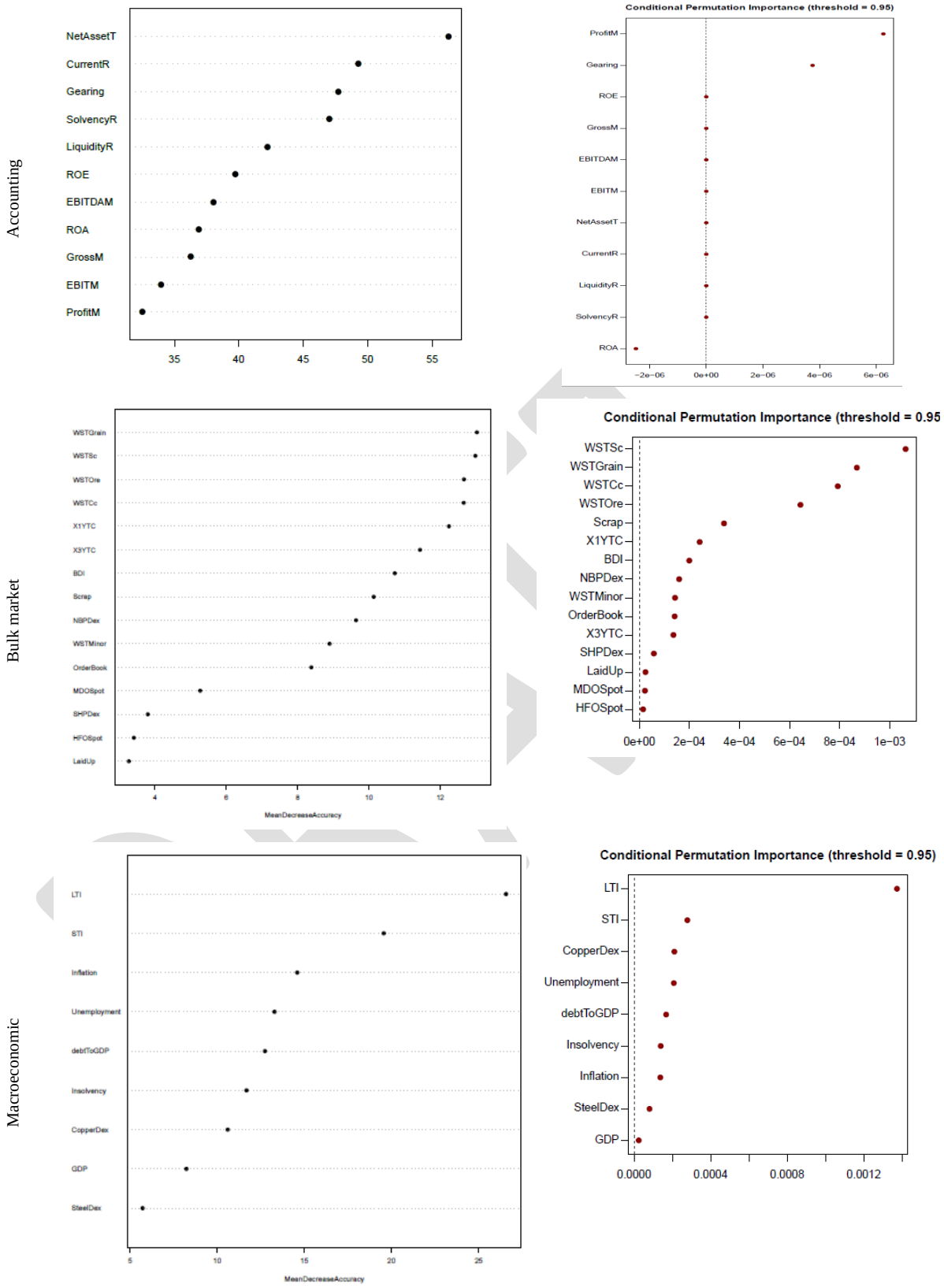


Figure 8: Unconditional and conditional permutation tables for each feature set

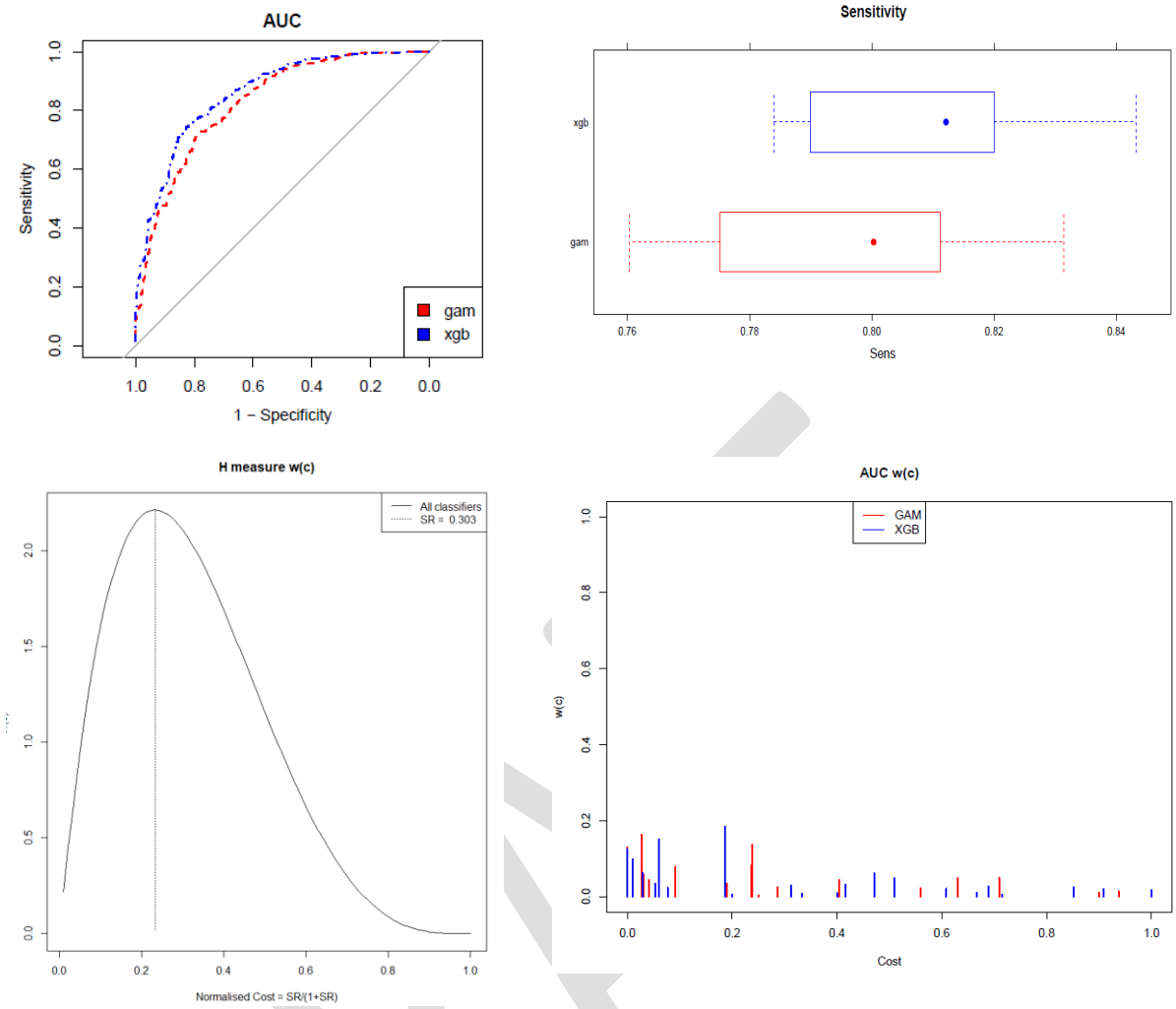


Figure 9: Classifier performance overview – Model 4

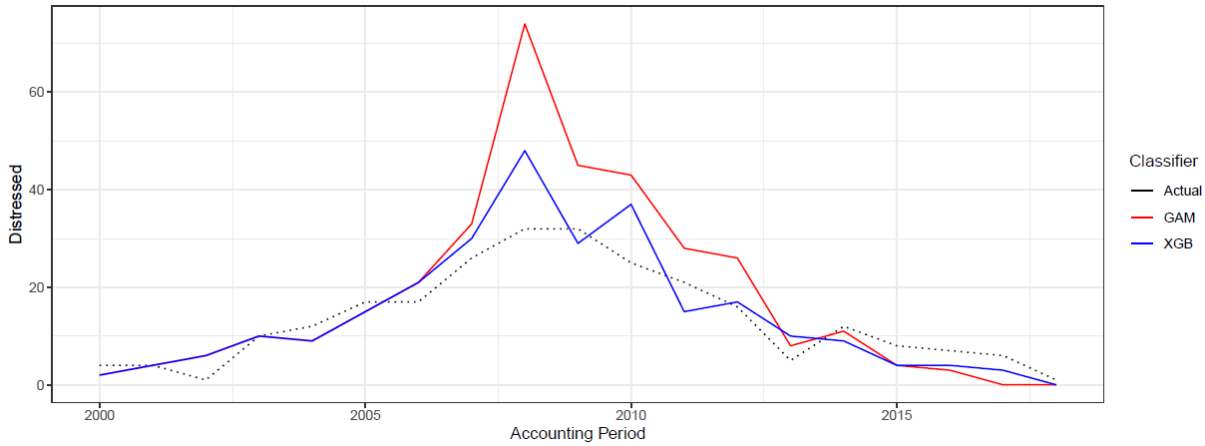


Figure 10: Aggregated distress predictions – Model 4

DRAFT

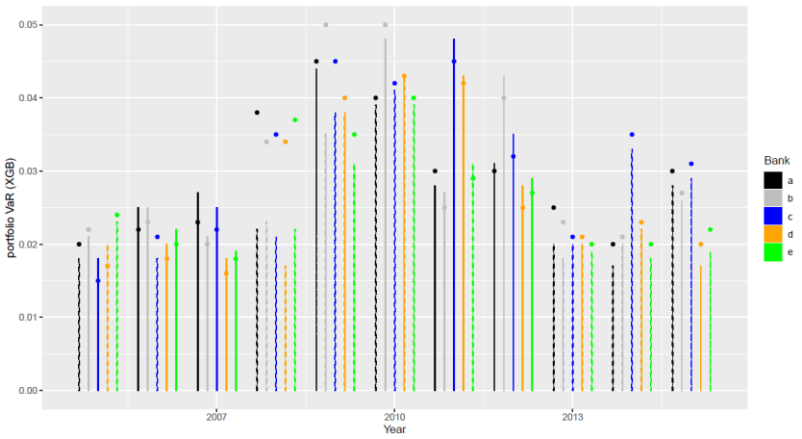


Figure 11: VaR – Model estimations v actuals

DRAFT

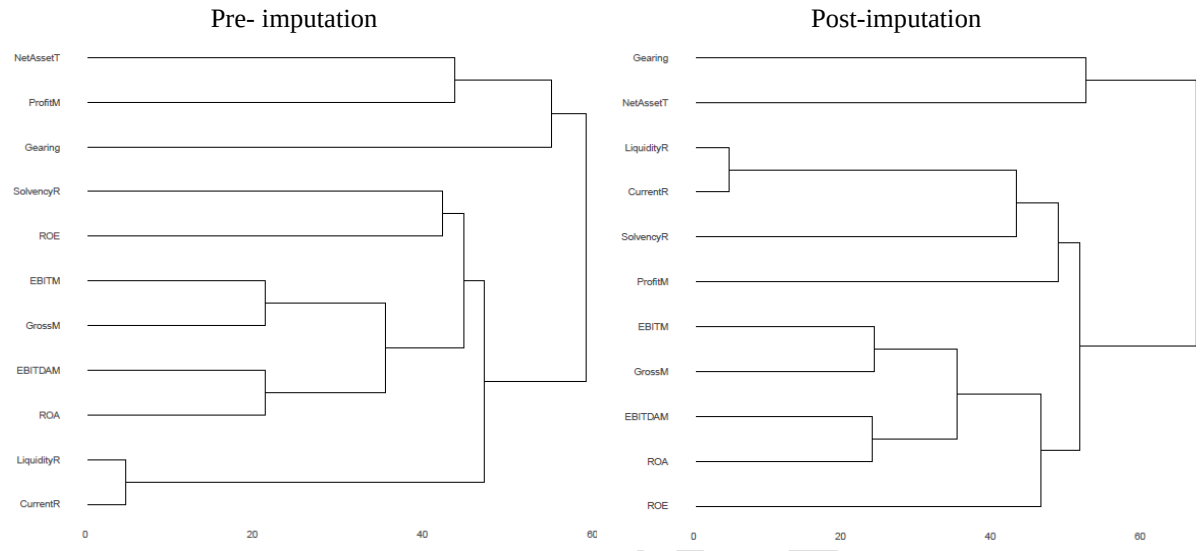


Figure 12: Original and imputed variable clusters - pre and post imputation

DRAFT

Appendix A - Methodology

This appendix is intended to provide further details on the methodological reasoning supporting this study.

Missing value treatment – Complete case and multivariate imputation

Under the MAR assumption, we can use information in the training set to estimate the values of other predictors. The now widely accepted random forest multiple imputation (MI) model was selected for the MAR data in our test case (Stekhoven and Bühlmann 2012; Shah et al. 2014; Van Buuren 2018). This method utilizes bootstrapped aggregation of multiple regression trees to combine predictions in order to both reduce overfitting and improve accuracy of predictions. Furthermore, Shah et al. (2014) conclude that this method is more efficient and produces narrower CIs compared to other MI models. We compare an RF imputation with the results from applying the "complete case" approach.

Pre-processing

Some advanced ML algorithms (e.g. tree-based models) have demonstrated robustness to skewness, kurtosis and outliers in contrast with generalised linear models, which are particularly sensitive to such issues. We used two pre-processing transformations to address these issues. Firstly, we used a variation of Box and Cox (1964) transformation, developed by Yeo and Johnson (2000), as Box-Cox requires input variables to be strictly positive. Secondly, the treatment of outliers, through the use of spatial sign transformation, from Serneels, De Nolf, and Van Espen (2006), was used as financial ratios values can exhibit heavily skewed distributions due to the presence of outliers.

Finally, classification algorithms are generally trained under the assumption that class ratios in the training data are balanced. However, in real world datasets this assumption is frequently violated. This was found to be the case with our dataset, with the "distressed" case being the minority case (consistent with World Bank and OEDC figures reflecting global long-term average commercial company default rates of around 4 to 6 percent). Our preliminary pre-processing indicated that down sampling produced the most optimal results following benchmark testing against "up" and SMOTE (Chawla et al. 2002) sampling.

Feature selection

The role of feature analysis and selection is essential for the identification of independent variables which contribute most to the dependent variable. In our methodology we implement RF (Breiman 2001) which are increasingly implemented for this task across many fields of research (Zhou, Zhou, and Li 2016b; Lakshmipadmaja and Vishnuvardhan 2018). Each node in an RF decision tree represents a condition on a feature which is designed to dissect the data so that similar response values are contained in a common set. We use conditional permutation as a feature importance metric in order to account for collinearity amongst variables.

Classification models – Theoretical background

Generalized 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 relax this assumption by accounting for the fact that some of the predictors exhibit a continuous, non-linear relationship with FD (Hastie and Tibshirani 1987). Furthermore, non-linear relationships are observed both below and above specific thresholds with respect to the adjusted financial ratios of shipping companies. This necessitates due account being taken of these non-linear relationships. Compared with other linear classifiers, GAMs demonstrate superior regularisation thus enabling them to more adequately address problems of overfitting. They also have an advantage over more complex models, by being more interpretable and as such, GAMs represent an acceptable solution between the interpretable, yet biased, linear models, and more complex, "black box" learning algorithms.

Our implementation of company FD prediction, utilising GAMs follows Berg (2007); Lohmann and Ohliger (2017); Christoffersen, Matin and Mølgaard (2018).

Classification model evaluation

The classification performance of each model/classification combination is carried out 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, the 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. 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:

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

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 models, and not only the final output of the classification. The stronger 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 most clearly distinguish classes. Furthermore, log loss is useful for comparing not only model output but on their individual probabilistic outcome.

Appendix B – Post imputation evaluation results

The effect of imputation on each individual financial covariate was examined first. The basic testing assumption was 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 were the Welch t-test P and the KS D values. For example, a 95% significance level indicates that when imputed covariates show a $p < .05$, the null hypothesis should be rejected. Also, KS D values should be close to 0.

Table 6: Imputation evaluation metrics

To summarise, the Welch's t-test and the KS-D test were used 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 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 cannot be rejected at the 5% level. The KS D values for the imputed data set indicate a close approximation to the validation data than dissimilarity, since all values in Table 6 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.

A visualisation of the hierarchical clustering of the missing data is provided in Figure 12**.** 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.

Figure 12 : Original and imputed variable clusters - pre and post imputation

The results summarised in Table 4, show that the complete case (CC) data set shows increased sensitivity and AUC results than that produced using RF MI data. This indicates improved out-of-sample generalisation performance using CC data. However, the log-loss values signal a greater misclassification error with the CC data. This is an indication that, as discussed previously, the removal of circa 72% of records (containing incomplete data) involved the possibility of introducing bias.

DRAFT