BANK OF ENGLAND

Staff Working Paper No. 831

Predicting bank distress in the UK with machine learning

Joel Suss and Henry Treitel

October 2019

BANK OF ENGLAND

Staff Working Paper No. 831
Predicting bank distress in the UK with machine learning

Joel Suss[1] and Henry Treitel[2]

Abstract

Using novel data and machine learning techniques, we develop an early warning system for bank distress. The main input variables come from confidential regulatory returns, and our measure of distress is derived from supervisory assessments of bank riskiness from 2006 through to 2012. We contribute to a nascent academic literature utilising new methodologies to anticipate negative firm outcomes, comparing and contrasting classic linear regression techniques with modern machine learning approaches that are able to capture complex non-linearities and interactions. We find the random forest algorithm significantly and substantively outperforms other models when utilising the AUC and Brier Score as performance metrics. We go on to vary the relative cost of false negatives (missing actual cases of distress) and false positives (wrongly predicting distress) for discrete decision thresholds, finding that the random forest again outperforms the other models. We also contribute to the literature examining drivers of bank distress, using state of the art machine learning interpretability techniques, and demonstrate the benefits of ensembling techniques in gaining additional performance benefits. Overall, this paper makes important contributions, not least of which is practical: bank supervisors can utilise our findings to anticipate firm weaknesses and take appropriate mitigating action ahead of time.

Key words: Machine learning, bank distress, early warning system.

JEL classification: C14, C33, C52, C53, G21.

(1) Bank of England and London School of Economics. Email: joel.suss@bankofengland.co.uk
(2) Bank of England. Email: henry.treitel@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are extremely grateful to Andreas Joseph and Sebastian De-Ramon for their encouragement, advice and valuable help throughout. We are also grateful to Helen Ainsworth, Marco Bardoscia, David Bholat, Marcus Buckmann, Bill Francis, Periklis Gogas, Marc Hinterschweiger, Abu Karbhari, Kate Laffan, James Proudman, Misa Tanaka, Arthur Turrell and seminar participants at the Bank of England and Financial Conduct Authority.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email publications@bankofengland.co.uk

ISSN 1749-9135 (on-line)

Section 1: Introduction

The Bank of England's Prudential Regulation Authority (PRA) is charged with ensuring the safety and soundness of banks and building societies in the UK. It is therefore integral that its supervisors, as with microprudential regulators elsewhere, are able to anticipate firm distress. This research paper aims to support this endeavour, developing an early warning system that can aid in anticipating distress.

Our work builds on the existing academic literature in a number of key ways. First, we utilise novel data not available elsewhere, namely confidential supervisory assessments of firm risk and regulatory returns data, for a sample of UK regulated firms between 2006 and 2012. This allows us to model distress – which we define as those firms that were rated 'high risk' by bank supervisors – rather than outright failure, as is typical in previous work. This approach has a number of key advantages; most notably, it is better aligned to the practical needs of regulatory bodies looking to intervene far ahead of failure. Moreover, from an academic perspective, it avoids the (surprisingly) difficult question of delineating what constitutes failure – is it simply actual bankruptcies and liquidations, or is it inclusive of distressed mergers and government bailouts or some technical threshold? The former excludes notable crises, for example HBOS and RBS. The latter requires drawing an arbitrary line that can be difficult to justify. Finally, our dataset contains a wider range of firms, including smaller, unlisted organisations that are typically excluded from other analyses due to data availability.

Second, we contribute to the literature on early warning systems by going beyond conventional modelling techniques, utilising methods from the machine learning literature alongside more traditional approaches. Conventional approaches, such as logistic regression models, are unable to account for complex interactions and non-linearities, thereby tending to perform worse than their more flexible machine learning counterparts. In this paper, we compare pooled logistic regression (the workhorse in the early warning system literature), with a linear random effects model, the k nearest neigbours (KNN) algorithm, two classfication tree ensembles (random forest and boosting), and a support vector machine (SVM). Our base models predict firm distress one year out, lagging each of the predictors by four quarters. In comparing and contrasting these six different approaches, we build on a growing body of work that examines the benefits of new methodological approaches for predicting bank distress (see, for example, Boyacioglu, Kara, and Baykan (2009); Iturriaga and Sanz (2015); Le and Viviani (2018); Gogas, Papadimitriou, and Agrapetidou (2018); Carmona, Climent, and Momparler (2018)).

In order to measure performance, we estimate out-of-sample predicted probabilities using a unique cross validation design that accounts for various potential sources of bias. In particular, and unlike most other research in this area, we account for the dependency structure in our data by block-randomising between training and test samples by both bank and quarter. Failure to do so results in 'data leakage' and leads to overly optimistic out-of-sample performance estimates (Roberts et al. 2017; Kaufman et al. 2012). Moreover, where hyperparameters need to be selected, we perform block-randomised cross validation within each fold (Cawley and Talbot 2010). Finally, we repeat the entire cross validation exercise ten times to account for the variability that arises due to the specific

initial random split performed (Bouckaert and Frank 2004). Altogether, while computationally costly, our procedure produces more reliable performance estimates than previous work applying machine learning techniques to bank failure, allowing us to be more confident in the generalisability of our results.

We find that the random forest algorithm significantly outperforms the other approaches examined in terms of the area under the ROC curve (AUC) and Brier score. We then assess performance at relevant decision thresholds under the assumption that regulators prefer reducing false negative error rates (missing actual cases of distress) to false positive error rates (wrongly predicting distress). This is because an early warning system that fails to set the alarm when it should can lead to costly consequences. We therefore alter the relative cost of these two types of error, finding that the random forest does increasingly better relative to other models as we increase the cost of the former versus the latter.

However, the performance advantage of the random forest algorithm comes with a transparency cost relative to the pooled logit model which, depending on the requirements of regulators, could outweigh the benefits. For example, supervisors might wish to understand how different mitigating actions, such as increases in capital or liquidity buffers, affect a bank's probability of distress one year out. This sort of analysis is substantially more straightforward for linear statistical models. In order to remediate this relative lack of transparency, we utilise recently developed techniques to understand the drivers of the random forest algorithm, computing and aggregating Shapley values to provide a measure of the relative importance of each variable in driving predictions (Strumbelj and Kononenko 2014; Lundberg and Lee 2017), and utilising a novel statistical framework for machine learning models: Shapley regressions (Joseph 2019).

The analysis reveals, unsurprisingly given our period of study, that lagged macroeconomic variables are very important for predicting distress. For the random forest, a measure of average real UK earnings is the single most important variable. In contrast, average earnings is not as important, relatively speaking, for the pooled logit model, a fact that we explain by demonstrating its elevated interaction strength as measured by the H-statistic (defined as the share of total variance explained by a variable's interaction with all other variables; (Friedman and Popescu 2008)). In terms of firm-level financial ratios, we find a bank's sensitivity to market risk (ratio of trading book to total assets), capital buffer, and net interest margin to be significant in explaining the random forest predictions.

Lastly, we also examine the performance benefits of different straightforward ensemble techniques, i.e. we evaluate how different combinations of the six different models perform. We find that a stacked ensemble using a linear regression as the second-level model does a better job than the simple averaging ensembles and the random forest alone, and the improvement relative to the random forest is significant in both a substantive and statistical sense.

The rest of the paper is organised as follows: Section 2 details the data utilised, Section 3 provides the methods used to predict bank distress, Section 4 presents the results, Section 5 provides robustness checks, and Section 6 concludes.

Section 2: Data

2.1 Definition of distress

Contrary to the vast majority of early warning system research, we do not model outright firm failure. Instead, we use a novel source of data to define distress: confidential supervisory assessments on the riskiness of UK banks and building societies from Q3 2006 to Q4 2012. This information comes from the Financial Services Authority (FSA) Arrow scores database and contains assessments of a variety of different categories of risk.[3] On each element, banks were scored at any one of 10 different levels from 'Low' (meaning low risk to the achievement of the FSA's objectives) to 'High'. In our analysis, we model the summary score given to each regulated bank, known as the Total Probability score, and treat a bank as being in a realised state of distress if its Total Probability score was in any of the 'High' notches, i.e. if it scored 8 or above in the current quarter.

The scores were the result of periodic assessments by supervisors which were reviewed, challenged and approved by panels involving senior management and risk review functions. The data are quarterly, although scores were not typically reviewed officially this frequently. Standard practice was for larger and riskier banks to have more frequent official reviews than smaller banks[4], however scores would be officially updated more frequently than the standard as circumstances required, for example as a result of a sharp, unexpected movement in a bank's position. The Arrow scores database also provides supervisory assessments in between official updates. These unofficial scores represent the interim views of the front-line supervisors and were updated at their discretion. We incorporate this interim information into our response variable to mitigate the possibility that official score reviews were not happening as frequently as warranted for smaller, less-risky banks.

Figure 1 below provides the quarterly distribution of Arrow Total Probability scores depicted through boxplots. There is a clear inflation of scores as the financial crisis hits, indicating that the bulk of assessments were on the higher end of the risk scale post-crisis. The elevated scores persist throughout our period given the fragility and uncertainty pervasive at the time, as well as the onset of the Euro crisis beginning in 2010.[5]

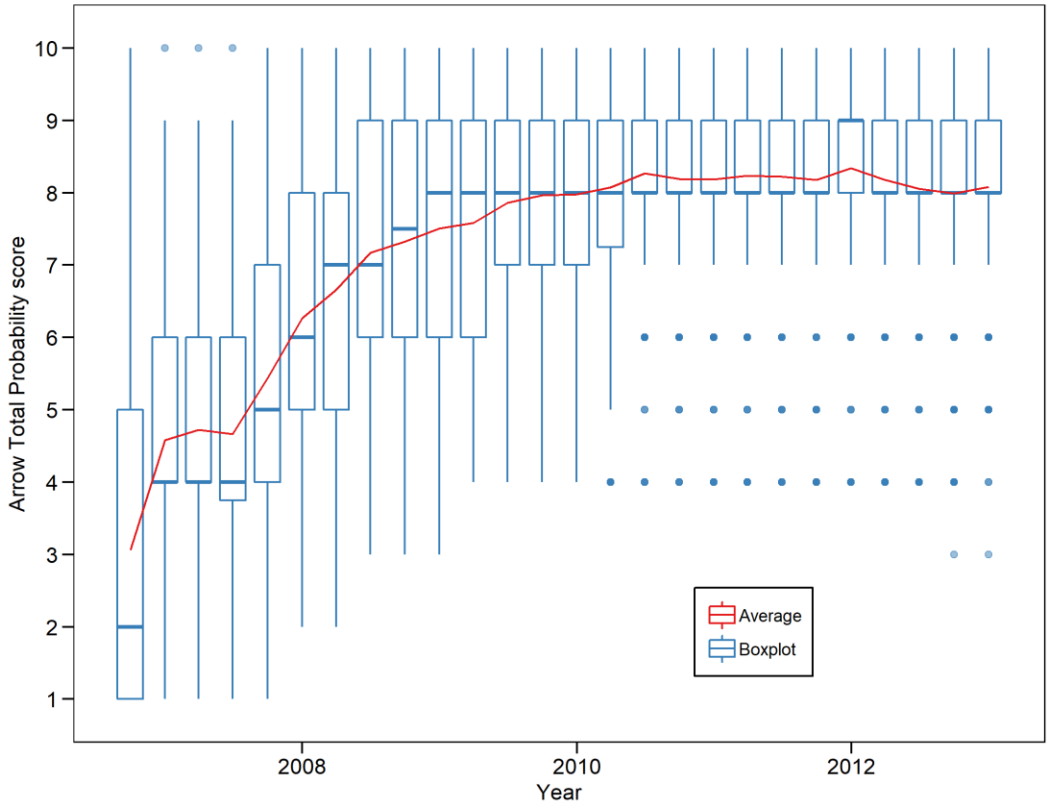
---

[3] Arrow stands for Advanced Risk Responsive Operating frameWork. The categories are: Environmental Risks; Customer Products and Markets; Business Process; Prudential; Customer Product Market Controls; Financial and Operating Controls; Prudential Risk Controls; Control Functions; Management, Governance and Culture; Excess Capital and Liquidity; Business Risks; Controls; Oversight and Governance; Operating; Financial Soundness; and Total Probability (the aggregate of all the above elements). Some of the categories are of little prudential relevance because the FSA was a conduct regulator as well as a prudential regulator, for example categories relating to risks of consumer detriment or financial crime.

[4] The FSA's Supervisory Enhancement Programme, begun in early 2008, reduced the maximum period between ARROWs for high impact banks from 3 years to 2. There would typically be an interim ARROW roughly halfway through this period.

[5] Our sample of banks contains both UK-domiciled and subsidiaries of internationally-headquartered banks. Many banks in this latter category were especially vulnerable to risks related to the Euro crisis.

Figure 1: Distribution of Arrow Total Probability scores by quarter



Note: Q32006 - Q42012. Total observations over the period equals 3,181 for 170 banks, with average distress equal to 56.7%.

Using the Arrow Total Probability score as our outcome measure, as opposed to the more typical binary indicator of bank failure, has a number of advantages:

- First, bank distress events can and do occur without there being any manifestation of failure, technical or otherwise, or negative media attention. These cases are typically uncovered by regulators and nipped in the bud without the market or public ever becoming aware (supervision, for this reason, is often a thankless job);
- Second, and related to the first, one of our aims is to provide microprudential regulators with a practical early warning system of distress, and so modelling a risk assessment framework that has actually been used by regulators will be better aligned and more comprehensible to them; and finally,
- Supervisors prefer to intervene far in advance of a firm actually failing. Predicting the probability of distress rather than failure is thus an important and useful distinction, potentially allowing regulators the opportunity to identify cases before capital ratios drop below a critical level or negative media reports surface, for example.

Moreover, modelling a binary indicator of failure is not feasible for the UK because of the lack of actual failures over the course of the last several decades.[6] Using failure as the definition of distress typically involves utilising a sample of US firms due to the number of observed failures in that country in recent decades (Cleary and Hebb 2016; Gogas, Papadimitriou, and Agrapetidou 2018), or by including observations from multiple countries, for example, bringing together outright failures from across Europe (Betz et al. 2014), or a number of related national economies (Bongini, Claessens, and Ferri 2001; Männasoo and Mayes 2009).

Scholars going beyond outright failure typically include accounting definitions of distress, for example a firm's capital falling below a certain threshold (Cole and White 2012). Swicegood and Clark (2001) partition banks into quintiles, with the lowest performing 20% (in terms of profitability) considered underperforming and given a binary indicator value of 0. Other studies use sentiment or key-word searches in news reports as alternative measures of distress, for instance Poghosyan and Čihak (2011) use news reports on firms which mentioned particular words, for example 'rescue' and 'liquidity support', to identify 79 distress events for 54 EU banks over the period 1997-2008. The authors recognise that their relatively broad definition of distress does not capture cases where a bank goes through stress while evading media attention, as might be common with smaller institutions or those that are not listed and do not have the concomitant disclosure requirements. Our data allows us to incorporate these sorts of institutions and to use a definition of distress based on front-line expert judgment.

There are, however, at least two potential issues regarding use of the Total Probability score as our response variable. First, the FSA no longer exists, having been disbanded in 2012 following the financial crisis. It is therefore relevant to ask whether these scores were competently assigned in the lead up to the crisis and accurately reflected reality. Second, and related to the first, the period from 2006 to 2012 saw changes in supervisory approach. In particular, during the beginning of our time period the FSA's approach was often described as "light touch regulation", a reflection of the prevailing views about the economic environment and appropriate stringency of regulation.[7] After the run on Northern Rock in 2007, the FSA initiated a Supervisory Enhancement Programme (SEP) which involved re-training and hiring of additional staff and the introduction of more intrusive regulation, which amounted to an admission that less intrusive regulation had been a mistake.[8]

---

[6] Augmenting a list of outright bankruptcies and liquidations with instances of distressed mergers, state bailouts and capital injections by existing shareholders and/or noteholders still does not yield a sufficient number; our attempt at compiling such a list for UK banks and building societies for which we have regulatory returns data yielded only 21 cases since 2001.

[7] *The Turner Review* – the FSA's response to financial crisis – described the "light touch" tag as a caricature and pointed out that the approach was based on a number of assumptions which made a more intrusive approach to supervision less likely. Key examples were faith in markets to correct themselves and in firms' management to follow viable business models and maintain robust controls over their risks (FSA 2009).

[8] The SEP was initiated in March 2008 after the FSA Internal Audit report into regulation of Northern Rock. *The Turner Review* described the FSA's new approach as "more intrusive, more systemic", and the programme was considered 90% complete by mid-2009 according to the FSA. In the Annex, we re-run the analysis,

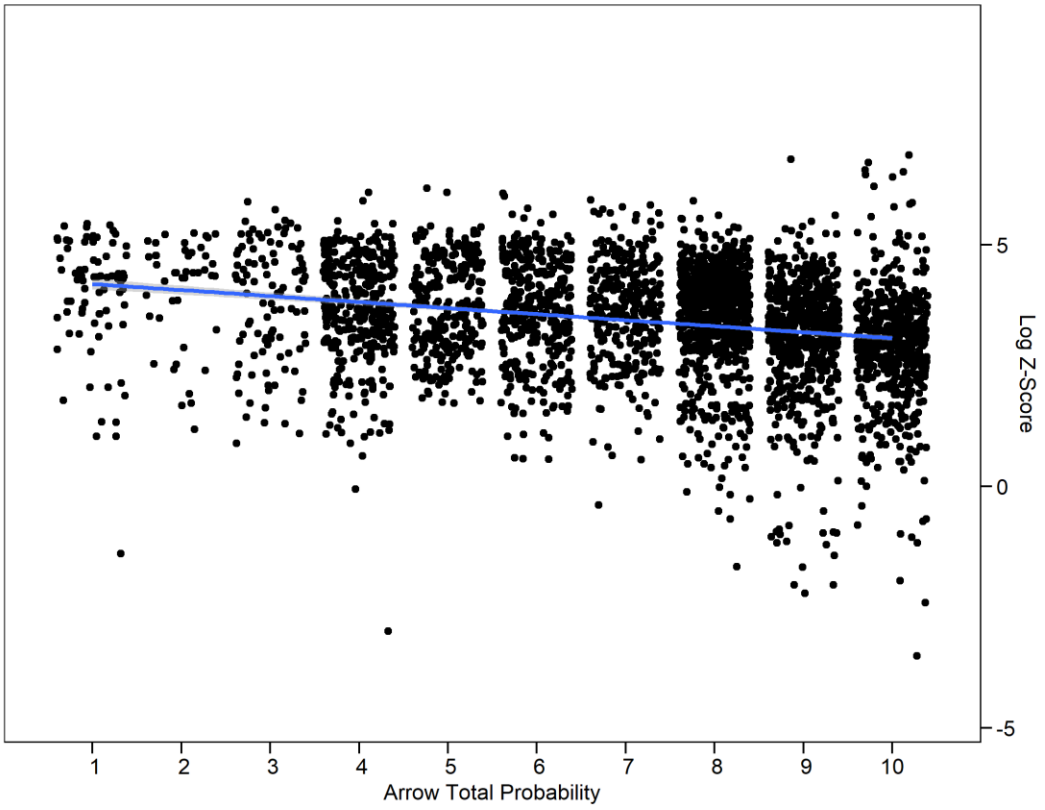
In order to address these concerns, we reviewed a number of reports into supervision following large UK failures (RBS, Northern Rock and HBOS), as well as *The Turner Review* which details lessons learnt and regulatory reforms. These reports, while critical of regulatory deficiencies in hindsight – in particular, insufficient focus on liquidity and capital – generally agree that FSA supervisors competently identified key risks. The exception to this is Northern Rock, where an FSA Internal Audit report found that there was a failure in identifying and following-up on key risks (Division 2008). However, a case-by-case examination of the Arrow scores reveals an acceptable categorisation: Northern Rock had been categorised as in distress at the time of the run in Q3 2007. We replicated this analysis for the other firms that failed, RBS and HBOS, finding appropriate categorisation as well. In essence, therefore, a case-by-case examination of the data validates the measure as reflecting realised distress.

We also quantitatively assess the validity of the Arrow Total Probability score by correlating it with the Z-score. The Z-score is a commonly used measure of distance to default, measuring the number of standard deviations asset returns have to decline to offset a bank's equity capital ratio. Empirical work finds there to be a clear relationship between firm distress and the Z-score; for instance, Chiaramonte et al. (2016) examine the relationship between Z-scores and US bank failures between 2004-2012, finding that Z-scores together with time-fixed effects are able to predict failures with a 76% accuracy. A lower Z-score implies a higher probability of insolvency, and so we expect there to be a negative correlation between the Arrow Total Probability score and the log of the Z-score. The below scatter plot shows that there is indeed a negative and significant relationship between the two variables, with a correlation coefficient of -0.24 ($p < 0.01$).

---

restricting our sample to observations from mid-2009 onwards in order to mitigate the possibility that pre-SEP supervisory scores were not of similar quality.

Figure 2: Relationship Between Z-Score and Arrow Total Probability



Note: The figure provides the correlation for the period 2006-2012. The Z-score is calculated as: $Z_{it} = (r_{it}^A + k_{it})/\sigma_{it\tau}^A$*, where* r_{it}^A *is the overall asset return for bank i at time t,* k_{it} *is the total capital ratio and* $\sigma_{it\tau}^A$ *is the standard deviation of asset returns calculated over* τ *periods. A 16 quarter window was used to calculate the Z-score, see de-Ramon, Francis, and Straughan (2018) for more detail on this measure. Total Probability is a discrete measure of risk. This scatter plot allows some random movement to display more clearly the number of firms in each category.*

2.2 Predictors

We use both firm-level and macroeconomic data as predictors. Our source of firm-level explanatory data comes from the Historical Banking Regulatory Database (HBRD). This dataset brings together information from regulatory returns going back to 1989. Much of this information is privileged, providing a more detailed, more frequent and more comprehensive picture of firm finances than is available from more conventional sources.[9]

[9] HBRD brings together firm data from a wide range of different regulatory reporting systems which had been in operation between 1989 and 2013. It amalgamated financial data submitted to different regulators which supervised banks and building societies over that period; namely, the Bank of England (including the PRA), the Building Societies Commission, and the FSA. In terms of coverage by time period, regulator and type of institution, it is the most comprehensive database so far made available on UK-incorporated deposit-takers. It

Utilising HBRD data, we computed quarterly financial ratios for the firms in our sample which could be categorised into the CAMELS schema, standing for Capital, Asset quality, Management, Earnings, Liquidity, and Sensitivity to market risk.[10] We also include balance sheet item growth rates and a measure of firm size (log of total assets).[11] We supplemented the firm-level data with macroeconomic variables (UK-focused year on year changes).[12] All predictors are lagged by four quarters in order to provide a probability of distress one year out.

Altogether, we started with 55 predictors, each chosen following a review of previous literature on the determinants of bank distress or based on the domain knowledge of subject matter experts at the Bank of England.[13] This set was reduced to a final count of 31 variables following removal in light of a Variance Inflation Factor (VIF) procedure to identify multicollinearity among the predictors, or because of poor coverage during the time period examined. The VIF statistic is computed for each variable, for example:

$$VIF_1 = \frac{1}{1 - R_1^2}$$

For $X_1 = \alpha_0 + \beta_2 X_2 + \ldots + \beta_k X_k$

Essentially, the square root of the VIF statistic indicates how much larger the coefficient standard error is compared with what it would be if that variable were uncorrelated with the other model predictors. The VIF procedure employed was as follows: the pooled logistic model was fit with all 55 predictors and VIF statistics were computed for each predictor. The variable with largest VIF subject to that value being greater than or equal to 10 was removed from the model (10 is a commonly used rule of thumb, see for example James et al. (2013) and Kutner et al. (2005)). This was repeated until no variable remained with a VIF statistic of greater than or equal to 10. Table 1 provides summary statistics for each remaining variable following the VIF procedure.

---

covers a wide range of variables, such as profitability, balance sheet size and composition, regulatory capital (including information on capital requirements), asset quality, and liquidity. For more details on HBRD, see de-Ramon, Francis, and Milonas (2018).

[10] There is a vast literature examining the determinants of bank distress using financial ratios based on the CAMELS typology. See, for example, Lane, Looney, and Wansley (1986); Bongini, Claessens, and Ferri (2001); Cole and Gunther (1995); Whalen, Thomson, and others (1988); Wheelock and Wilson (2000); Coen, Francis, and Rostom (2017).

[11] A number of studies look at market price-based indicators as deteriminants of distress with mixed results (Flannery 1998; Bongini, Claessens, and Ferri 2001; Curry, Elmer, and Fissel 2003; Čihák 2007). However, our sample includes many non-listed firms or firms that do not have any market-traded instruments, so we do not include any market-based variables.

[12] Arena (2008); Betz et al. (2014); Tinoco and Wilson (2013); Mare (2015) amongst other studies show the importance of macroeconomic variables.

[13] The selection of finacial ratios on the basis of domain knowledge has been shown to improve classifier performance relative to raw accounting variables, for example (Zhao, Sinha, and Ge 2009).

Table 1: Summary statistics for predictors

	N	Mean	Std Dev	Median	Min	Max	Skew
Asset growth	3717	8.45	16.06	7.20	-23.25	38.75	0.21
Average risk weight	3716	49.29	21.72	45.30	1.45	100.25	0.63
Capital buffer	3694	4.56	5.42	2.20	-8.85	18.75	1.69
Core deposit ratio	3711	59.85	32.72	74.33	0.00	97.27	-0.68
Deposit growth	3688	8.23	18.21	6.60	-24.85	41.15	0.19
Earning assets to total	3418	97.48	2.95	98.86	90.12	102.19	-1.43
Efficiency ratio	3624	67.64	24.90	67.40	7.85	120.25	0.00
Interest expense	3373	2.90	1.67	2.98	-1.48	8.80	0.41
Broad liquidity ratio	3685	10.65	11.23	6.67	0.00	38.95	1.12
Narrow liquidity ratio	3527	3.60	4.73	1.28	0.00	16.82	1.49
Loan growth	3594	9.02	20.40	7.30	-28.85	45.95	0.14
Loans to deposits	3678	63.15	38.31	72.40	0.00	168.26	0.22
Loans to retail deposits	3580	85.90	52.88	81.73	0.00	191.89	0.54
Net interest margin	3373	1.53	1.01	1.33	-1.08	4.77	1.42
Non-interest income	3660	1.04	1.31	0.50	-2.25	4.16	1.20
Provisions to loans	3592	0.76	1.41	0.20	-3.20	5.60	2.51
Pre-tax net income	3660	0.51	0.85	0.42	-1.13	2.47	0.47
Retained profit	3703	2.47	5.60	1.60	-8.25	13.75	0.20
ROE	3646	4.62	7.01	4.29	-9.43	20.16	0.08
Size	3717	7.01	2.30	6.60	2.00	14.50	0.94
Solvency	3694	177.41	79.27	143.88	0.40	365.85	1.41
T1 capital	3716	20.17	13.14	15.40	0.00	52.75	1.45
T1 growth	3704	5.83	11.12	4.30	-16.05	26.75	0.25
Trading book assets	3717	6.40	20.10	0.00	0.00	99.67	3.52
Trading income to NOI	3565	1.25	2.42	0.09	-2.88	4.79	0.29
Unsecured assets	3683	19.22	22.17	10.40	-31.40	86.60	1.21

Note: Outliers are considered to be observations 1.5 times the interquartile range below the 25th or above the 75th percentile and are capped at the limit values.

For capital, we include four variables: the tier 1 capital ratio, retained profit to equity, the buffer of capital a firm holds (difference between its total regulatory capital held and capital requirements divided by total assets), and the solvency ratio (defined as total regulatory capital divided by a firm's capital requirements). These latter two ratios are, owing to the capital requirements element, confidential and unique to our dataset.

For asset quality, we include provisions for non-performing loans divided by loans, and average risk weight. To proxy for the quality of a firm's management, we use the efficiency ratio, defined as the ratio of total overhead costs to the sum of net-interest and other non-interest income, and the proportion of unsecured assets to total assets. We include six variables for earnings: net interest margin, non-interest income over total assets, pre-tax net income as a proportion of total assets, return on equity, the proportion of earning

assets over total assets, and the ratio of interest expense to earning assets. For liquidity, we use the ratio of loans to assets, loans to retail deposits, broad and narrow liquidity ratios[14], and the core deposit ratio (defined as all deposits excluding financial institutions over total deposits). We use the proportion of net operating income which is derived from trading income and the proportion of total assets which are in the trading book as proxies for sensitivity to market risk.

Table 1 also provides descriptive statistics for other variables outside the CAMELS schema which have been shown to be important predictors in previous studies, in particular asset growth, loan growth, deposit growth, and T1 capital growth (Fahlenbrach, Prilmeier, and Stulz 2017). Finally, for macroeconomic variables we include UK GDP, unemployment, inflation, real average earnings, and FTSE all share index, all as year-on-year changes. Definitions and sources for the macroeconomic variables are included in the Annex.

For modelling purposes, we only include observations which are not missing any predictor value, leaving us with a total of N = 3,181 observations over 26 quarters, with an average of 122 observations per quarter, standard deviation of 9, and a minimum and maximum of 97 and 137 respectively.

Section 3: Methodology

In predicting bank distress events, scholars have typically employed classical statistical models, in particular logistic regression (Martin 1977; Betz et al. 2014; Coen, Francis, and Rostom 2017; Tinoco and Wilson 2013; Cole and White 2012; DeYoung and Torna 2013; Oet et al. 2013), as well as Cox proportional hazards models (Lane, Looney, and Wansley 1986; Whalen 1991; Wheelock and Wilson 2000; Shumway 2001; Gomez-Gonzalez and Kiefer 2009). Hazards models predict the timing of failure events rather than the probability and, like logistic regression, assume a linear functional form. In the US, microprudential regulatory bodies have been known to utilise early warning systems based on linear models, in particular the Federal Deposit Insurance Corporation (Collier et al. 2003) and the Federal Reserve through its System to Estimate Examination Ratings (Jagtiani et al. 2003).

An important shortcoming of these approaches is that the relationships in question might be highly non-linear and complex, and so assuming a linear functional form is likely to lead to an underperforming model. Indeed, as we show in section 4.6, complex interactions are important. In contrast, machine learning techniques are typically more flexible than linear approaches, automatically allowing for inherent non-linearities and thereby improving prediction accuracy in many applications (James et al. 2013).

The application of machine learning techniques to predicting bank distress is in its infancy, however. Recent years have seen important contributions, primarily in the operational research and systems application literature, that demonstrate the superiority of non-linear

[14] The narrow liquidity ratio refers to high quality liquid assets over total assets, while broad liquidity also includes credit to other financial institutions, debt securities and equity shares.

methodological approaches for predicting bank failure (Boyacioglu, Kara, and Baykan 2009; Iturriaga and Sanz 2015; Le and Viviani 2018; Gogas, Papadimitriou, and Agrapetidou 2018; Carmona, Climent, and Momparler 2018; Bell 1997; Swicegood and Clark 2001).[15] We contribute to this nascent body of work, comparing the performance of six different approaches: two within the linear family (pooled binary logistic regression and random effects logistic regression) and four that fall under the machine learning banner – k nearest neighbours (KNN), random forest, boosting, and support vector machines (SVM).

In what follows, we first provide a brief explanation of each technique, before going on to detail our approach for estimating and comparing performance.

3.1: Linear statistical models

The pooled binary logistic model is provided by:

$$log(\frac{Pr(y_i = HighRisk)}{1 - Pr(y_i = HighRisk)}) = \alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Where α is the model intercept, each x_i is a year-lagged predictor, and β represents the fixed model parameters determined through maximum likelihood estimation, reflecting the partial association between the lagged predictors and the log odds of distress. Backing out of the logistic formulation, the predicted probability of distress for each individual observation is computed by:

$$Pr(y_i = HighRisk) = \frac{exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

We call the above model 'pooled' because it does not account for the clustered nature of the dataset. In our case, the same firm is observed in more than one quarter, and firms are clustered by quarter. In such situations, there is typically a dependency between observations within firm and quarter, violating the assumption of independence that lies behind the linear model and leading to biased parameter estimates. In terms of predictive accuracy, accounting for the hierarchical nature of clustered data through a linear random effects model has been shown to improve upon the pooled model in other contexts (Afshartous and Leeuw 2005; Bouwmeester et al. 2013). The random effects binary logistic model is given as the following:

$$log(\frac{Pr(y_i = HighRisk)}{1 - Pr(y_i = HighRisk)}) = \alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \gamma_i + \tau_t$$

[15] Other studies examine machine learning techniques in predicting bankruptcy of corporations across a number of sectors, see P. R. Kumar and Ravi (2007), Barboza, Kimura, and Altman (2017) and Alaka et al. (2018) for reviews.

where γ_i is the time-invariant random firm effect and τ_i is the firm-invariant random time effect.[16] The intraclass correlation coefficient (ICC) measures the strength of intra-firm or intra-quarter correlation and is defined as (for intra-firm):

$$\frac{\sigma_\gamma^2}{(\sigma_\gamma^2 + \sigma_\epsilon^2)}$$

In our sample of data, the firm ICC is equal to 0.318 and the quarter ICC is equal to 0.295, suggesting substantive correlations within clusters that need to be accounted for in order to avoid biased parameter estimates. In some sense, because the random effects model accounts for the intra-group correlations, the parameter estimates are closer to the 'true' estimates of the relationship between our explanatory variables and the outcome measure relative to the pooled model. However, a problem arises when utilising this model to predict distress for out-of-sample firms (i.e. firms or quarters that are not part of our training dataset). Because they are not included in-sample, there is no random effect estimate that can be plugged in to the model equation. As such, we assume the random effect term is equal 0 for out-of-sample firms and quarters, an assumption that is likely to have negative implications for prediction accuracy. It is because of the need to predict out-of-sample that academics have tended towards the pooled logistic model in early warning system applications (see, for example, Coen, Francis, and Rostom (2017)).[17]

Regardless of model chosen, whether linear or under the machine learning umbrella, the presence of substantive intra-firm and intra-quarter correlation is also important in the context of estimating out-of-sample performance. We account for this by double block randomising between training and test sets by firm and quarter (for further details of our cross-validation procedure, please see Section 3.6).

3.2: KNN

A k nearest neighbours classifier computes the distance[18] between a target observation and the *k* nearest sample data observations, where *k* is termed a *hyperparameter* which is chosen through cross-validation. The test observation is then classified into a category based on the actual *y* values of the *k* nearest observations, utilising some threshold voting rule.[19] Thus, for example, if we set the threshold voting rule to be greater than or equal to 50% and the number of neighbours is set to three, we classify our target observation as high risk if at least two of the three nearest neighbours are in the distress category. The KNN 'votes' in this example equals 66.7%, i.e. the proportion of neighbouring observations

---

[16] We assume that the random effects follow a normal distribution, although mild to moderate violations of this assumption have been shown to have only a small effect on prediction accuracy (McCulloch and Neuhaus 2011).

[17] Finkelman, French, and Kimmel (2016) and Ni et al. (2018) provide alternative approaches to overcome this limitation with dynamic random effects models, incorporating information on clusters from previous model fits for future predictions.

[18] Typically calculated as the Euclidean distance, although other distances are sometimes used.

[19] We implement the KNN algorithm using the `class` package in R (Venables and Ripley 2002) and scale each of our predictors.

that are in the distress category. This number is taken to be the probability of distress for our target observation but does not equate with this meaning in the same way as for the logistic regression models.[20]

Relative to linear models, KNN is a very flexible approach that is inherently able to account for complex relationships between variables. On the downside, however, KNN is completely opaque – there are no estimated parameters which enables us to interrogate how specific variables relate to our outcome of interest.

3.3: Random forest

The random forest algorithm is a widely utilised machine learning technique which excels at a wide variety of prediction problems (Breiman 2001; Fernandez-Delgado et al. 2014). It involves the bootstrap aggregation (or bagging) of individual decision trees, with a random number of predictors chosen at each tree node to de-correlate the trees.[21] In other words, random forests are 'ensembles' of individual trees. As we will discuss in more depth in Section 4.8, bringing together diverse models which are also good at predicting our outcome of interest will likely improve out-of-sample performance. For greater detail on decision trees, please refer to the Annex.

As with KNN, random forests provide us with what we will term a 'probability of distress' that is not equivalent to the fitted probability of a logistic regression. In a random forest, each tree classifies individual observations based on a set of decision rules. The predicted probability of distress is taken as the proportion of trees in the forest that classify a given observation in the high risk category.[22] In terms of opacity, random forests are an intermediate step between linear models and KNN. While individual classification trees are very transparent – each decision rule is explicit – the aggregation of trees reduces our ability to determine drivers.

3.4: Boosting

Boosting is another ensemble approach which works well with decision trees. Instead of aggregating a number of different trees as with random forests, boosting involves growing trees sequentially; new trees are fit with the previous tree's residuals as the outcome measure rather than y_i.[23] In this way, new trees improve on previous ones.

As with random forests, the boosting ensemble output is a vote on which state a firm is in based on the classification of each tree. Unlike random forests, the number of trees used is a hyperparameter that needs to be selected through cross-validation (too many trees could lead to overfitting). There are two other hyperparameters to choose in the context of

---

[20] We examine whether the KNN predicted probabilities are well calibrated in Section 4.2.
[21] We utilise the `randomForest` package in R (Liaw and Wiener 2002).
[22] Previous research has found the probability outputs of random forests to be well calibrated (Niculescu-Mizil and Caruana 2005). We examine whether this is also true in our case in Section 4.2.
[23] We utilise the `gbm` package in R to implement the gradient boosting algorithm (Friedman 2001); other decision tree boosting algorithms, such as adapative (Freund and Schapire 1997), implement different techniques for learning from previous decision trees and weighting trees for probability votes.

boosting: the depth of each tree (i.e. the number of splits in each tree), and a shrinkage parameter (also known as the learning rate) which specifies the contribution of each tree to the outcome. In terms of transparency, boosting approaches are in an intermediate place alongside random forests.

3.5: SVM

Support vector machines are generalisations of the maximal margin classifier to accommodate non-linearities. The maximal margin classifier is a technique for separating two classes with a hyperplane (when there are more than two predictors). Rather than add polynomial or interaction terms to the set of predictors, SVMs uses kernel functions to allow for non-linearities (kernels are an efficient computational means of accomplishing this, the technical details of which can be found in James et al. (2013)). We utilise the radial basis function which has been shown to outperform other kernel functions in previous work on bankruptcy prediction (in particular, linear, sigmoid and polynomial functions; Min and Lee (2005).[24]

As with the other machine learning techniques, there are hyperparameters which need to be selected which control SVMs; in particular, cost – the amount of slack you allow the separating hyperplane in terms of observations being on the wrong side – and γ, which affects the complexity of the radial kernel function. In terms of opacity, the fact that we are using a non-linear kernel function means that we are unable to understand the relationship between individual predictors and the predicted probabilities.

The output of an SVM prediction for any given test observation is in the range of [-1,1]. By utilising the Platt scaling technique (Platt 1999), the decision values are translated into probabilities.

3.6: Estimating performance

In order to assess the predictive performance of each of the abovementioned classifiers, we adopt cross-validation – a technique common in the machine learning literature. This procedure splits the dataset between a training and test sample, fitting each model using only training data and evaluating performance on the excluded test sample. In other words, we evaluate the performance of each model on data that was not used to build the model. In this way, our estimates more closely approximate performance if it were actually applied in the real world, whereby the model will be asked to make a prediction for a previously unseen observation. This approach prevents optimistic performance evaluations, particularly common in the context of flexible machine learning algorithms where overfitting on training data can be a problem.

In a typical cross-validation exercise, observations are randomly split between training and test sets. However, because our dataset is clustered, a simple random split results in flattering out-of-sample estimates owing to to the strong correlation between observations within quarter and firm (Brenning and Lausen 2008; Adler et al. 2011; Roberts et al. 2017).

---

[24] We utilise the `e1071` package in R to implement SVM (Meyer et al. 2019).

This 'data leakage' problem arises, roughly speaking, because different machine learning techniques can *learn* a specific firm or quarter from the training observations and identify it successfully in the test set due to the dependence between them.

As such, we perform double-block randomisation to split observations between training and test sets.[25] The first step of the procedure is to split the full dataset into five mutually exclusive folds based on a random partitioning of quarters. Within four-fifths of this, we randomly split the data into another five mutually exclusive parts, this time randomising by firm. We use four folds of this second split to train our models, and utilise the intersection of the fifth part left out of both the quarter and firm split as our test set. We repeat this for every fold of the firm split successively, with each fifth being left out in turn. Once the firm five-fold cross-validation is complete, we then leave out a different fifth of the quarter split and repeat. Figure 3 lays out how this works visually.

Within each training sample, we employ nested five-fold cross-validation in order to choose hyperparameters (Min and Lee 2005). This laborious and computationally costly approach means that they are selected using training data only, avoiding another source of data leakage. Finally, we also repeat the entire procedure ten times to account for the variability in performance estimates that results from the specific random split performed at the outset (Bouckaert and Frank 2004).[26,27]

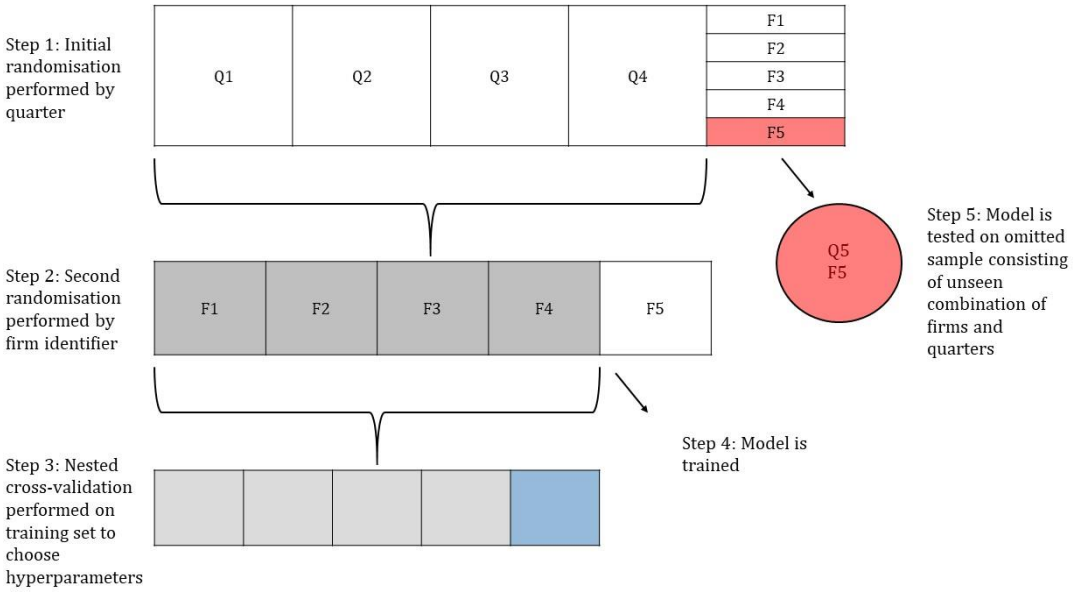
We thus have 250 performance estimates for each model on truly unseen data – on quarters and firms that are not part of the training data. Taking the average provides us with an unbiased estimate of performance (the results will be presented in Section 4). The procedure undertaken in this study is more rigorous than those in related pieces of work. For instance, many studies tend to perform only one random split between a training and test sample rather than cross-validation (Boyacioglu, Kara, and Baykan 2009; Le and Viviani 2018; Gogas, Papadimitriou, and Agrapetidou 2018; Swicegood and Clark 2001), while others perform cross-validation but fail to account for intra-group clustering (Carmona, Climent, and Momparler 2018). Other studies avoid the issues associated with intra-quarter clustering, as well as another potential pitfall known a 'look ahead bias'[28], by performing a rolling window forecast, where the test data is always future periods of time. However, these studies tend to ignore other relevant clusters and forms of data leakage. For example, Betz et al. (2014) evaluate the performance of a pooled logistic regression

---

[25] Even though the random effects model accounts for the intra-firm and intra-quarter correlation, block cross-validation by firm might still yield overly optimistic performance estimates (Roberts et al. 2017).
[26] We also scale predictors for the KNN and SVM at the point of training each model, as scaling on the overall dataset before splitting is another potential source of data leakage.
[27] Stratification is not enforced given our method of sampling, which leads to imbalances between test and training set in terms of the proportion of actual distress cases. However, we judge these imbalances to be minor. The average prevalence of distress across the 250 test samples is very close to the overall sample proportion at 0.535, while the standard deviation is 0.12.
[28] In our context, we are not looking to evaluate the performance of each model as if it were built at different points in time using only data available up to that point, rather we are seeking to make the best use of all the data we currently have to understand the relationship between the lagged predictors and distress. Nevertheless, to deal with this as a potential source of optimism due to selection effects as poor performers drop out of sample, we conduct a rolling window forecast as a robustness check in Section 5.

model for predicting bank distress using this approach, but they fail to account for other sources of dependency that might lead to over-optimistic results, such as intra-country and intra-firm correlation. Iturriaga and Sanz (2015) estimate out-of-sample error on one future hold-out period but fail to account for intra-firm dependency across time.

Figure 3: Double-block cross-validation procedure



Section 4: Results

4.1: AUC

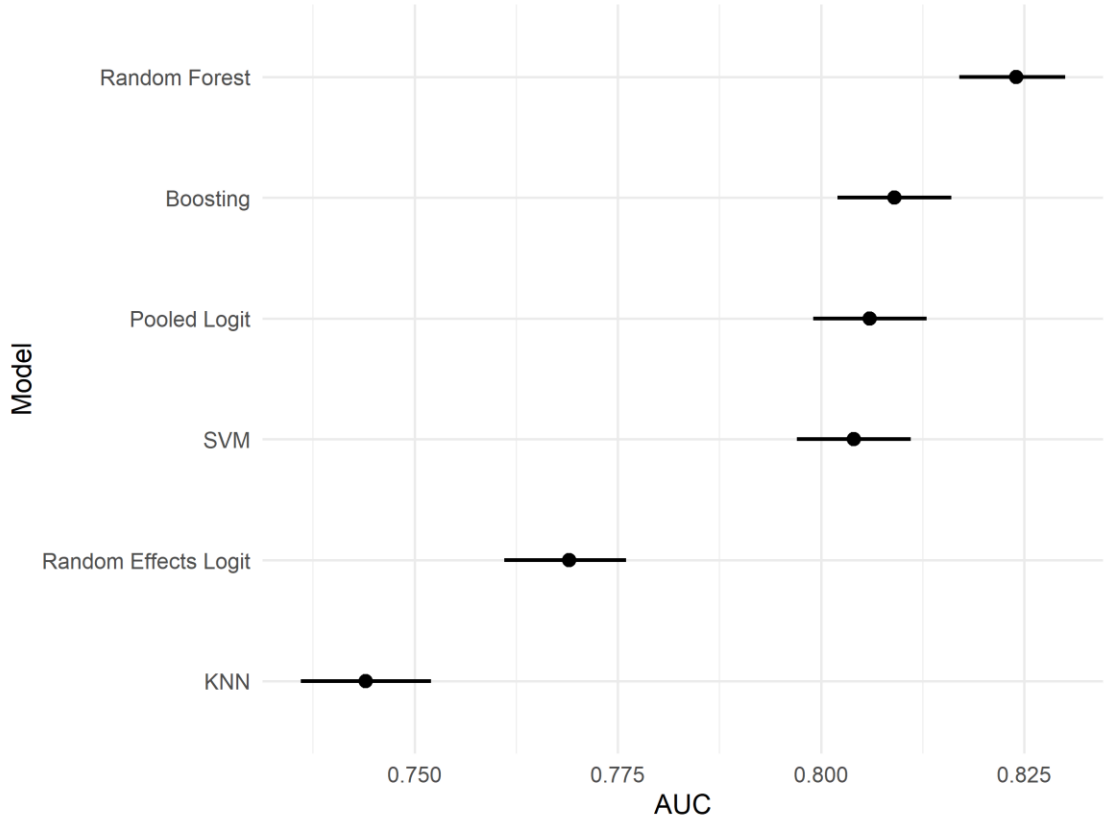
A general way of summarising performance and comparing models is by calculating the area under the ROC curve (AUC). A ROC (or Receiver Operating Characteristic) curve plots the trade-off between the True Positive (TP) and False Positive (FP) error rates at every possible classification threshold. The AUC summarises each ROC curve and is therefore threshold-invariant. Moreover, the AUC is scale-invariant – only the relative order of predictions matters. This latter property is particularly important for choosing between models when predicted probabilities of distress may not be well calibrated (Section 4.2 examines whether this is the case in our context). Finally, the AUC also represents the probability that a randomly selected distressed observation has a higher predicted probability of distress than a randomly selected non-distressed observation (see Fawcett (2006) for an overview of the AUC and its properties).

Figure 4 provides central estimates and 95% confidence interval bands on the AUCs using the DeLong et al. (1988) method.[29] As indicated by the interval estimates, the random

[29] The fact that we are resampling from the same data set through cross-validation invalidates a simple Z-test statistic calculation. To deal with this, E. DeLong, DeLong, and Clarke-Pearson (1988) derive a non-parametric

forest outperforms all the other approaches on this metric at the 95% level of confidence. The boosting, pooled logit and SVM approaches are indistinguishable from one another in a statistical sense given that their confidence interval estimates overlap.

Figure 4: AUC estimates and confidence intervals



4.2: Calibration of predicted probabilities

The AUC estimates in Figure 4 provide a threshold- and scale-invariant ranking of the different models, pointing to the random forest algorithm as the superior choice in this respect. However, in our context we are also interested in the predicted probabilities of distress. This is because the predicted probability of an early warning model has important decision-making value for regulators, for example in relation to allocating resources and taking mitigating action. We therefore need to evaluate the overall *calibration* of the fitted probabilities for each model. In other words, we examine the quality of the predicted probabilities given our knowledge of the outcome for each test observation.[30]

---

estimate of the AUC variance. The same issue arises when we compare error rates at the any given classification threshold.

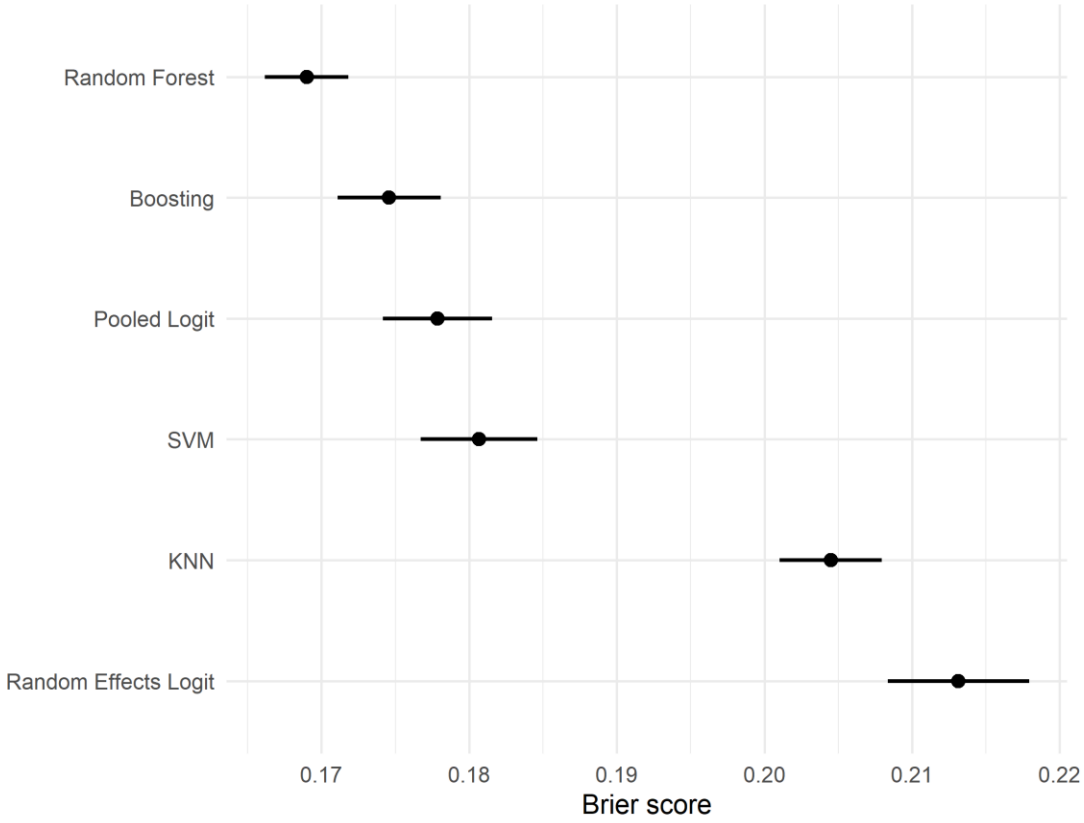
[30] As explained in Section 3, the probability of distress provided by different machine learning techniques is not equivalent to that provided by logistic regression, and there are certain known calibration issues with machine learning techniques such as boosting and SVM (Niculescu-Mizil and Caruana 2005).

To do so we use an intuitive and widely used metric to assess prediction accuracy: the Brier score, defined as:

$$B = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2$$

Where n equals the total number of predicted probabilities, p_i is the prediction for observation i and y_i is the actual outcome. The Brier score penalises predictions the further away they are from reality. A well calibrated classifier is one with a Brier score closest to zero. Figure 5 provides the estimates and 95% confidence intervals for each model. The random forest is once again the best performing, having the lowest Brier score, albeit the difference relative to next closest model (boosting) is only marginally significant ($p < 0.1$).

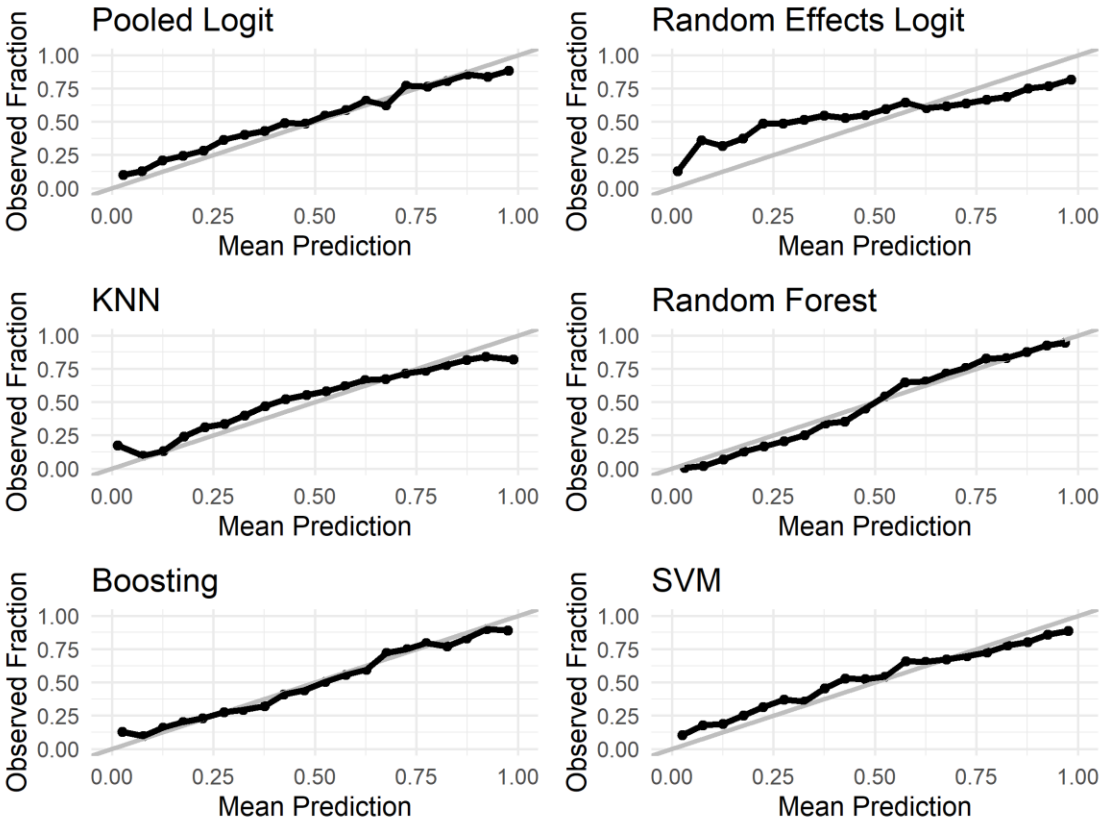
Figure 5: Brier score



We also qualitively evaluate the accuracy of the probabilities in different segments of the probability range. To do so we construct reliability plots, binning each predicted probability in one of 20 bins by intervals of 5 percentage points. For each bin, we calculate the actual proportion of distressed banks. Plotting the predicted versus actual proportion of distressed cases gives us a sense of how reliable the model outputs are per bin, with wide divergence from the diagonal line indicative of a poorly calibrated model and vice versa. Figure 6 provides the plots for each model. The random forest, boosting, pooled logit and SVM models all demonstrate well calibrated probabilities at each, whereas random

effects logit model departs from the diagonal line in substantive ways for most of the range and the KNN algorithm appears relatively poorly calibrated at the extremes.

Figure 6: Reliability plots



4.3: False negative and false positive error rates

We now turn to two different performance metrics that are relevant to decision-makers of an early warning system: the false negative (FN) and false positive (FP) error rates. The FN rate – the proportion of actual high risk firms that are predicted to be low risk – is likely to be the more important out of the two from a regulatory perspective. An early warning system that fails to set the alarm when it should, particularly for large, systemically important institutions, can have deleterious consequences. Of course, in order to minimise the FN rate we can just lower the predicted probability threshold by which we classify low and high risk firms. For example, we could set the threshold to 5% which means that each model would classify almost all test cases as high risk firms, including many that turn out to be low risk. There would thus be an elevated FP rate, which is likely to be unacceptable. Table 2 lays out the different possibilities. The trade-off between FNs and FPs is a subject we return to in Section 4.4 when we specify different relative costs between these two types of errors. The FN rate is calculated as $\frac{FN}{FN+TP}$ and the FP rate as $\frac{FP}{FP+TN}$.

Figure 7 provides the FNR and FPR for the top four models as we vary the decision threshold (we drop KNN and the random effects logit due to relative underperformance). As these curves demonstrate, none of the techniques dominate throughout the decision

space. The random forest does better on the FNR and FPR relative to the others at low and high thresholds respectively, while performing the worst in the opposite direction. The boosting ensemble and pooled logit tend to be in between the others for most thresholds on both metrics, while SVM does the worst on the FNR and the best on FPR at low thresholds and the reverse at high thresholds.

So how do we choose based on these metrics? If we assume a regulatory preference for a low FNR, a relatively low decision threshold should be chosen. For example, a regulator might consider a FNR above 20% to be intolerable regardless of the FPR. This would require a threshold approximately at or below 50%. The two first columns of Table 3 provide the performance estimates for each model when the classification threshold is set at 50%. Looking at the FNR, the random forest algorithm outperforms all the other approaches with an error rate of 16.7%. The other decision tree ensemble is close behind at 17.2%, followed by the pooled logit model which has a FN rate of 20%. In terms of the FPR, the SVM algorithm is the top performer (30.5%) followed by the pooled logit model (33.2%) and the random forest at (33.6%).

In order to do better in terms of the FN error rate, we can lower the threshold from 50% to 25%. The random forest once again leads the way, only misclassifying 2.2% of the actual high risk cases at this threshold, followed by the SVM (8.8%) and the boosting technique 6.1%. At this threshold, the SVM has the lowest FPR (52.1%).

Table 2: Model prediction possibilities

		Predicted	
		True	False
Actual	True	True positive (TP)	False negative (FN)
	False	False positive (FP)	True negative (TN)

Figure 7: FNR and FPR rates

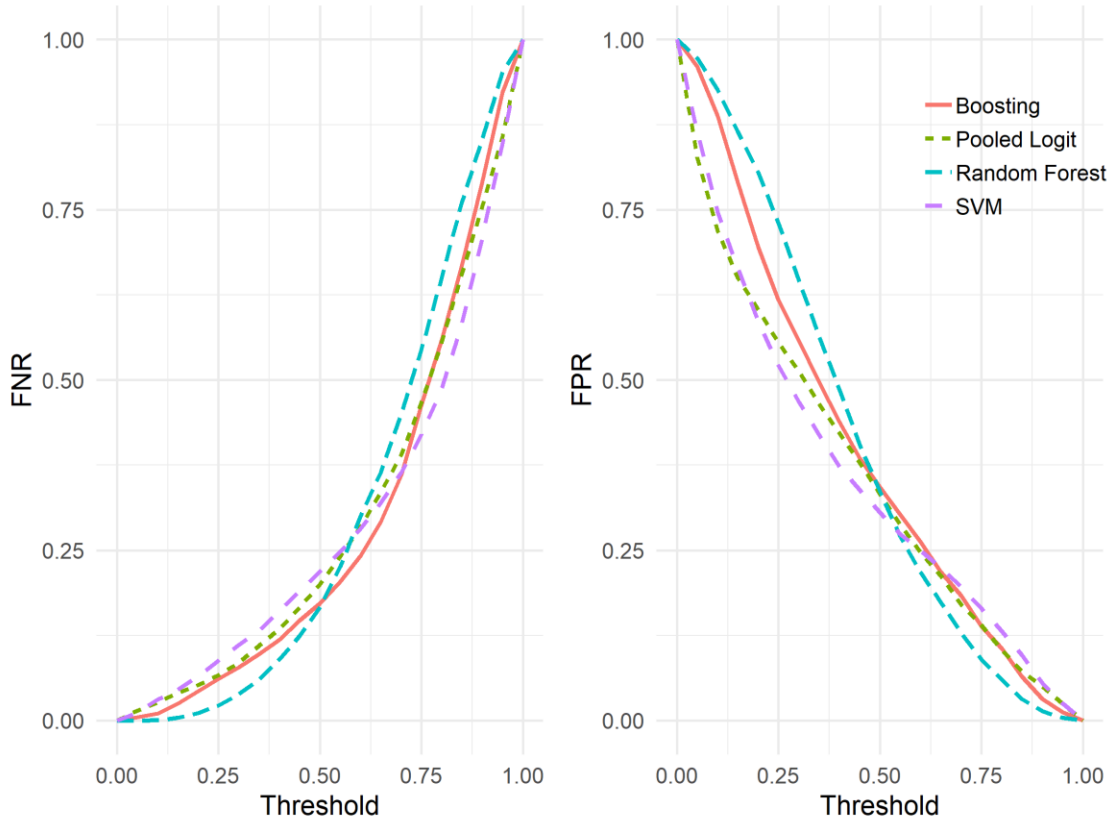


Table 3: Performance on FNR and FPR at select thresholds

	50%		25%	
	FNR	FPR	FNR	FPR
Pooled Logit	0.200	0.332	0.067	0.555
Random Forest	0.167	0.336	0.022	0.730
Boosting	0.172	0.343	0.061	0.618
SVM	0.219	0.305	0.088	0.521

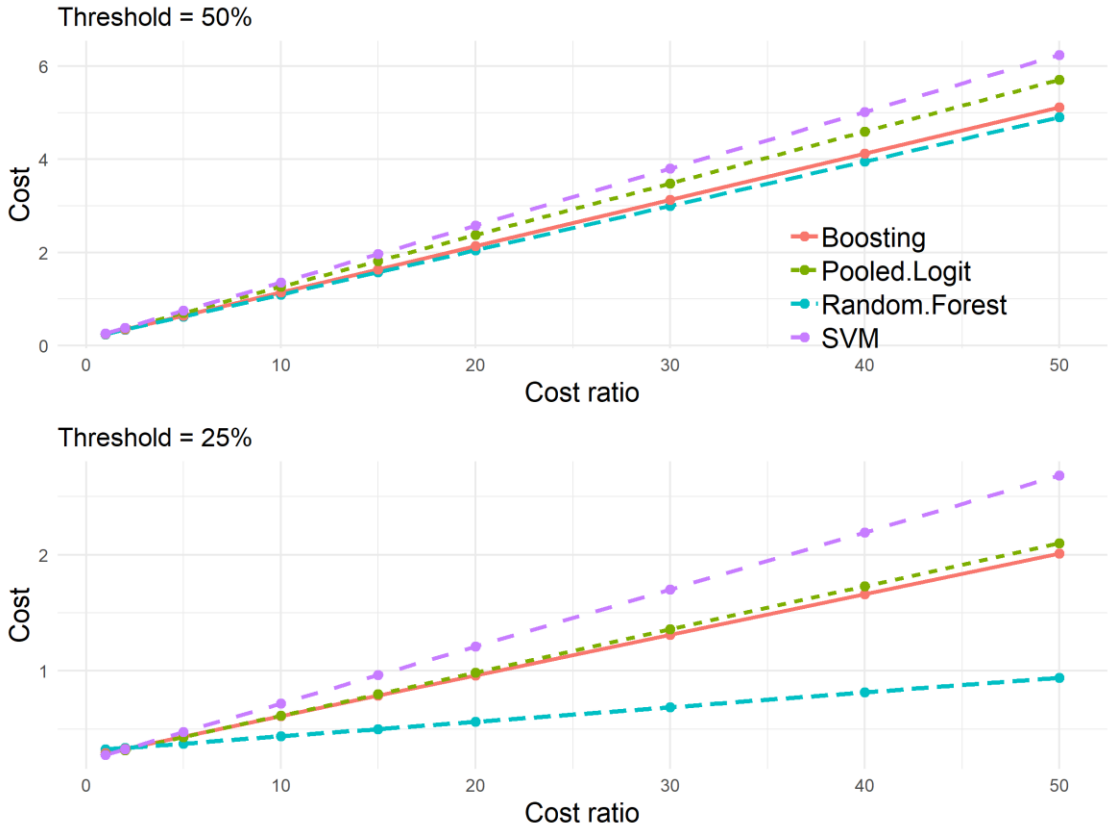
4.4: Relative misclassification cost

Given a preference for avoiding FN relative to FP errors, we now examine scenarios where the relative cost of the two types of errors is weighted to such that the former are more costly. Following Swicegood and Clark (2001), we calculate the relative misclassification cost as:

$$RMC = \alpha_i(p_2c_2) + (1 - \alpha_i)(p_1c_1)$$

where α_i is equal to the predicted probability of being in distress for firm i, p_1 is the probability of type 1 (FP) error, p_2 for Type 2 (FN) error, and c is the respective cost of each type of error. Figure 8 looks at the relative cost between the tree ensembles – random forest and boosting – the pooled logit model and SVM at a threshold of 50% when we increase the cost ratio (C2:C1) from 1:1 to 2:1, 5:1, 10:1, 15:1, 20:1, 30:1, 40:1 and 50:1. It demonstrates that the models have similar misclassification costs when we weigh FP and FN errors the same – the random forest and boosting cost is only slightly lower than the pooled logit model – but as we increase the relative cost of FN errors we see that the random forest misclassification cost increases at a slower rate to the other approaches. This slower increase in misclassification cost as the cost ratio increases is even more pronounced when the classification threshold is reduced from 50% to 25%. This suggests that the random forest approach is superior as we increase the relative importance of reducing FN errors.

Figure 8: Relative misclassification cost at different decision thresholds



Note: Misclassification cost as we increase the relative cost between FN and FP

4.5: Shapley values

Machine learning techniques such as the random forest algorithm improve prediction performance relative to linear models in many applications, including the one presented in this paper. The downside, however, is that transparency is sacrificed to some extent, being relatively less able to understand the factors driving the output. Having established the

superiority of the random forest with respect to key performance metrics, in this section we examine the underlying drivers of the model compared with the benchmark pooled logit model using a recently developed technique for interpreting machine learning predictions: Shapley values (Strumbelj and Kononenko 2014; Lundberg and Lee 2017; Joseph 2019).

Shapley values are computed per test set observation, providing the average marginal contribution of each predictor value to the difference between the actual prediction and the mean prediction of the associated training set.[31] The technique, which has its origins in cooperative game theory (Shapley 1953), has a number of properties which distinguish it from other interpretation techniques; in particular, it combines efficiency, missingness, symmetry, strong monotonicity and additivity (for technical detail on Shapley values and their properties, please see Joseph (2019); for an intuitive explanation and example, see Molnar (2019)).

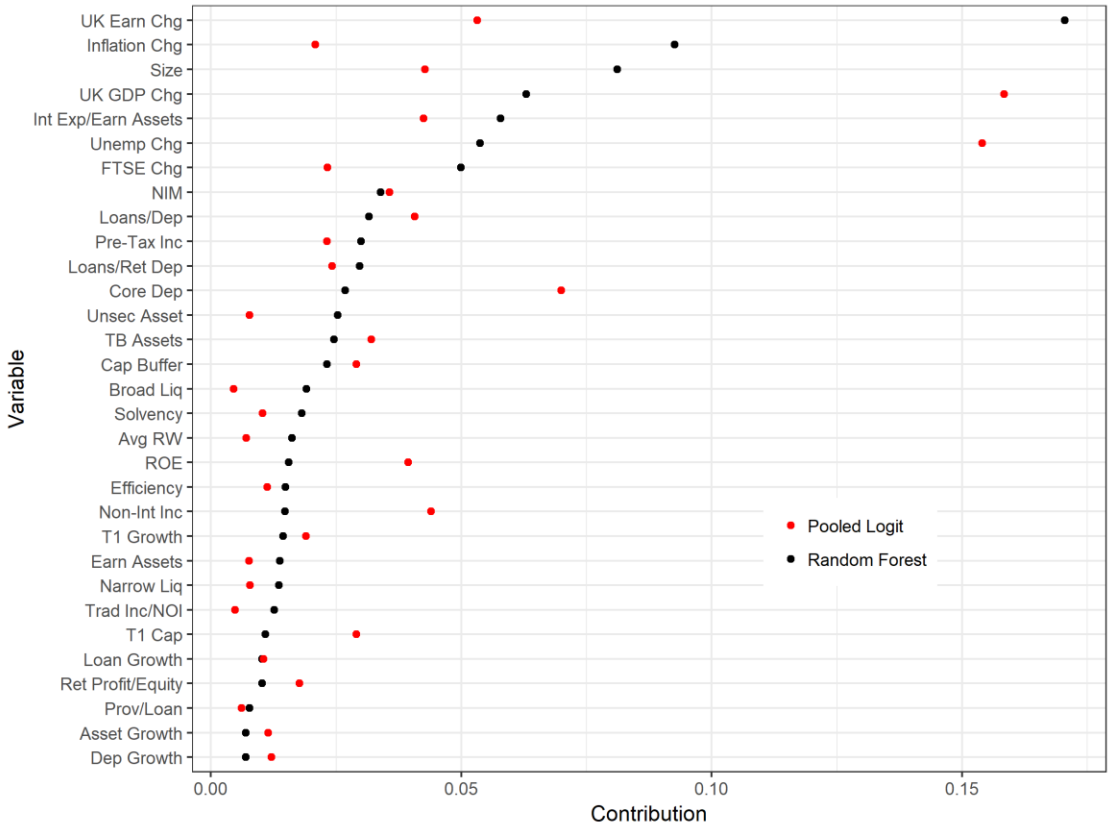
In order to provide a global interpretation, we follow Bluwstein et al. (2019) and compute the average absolute Shapley values per predictor for all test observations. In order to more easily compare the values across predictors, Figure 9 displays the normalised values so that they sum up to 1. The results are ordered by importance for the random forest (black), with the pooled logit (red) model included for comparison.

Figure 9 shows that there is substantial disagreement between the random forest and logit model in terms of Shapley values. In particular, the measure of UK average real earnings is the most important predictor for the random forest, but far less so for the pooled logit model. The disagreement between the two approaches is due to the inherent ability of the random forest to capture interactions between predictors, a subject we investigate in detail in Figure 10. In general, and unsurprisingly given our sample period, macroeconomic variables are important for both models, amounting to five out of the top seven variables for the random forest. In terms of firm-level financial ratios, a firm's interest expense to earning assets ratio, NIM, and loans to deposit are the top three for the random forest. For the pooled logit, core deposit ratio, non-interest income to assets, and interest expense to earning assets are the top three.

---

[31] This feature value contribution, or Shapley value, is calculated by examining all possible coalitions of feature values. In mathematical notation, the Shapley value for a given observation for predictor k is: $\phi_k = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_k\}} \frac{|S|!(p-|S|-1)!}{p!} (f(S \cup \{x_k\}) - f(S))$, where p is the total number of predictors, S is a subset of the predictors and $(f(S \cup \{x_k\}) - f(S))$ is how much the value of variable k adds to predictive value of the coalition of predictors S. Note that the computational cost grows exponentionally with every additional feature, making an approximation calculation necessary in our case, estimated by Monte-Carlo sampling (Strumbelj and Kononenko 2014). We utilise the R package `iml` to carry this out.

Figure 9: Mean absolute Shapley values for random forest and pooled logit



4.6: Interactions

What explains the discrepancy between the random forest and pooled logit Shapley values? Figure 11 provides an indication of the interaction strength for each variable using the H-statistic (Friedman and Popescu 2008), defined as the share of total variance explained by a given predictor's interaction with all of the other predictors in the model.[32,33] Figure 11 indicates that the average earnings variable has the largest overall interaction strength. This result highlights the importance of interactions in the data, demonstrating why the random forest outperforms the pooled logit model - the former automatically incorporates the non-linearities that are ignored by the linear model.[34]

---

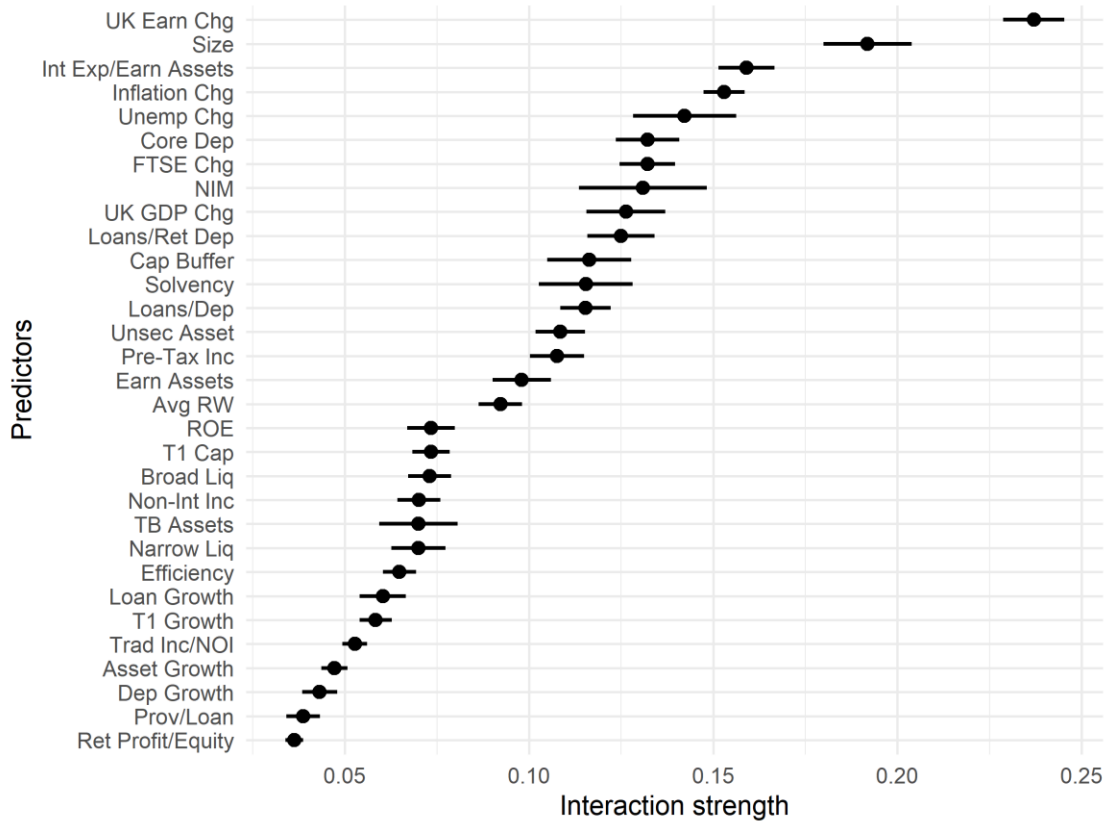
[32] We opt to compute the H-statistic rather than Shapley interactions because the latter are restricted to pairwise interactions whereas the H-statistic provides the strength of interaction between a variable and all other predictors.

[33] The H-statistic is computationally expensive, however, requiring us to sample a subset of the data to arrive at an estimate. Due to the variance associated with this sampling procedure, we repeat the calculation 25 times and averaged the results to arrive at a stable point estimate and to compute confidence bounds. See Molnar 2019 for the theoretical underpinning of this model-agnostic measure. We implement the H-statistic using the R package `iml`.

[34] Note that the H-statistic here is caculated after fitting a random forest on the full dataset, with 5 random variables selected at each node (this is the mode hyperparameter when looking across the distribution of training sets), as opposed to computing on test data only. It isn't clear whether computing it on training or

We next examine which variables account for the overall interaction strength of average earnings (Figure 12). We see that average earnings most strongly interacts with a change in unemployment and the FTSE all share index, followed by a firm's ratio of loans to retail deposits, core deposit ratio, NIM and solvency ratio.

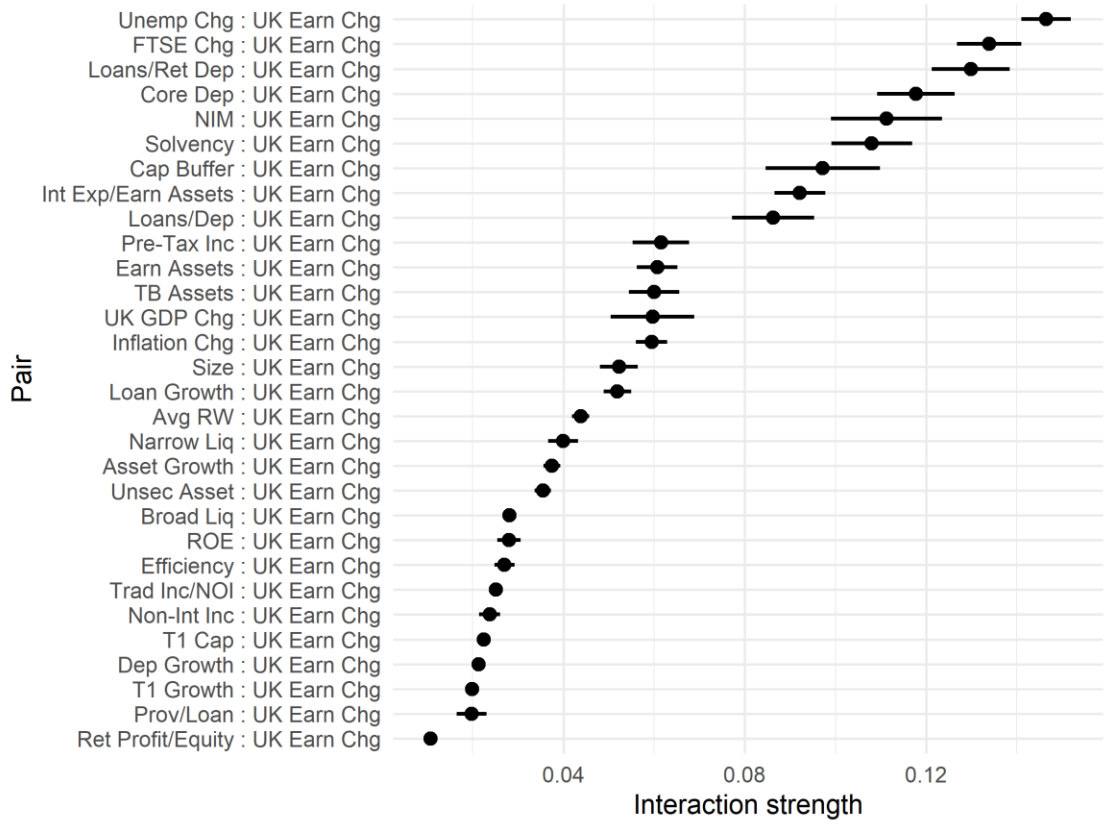
Figure 11: Interaction strength by variable



Note: H-statistic for each predictor with all other predictors. The error bars represent the 95% confidence interval.

test data is more appropriate (Molnar 2019). We opt for the former here since we are interested in the strength of interaction for a single, fixed random forest.

Figure 12: Interaction between average earnings and other model predictors



Note: Pairwise interaction strength for all variables with average earnings.

4.7: Shapley regression

We now turn to providing a rigorous statistical analysis of the outputs of our random forest utilising the Shapley regression framework put forward by Joseph (2019). This involves regressing our measure of firm distress on the Shapley values in order to ascertain the significance of each predictor in explaining bank distress. The Shapley regression model is defined as:

$$log(\frac{Pr(y_i = HighRisk)}{1 - Pr(y_i = HighRisk)}) = \alpha_0 + \beta_1\phi_{i1} + \ldots + \beta_p\phi_{ip}$$

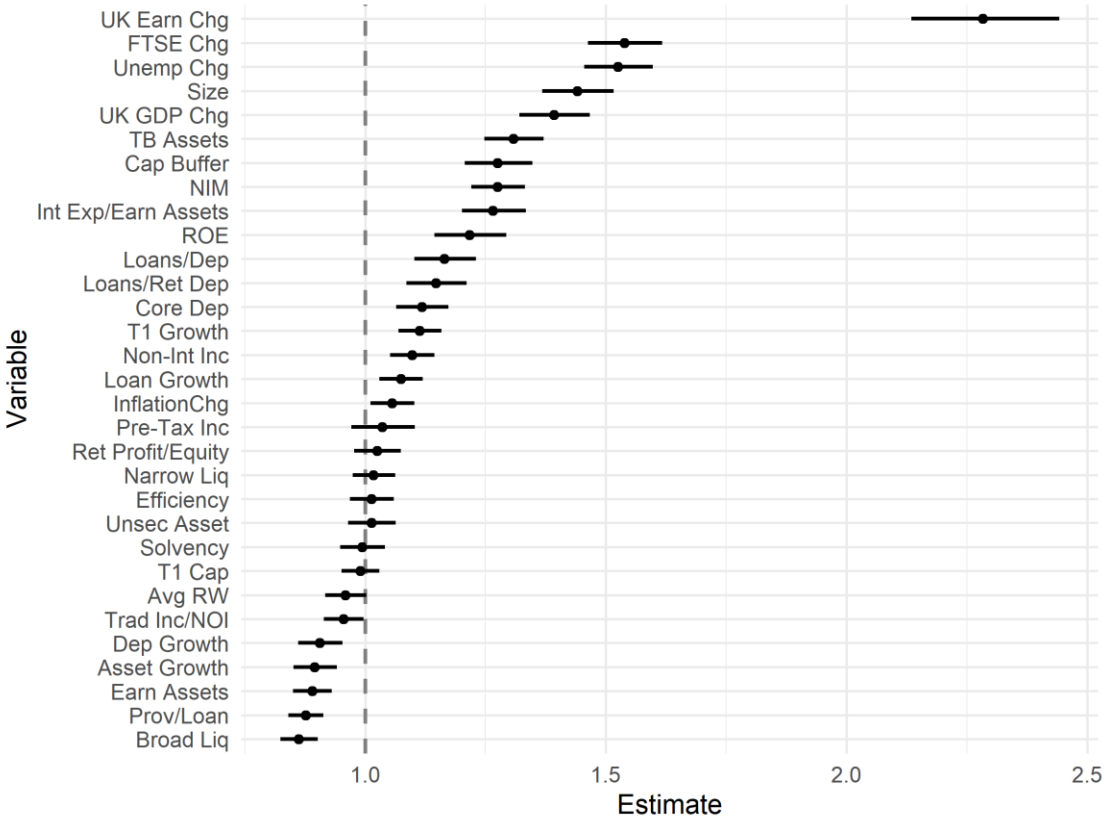
This equation is analagous to the pooled logit model but with Shapley values per predictor replacing the actual predictor values. The binary distress outcome for observation i is y_i, ϕ_{ip} is the Shapley value for predictor p and observation i, β_p is the parameter for ϕ_{ip}, and α_0 is the constant term. As with the pooled logit model, the response variable is taken as the log transformed odds of distress to constrain the fitted probabilities to between 0 and 1.

Figure 13 provides the scaled and exponentiated coefficients from the Shapley regression for the random forest. Importantly, whether a coefficient is above or below 1 (i.e. the sign of association) does not indicate whether there is a positive or negative relationship

between the predictor and distress. Instead, given the relationship between Shapley values and predicted probabilities, coefficients which are above 1 indicate a significant relationship, whereas below 1 is considered non-significant (even if the interval does not contain 1 (Joseph 2019; Bluwstein et al. 2019)). To incorporate direction of association we utilise the sign from the pooled logistic regression for each variable. Table 4 provides the sign alongside the Shapley and pooled logit regression coefficients.

Each coefficient in the Shapley regression represents the factor at which the predicted odds of distress is multiplied by when we increase the Shapley value for that predictor by one standard deviation. For example, if we increase the average earnings Shapley value by one standard deviation (or by 0.095), we would expect the odds that a firm will be in distress in a year's time to increase by 128%. In terms of a firm's capital buffer, an increase in the Shapley value by one standard deviation (0.018), would decrease the expected odds of distress in a year's time by 28%.

Figure 13: Shapley regression coefficients for the random forest



Note: Coefficients are standardised and exponentiated. The error bars represent the 95% confidence interval. If the interval estimate is above 1, the predictor is statistically significant.

Table 4: Shapley regression table

	Random Forest				Pooled logit regression	
	exp(Est.)	p	Sign	Shapley contribution	exp(Est.)	p
UK average earnings	2.283	0.000	-	0.170	0.670	0.000
FTSE All Share	1.538	0.000	+	0.050	1.321	0.002
Unemployment	1.525	0.000	-	0.054	0.199	0.000
Size	1.440	0.000	+	0.081	1.535	0.000
UK GDP	1.392	0.000	-	0.063	0.133	0.000
TB assets	1.308	0.000	+	0.025	1.640	0.000
Capital buffer	1.275	0.000	-	0.023	0.706	0.008
NIM	1.275	0.000	-	0.034	0.680	0.000
Interest expense	1.266	0.000	-	0.058	0.660	0.000
ROE	1.216	0.000	-	0.016	0.639	0.000
Loans deposit	1.164	0.000	+	0.032	1.475	0.000
Loans retail deposit	1.147	0.000	+	0.030	1.316	0.009
Core dep ratio	1.118	0.000	+	0.027	1.973	0.000
T1 growth	1.113	0.000	+	0.014	1.218	0.000
Non-interest income	1.097	0.000	+	0.015	1.575	0.000
Loan growth	1.074	0.001	+	0.010	1.081	0.182
Inflation	1.056	0.013	+	0.093	1.203	0.001
Pre-tax income	1.035	0.289	-	0.030	0.784	0.019
Retained profit	1.025	0.309	+	0.010	1.216	0.006
Narrow liq. ratio	1.017	0.437	+	0.014	1.031	0.659
Efficiency ratio	1.013	0.578	-	0.015	0.879	0.068
Unsecured asset	1.013	0.612	+	0.025	1.035	0.656
Solvency	0.994	0.789	+	0.018	1.056	0.652
T1 capital ratio	0.990	0.601	+	0.011	1.389	0.001
Avg. risk-weight	0.959	0.061	+	0.016	1.015	0.880
Trading income	0.955	0.033	-	0.013	0.973	0.625
Deposit growth	0.906	0.000	+	0.007	1.132	0.192
Asset growth	0.895	0.000	-	0.007	0.884	0.222
Earning assets	0.890	0.000	+	0.014	1.069	0.297
Provisions	0.876	0.000	+	0.008	1.067	0.227
Broad liq. ratio	0.862	0.000	+	0.019	1.017	0.772

Note: The left hand side of the table provides estimated Shapley regression coefficients (scaled and exponentiated) for the random forest model, alongside the associated p-value, sign (taken from the sign of the pooled logit regression coefficients), and the average absolute Shapley value for each predictor. The right hand side of the table provides estimated coefficients for the pooled logit model (scaled and exponentiated) and p-values.

4.8: Ensembles

The results above demonstrate that the two different ensemble models (bagging for the random forest, and the boosting approach) outperform the other models examined in terms of AUC and Brier score. This is no accident; bringing together diverse and accurate models will often be superior to individual models on their own (Dietterich 2000; Z. H. Zhou 2012; James et al. 2013). The application of ensemble techniques to bankruptcy prediction is a young and exciting area of research (Ravi et al. 2008; Nanni and Lumini 2009; Jardin 2016; Jardin 2018; Alaka et al. 2018). In this section, we bring together the predictions of all our models in different ways in an attempt to improve performance.

We perform three different combinations: a simple average of all six models, a simple average of the top four performing models (random forest, boosting, pooled logit and SVM), and a stacked procedure with a linear regression model at the second-level. Essentially these differ in terms of the weight put on the predicted probability derived from the different models, with the first assigning equal weight, the second assigning weights of zero to the two worst-performing models and equal weight to the remaining four, and the stacking approach assigning weights equal to the parameters of the second-level linear regression model.[35] The stack regression model is as follows:

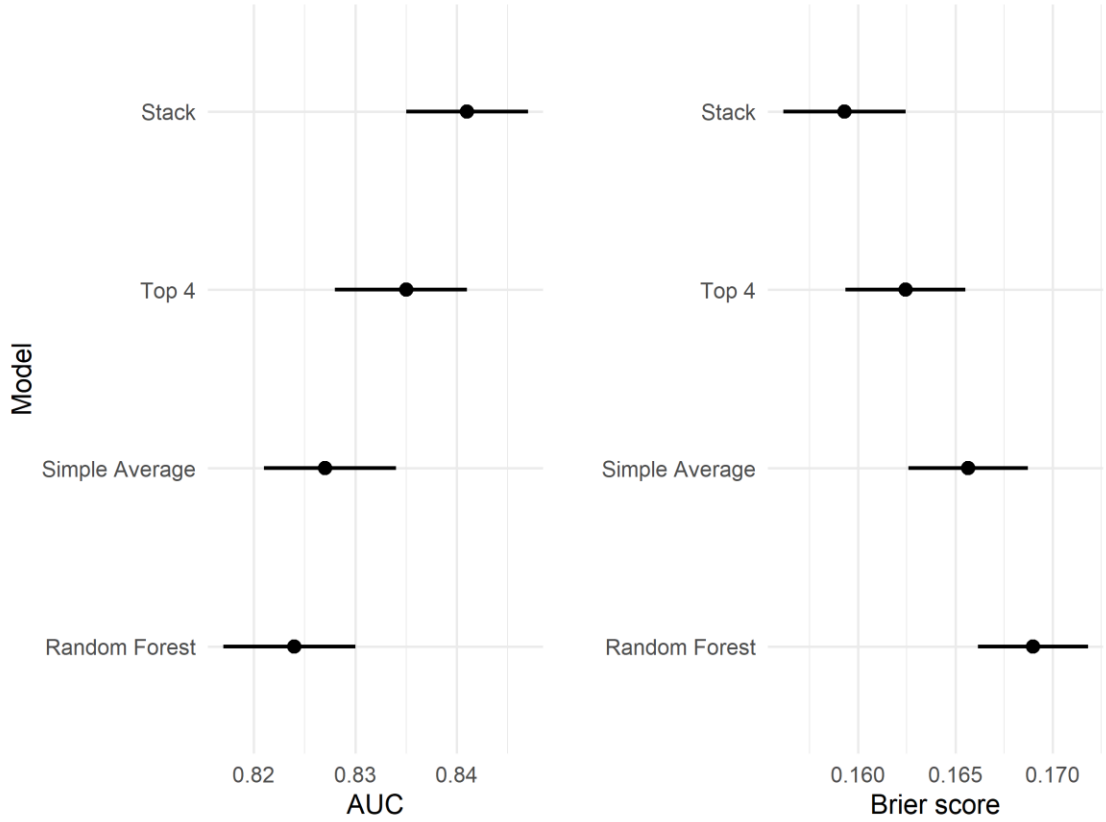
$$y_i = \alpha_0 + \beta_1 p_{i1} + \ldots + \beta_6 p_{i6}$$

Where y_i is the distressed indicator, α_0 is the intercept, p_{i1} is the predicted probability for model 1 and test observation i, and β_i is the coefficient or weight placed on the output of model 1. Figure 14 provides the AUC calculations and 95% confidence interval limits for the ensembles, alongside, for comparative purposes, the random forest.

The results indicate that the simple average ensemble does about the same as the random forest on its own, while the top 4 and stacking ensembles improve upon the random forest by 1.1 and 1.7 percentage points respectively. The interval estimates indicate that the stacking approach lower limit exceeds the upper limit of the random forest, demonstrating that this simple ensembling method significantly and substantively improves our ability to predict distress one year out. Similarly for the Brier score, the stacked procedure significantly outperforms the random forest with a difference of 0.97 percentage points.

[35] The stacking approach differs from other general ensembling techniques – namely, boosting and bagging – by creating a new data set from the outputs of the base (or first-level) models. In our case, the second-level data set consists of the predicted probabilities of each of the six original models as predictor observations and our measure of distress remains as the outcome variable. We utilise a simple linear regression as the second-level model but this could be in practice any statistical or machine learning approach (see Z. H. Zhou (2012) for details on stacking and other ensembling techniques more generally).

Figure 14: Ensemble AUC and Brier score estimates and confidence intervals



Section 5: Robustness checks

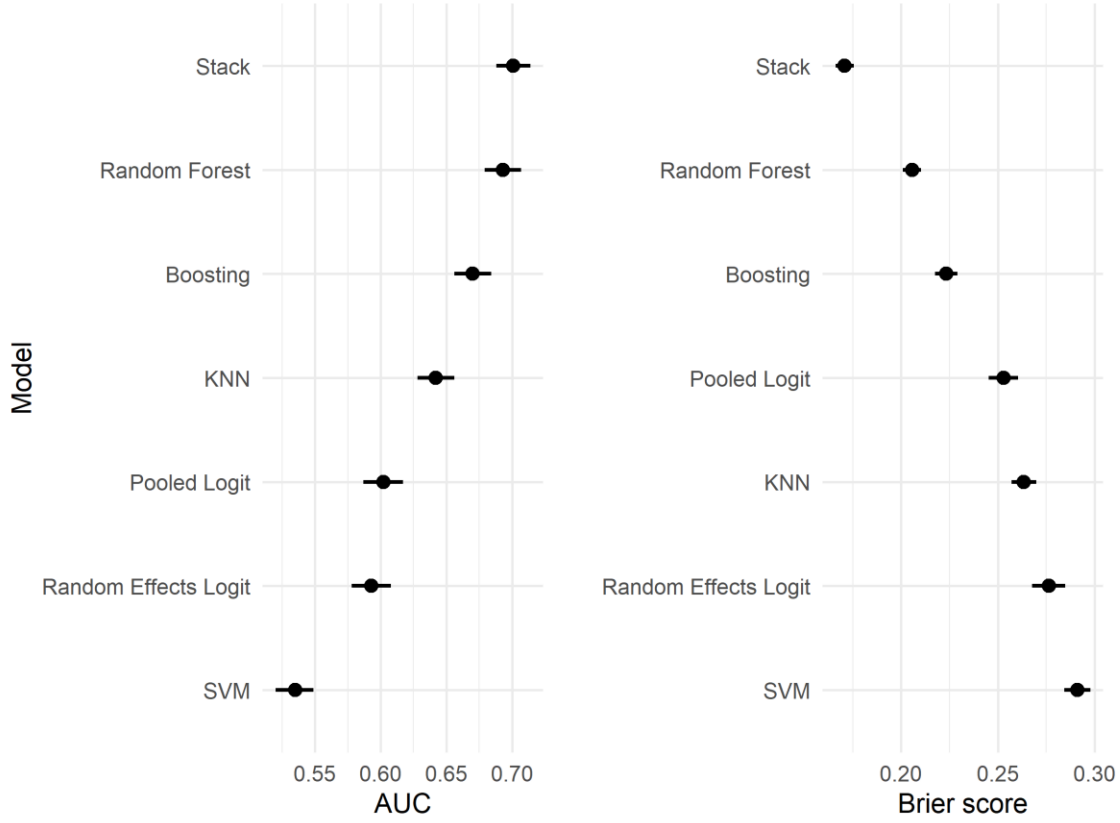
In this section, we investigate the robustness of our results to a different cross-validation procedure, implementing a rolling forecast window. We also investigate whether our results are sensitive to the choice of predictor lag structure.

The rolling forecast is constructed such that our test sample is always future data with respect to the training set. We make use of a fixed window of eight quarters for training and a horizon of four quarters for testing. Given a total of 26 quarters for the full sample, we effectively have 15 'time slices' – i.e. 15 different combinations of eight quarter training sets and four quarter test sets. Within each time slice we adopt ten-fold cross-validation, randomising at the firm-level, to estimate out-of-sample performance. Figure 15 provides the AUC and confidence intervals for each model when implementing this procedure. It shows that the random forest once again outperforms all the other techniques, providing us with confidence in our baseline results.

However, the overall performance is substantively reduced relative to our preferred cross-validation procedure. This is due to the reduced training sample in each fold relative to our baseline procedure, as well as the more severe class imbalance between train and test sets, with a difference in mean levels of distress of 18.4 percentage points. The estimates are also more uncertain in the rolling forecast window owing to a smaller number of overall

test observations (7040), leading the difference in AUCs between the random forest and boosting approach to be statistically insignificant. We also include the stack ensemble in Figure 15, demonstrating this ensemble of all the models once again outperforms each standalone technique, albeit the AUC interval estimate of the stack overlaps with that of the random forest.

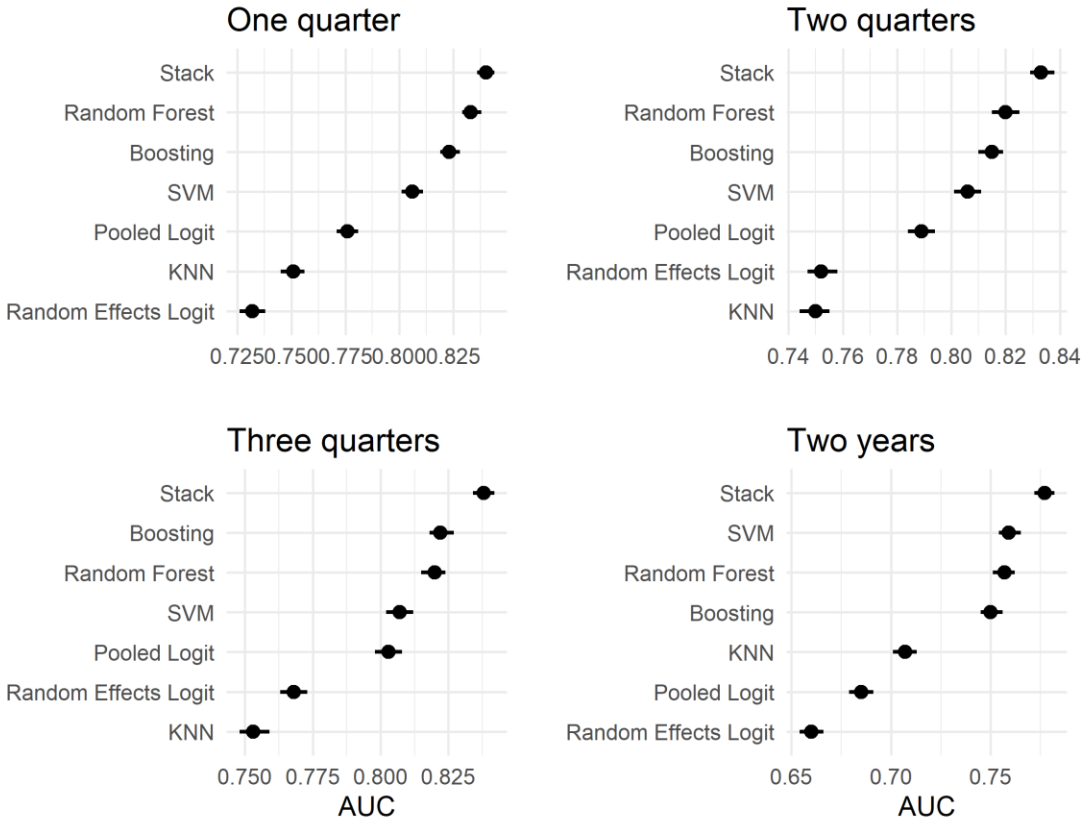
Figure 15: AUC and Brier score for rolling forecast



Note: AUC and Brier score metrics for each model with 95% confidence intervals. Estimates are based on a total of n = 7040 test observations and a rolling window forecast.

We next examine whether our results are sensitive to the choice of one year as the baseline forecasting horizon – i.e. we alter the predictor lag structure. Figure 16 shows how the different models perform in terms of AUC at four different horizons: one quarter, two quarters, three quarters, and two years. The random forest is still the best performing approach, albeit it is only significantly better than the other three when predictors are lagged by one quarter, whereas it is tied as the top performing for the other horizons. Figure 16 also provides the AUC for the stacked ensemble, showing this once again improves performance substantively relative to the standalone approaches in these alternative horizons.

Figure 16: AUC at different forecasting horizons



Section 6: Conclusion

In this paper, we utilise novel data and machine learning techniques to build an early warning system for UK bank distress. We compare a number of machine learning and classical statistical techniques, implementing a rigorous, double-block randomised cross-validation procedure to evaluate out-of-sample performance. We find the random forest algorithm to be superior in terms of ranking test observations (i.e. maximising AUC), while also having relatively better calibrated probabilities than the other techniques (i.e. minimising the Brier score). We also examine performance at two different decision thresholds, 50% and 25%, and vary the relative cost of misclassification between FN and FP errors, demonstrating the random forest to have lower cost as the weight changes in favour of the former over the latter.

The performance results indicate that the random forest should be used to build an early warning system. In order to improve the algorithm's transparency, we examine the drivers of the model's predicted probabilities, utilising an aggregation of Shapley values per test set observation and Shapley regression framework (Joseph 2019). The results of this reveal the drivers of the random forest to be qualitatively different from the pooled logit regression, a fact we explain by investigating the interaction strength (H-statistic) for each explanatory variable. The Shapley regression reveals the importance of macroeconomic variables (especially year-on-year change of average UK real earnings), and a firm's

sensitivity to market risk (ratio of trading book to total assets), capital buffer and net interest margin. Finally, we also perform simple ensembling techniques to combine all the model outputs, demonstrating substantive and statistically significant improvements relative to the random forest on its own.

Future research might extend this analysis in a number of ways. First, scholars might seek to incorporate additional data beyond financial ratios and macroeconomic variables. For example, textual data which sheds light on the quality of a firm's management and governance, or metrics which capture aspects of a firm's cultures would enrich the set of input variables (see Graham et al. (2017) for a review of the literature on corporate culture). Second, we have performed only simple ensembling techniques to gain additional performance benefits. Future work might delve into more complex configurations of diverse underlying models to reap substantive improvements. Third, this paper's analysis relies on data from a highly unusual period in economic history. Future research might seek to establish whether the documented relationship between input variables and measures of distress persist in relatively benign economic environments. It is likely that in such periods macroeconomic variables are less important in predicting firm distress (which would also obviously be less common occurences), and so an early warning system might be better if it were based on data which encompasses more or all of an economic cycle.

Overall, this research paper demonstrates the practical benefits of machine learning and ensembling methods for providing regulators with advance warning of firm distress. Supervisors can apply these findings in order to aid in anticipating problems before they occur, thereby helping them in their mission to keep financial institutions safe and sound.

Annex

Macroeconomic data definitions

The below table provides definitions and sources for the macroeconomic variables we introduce as predictors.

Table 5: Macroeconomic variable definitions and sources

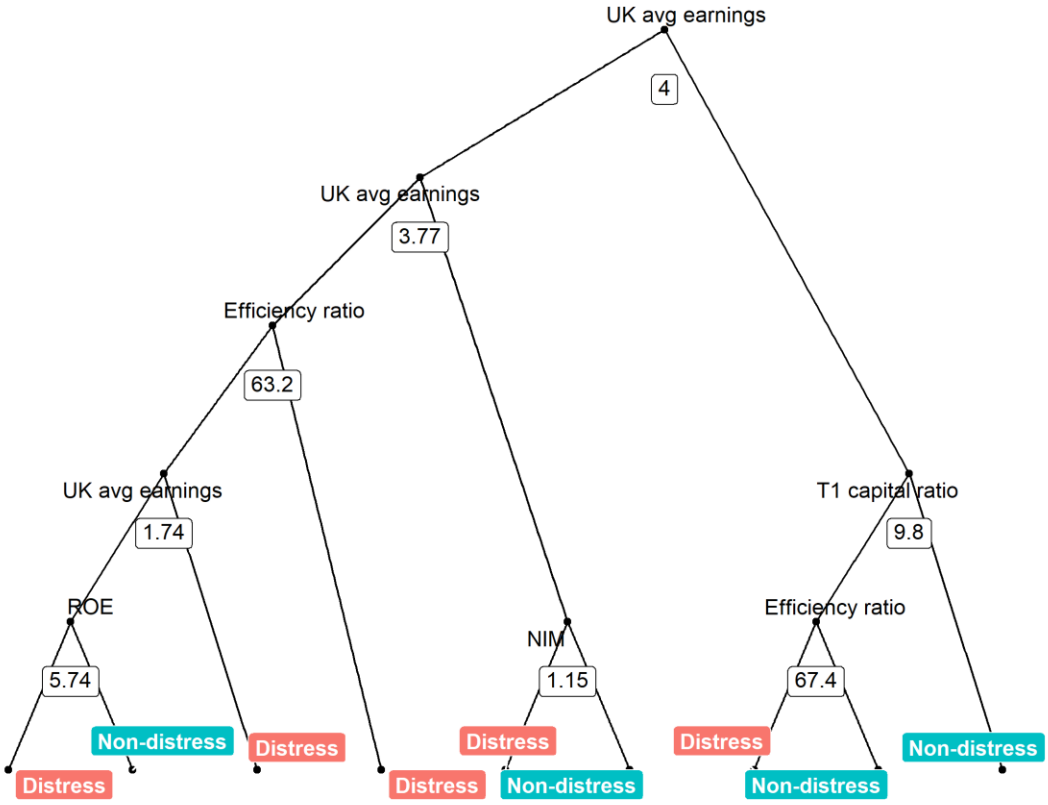
Variable	Definition	Source
FTSE All Share YoY change	Index of approximately 600 listed companies on the London Stock Exchange (at least 98% of total capital value of listed companies)	Refinitiv Eikon
UK inflation YoY change	Consumer price inflation index year on year change	Office for National Statistics
UK average real earnings YoY change	Weekly average real earnings	Office for National Statistics
UK real GDP YoY change	Gross domestic product (seasonally adjusted)	Office for National Statistics
UK unemployment YoY change	Number of unemployed (aged 16 and over, seasonally adjusted)	Office for National Statistics

Decision trees

An example of an individual decision tree fit on a subset of our overall sample can be seen in Figure 17.[36] A decision tree splits the data successively in two parts, with each selection chosen in order to optimise a specific criteria of interest. In our case, we are interested in reducing the Gini index – a metric for quantifying the class homogeneity of the resulting parts of a split. In Figure 17, the first split occurs at the top: the best way of initially splitting the data in two is to separate observations by whether the UK saw year-on-year growth above or below 4% in average real earnings the previous year (since all variables are lagged by four quarters in our base model). The ends of the trees at the bottom of Figure 17, known as the terminal nodes, provide the prediction rule for each observation. For the terminal node all the way to the right, a bank which has an above 9.8% T1 capital ratio when the UK economy sees above 4% year-on-year average real earnings growth is predicted to not be in distress. Essentially, what a decision tree provides is a series of decision rules for assigning predictions.

[36] With thanks to Shirin Glander for the R code used to produce this dendrogram.

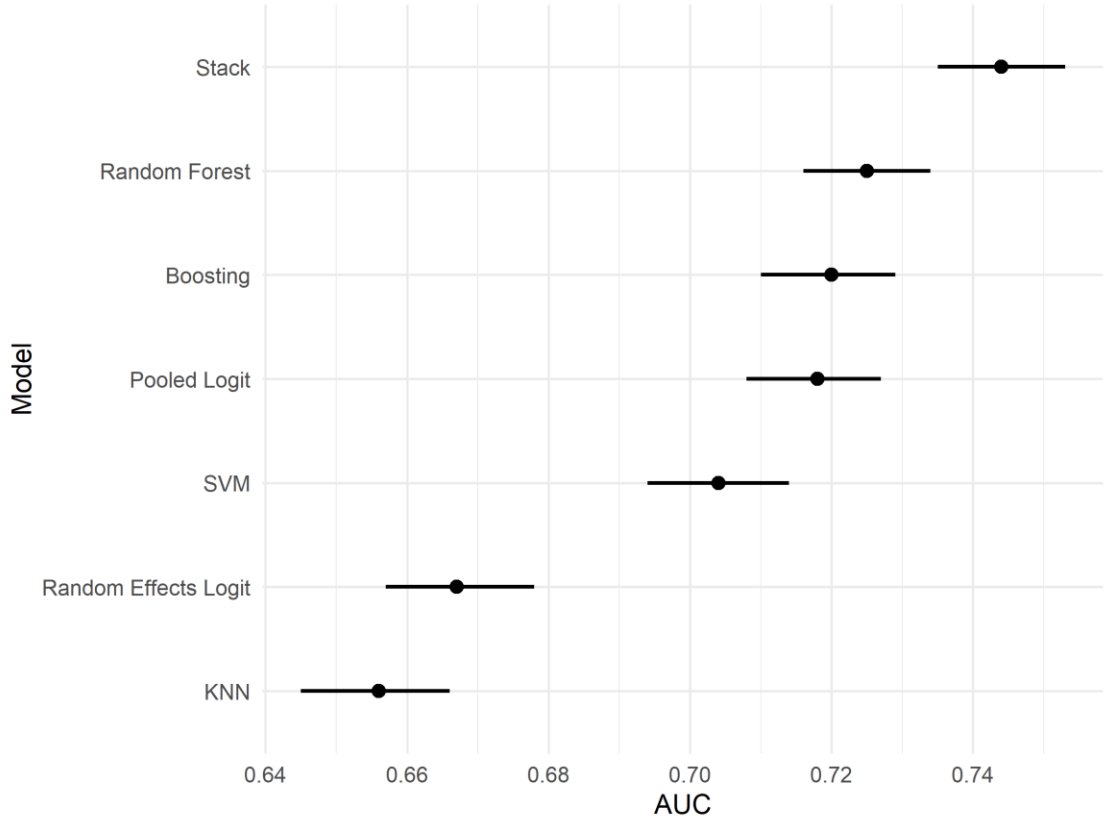
Figure 17: Individual decision tree fit on a subset of the overall sample



Additional robustness check

The FSA's implemented a a Supervisory Enhancement Programme (SEP) following the run on Northern Rock. Figure 18 provides the AUCs and 95% confidence intervals when we restrict our sample to the period after the SEP was largely complete, from mid-2009 onwards. The figure shows that the results are generally the same: the random forest remaine the top standalone performer, albeit not significantly different than the boosting and pooled logit models, and the stack ensemble once again outperforms each of the standalone approaches. Relative to the baseline results in Figure 4, the AUCs are substantively smaller. This weakening of performance is a result of a much smaller overall sample for the different approaches to train on – restricting the sample leaves us with 1796 total observations.

Figure 18: AUCs for models post-SEP



References

Adler, Werner, Alexander Brenning, Sergej Potapov, Matthias Schmid, and Berthold Lausen. 2011. "Ensemble Classification of Paired Data." *Computational Statistics & Data Analysis* 55 (5). Elsevier: 1933–41.

Afshartous, David, and Jan de Leeuw. 2005. "Prediction in Multilevel Models." *Journal of Educational and Behavioral Statistics* 30 (2). Sage Publications Sage CA: Los Angeles, CA: 109–39.

Alaka, Hafiz A, Lukumon O Oyedele, Hakeem A Owolabi, Vikas Kumar, Saheed O Ajayi, Olugbenga O Akinade, and Muhammad Bilal. 2018. "Systematic Review of Bankruptcy Prediction Models: Towards a Framework for Tool Selection." *Expert Systems with Applications* 94. Elsevier: 164–84.

Arena, Marco. 2008. "Bank Failures and Bank Fundamentals: A Comparative Analysis of Latin America and East Asia During the Nineties Using Bank-Level Data." *Journal of Banking & Finance* 32 (2). Elsevier: 299–310.

Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. "Machine Learning Models and Bankruptcy Prediction." *Expert Systems with Applications* 83. Elsevier: 405–17.

Bell, Timothy B. 1997. "Neural Nets or the Logit Model? A Comparison of Each Model's Ability to Predict Commercial Bank Failures." *Intelligent Systems in Accounting, Finance & Management* 6 (3). Wiley Online Library: 249–64.

Betz, Frank, Silviu Oprică, Tuomas A Peltonen, and Peter Sarlin. 2014. "Predicting Distress in European Banks." *Journal of Banking & Finance* 45. Elsevier: 225–41.

Bluwstein, Kristina, Buckmann Marcus, Joseph Andreas, Kang Miao, Kapadia Sujit, and Ozgur Simsek. 2019. "Credit Growth, the Yield Curve and Financial Crisis Prediction: Evidence from a Machine Learning Approach." *Bank of England Staff Working Paper Series*.

Bongini, Paola, Stijn Claessens, and Giovanni Ferri. 2001. "The Political Economy of Distress in East Asian Financial Institutions." *Journal of Financial Services Research* 19 (1). Springer: 5–25.

Bouckaert, Remco R, and Eibe Frank. 2004. "Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 3–12. Springer.

Bouwmeester, Walter, Jos WR Twisk, Teus H Kappen, Wilton A van Klei, Karel GM Moons, and Yvonne Vergouwe. 2013. "Prediction Models for Clustered Data: Comparison of a Random Intercept and Standard Regression Model." *BMC Medical Research Methodology* 13 (1). BioMed Central: 19.

Boyacioglu, Melek Acar, Yakup Kara, and Ömer Kaan Baykan. 2009. "Predicting Bank Financial Failures Using Neural Networks, Support Vector Machines and Multivariate Statistical Methods: A Comparative Analysis in the Sample of Savings Deposit Insurance

Fund (Sdif) Transferred Banks in Turkey." *Expert Systems with Applications* 36 (2). Elsevier: 3355–66.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1). Springer: 5–32.

Brenning, Alexander, and Berthold Lausen. 2008. "Estimating Error Rates in the Classification of Paired Organs." *Statistics in Medicine* 27 (22). Wiley Online Library: 4515–31.

Carmona, Pedro, Francisco Climent, and Alexandre Momparler. 2018. "Predicting Failure in the Us Banking Sector: An Extreme Gradient Boosting Approach." *International Review of Economics & Finance*. Elsevier.

Cawley, Gavin C, and Nicola LC Talbot. 2010. "On over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation." *Journal of Machine Learning Research* 11 (Jul): 2079–2107.

Chiaramonte, Laura, Hong Liu, Federica Poli, and Mingming Zhou. 2016. "How Accurately Can Z-Score Predict Bank Failure?" *Financial Markets, Institutions & Instruments* 25 (5). Wiley Online Library: 333–60.

Cleary, Sean, and Greg Hebb. 2016. "An Efficient and Functional Model for Predicting Bank Distress: In and Out of Sample Evidence." *Journal of Banking & Finance* 64. Elsevier: 101–11.

Coen, Jamie, William Francis, and May Rostom. 2017. "The Determinants of Uk Credit Union Failure." *Bank of England Staff Working Paper Series No. 658*.

Cole, Rebel A, and Jeffery W Gunther. 1995. "Separating the Timing and Likelihood of Bank Failure." *Journal of Banking & Finance* 19 (6): 1073–89.

Cole, Rebel A, and Lawrence J White. 2012. "Déjà Vu All over Again: The Causes of Us Commercial Bank Failures This Time Around." *Journal of Financial Services Research* 42 (1-2). Springer: 5–29.

Collier, Charles, Sean Forbush, Daniel A Nuxoll, and John O'Keefe. 2003. "The Scor System of Off-Site Monitoring: Its Objectives, Functioning, and Performance." *FDIC Banking Rev.* 15. HeinOnline: 17.

Curry, Timothy J, Peter J Elmer, and Gary S Fissel. 2003. "Using Market Information to Help Identify Distressed Institutions: A Regulatory Perspective." *FDIC Banking Rev.* 15. HeinOnline: 1.

Čihák, Martin. 2007. "Systemic Loss: A Measure of Financial Stability." *Czech Journal of Economics and Finance* 57 (1-2): 5–26.

de-Ramon, Sebastian, William Francis, and Kristoffer Milonas. 2018. "An Overview of the Uk Banking Sector Since the Basel Accord: Insights from a New Regulatory Database." *Bank of England Staff Working Paper No. 652*.

de-Ramon, Sebastian, William Francis, and Michael Straughan. 2018. "Bank Competition and Stability in the United Kingdom." *Bank of England Staff Working Paper No. 748*.

DeLong, Elizabeth, David DeLong, and Daniel Clarke-Pearson. 1988. "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics* 44 (3): 837–45.

DeYoung, Robert, and Gökhan Torna. 2013. "Nontraditional Banking Activities and Bank Failures During the Financial Crisis." *Journal of Financial Intermediation* 22 (3). Elsevier: 397–421.

Dietterich, Thomas G. 2000. "Ensemble Methods in Machine Learning." In *International Workshop on Multiple Classifier Systems*, 1–15. Springer.

Division, FSA Internal Audit. 2008. "The Supervision of Northern Rock: A Lessons Learned Review." https://www.fca.org.uk/publication/corporate/fsa-nr-report.pdf.

Fahlenbrach, Rüdiger, Robert Prilmeier, and René M Stulz. 2017. "Why Does Fast Loan Growth Predict Poor Performance for Banks?" *The Review of Financial Studies* 31 (3). Oxford University Press: 1014–63.

Fawcett, Tom. 2006. "An Introduction to Roc Analysis." *Pattern Recognition Letters* 27 (8). Elsevier: 861–74.

Fernandez-Delgado, Manuel, Eva Cernadas, Senen Barro, and Dinani Amorim. 2014. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" *The Journal of Machine Learning Research* 15 (1): 3133–81.

Finkelman, Brian S, Benjamin French, and Stephen E Kimmel. 2016. "The Prediction Accuracy of Dynamic Mixed-Effects Models in Clustered Data." *BioData Mining* 9 (1). BioMed Central: 5.

Flannery, Mark J. 1998. "Using Market Information in Prudential Bank Supervision: A Review of the U.S. Empirical Evidence." *Journal of Money, Credit and Banking* 30 (3). Ohio State University Press: 273–305.

Freund, Yoav, and Robert E Schapire. 1997. "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting." *Journal of Computer and System Sciences* 55 (1). Elsevier: 119–39.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*. JSTOR, 1189–1232.

Friedman, Jerome H, and Bogdan E Popescu. 2008. "Predictive Learning via Rule Ensembles." *The Annals of Applied Statistics* 2 (3). Institute of Mathematical Statistics: 916–54.

FSA. 2009. "The Turner Review: A Regulatory Response to the Global Banking Crisis." http://www.fsa.gov.uk/pubs/other/turner_review.pdf.

Gogas, Periklis, Theophilos Papadimitriou, and Anna Agrapetidou. 2018. "Forecasting Bank Failures and Stress Testing: A Machine Learning Approach." *International Journal of Forecasting* 34 (3): 440–55. doi:https://doi.org/10.1016/j.ijforecast.2018.01.009.

Gomez-Gonzalez, Jose E, and Nicholas M Kiefer. 2009. "Bank Failure: Evidence from the Colombian Financial Crisis." *The International Journal of Business and Finance Research* 3 (2): 15–31.

Graham, John R, Campbell R Harvey, Jillian Popadak, and Shivaram Rajgopal. 2017. "Corporate Culture: Evidence from the Field." National Bureau of Economic Research.

Iturriaga, Félix J López, and Iván Pastor Sanz. 2015. "Bankruptcy Visualization and Prediction Using Neural Networks: A Study of Us Commercial Banks." *Expert Systems with Applications* 42 (6). Elsevier: 2857–69.

Jagtiani, Julapa, James Kolari, Catharine Lemieux, and Hwan Shin. 2003. "Early Warning Models for Bank Supervision: Simpler Could Be Better." *Economic Perspectives* 27 (3). Federal Reserve Bank of Chicago: 49–61.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

Jardin, Philippe du. 2016. "A Two-Stage Classification Technique for Bankruptcy Prediction." *European Journal of Operational Research* 254 (1). Elsevier: 236–52.

———. 2018. "Failure Pattern-Based Ensembles Applied to Bankruptcy Forecasting." *Decision Support Systems* 107. Elsevier: 64–77.

Joseph, Andreas. 2019. "Shapley Regressions: A Framework for Statistical Inference on Machine Learning Models." *Bank of England Staff Working Paper No. 784*.

Kaufman, Shachar, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. "Leakage in Data Mining: Formulation, Detection, and Avoidance." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6 (4). ACM: 15.

Kumar, P Ravi, and Vadlamani Ravi. 2007. "Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques–A Review." *European Journal of Operational Research* 180 (1). Elsevier: 1–28.

Kutner, Michael H, Christopher J Nachtsheim, John Neter, William Li, and others. 2005. *Applied Linear Statistical Models*. Vol. 103. McGraw-Hill Irwin Boston.

Lane, William R, Stephen W Looney, and James W Wansley. 1986. "An Application of the Cox Proportional Hazards Model to Bank Failure." *Journal of Banking & Finance* 10 (4). Elsevier: 511–31.

Le, Hong Hanh, and Jean-Laurent Viviani. 2018. "Predicting Bank Failure: An Improvement by Implementing a Machine-Learning Approach to Classical Financial Ratios." *Research in*

International Business and Finance 44: 16–25. doi:https://doi.org/10.1016/j.ribaf.2017.07.104.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*, 4765–74.

Mare, Davide Salvatore. 2015. "Contribution of Macroeconomic Factors to the Prediction of Small Bank Failures." *Journal of International Financial Markets, Institutions and Money* 39. Elsevier: 25–39.

Martin, Daniel. 1977. "Early Warning of Bank Failure: A Logit Regression Approach." *Journal of Banking & Finance* 1 (3). Elsevier: 249–76.

Männasoo, Kadri, and David G Mayes. 2009. "Explaining Bank Distress in Eastern European Transition Economies." *Journal of Banking & Finance* 33 (2). Elsevier: 244–53.

McCulloch, Charles E, and John M Neuhaus. 2011. "Prediction of Random Effects in Linear and Generalized Linear Models Under Model Misspecification." *Biometrics* 67 (1). Wiley Online Library: 270–79.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2019. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), Tu Wien*. https://CRAN.R-project.org/package=e1071.

Min, Jae H, and Young-Chan Lee. 2005. "Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters." *Expert Systems with Applications* 28 (4). Elsevier: 603–14.

Molnar, Christoph. 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/. https://christophm.github.io/interpretable-ml-book/.

Nanni, Loris, and Alessandra Lumini. 2009. "An Experimental Comparison of Ensemble of Classifiers for Bankruptcy Prediction and Credit Scoring." *Expert Systems with Applications* 36 (2). Elsevier: 3028–33.

Ni, Haifang, Rolf HH Groenwold, Mirjam Nielen, and Irene Klugkist. 2018. "Prediction Models for Clustered Data with Informative Priors for the Random Effects: A Simulation Study." *BMC Medical Research Methodology* 18 (1). BioMed Central: 83.

Niculescu-Mizil, Alexandru, and Rich Caruana. 2005. "Predicting Good Probabilities with Supervised Learning." In *Proceedings of the 22nd International Conference on Machine Learning*, 625–32. ACM.

Oet, Mikhail V, Timothy Bianco, Dieter Gramlich, and Stephen J Ong. 2013. "SAFE: An Early Warning System for Systemic Banking Risk." *Journal of Banking & Finance* 37 (11). Elsevier: 4510–33.

Platt, John. 1999. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." *Advances in Large Margin Classifiers* 10 (3). Cambridge, MA: 61–74.

Poghosyan, Tigran, and Martin Čihak. 2011. "Determinants of Bank Distress in Europe: Evidence from a New Data Set." *Journal of Financial Services Research* 40 (3). Springer: 163–84.

Ravi, Vadlamani, H Kurniawan, Peter Nwee Kok Thai, and P Ravi Kumar. 2008. "Soft Computing System for Bank Performance Prediction." *Applied Soft Computing* 8 (1). Elsevier: 305–15.

Roberts, David R, Volker Bahn, Simone Ciuti, Mark S Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40 (8). Wiley Online Library: 913–29.

Shapley, Lloyd S. 1953. "A Value for N-Person Games." *Contributions to the Theory of Games*, no. 28. Princeton University Press: 307–17.

Shumway, Tyler. 2001. "Forecasting Bankruptcy More Accurately: A Simple Hazard Model." *The Journal of Business* 74 (1). JSTOR: 101–24.

Strumbelj, Erik, and Igor Kononenko. 2014. "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge and Information Systems* 41 (3). Springer: 647–65.

Swicegood, Philip, and Jeffrey A Clark. 2001. "Off-Site Monitoring Systems for Predicting Bank Underperformance: A Comparison of Neural Networks, Discriminant Analysis, and Professional Human Judgment." *Intelligent Systems in Accounting, Finance & Management* 10 (3). Wiley Online Library: 169–86.

Tinoco, Mario Hernandez, and Nick Wilson. 2013. "Financial Distress and Bankruptcy Prediction Among Listed Companies Using Accounting, Market and Macroeconomic Variables." *International Review of Financial Analysis* 30. Elsevier: 394–419.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Whalen, Gary. 1991. "A Proportional Hazards Model of Bank Failure: An Examination of Its Usefulness as an Early Warning Tool." *Economic Review* 27 (1): 21–31.

Whalen, Gary, James B Thomson, and others. 1988. "Using Financial Data to Identify Changes in Bank Condition." *Economic Review* 24 (2). Federal Reserve Bank of Cleveland: 17–26.

Wheelock, David C, and Paul W Wilson. 2000. “Why Do Banks Disappear? The Determinants of Us Bank Failures and Acquisitions.” *Review of Economics and Statistics* 82 (1). MIT Press: 127–38.

Zhao, Huimin, Atish P Sinha, and Wei Ge. 2009. “Effects of Feature Construction on Classification Performance: An Empirical Study in Bank Failure Prediction.” *Expert Systems with Applications* 36 (2). Elsevier: 2633–44.

Zhou, Zhi Hua. 2012. *Ensemble Methods: Foundations and Algorithms*. Chapman; Hall/CRC.