DOI: 10.1111/eufm.12369

ORIGINAL ARTICLE

EUROPEAN FINANCIAL MANAGEMENT WILEY

# Modelling failure rates with machine-learning models: Evidence from a panel of UK firms

**Georgios Sermpinis**[1] | **Serafeim Tsoukas**[1] | **Yiqun Zhang**[2]

[1]Adam Smith Business School, University of Glasgow, Glasgow, UK

[2]School of Insurance, Central University of Finance and Economics, Beijing, China

**Correspondence**
Georgios Sermpinis, Adam Smith Business School, University of Glasgow, Glasgow G12 8QQ, UK.
Email: georgios.sermpinis@glasgow.ac.uk

## Abstract

In this study, we investigate the ability of machine-learning techniques to predict firm failures and we compare them against alternatives. Using data on business and financial risks of UK firms over 1994–2019, we document that machine-learning models are systematically more accurate than a discrete hazard benchmark. We conclude that the random forest model outperforms other models in failure prediction. In addition, we show that the improved predictive power of the random forest model relative to its counterparts persists when we consider extreme economic events as well as firm and industry heterogeneity. Finally, we find that financial factors affect failure probabilities.

**KEYWORDS**

business closures, finance, financial ratios, machine-learning models, random forest

**JEL CLASSIFICATION**

G17, G33, C25, E37

We are grateful to the editor (John Doukas) and one anonymous reviewer for insightful comments and suggestions. We acknowledge the participants at the 2021 International Conference of the French Finance Association and the 2022 International Conference on Macroeconomic Analysis and International Finance for their useful comments. Any remaining errors are our own.

# 1 | INTRODUCTION

It is well accepted that timely detection and accurate prediction of firm failures are essential to firm managers, market participants and policymakers. Managers, as insiders, can incorporate reliable and efficient failure predictions into their internal performance evaluations to check management performance and construct early warning mechanisms so they can implement remedial actions (Geng et al., 2015). Moreover, accurate failure prediction can lower the probability that a firm's outsiders (e.g., investors and creditors) become exposed to default risks and losses. Failure prediction can also encourage policymakers to create regulations or policies that can stabilize the financial markets.

Improvement of prediction techniques has become a pressing issue for academics and practitioners in light of extreme economic events such as the most recent global financial crisis (GFC) or the United Kingdom's decision to leave the European Union in the 23 June 2016 referendum (Brexit). Both events are characterized by heightened economic and policy uncertainty with implications for firms' real activities, as well as for trade, immigration and regulation (Bloom et al., 2019; van Reenen, 2016). The empirical literature confirms that uncertainty affects firms due to declines in demand and supply or in the extreme, corporate bankruptcy. Surprisingly, however, there is limited empirical evidence regarding the most appropriate modelling strategy and the characteristics that affect firm closures during the recent global financial crisis and Brexit.

The purpose of this paper is to take a deeper look at firms' failures both in tranquil and crisis times, paying attention to firm and industry heterogeneity and improved methods of firms' exit assessments. Analysis of bankruptcy prediction techniques has a long pedigree (see Altman, 1968; Chava & Jarrow, 2004; Dimitras et al., 1996; Doumpos et al., 2017; Kumar & Ravi, 2007; Ohlson, 1980; Shumway, 2001) and seeks to model default likelihood using a set of business and financial risks (Kumar & Ravi, 2007). Such studies typically apply firm-specific financial ratios and other publicly available information in reduced-form models and find that firms' chances of survival correspond strongly to a number of balance sheet indicators and macroeconomic conditions.

However, the conventional approach in the literature has important drawbacks because it relies on strict assumptions such as linearity, normality and pre-existing functional forms and the selection of covariates depends on researchers' knowledge. If these assumptions are violated, the statistical models produce biased estimates which may further reduce the models' predictive power. In addition, the conventional approach might work well in tranquil periods, but they are less sensitive to financial datasets characterized by heightened uncertainty and financial distress. In other words, the performance of the conventional approach not only relies on the mathematical algorithm but the quality of data set. As computing techniques advance, reducing the interference from superabundant outliers in the data set, simplifying the constructed model and evaluating the predictive ability of predictors with high explanatory ability have gained momentum in modelling corporate bankruptcy.

Our study makes three important contributions to the literature. First, we rely on machine learning (ML) tools, which are gaining ground in economic and finance applications (see, e.g., Akyildirim et al., 2021; Athey et al., 2019; Aziz et al., 2021; Barboza et al., 2017; Knaus et al., 2021; de Moor et al., 2018). We compare the performance of these models with the golden standard of default prediction studies—the discrete hazard (DH) model. These modelling techniques have different properties and limitations but are the most promising ML classifiers. More specifically, we cover the ML models most widely applied in bankruptcy prediction

(Lin et al. 2011) and the top-performing ML in data mining (Bergstein et al., 2008). This allows us to investigate the value of the ML models' properties in a rich data set characterized by structural breaks and extreme economic events. We identify the conditions under which these models gain or lose accuracy in corporate failures. For this purpose, we explore the relative importance of several time-varying covariates from an extensive set of balance sheet indicators and macroeconomic explicators in the related empirical literature. We further improve the literature by going beyond measuring models' predictive performance. To do so we report the mean decrease in accuracy to gauge variables' importance in predicting firm failures. Therefore, we provide a sparse representation of failure predictors that market participants, financial institutions and governments can readily use.

Our second contribution is that contrary to the literature, which looks at how well ML techniques predict corporate bankruptcies using listed firms, we focus on a large panel of mainly unquoted firms. There are strong reasons to suppose that a segment of firms is more likely to face problems of asymmetric information and is, therefore, more likely to be affected by higher liquidity risk and financing constraints, especially during periods of financial distress. These firms are typically small, less well-known and lack a track record. Focusing on unlisted firms allows us, therefore, to provide a crisp comparison of how ML models predict failures of different types of firms. The richness of our panel enables us to take into account two dimensions of firm heterogeneity (age and size), aimed at measuring the degree of financing constraints faced by firms. In addition, we recognize that the likelihood of firm survival is shaped by technology. Therefore, we allow for the fact that firms operating in different industrial groups might respond to changes in economic conditions disproportionately. This is also relevant because bankruptcy contagion takes place among industry peers (Chang et al., 2020). Exploiting heterogeneity at the firm and industry level is an important contribution in light of the fact that the firms in our sample are heterogeneous and are unlikely to be affected by changes in the economic uncertainty in the same way.

Third, our study spans two important episodes in the United Kingdom's recent economic history. Specifically, although the 2007–2009 crisis was a global phenomenon, Brexit is a major event with large-scale implications across the political, social and economic spectrum (Davies & Studnicka, 2018). The GFC generated heightened uncertainty about the markets, which subsided reasonably quickly, but Brexit's uncertainty was persistent and remained elevated even 3 years after the initial shock (Bloom et al., 2019). Because these two major negative episodes have different causes, scopes and implications, we separate our rich historical data set into subsamples to assess the choice of predictors for bankruptcy and the predictive ability of our models during the crisis and tranquil times. To the best of our knowledge, this channel is yet to be documented.

Previewing our main results, we find evidence that ML models with a large number of covariates are systematically more accurate than the benchmark model in the literature. Having the ML models as the starting point, interest lies in identifying the model that outperforms all other alternatives in terms of point statistics. We provide compelling evidence that the random forest model consistently beats the benchmarks and other ML tools in failure prediction. The model's superior performance stems from its variable selection mechanism and its ability to capture wider underlying patterns and relationships in the data set (Medeiros et al., 2021). In addition, we find that the model's predictive ability differs when we account for crisis/noncrisis periods and for firm- and industry-level heterogeneity. Finally, several balance sheet indicators contain information regarding firms' chances of failure. Firm-specific uncertainty is one of the key variables that plays a more potent role in young firms and during extreme economic periods. Our results are robust to various sensitivity checks.

The rest of the work is organized as follows. We document the relevant literature on failure prediction for firms and implications for managers in Section 2. Following that, we discuss data and summary statistics in Section 3. Section 4 introduces in detail the methodologies we use. In Section 5 we report the empirical results and robustness tests, with Section 6 concluding this study.

## 2 | RELATED LITERATURE

Predicting bankruptcies among general businesses and financial institutions has been on the top of the research agenda since the Great Depression (see Bellovary et al., 2007 for a detailed review). The seminal contribution of Altman (1968) prompted researchers to use multivariate analysis to predict firm bankruptcies. Following that, Ohlson (1980) proposes the O-score model in bankruptcy research. Both are classical models based on several accounting-based financial ratios to group failed and non-failed companies.

To produce more consistent estimates and more efficient out-of-sample predictions, Shumway (2001) employs a DH model with market-driven and accounting variables from previous studies; that study confirms that the predictive ability of DH outperforms discriminant analysis and logit models. The Shumway model has been modified in various ways to establish the most important predictive variables. Specifically, Chava and Jarrow (2004) investigate how industry effects influence bankruptcy prediction by extending Shumway's model. They achieve relatively higher forecasting accuracy than previous studies. Similarly, Campbell et al. (2008) introduce additional market variables to achieve a noticeable improvement in the corporate bankruptcy forecast model.[1] Other papers, including Bharath and Shumway (2008) examine the accuracy and the contribution of the Merton distance to the default model, which is based on Merton's (1974) bond pricing model. They find that the predictive power of distance to default is significant in determining corporate failures. Bonfim (2009) advocates accounting for macroeconomic conditions in assessing bankruptcy probabilities. In addition, survival analysis related to a hazard model has gradually become an important methodology for predicting failed events in finance because they capture the timing of alternative outcomes in the work (Beaver et al., 2005; Ding et al., 2012; Duffie et al., 2007). Finally, Altman et al. (2017) provide a review on the prediction of corporate failures using cross-country studies. Their study highlights that accounting-based models have good predictive ability for most countries with improvements in accuracy when country-specific estimation is carried out that incorporates additional variables.

On the methodological side, conventional models, such as the ones described earlier, are unable to identify complex patterns in millions of data points to make accurate inferences and predictions. Therefore, the literature adopts nonlinear numerical methodologies to deal with this challenge. Specifically, the support vector machine (SVM), recently introduced in default risk analysis, performs better than competing models (Chen et al., 2011; Härdle & Simar, 2012; Härdle et al., 2009). Improvements in computer technology have led to a significant reduction in computation costs and the literature investigates credit default and firm bankruptcies using variable-selection techniques or ensemble techniques.[2] For example, Tian et al. (2015) use the least absolute shrinkage and selection operator (LASSO) to evaluate the probability of bankruptcy using a comprehensive sample of US firms. The authors conclude that accuracy in

[1] Agarwal and Taffler (2008) show that there is little difference in the predictive accuracy of accounting-based and market-based models, especially for UK firms.

[2] For a recent bibliographic review on ML techniques and prediction of corporate failures, see Kim et al. (2020).

the out-of-sample prediction is superior to previous studies in estimating default by combining reduced-form models with the LASSO procedure.

Other studies use newer statistical models for forecasting corporate bankruptcy. In particular, Olson et al. (2012) provide a comparative analysis of data mining methods for bankruptcy prediction. The authors find that decision tree algorithms are more straightforward to implement and are relatively more accurate compared to neural networks and support vector machines. More recently, Traczynski (2017) breaks ground from previous studies by developing a Bayesian model-averaging approach to analysing firm bankruptcies and default predictability. The study shows that this method is out-of-sample performance superior to other common models. Finally, Barboza et al. (2017) show that ML models have, on average, approximately 10% higher accuracy in relation to traditional models.

# 3 | DATA AND SUMMARY STATISTICS

## 3.1 | Data sources

We construct our data set from the profit and loss and balance sheet data gathered by Bureau Van Dijk Electronic Publishing in the FAME database. We use the FAME August 2019, October 2010, October 2008 and February 2005 editions, as well as archived FAME 1998. In line with Guariglia et al. (2016) and Görg and Spaliara (2018), we take this approach to track the status of firms continuously and address potential attrition bias because FAME records firms within the last 5 years.[3] Our data set covers 1994–2019.

Our database includes a majority of firms (99%) not in the public market or alternative exchanges such as the alternative investment market and the off-exchange market. This is an attractive characteristic of the data set because unlisted firms are likely to suffer more from a high degree of information asymmetry compared to public firms; hence they are the most affected during extreme economic events. Following common selection criteria in the literature, we exclude companies that do not have complete records on our explanatory variables, as well as firm-years with negative sales and assets. To control for the potential influence of outliers, we winsorize the regression variables at the 5th and 95th percentiles. Finally, to prevent double-counting firms and subsidiaries or operations abroad, we keep consolidated firms. Our combined panel has an unbalanced structure containing 66,165 annual observations (firm-years) on 14,825 UK firms.[4]

## 3.2 | Choice of explanatory variables

Previous research on failure prediction accounts for both financial and business risk. Accordingly, the existing failure-prediction literature guides the selection of independent variables in our study. We present the expected relationship between applied predictors and firm failure in Supporting Information: Appendix Table A1, which provides a detailed description of the variables in this study.

[3] For example, a firm that existed before 2006 may be omitted if only the 2010 version of FAME is used. Thus, our data set tracks firm that exits up to the earlier part of the sample period.

[4] See online Supporting Information: Appendix A for details about our sample selection criteria.

### 3.2.1 | Firm-specific variables

We rely on four variables that are components of the *Z*-score created in Altman (1968). These are working capital to total assets (*WC/TA*), earnings before interest and taxes to total assets (*EBIT/TA*), retained earnings to total assets (*RE/TA*) and sales to total assets (*S/TA*). In addition, we incorporate three accounting variables from Ohlson (1980) and Traczynski (2017), namely, the ratio of net income to total assets (*NI/TA*), total liabilities to total assets (*TL/TA*) and current assets to current liabilities (*CA/CL*). Next, we use an efficiency indicator, measured as gross profits to total assets (*GP/TA*) in Psillaki et al. (2010) and Görg and Spaliara (2018). We capture firms' ability to pledge collateral via the ratio of tangible assets to total assets (*TAN/TA*), in line with Bonfim (2009), Psillaki et al. (2010) and Farinha et al. (2019). In addition, we measure firm age (*AGE*), defined as the logarithm of the difference between the current year and the date of incorporation; we measure growth prospects using growth in sales (*GRS*). Finally, in line with Byrne et al. (2016), we employ a measure of firm-specific uncertainty (*F_UNC*), by estimating an AR(1) model of sales augmented with time and industry-specific dummies and then taking the standard deviation of the firm's total real sales in the 3 years preceding and including year $t$.

### 3.2.2 | Macroeconomic indicators

We consider a list of macroeconomic factors that measure different aspects of the aggregate economy's performance; they may influence the probability of failure. Specifically, the growth rate of real gross domestic product (GDP) (*RGDPGR_UK* and *RGDPGR_US*) captures the aggregate business cycle in the United Kingdom and the United States, respectively. The interest rate (*RINTR*) is the yield on 10-year Treasury bonds in the United Kingdom minus the annual rate of inflation (*CPI*). The real effective exchange rate in the United Kingdom (*REER*) and the volatility of the real effective exchange rate (*REER_VOL*) are included. The stock market performance is the FTSE 100 return, which calculates logarithm returns on the FTSE 100 index (*LNRET*). *VOL* represents the volatility of the stock price index for the United Kingdom. Aggregate economic activity is a coincident indicator in the United Kingdom (*CIEA*). Finally, we measure policy uncertainty (*POL_UNC*), which likely plays a major role in the Brexit period. This variable is from Baker et al. (2016) and uses a 50% weight on a news-based component from the *Financial Times* and *The Times* newspapers (i.e., the mention of policy-relevant terms), as well as a 50% weight on Consensus Economics CPI and budget deficit forecaster disagreement.[5]

### 3.2.3 | Summary statistics

Table 1 reports summary statistics related to the explanatory variables in the empirical models. In columns (1)–(4), we present statistics splitting the sample between non-failed and failed

[5] We carry out Im–Pesaran–Shin (2003) unit root tests for all the series included. The statistics for all variables reject the unit root hypothesis and we conclude that the variables are stationary (see Supporting Information: Appendix Table D11). Hence, any shock affecting the variables is likely to be temporary and we should not be concerned over spurious regression.

**TABLE 1** Descriptive statistics

This table reports sample means. Standard deviations are in parentheses. *Fail* is a dummy that equals 1 in a given year if the firm is recorded as failed in that year and 0 otherwise. *GFC* is a dummy representing the recent financial crisis. It equals 1 in the years 2007–2009 and 0 otherwise. *Brexit* is a dummy representing the Brexit vote. It equals 1 in the years 2016–2019 and 0 otherwise. *Diff.* is the $p$ value of the test statistic for the equality of means. Abbreviations: *AGE*, firm age; *CA/CL*, ratio of current assets to current liabilities; *EBIT/TA*, ratio of earnings before interest and taxes to total assets; *F_UNC*, firm-specific uncertainty; *GP/TA*, ratio of gross profits to total assets; *GRS*, growth in sales; *NI/TA*, ratio of net income to total assets; *RE/TA*, ratio of retained earnings to total assets; *S/TA*, ratio of sales to total assets; *TAN/TA*, ratio of tangible assets to total assets; *TL/TA*, ratio of total liabilities to total assets; *WC/TA*, ratio of working capital to total assets.

| Variables | Full sample | Fail = 1 | Fail = 0 | *Diff.* | *GFC* = 1 | *GFC* = 0 | *Diff.* | Brexit = 1 | Brexit = 0 | *Diff.* |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| *S/TA* (%) | 16.762 | 17.408 | 16.686 | 0.000 | 16.847 | 16.757 | 0.369 | 16.015 | 16.862 | 0.000 |
| | (6.37) | (6.36) | (6.36) | | (6.52) | (6.36) | | (6.40) | (6.36) | |
| *WC/TA* (%) | 28.528 | 22.011 | 29.297 | 0.000 | 27.289 | 28.613 | 0.000 | 29.574 | 28.388 | 0.000 |
| | (20.44) | (20.89) | (20.25) | | (21.06) | (20.40) | | (21.41) | (20.31) | |
| *RE/TA* (%) | 3.138 | 1.634 | 3.316 | 0.000 | 3.036 | 3.145 | 0.285 | 3.537 | 3.085 | 0.000 |
| | (6.43) | (6.56) | (6.39) | | (6.77) | (6.40) | | (6.39) | (6.43) | |
| *EBIT/TA* (%) | 7.691 | 5.110 | 7.996 | 0.000 | 7.228 | 7.723 | 0.000 | 7.740 | 7.685 | 0.583 |
| | (8.35) | (8.26) | (8.31) | | (8.63) | (8.33) | | (7.79) | (8.42) | |
| *TL/TA* (%) | 56.903 | 63.530 | 56.121 | 0.000 | 56.150 | 56.955 | 0.021 | 51.475 | 57.627 | 0.000 |
| | (21.92) | (21.91) | (21.79) | | (21.64) | (21.94) | | (21.21) | (21.91) | |
| *CA/CL* (%) | 18.567 | 16.137 | 18.854 | 0.000 | 19.423 | 18.508 | 0.000 | 20.691 | 18.283 | 0.000 |
| | (10.15) | (8.59) | (10.28) | | (10.46) | (10.12) | | (10.88) | (10.01) | |
| *NI/TA* (%) | 5.503 | 3.274 | 5.767 | 0.000 | 5.142 | 5.528 | 0.000 | 6.140 | 5.418 | 0.000 |
| | (6.83) | (6.91) | (6.77) | | (7.07) | (6.81) | | (6.77) | (6.83) | |
| *GP/TA* (%) | 45.324 | 45.459 | 45.308 | 0.605 | 45.715 | 45.298 | 0.256 | 42.526 | 45.697 | 0.000 |
| | (23.08) | (24.16) | (22.95) | | (23.44) | (23.06) | | (21.47) | (23.27) | |

**TABLE 1** (Continued)

| **Variables** | **Full sample** | **Fail = 1** | **Fail = 0** | ***Diff.*** | ***GFC* = 1** | ***GFC* = 0** | ***Diff.*** | **Brexit = 1** | **Brexit = 0** | ***Diff.*** |
|---|---|---|---|---|---|---|---|---|---|---|
| | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** | **(7)** | **(8)** | **(9)** | **(10)** |
| *AGE* | 2.995 | 2.890 | 3.007 | 0.000 | 2.998 | 2.994 | 0.780 | 3.114 | 2.979 | 0.000 |
| | (0.73) | (0.76) | (0.73) | | (0.75) | (0.73) | | (0.68) | (0.74) | |
| *TAN/TA* (%) | 25.286 | 24.885 | 25.334 | 0.031 | 23.093 | 25.437 | 0.000 | 23.604 | 25.510 | 0.000 |
| | (16.44) | (17.17) | (16.36) | | (16.29) | (16.45) | | (15.91) | (16.50) | |
| *GRS* | 2.577 | 0.910 | 2.774 | 0.000 | 0.081 | 2.747 | 0.000 | 3.928 | 2.397 | 0.000 |
| | (16.96) | (17.75) | (16.86) | | (17.34) | (16.93) | | (14.59) | (17.25) | |
| *F_UNC* | 0.130 | 0.141 | 0.129 | 0.000 | 0.123 | 0.131 | 0.000 | 0.108 | 0.133 | 0.000 |
| | (0.09) | (0.10) | (0.09) | | (0.08) | (0.09) | | (0.08) | (0.09) | |
| No. of firms | 14,825 | 2015 | 12,810 | | 2755 | 12,070 | | 3910 | 10,915 | |
| Observations | 66,165 | 6987 | 59,178 | | 4217 | 61,948 | | 7784 | 58,381 | |

firms to measure any differences across operating statuses. We test the equality of means across the above-mentioned groups and report the corresponding $p$ values in the final columns of the table.[6] When comparing failing and surviving firms, we observe, as expected, that the latter group of firms have better financial characteristics, as measured by the balance sheet indicators. Also, they are older, as indicated by their date of incorporation. Moreover, non-failed firms have higher growth rates and less sales uncertainty. The results from equality tests suggest significant differences between the two groups, which indicates a correlation between better financial health and a lower risk of failure. In other words, there is considerable cross-sectional variation in the probability of a firm failing. This motivates our study to consider how firm heterogeneity affects failure predictions for candidate models.

In columns (5)–(10) of Table 1, we compare the GFC and Brexit periods to calmer times. During the GFC, firms display worse balance sheet indicators, such as higher leverage and lower profitability. This pattern does not hold for the Brexit period, which is characterized by better firm fundamentals. This observation supports the idea that the episodes are very different and have different implications for firm performance. The statistics suggest that firm characteristics are very different in crisis versus noncrisis periods and hence we should separate our sample into subsample periods to check the predictive ability of the models and the chosen explicators.

## 4 | METHODOLOGY

In this section, we present models that predict the failure of UK firms.[7] In line with Bunn and Redwood (2003) and Guariglia et al. (2016), we define a firm as failed in a given year when its status is that of receivership, liquidation or dissolution.

### 4.1 | DH model

The DH model is widely used as a benchmark in bankruptcy-prediction studies (see among others, Beaver et al., 2005; Campbell et al., 2008; Chava & Jarrow, 2004; Ding et al., 2012; Duffie et al., 2007; Shumway, 2001; Tian et al., 2015; Traczynski, 2017). Compared to static models, the DH model using time-varying panel data produces consistent and unbiased estimates that consider potential duration dependence (Shumway, 2001).

Let us define $Y_{s,t}$ as the dependent binary variable that equals 1 if at year $t$ a firm fails and 0 otherwise. The DH 1-year-ahead prediction of firm $s$ failing is given by the following equation:

[6] The average failure rate in our sample is 10.6%, which is comparable with previous UK studies (e.g., Guariglia et al., 2016).

[7] For details about the selection of hyperparameters and alternative shrinkage regression methods see online Supporting Information: Appendix C. In online Supporting Information: Appendix D, we show the performance of a set of additional ML models (the Multilayer Perceptron, the Recurrent Neural Network, Genetic Algorithms, bagging and boosting) that display promising results in other fields of science. As can be seen, the performance of these models is inferior to those presented in our main manuscript. Further, we show the computational cost of our main models in online Supporting Information: Appendix F.

$$Pr\left(Y_{s,t} = 1|Y_{s,t-1} = 0, X_{s,t-1}\right) = \frac{e^{\beta_0+\beta' x_{s,t}}}{1 + e^{\beta_0+\beta' x_{s,t}}}, \tag{1}$$

where $s = 1, 2..., S$ refers to firms and $t = 1, 2..., T$ stands for the time period. $X_{s,t}$ is a vector of the time-varying explanatory variables including firm-specific and macroeconomic variables for each firm $s$ at time $t$, $\beta$ is a vector of covariate effect parameters and $\beta_0$ is a scalar parameter. The parameter estimates can be achieved by maximizing the log-likelihood function[8]:

$$l(\hat{\beta}) = \sum\left(Y_{s,t} \times \ln\left(\frac{1}{1 + \exp(-(\beta_0 + \beta' X_{s,t}))}\right) + (1 - Y_{s,t}) \times \ln \left(\frac{\exp(-(\beta_0 + \beta' X_{s,t}))}{1 + \exp(-(\beta_0 + \beta' X_{s,t}))}\right)\right). \tag{2}$$

## 4.2 | The Bayesian model averaging-DH (BMA-DH)

In our study, BMA is combined into the DH model to capture the parameter uncertainty in a model through the prior distribution and solve model uncertainty using the posterior parameters in Bayes' theorem. Suppose that $M = (M_1, ..., M_m)$ is a collection of candidate DH models, each based on a subset of the full set of predictors, $X_{s,t}$. Following Equation (2), we estimate the parameters $\beta_{M_i}$ for each model $M_i$ for $i = 1...m$. The $\hat{\beta}_M$ for all the candidate models, we can estimate $M$ as:

$$\hat{\beta}_M = \sum_{i=1}^{m} \hat{\beta}_{M_i} \times p(M_i|\boldsymbol{y}), \tag{3}$$

where $p(M_i|\boldsymbol{y})$ is the posterior probability that model $M_i$ given data $\boldsymbol{y}.p(M_i|\boldsymbol{y})$ can be computed by Bayes' rule as:

$$p(M_i|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M_i)p(M_i)}{\sum_{i=1}^{m} p(\boldsymbol{y}|M_i)p(M_i)} \tag{4}$$

and $p(\boldsymbol{y}|M_i)$ is calculated by the integral:

$$p(\boldsymbol{y}|M_i) = \int f(\boldsymbol{y}|\beta_{M_i}, M_i) \times f(\beta_{M_i}|M_i) d\beta_{M_i}, \tag{5}$$

where $f(\boldsymbol{y}|\beta_{M_i}, M_i)$ is the likelihood of the data conditional on the model $M_i$ and $f(\beta_{M_i}|M_i)$ is the prior distribution of $\beta_{M_i}$. The log-likelihood function of a model $M_i$ in the BMA version of DH models can be written as:

[8] For a detailed description of the DH model, refer to Shumway (2001), Tian et al. (2015) and Traczynski (2017).

$$\ln\left(f\left(y|\beta_{0,M_i},\beta_{M_i},M_i\right)\right) = \sum\left(Y_{s,t}\times\ln\left(\frac{1}{1+\exp\left(-\left(\beta_{0,M_i}+\beta'_{M_i}X_{s,t,M_i}\right)\right)}\right) + (1-Y_{s,t})\times\ln\left(\frac{\exp\left(-\left(\beta_{0,M_i}+\beta'_{M_i}X_{s,t,M_i}\right)\right)}{1+\exp\left(-\left(\beta_{0,M_i}+\beta'_{M_i}X_{s,t,M_i}\right)\right)}\right)\right), \quad (6)$$

where $X_{s,t,M_i}$ contains the predictors in model $M_i$, which is part of $X_{s,t}$ in Equation (1). $\beta_{0,M_i}$ and $\beta_{M_i}$ are the corresponding estimated parameter vectors in model $M_i$.

## 4.3 | LASSO with DH (LASSO-DH) model

The LASSO technique, initially proposed by Tibshirani (1996), simultaneously performs predictors' selection and regression. It chooses variables by forcing some coefficients to zero and shrinking others by adding the penalty function $\sum_{q=1}^{p}|\beta_q|\leq\sigma$ into the model estimation. The constraint $\sigma$ is a user-specified tuning parameter and $q = 1, 2. ..p$ indicates the number of surviving predictors with nonzero estimated coefficients. LASSO-DH can be expressed as:

$$\hat{\beta} = argmax_{\beta}\left(\ell(\beta \mid Y_{s,t}, X_{s,t}) - \lambda\sum_{q=1}^{p}|\beta_q|\right), \quad (7)$$

where $\ell(\beta \mid Y_{s,t}, X_{s,t})$ is the same as Equation (2) and $\lambda$ is a tuning parameter that controls the shrinkage.

## 4.4 | Naive Bayes (NB) classifier

NB is a probabilistic ML classifier based on the Bayes theorem with the assumption of conditional independence between every pair of predictors given the value of the class variable. It is a computationally simple method that can handle missing values in the data set and irrelevant predictors. Although the assumption of independent predictors rarely holds in finance and economics datasets, NB can still provide accurate forecasts (Sarkar & Sriram, 2001). The NB classifier can be defined as[9]:

$$Y_{s,t} = argmax(p(Y_{s,t}|X_{s,t})) = argmax\left(p(X_{s,t}|Y_{s,t})\times\frac{p(Y_{s,t})}{p(X_{s,t})}\right), \quad (8)$$

[9] For details on the exposition of the NB classifier, see Rish (2001).

where $p(Y_{s,t}|X_{s,t})$ is the conditional probability given the predictor's vector $X_{s,t}$, $p(X_{s,t}|Y_{s,t})$ is equal to the posterior probability of the vector $X_{s,t}$ conditioned on a specific class $Y_{s,t}$, $p(Y_{s,t})$ is the prior probability of $Y_{s,t}$ and $p(X_{s,t})$ is the prior probability of the vector $X_{s,t}$. Because $p(X_{s,t})$ is constant given the input, using the conditional independence assumption, $argmax\left(p(X_{s,t}|Y_{s,t}) \times \frac{p(Y_{s,t})}{p(X_{s,t})}\right)$ in Equation (8) converts into $argmax(p(X_{s,t}|Y_{s,t})^*p(Y_{s,t}))$. Hence, Equation (8) can be written as:

$$Y_{i,t} = argmax(p(X_{s,t}|Y_{s,t})^*p(Y_{s,t})) \propto argmax(p(X_{1,t}, ..., X_{n,t}|Y_{s,t})^*p(Y_{s,t})). \tag{9}$$

## 4.5 | k-Nearest neighbour (k-NN) classifier

k-NN is a nonparametric and nonlinear classifier based on the premise that pieces of past time series have patterns that might resemble pieces of future time series. It uses the Euclidean distance to locate similar patterns of past behaviour, and, based on them, it forecasts the immediate future. The k-NN uses only local information and does not require any assumptions on the data set. As k-NN is based on predictors' similarity, it is ideal for applications such as ours, where default firms might share the same past characteristics. In this study, following the description in Murphy (2012), the probability of a specific classification for an object in the k-NN classifier can be written as:

$$p(Y_{s,t}|X_{s,t}, D, k) = \frac{1}{k} \sum_{s \in N_k(X_{s,t}, D)} \mathbf{1}(Y_{s,t}), \tag{10}$$

where $N_k(X_{s,t}, D)$ is a specific set of $k$ observations in the training data that are the closest to $X_{s,t}$ based on their Euclidean distance $D$, $X_{s,t}$ is the predictors' set, $k$ is a predefined parameter and $\mathbf{1}(e)$ is the indicator function defined as follows:

$$\mathbf{1}(e) = \begin{cases} 1, & \text{if} \quad e \quad \text{is} \quad \text{true}, \\ 0, & \text{if} \quad e \quad \text{is} \quad \text{false}. \end{cases} \tag{11}$$

Following Rodriguez et al. (2010), we select $k$ based on a 10-fold cross-validation.

## 4.6 | Support vector machines

SVM is a class of ML models introduced by Boser et al. (1992) that is widely used in modern classification and regression. Its aim is to project nonlinear separable samples onto another higher dimensional space, with the assistance of a kernel function, where the data points can be distinctly classified. The SVM algorithm tries to maximize the margin between the data points and the hyperplane by minimizing the loss function. The classification prediction takes the form:

$$sign\left(\sum_{s=1}^{F} \alpha_s Y_{s,t} K(X, X_{s,t}) + b\right), \tag{12}$$

where $F$ is the number of firms in the training sample, $\alpha$ are Lagrange multipliers, $b$ is the bias and $K$ is the kernel function. In our study, like kernel, we apply the radial basis function. The parameters of the radial basis function are selected based on a 10-fold cross-validation.

### 4.7 | Random forest (RF) model

RF is an ensemble learning method for classification tasks, which contains the collection of decision trees. RF is a way of bagging multiple, deep, decision trees trained on different parts of the same training set, with the goal of reducing the variance (Breiman, 2001). The RF model does not require statistical assumptions and handles multicollinearity. Breiman (2001) indicates that random inputs and random features tend to produce better results in RF models. Suppose that there are $d = 1, 2, \ldots, D$ bootstrap samples. In each $d$ sample, $\omega$ predictors are randomly selected from the vector $X_{s,t}$ to produce the ensemble of trees $\{T_d\}_1^D$. The classification prediction in RF model[10] can be written as:

$$\hat{Y_{s,t}}_{rf}^{D} = majority \quad vote \quad \{\hat{C}_b(X_{s,t})\}_1^D, \tag{13}$$

where $\hat{C}_b(X_{s,t})$ is the classification prediction of the $d$th RF tree.

## 5 | EMPIRICAL RESULTS

### 5.1 | Full sample

To measure the predictive performance of all competing models, we calculate the area under the receiver operating characteristic curve (AUC), the Brier score and type I and II errors (see Bharath & Shumway, 2008; Chava & Jarrow, 2004; Shumway, 2001; Tian et al., 2015; Traczynski, 2017).[11] For the out-of-sample predictions, we employ an expanding-window method based on the past and current information available up to time $T$. It allows us to include successive observations in the initial sample before forecast of the next one-step-ahead prediction of firm failure while keeping the start date of the sample fixed.

By this method, we forecast future failure $\hat{f}_{t+1}$, $\hat{f}_{t+2}$, and so forth. The initial estimation window is 1994 to 2015 and the first prediction date is 2016. We then increase $T$ by 1 each year until $T$ reaches 2019. For AUC the highest value the better is considered the model. The opposite is true for the Brier score and type I and II errors where the lowest scores suggest better models. These measures do not indicate statistical significance. For this reason, we compare the difference between two AUCs using the DeLong test (DeLong et al., 1988). The null hypothesis of the DeLong test for two models, A and B, is that the AUCs of the two models are not statistically different. We apply this procedure to examine the gain of the ML models over DH. In other words, we apply the DeLong test in pairs of models between an ML method

[10] See Hastie et al. (2008) for an excellent overview of the RF model.

[11] For a detailed description of the metrics, please see online Supporting Information: Appendix B.

and the DH. In Table 2, we report the above-mentioned statistics for the out-of-sample predictions of all candidate models.

Starting with the analysis of the AUC measure, we note that the RF outperforms all models. Using the RF model, the AUC value is around 75%, which is the highest one among all AUC values for candidate models. There exists a 12% increase in AUC values relative to the DH model, which suggests that RF clearly outperforms its DH benchmark. This gain in AUC values is related to less misclassification, which is a meaningful improvement in predictive ability. Campbell et al. (2008) and Traczynski (2017) argue that a 1% difference in predicted default probabilities is considerable for a firm, affecting its performance in the stock market. Concerning the other ML models, we note that k-NN, NB and SVM present performance that is worse than the linear DH benchmark. On the other hand, applying BMA in conjunction with DH marginally improves the DH. LASSO-DH has a statistically better AUC compared to simple DH. Our Brier scores realizations are consistent with the AUC analysis. The type I errors of our best model, the RF, are lower than their counterparts. The type II errors are close to the ones of BMA-DH and LASSO-DH.

In Table 3, we present the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model (Bharath & Shumway, 2008; Chava & Jarrow, 2004; Shumway, 2001; Tian et al., 2015; Traczynski, 2017). We opt for the decile method because depending on the rank order of firm-years may not be significantly affected by small changes in the predicted probabilities of failure, given they are unlikely to change the decile in which a firm-year lies in the distribution. The lowest probabilities of failure are in the 10th decile and the highest probabilities are in the first decile. Thus, a high proportion in high-probability deciles suggests improved accuracy of out-of-sample prediction.

**TABLE 2** Accuracy tests during 1994–2019

This table reports the out-of-sample AUC, Brier score and type I and II errors of all models. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The bold values are the highest AUC and lowest Brier scores. The ***, ** and * marks denote that the DeLong test's null hypothesis of no difference in AUCs of a machine-learning model and DH is rejected at the 1%, 5% and 10% significance level, respectively.

| | AUC (%) | Brier score | Type I error (%) | Type II error (%) |
|---|---|---|---|---|
| DH | 63.29 | 0.0124 | 36.96 | 41.73 |
| BMA-DH | 63.70 | 0.0128 | 38.31 | 37.01 |
| LASSO-DH | 64.39** | 0.0130 | 39.92 | 36.22 |
| NB | 59.97 | 0.0261 | 28.46 | 55.91 |
| k-NN | 56.23*** | 0.0334 | 45.25 | 41.73 |
| SVM | 54.19** | 0.0256 | 37.85 | 51.18 |
| RF | **74.74*****| **0.0022** | 24.71 | 38.58 |

**TABLE 3** Defaults by out-of-sample prediction decile during 1994–2019

This table reports the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. Abbreviations: AUC, area under receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The bold value denotes the higher AUC value.

| Decile | Discrete hazard (DH) | BMA-DH | LASSO-DH | NB | k-NN | SVM | RF |
|---|---|---|---|---|---|---|---|
| 1 | 22.05 | 19.69 | 24.41 | 20.47 | 14.17 | 11.02 | 34.65 |
| 2 | 10.24 | 17.32 | 12.60 | 16.54 | 10.24 | 14.96 | 18.90 |
| 3 | 14.96 | 14.96 | 16.54 | 7.09 | 11.81 | 10.24 | 13.39 |
| 4 | 13.39 | 11.81 | 9.45 | 7.09 | 12.60 | 14.17 | 10.24 |
| 5 | 6.30 | 5.51 | 6.30 | 9.45 | 13.39 | 11.02 | 5.51 |
| 6–10 | 33.07 | 30.71 | 30.70 | 39.37 | 37.81 | 38.58 | 17.31 |
| AUC (%) | 63.29 | 63.70 | 64.39 | 59.97 | 56.23 | 54.19 | **74.74** |

We show that all models display the highest percentage in the first decile compared with the rest of the deciles, with RF presenting the best predictive accuracy in the first decile. We can predict almost all failed events in the first five deciles for all competing models.[12] This suggests that all candidate models have predictive accuracy for firm failure.

Our results highlight the value of ML methods in predicting firms' probability of failure. RF clearly outperforms all other models in terms of statistical accuracy, which is consistent with previous evidence that points to the superiority of RF among a number of learning algorithms in different settings (Varian, 2014). It can also handle large datasets with higher dimensionality and identify the most important variables. These properties are highly beneficial in our data set, where the true functional form between our inputs and output is unknown and our out-of-sample includes the Brexit referendum as well as the associated noise, outliers and turbulence. We also note that DH proves a tough benchmark for NB, k-NN and SVM. In addition, SVM is unable to handle large datasets such as ours and NB is based on the assumption that the predictors are independent (a property violated in our data set) (Han et al., 2011). These elements can explain the underperformance of NB, k-NN and SVM compared to DH, which confirms its popularity in the related literature as a robust failure-prediction approach. LASSO and BMA combined with DH act as forecast combination techniques. LASSO-DH manages to provide statistically better forecasts, which we attribute to the models' shrinkage properties.

[12] According to Table 3, we observe that the 66.94% of the DH model predictions are correct in the top five deciles, which is the sum of the percentage of actual bankruptcies observed in the first five deciles (22.05% + 10.24% + 14.96% + 13.39% + 6.30% = 66.94%). Hence, we calculate that the BMA-DH, ADALLASO-DH, NB, k-NN, SVM and RF predictions are correct at 69.29%, 69.30%, 60.64%, 62.21%, 61.41% and 82.69% of the time in the top five deciles, respectively.

## 5.2 | Splitting the sample into subperiods

The preceding analysis employs the full time period (1994–2019), which spans the global financial crisis and the Brexit referendum. These extreme events caused uncertainty and stressed market conditions with implications for firm failures. To examine how these adverse economic events, affect the performance of our models, we drop the years 2017–2019 and construct a pre-Brexit period. We also explore the GFC (1994–2009) and the pre-GFC (1994–2007). The GFC and the related financial turbulence have different characteristics and natures compared to the effects of Brexit on UK firms. Examining these subperiods can act as a robustness check to our main results and reveal whether the different natures of the two crises have any impact on our models' performance.

### 5.2.1 | Pre-Brexit period (1994–2016)

In this exercise, the in-sample period spans from 1994 to 2014, with 2015 and 2016 acting as the out-of-sample period. Table 4 presents the AUC values, DeLong test statistics, Brier scores, type I and II errors for each model. We report the percentage of firm failures in each decile for each candidate model in Table 5.

We note that the performance of all models is robust in this subsample, as RF still outperforms all other algorithms under study. The decile analysis continues to report that the highest percentages of failed events are in the first deciles for all models. It is interesting to note that NB, k-NN and SVM are substantially more accurate compared to our main sample period. Models such as the NB and SVM have less noise to handle and the corresponding AUC value in each model is 5% and 4% higher, respectively. Their accuracy is now on par with the DH model. RF presents the lowest type I error and BMA-DH has the lowest type II error.

**TABLE 4** Accuracy tests during 1994–2016

This table reports the out-of-sample AUC, Brier score and type I and II errors of all models. The years 1994–2014 act as in-sample and the years 2015–2016 are out-of-sample in a rolling forward exercise. The bold values are the highest AUC and lowest Brier scores. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The ***, ** and * marks denote that the DeLong test's null hypothesis of no difference in AUCs of a machine-learning model and DH is rejected at the 1%, 5% and 10% significance level, respectively.

| | AUC (%) | Brier score | Type I error (%) | Type II error (%) |
|---|---|---|---|---|
| DH | 64.48 | 0.0442 | 41.32 | 39.62 |
| BMA-DH | 64.48 | 0.0443 | 41.62 | 39.15 |
| LASSO-DH | 64.57 | 0.0430 | 41.37 | 42.45 |
| NB | 65.26 | 0.0461 | 30.67 | 43.87 |
| k-NN | 60.06* | 0.0429 | 42.96 | 45.28 |
| SVM | 58.35** | 0.0407 | 38.60 | 48.11 |
| RF | **71.41*** ** | **0.0047** | 22.81 | 42.92 |

TABLE 5 Defaults by out-of-sample prediction decile during 1994–2016

This table reports the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model. The years 1994–2014 act as in-sample and the years 2015–2016 are out-of-sample in a rolling forward exercise. The bold value denotes the higher AUC value. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest.

| Decile | Discrete hazard (DH) | BMA-DH | LASSO-DH | NB | k-NN | SVM | RF |
|---|---|---|---|---|---|---|---|
| 1 | 26.42 | 27.36 | 26.42 | 25.00 | 16.98 | 16.04 | 27.83 |
| 2 | 11.79 | 10.85 | 13.21 | 17.45 | 14.15 | 11.32 | 21.70 |
| 3 | 8.49 | 8.96 | 7.08 | 12.26 | 11.32 | 13.21 | 10.85 |
| 4 | 11.32 | 10.85 | 9.91 | 8.49 | 8.49 | 12.26 | 8.49 |
| 5 | 8.49 | 8.96 | 11.32 | 5.66 | 8.49 | 8.96 | 8.96 |
| 6–10 | 33.49 | 33.01 | 32.07 | 31.13 | 40.56 | 38.22 | 22.17 |
| AUC (%) | 64.48 | 64.48 | 64.57 | 65.26 | 60.06 | 58.35 | **71.41** |

TABLE 6 Accuracy tests during 1994–2007

This table reports the out-of-sample AUC, Brier score and type I and II errors of all models. The years 1994–2004 act as in-sample and the years 2005–2007 are out-of-sample in a rolling forward exercise. The bold values are the highest AUC and lowest Brier scores. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The ***, ** and * marks denote that the DeLong test's null hypothesis of no difference in AUCs of a machine-learning model and DH is rejected at the 1%, 5% and 10% significance level, respectively.

| | AUC (%) | Brier score | Type I error (%) | Type II error (%) |
|---|---|---|---|---|
| DH | 61.51 | 0.1842 | 42.32 | 40.80 |
| BMA-DH | 61.58 | 0.1833 | 42.01 | 40.93 |
| LASSO-DH | 61.46 | 0.1815 | 43.67 | 38.80 |
| NB | 60.01*** | 0.3561 | 46.73 | 38.93 |
| k-NN | 59.89 | 0.1772 | 43.38 | 42.60 |
| SVM | 58.87*** | 0.1861 | 40.18 | 47.13 |
| RF | **66.44*** ** | **0.1652** | 39.21 | 37.40 |

### 5.2.2 | The global financial crisis period

In this section, we explore two subsamples related to the global financial crisis. The first subsample covers the years before the crisis (1994–2007) with the years 2005–2007 as out-of-sample. The second one includes 1994–2006; the out-of-sample years are 2007–2009. Tables 6

**TABLE 7** Accuracy tests during 1994–2009

This table reports the out-of-sample AUC, Brier score and type I and II errors of all models. The years 1994–2006 act as in-sample and the years 2007–2009 are out-of-sample in a rolling forward exercise. The bold values are the highest AUC and lowest Brier scores. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The ***, ** and * marks denote that the DeLong test's null hypothesis of no difference in AUCs of a machine-learning model and DH is rejected at the 1%, 5% and 10% significance level, respectively.

| | AUC (%) | Brier score | Type I error (%) | Type II error (%) |
|---|---|---|---|---|
| DH | 54.19 | 0.2230 | 42.96 | 54.42 |
| BMA-DH | 59.58*** | 0.1712 | 42.75 | 43.48 |
| LASSO-DH | 58.60*** | 0.1830 | 42.62 | 45.31 |
| NB | 59.12*** | 0.2290 | 30.08 | 61.44 |
| k-NN | 55.93 | 0.1527 | 35.00 | 60.62 |
| SVM | 54.07 | 0.1426 | 23.68 | 69.64 |
| RF | **66.37*****| **0.1412** | 41.34 | 36.55 |

**TABLE 8** Defaults by out-of-sample prediction decile during 1994–2007

This table reports the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model. The years 1994–2004 act as in-sample and the years 2005–2007 are out-of-sample in a rolling forward exercise. The bold value denotes the higher AUC value. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest.

| Decile | DH | BMA-DH | LASSO-DH | NB | k-NN | SVM | RF |
|---|---|---|---|---|---|---|---|
| 1 | 15.87 | 16.27 | 15.93 | 15.27 | 14.13 | 12.13 | 18.27 |
| 2 | 13.60 | 13.13 | 14.00 | 13.80 | 14.20 | 12.33 | 14.80 |
| 3 | 12.60 | 12.40 | 12.07 | 10.73 | 12.53 | 13.07 | 12.80 |
| 4 | 10.53 | 11.60 | 10.60 | 11.87 | 9.60 | 12.20 | 11.93 |
| 5 | 10.73 | 9.93 | 10.67 | 9.40 | 10.07 | 9.73 | 10.27 |
| 6–10 | 36.67 | 36.66 | 36.74 | 38.93 | 39.47 | 40.53 | 31.93 |
| AUC (%) | 61.51 | 61.58 | 61.46 | 60.01 | 59.89 | 58.87 | **66.44** |

and 7 present the related failure–accuracy metrics and Tables 8 and 9 present the percentage of firm failures in each decile for each candidate model in each subsample.

We observe a similar picture regarding the performance of our models. RF displays the best performance. It is interesting to note the imbalance of k-NN, SVM and NB models in the GFC period on the percentages of type I and type II errors. They present low type I errors but high type II errors, which is a further indication of their inability to map the data set. BMA and LASSO with DH offer an advantage during the crisis but not before. All models (except RF)

**TABLE 9** Defaults by out-of-sample prediction decile during 1994–2009

This table reports the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model. The years 1994–2006 act as in-sample and the years 2007–2009 are out-of-sample in a rolling forward exercise. The bold value denotes the higher AUC value. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest.

| Decile | DH | BMA-DH | LASSO-DH | NB | k-NN | SVM | RF |
|---|---|---|---|---|---|---|---|
| 1 | 11.85 | 14.86 | 14.59 | 18.23 | 13.49 | 10.30 | 19.60 |
| 2 | 10.85 | 13.95 | 12.12 | 10.94 | 9.75 | 12.76 | 14.31 |
| 3 | 10.67 | 11.49 | 12.67 | 8.11 | 8.66 | 12.22 | 13.49 |
| 4 | 10.12 | 10.21 | 9.66 | 12.58 | 13.49 | 9.48 | 11.12 |
| 5 | 9.66 | 11.12 | 11.21 | 10.39 | 11.39 | 12.49 | 10.39 |
| 6–10 | 46.87 | 38.37 | 39.75 | 39.74 | 43.21 | 42.75 | 31.09 |
| AUC (%) | 54.19 | 59.58 | 58.60 | 59.12 | 55.93 | 54.07 | **66.37** |

perform considerably better before the crisis than in the sample that covers it. RF seems unaffected by the extreme economic event in our data set. Our findings are consistent with those from the pre-Brexit and Brexit samples.

### 5.2.3 | Discussion

For the full sample and our three subsamples, the ranking of our models in terms of statistical accuracy is robust. RF presents the best performance for the statistical measures retained. It is impressive that, unlike other models, its improved accuracy relative to its counterparts, persists in the crisis periods. Concerning DH, it can beat more complicated methods such as SVM. Nevertheless, BMA and LASSO can improve their performance when there is turbulence in the data set. SVM, k-NN and NB present a volatile accuracy that depends on the characteristics of the data set.

## 5.3 | The role of firm-level heterogeneity

We differentiate old from young firms, whereby the latter lack track records or reputations. Young firms are less likely to weather economic and financial downturns and therefore face higher liquidation risk (Guariglia et al., 2016). In doing so, we take into account firms' relative age to separate firms that are likely more financially constrained from those that are not. We use median firm age as a cut-off in keeping with normal practice in the literature.[13]

[13] To ensure that our results are not driven by the way that we split our sample, we experiment with alternative cut-off points in the age distribution. Specifically, young firms are those in the bottom percentile of the age distribution and old firms are in the top percentile of the age distribution. The results, not reported for brevity, are similar to those obtained using the median sample splitting criterion and are available upon request.

As such, we estimate our models for two subgroups: young (old) firms whose ages are below (above) the median of the age distribution. Our goal is to assess whether the bankruptcy models account for financial constraints in their methodologies. The in-sample and out-of-sample periods are the same as in the full sample. Tables 10 and 11 report the accuracy tests and the percentage of failed events for young companies.

Based on Table 10, RF continues to significantly outperform other candidates according to both AUC and Brier scores. Compared to DH, the AUC value of RF is 11.5% higher. BMA and LASSO marginally improve the performance of DH, but the remaining three have lower classification accuracy. All results are also supported by decile methods. Tables 12 and 13 present the related results for old firms.

Comparing across columns in Tables 12 and 13 allows us to investigate the specific influence of older firms on each forecasting model. Our results are consistent with previous findings in the sense that our best model and the ranking of its benchmarks remain the same. The predictive power of the DH model improves only marginally by adding BMA and LASSO. Behind RF, the second-highest correct prediction rate in the first decile is in the NB classifier; in the top five deciles, it is the DH models.

### 5.3.1 | Discussion

When we allow for firm heterogeneity, we show that the predictions are more accurate for older firms because the AUC scores of all models are lower for their younger counterparts. This is an important finding, as the operation of young firms, for whom access to external finance is expensive, depends critically on balance sheet health and external financial conditions. Modelling their failure is a more tenacious task compared to older, well-established firms that are less risky, have management experience, track-record reputation and have sound financial statements to counter the financial turmoil that our data set covers (Robb, 2002). Our results can help policymakers and managers better assess credit risk and financing needs for younger firms. This is particularly helpful for financially constrained firms.

## 5.4 | Industry-level analysis

Next, we take into account the role of technology in determining the chances of firm failures. Previous studies show that firms that engage in more innovative activities experience a higher likelihood of survival (e.g., Audretsch, 1995; Esteve-Pérez & Mañez-Castillejo, 2008). The rationale is that firms that invest in innovation can adapt to changes in the business environment and to better respond to their customers' changing requirements.[14] Motivated by this consideration, we split the sample into high- and low-technology firms based on a two-digit Standard industrial classification of economic activities. The in-sample and out-of-sample periods are the same as in the full sample. Tables 14 and 15 report the accuracy tests and the percentage of failed events for high-tech companies.

[14] A counterargument is that firms operating in high-tech industries are likely to face lower chances of survival. As technological uncertainty rises, the probability that a firm will be able to produce a viable product will deteriorate, along with their chances of survival.

**TABLE 10** Accuracy tests during 1994–2019 for young firms

This table reports the out-of-sample AUC, Brier score and type I and II errors of all models. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. The bold values are the highest AUC and lowest Brier scores. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The ***, ** and * marks denote that the DeLong test's null hypothesis of no difference in AUCs of a machine-learning model and DH is rejected at the 1%, 5% and 10% significance level, respectively.

| | AUC (%) | Brier score | Type I error (%) | Type II error (%) |
|---|---|---|---|---|
| DH | 57.70 | 0.0155 | 36.61 | 53.73 |
| BMA-DH | 59.15* | 0.0158 | 36.02 | 44.78 |
| LASSO-DH | 59.09 | 0.0164 | 40.45 | 41.79 |
| NB | 55.90 | 0.0324 | 32.94 | 53.73 |
| k-NN | 48.16** | 0.0427 | 47.38 | 55.22 |
| SVM | 53.60 | 0.0258 | 38.49 | 47.76 |
| RF | **69.20*** ** | **0.0141** | 26.52 | 43.28 |

**TABLE 11** Defaults by out-of-sample prediction decile during 1994–2019 for young firms

This table reports the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. The bold value denotes the higher AUC value. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest.

| Decile | DH | BMA-DH | LASSO-DH | NB | k-NN | SVM | RF |
|---|---|---|---|---|---|---|---|
| 1 | 14.93 | 16.42 | 17.91 | 16.42 | 7.46 | 23.88 | 34.33 |
| 2 | 14.93 | 13.43 | 16.42 | 13.43 | 13.43 | 11.94 | 11.94 |
| 3 | 13.43 | 14.93 | 10.45 | 13.43 | 4.48 | 5.97 | 10.45 |
| 4 | 7.46 | 11.94 | 10.45 | 7.46 | 8.96 | 10.45 | 5.97 |
| 5 | 10.45 | 4.48 | 7.46 | 2.99 | 10.45 | 1.49 | 14.93 |
| 6–10 | 38.81 | 38.82 | 37.31 | 46.28 | 55.23 | 46.28 | 22.40 |
| AUC (%) | 57.70 | 59.15 | 59.09 | 55.90 | 48.16 | 53.60 | **69.20** |

In Table 14, our main finding, which is that RF significantly outperforms other modes, is upheld. This is true when considering both AUC and Brier scores. Moreover, our results are supported by decile methods. Tables 16 and 17 present the outputs for low-tech firms. We find, once again, that RF displays superior predictive ability compared to the other models.

**TABLE 12** Accuracy tests during 1994–2019 for old firms

This table reports the out-of-sample AUC, Brier score and type I and II errors of all models. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. The bold values are the highest AUC and lowest Brier scores. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The ***, ** and * marks denote that the DeLong test's null hypothesis of no difference in AUCs of a machine-learning model and DH is rejected at the 1%, 5% and 10% significance level, respectively.

| | AUC (%) | Brier score | Type I error (%) | Type II error (%) |
|---|---|---|---|---|
| DH | 68.87 | 0.0102 | 35.03 | 33.33 |
| BMA-DH | 69.12 | 0.0105 | 32.96 | 36.67 |
| LASSO-DH | 69.23 | 0.0110 | 28.90 | 36.67 |
| NB | 63.28* | 0.0239 | 25.40 | 55.00 |
| k-NN | 60.13*** | 0.0258 | 51.89 | 30.00 |
| SVM | 59.02* | 0.0180 | 30.65 | 60.00 |
| RF | **73.12*** | **0.0022** | 26.22 | 38.33 |

**TABLE 13** Defaults by out-of-sample prediction decile during 1994–2019 for old firms

This table reports the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. The bold value denotes the higher AUC value. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest.

| Decile | DH | BMA-DH | LASSO-DH | NB | k-NN | SVM | RF |
|---|---|---|---|---|---|---|---|
| 1 | 20.00 | 21.67 | 20.00 | 25.00 | 11.67 | 10.00 | 31.67 |
| 2 | 25.00 | 25.00 | 28.33 | 13.33 | 8.33 | 15.00 | 20.00 |
| 3 | 15.00 | 15.00 | 15.00 | 15.00 | 21.67 | 13.33 | 15.00 |
| 4 | 8.33 | 6.67 | 5.00 | 5.00 | 11.67 | 16.67 | 5.00 |
| 5 | 10.00 | 5.00 | 6.67 | 6.67 | 8.33 | 11.67 | 6.67 |
| 6–10 | 21.67 | 26.66 | 25.00 | 34.99 | 38.33 | 33.33 | 21.66 |
| AUC (%) | 68.87 | 69.12 | 69.23 | 63.28 | 60.13 | 59.02 | **73.12** |

### 5.4.1 | Discussion

Technological intensity can shape the likelihood of firm survival (Audretsch, 1991). In this subsection, we allow for industry-level heterogeneity by separating firms according to their degree of technological innovation. The resulting models show that we achieve improved predictive ability for firms operating in high-tech industries. Importantly, the predictions for the low-tech group of firms do not deteriorate significantly compared to their counterparts. Our

TABLE 14 Accuracy tests during 1994–2019 for high-tech firms

This table reports the out-of-sample AUC, Brier score and type I and II errors of all models. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. The bold values are the highest AUC and lowest Brier scores. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The ***, ** and * marks denote that the DeLong test's null hypothesis of no difference in AUCs of a machine-learning model and DH is rejected at the 1%, 5% and 10% significance level, respectively.

| | AUC (%) | Brier score | Type I error (%) | Type II error (%) |
|---|---|---|---|---|
| DH | 63.86 | 0.0123 | 37.06 | 39.22 |
| BMA-DH | 71.86*** | 0.0012 | 36.60 | 38.24 |
| LASSO-DH | 65.40** | 0.0112 | 37.63 | 38.24 |
| NB | 62.22 | 0.0282 | 40.56 | 44.12 |
| k-NN | 57.69** | 0.0335 | 41.32 | 48.04 |
| SVM | 60.96 | 0.0238 | 36.43 | 43.14 |
| RF | **73.56*** ** | **0.0011** | 27.80 | 36.27 |

TABLE 15 Defaults by out-of-sample prediction decile during 1994–2019 for high-tech firms

This table reports the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. The bold value denotes the higher AUC value. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest.

| Decile | DH | BMA-DH | LASSO-DH | NB | k-NN | SVM | RF |
|---|---|---|---|---|---|---|---|
| 1 | 22.55 | 32.35 | 26.47 | 21.57 | 15.69 | 17.65 | 35.29 |
| 2 | 12.75 | 15.69 | 11.76 | 17.65 | 7.84 | 21.57 | 20.59 |
| 3 | 12.75 | 15.69 | 17.65 | 6.86 | 15.69 | 11.76 | 7.84 |
| 4 | 12.75 | 7.84 | 7.84 | 6.86 | 11.76 | 10.78 | 8.82 |
| 5 | 6.86 | 7.84 | 4.90 | 9.80 | 12.75 | 4.90 | 8.82 |
| 6–10 | 32.34 | 20.58 | 31.36 | 37.24 | 36.26 | 33.32 | 18.62 |
| AUC (%) | 63.86 | 71.86 | 65.40 | 62.22 | 57.69 | 60.96 | **73.56** |

findings highlight the importance for financial managers to invest in innovative products to improve their chances of survival.

## 5.5 | Variable importance for the RF model

Although accurate failure predictions are valuable, financial managers are also interested in the importance of the predictors that drive these potential outcomes. RF is the most accurate

TABLE 16 Accuracy tests during 1994–2019 for low-tech firms

This table reports the out-of-sample AUC, Brier score and type I and II errors of all models. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. The bold values are the highest AUC and lowest Brier scores. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest. The ***, ** and * marks denote that the DeLong test's null hypothesis of no difference in AUCs of a machine-learning model and DH is rejected at the 1%, 5% and 10% significance level, respectively.

| | AUC (%) | Brier score | Type I error (%) | Type II error (%) |
|---|---|---|---|---|
| DH | 61.59 | 0.0133 | 35.06 | 40.00 |
| BMA-DH | 69.19** | 0.0132 | 36.03 | 36.00 |
| LASSO-DH | 62.14 | 0.0141 | 34.68 | 44.00 |
| NB | 52.50 | 0.0163 | 47.00 | 52.00 |
| k-NN | 51.33 | 0.0348 | 39.82 | 60.00 |
| SVM | 66.59 | 0.0219 | 35.17 | 36.00 |
| RF | **72.68*** | **0.0028** | 32.20 | 36.00 |

TABLE 17 Defaults by out-of-sample prediction decile during 1994–2019 for low-tech firms

This table reports the percentage of failed events in each decile of the predicted distribution for the probability of failure and the corresponding AUC for each candidate model. The years 1994–2015 act as in-sample and the years 2016–2019 are out-of-sample in a rolling forward exercise. The bold value denotes the higher AUC value. Abbreviations: AUC, area under the receiver operating characteristic curve; DH, discrete hazard; BMA-DH, Bayesian model averaging; LASSO-DH, least absolute shrinkage and selection operator; NB, Naive Bayes; k-NN, k-nearest neighbour; SVM, support vector machine; RF, random forest.

| Decile | DH | BMA-DH | LASSO-DH | NB | k-NN | SVM | RF |
|---|---|---|---|---|---|---|---|
| 1 | 8.00 | 24.00 | 12.00 | 16.00 | 0.00 | 20.00 | 24.00 |
| 2 | 20.00 | 20.00 | 24.00 | 12.00 | 16.00 | 20.00 | 16.00 |
| 3 | 12.00 | 4.00 | 4.00 | 8.00 | 8.00 | 12.00 | 16.00 |
| 4 | 20.00 | 24.00 | 16.00 | 4.00 | 16.00 | 12.00 | 16.00 |
| 5 | 8.00 | 16.00 | 8.00 | 12.00 | 0.00 | 4.00 | 16.00 |
| 6–10 | 32.00 | 12.00 | 36.00 | 48.00 | 60.00 | 32.00 | 12.00 |
| AUC (%) | 61.59 | 69.19 | 62.14 | 52.50 | 51.33 | 66.59 | **72.68** |

classifier in all previous settings. However, it does not allow interpretation of coefficient estimates, because it is not possible to measure how much each variable contributes to the final split-up into all final nodes (de Moor et al., 2018). To this end, we report results on the mean decrease in accuracy to gauge the mean loss of accuracy when we exclude each specific predictor from the algorithm.[15] In Table 18, we present the top 10 variables in terms of importance for all our subsamples.

We find that the covariates that are meant to measure firms' financial health are by far the most important. Specifically, profitability indicators and the degree to which a firm is well collateralized behave as conjectured and strongly affect firms' chances of failure in line with prior literature (e.g., Traczynski, 2017). Notably, firm-specific uncertainty is a key factor affecting firm survival probabilities when the sample covers the Brexit event (column 1). We also observe that the importance of firm uncertainty persists in the sample that refers to young firms. The firm-specific uncertainty indicator (measured by the volatility of sales) addresses the concept that uncertainty varies over time and this process affects managers' response in an environment where access to financing is hard or prohibitively expensive. Hence, our finding is important as it confirms the impact of demand uncertainty on financially constrained firms' performance. To the best of our knowledge, this is the first paper to make this point and we document the role of sales volatility when it comes to firm failure using ML models. In other words, we show that uncertainty about future sales and demand conditions can generate fluctuations in firms' chances of failure. This can be explained as higher levels of uncertainty generate a temporary slowdown and bounce back as firms postpone their activities and projects and wait for uncertainty to subside (Bloom et al., 2007; Byrne et al., 2016). In summary, investors and managers assign slightly different weights to the variables under consideration, but there is consensus about the most important determinants of the firm probability of failure in tranquil and in crisis periods.

## 5.6 | Robustness tests

We conduct six additional tests of the results we report in the main section. We summarize these additional robustness tests below, but we do not report them due to space constraints. They are available upon request.

First, to consider the potential impact of autocorrelation (and hence history) of the explanatory variables in the models, we augment the current set of predictors by using the variables in levels and their first lags.[16] We present the results with the extended sample in online Supporting Information: Appendix Tables E1 and E2. If anything, we continue to observe that the RF model continues to beat all other methods. Hence, our main findings are robust to an extended set of predictors that account for past history.

Second, we check whether our results on firm heterogeneity are robust to using a different firm splitting criterion. As such, we rely on firms' size which is a classification related to the well-established empirical financing constraints literature (see e.g., Farinha et al., 2019). We use median firm size as a cut-off point and report accuracy tests and the percentile of failed firms in online Supporting Information: Appendix Tables E3–E6. We confirm our main findings. Our models perform better for larger firms and the RF model has the highest proportion of correct predictions. We conclude, therefore, that our models uphold their predictive ability when we conduct out-of-sample exercises using an alternative criterion for firm heterogeneity.

[15] We construct an alternative measure based on the Gini impurity. It measures the average decrease in Gini-impurity across the forest. Our results are robust to using both measures.

[16] We consider using deeper lags and our results remain intact.

**TABLE 18** Mean decrease accuracy predictors' importance

This table reports the top 10 predictors in terms of mean decrease in accuracy. The values in the parentheses are the relevant metrics. Abbreviations: *AGE*, firm age; *CA/CL*, ratio of current assets to current liabilities; *EBIT/TA*, ratio of earnings before interest and taxes to total assets; *F_UNC*, firm-specific uncertainty; *GP/TA*, ratio of gross profits to total assets; *GRS*, growth in sales; *NI/TA*, ratio of net income to total assets; *RE/TA*, ratio of retained earnings to total assets; *S/TA*, ratio of sales to total assets; *TAN/TA*, ratio of tangible assets to total assets; *TL/TA*, ratio of total liabilities to total assets; *WC/TA*, ratio of working capital to total assets.

| Data set | 1994–2019 | 1994–2016 | 1994–2009 | 1994–2007 | Young | Old | High-tech | Low-tech |
|---|---|---|---|---|---|---|---|---|
| 1 | *EBIT/TA*<br>(98.689) | *EBIT/TA*<br>(174.725) | *EBIT/TA*<br>(70.845) | *EBIT/TA*<br>(70.272) | *EBIT/TA*<br>(78.141) | *EBIT/TA*<br>(77.691) | *EBIT/TA*<br>(93.308) | *EBIT/TA*<br>(45.507) |
| 2 | *TL/TA*<br>(72.702) | *TL/TA*<br>(134.277) | *S/TA*<br>(48.662) | *S/TA*<br>(48.904) | *TL/TA*<br>(66.742) | *GP/TA*<br>(58.744) | *TL/TA*<br>(81.234) | *NI/TA*<br>(41.907) |
| 3 | GP/TA<br>(59.741) | *CIEA*<br>(123.987) | *TL/TA*<br>(48.301) | *TL/TA*<br>(42.301) | *F_UNC*<br>(53.476) | *TL/TA*<br>(58.454) | *GP/TA*<br>(73.372) | *REER*<br>(35.427) |
| 4 | *TAN/TA*<br>(58.883) | *GP/TA*<br>(118.809) | *GP/TA*<br>(45.010) | *GP/TA*<br>(40.694) | *S/TA*<br>(46.567) | *NI/TA*<br>(58.367) | *S/TA*<br>(68.640) | *WC/TA*<br>(33.420) |
| 5 | *F_UNC*<br>(58.188) | *S/TA*<br>(113.074) | *TAN/TA*<br>(43.212) | *TAN/TA*<br>(38.662) | *GP/TA*<br>(44.562) | *TAN/TA*<br>(54.721) | *NI/TA*<br>(59.273) | *TL/TA*<br>(30.794) |
| 6 | *S/TA*<br>(57.952) | *NI/TA*<br>(109.264) | *NI/TA*<br>(40.155) | *NI/TA*<br>(38.204) | *TAN/TA*<br>(43.456) | *S/TA*<br>(53.188) | *TAN/TA*<br>(57.784) | *CA/CL*<br>(29.138) |
| 7 | *NI/TA*<br>(51.229) | *REER*<br>(103.331) | *AGE*<br>(37.418) | *RE/TA*<br>(33.2438) | *RE/TA*<br>(42.421) | *CA/CL*<br>(45.412) | *CIEA*<br>(52.498) | *RE/TA*<br>(29.121) |
| 8 | *RE/TA*<br>(46.463) | *TAN/TA*<br>(99.242) | *RE/TA*<br>(34.785) | *AGE*<br>(32.8631) | *NI/TA*<br>(37.647) | *RE/TA*<br>(44.472) | *F_UNC*<br>(49.966) | *AGE*<br>(28.963) |
| 9 | *AGE*<br>(44.417) | *AGE*<br>(86.406) | *F_UNC*<br>(34.491) | *F_UNC*<br>(30.9323) | *CIEA*<br>(36.129) | *F_UNC*<br>(41.821) | *AGE*<br>(49.408) | GP/TA<br>(27.690) |
| 10 | *GRS*<br>(43.348) | *F_UNC*<br>(86.337) | *CIEA*<br>(33.082) | *CIEA*<br>(29.0717) | *GRS*<br>(34.960) | *WC/TA*<br>(36.025) | *RE/TA*<br>(45.237) | *TAN/TA*<br>(24.928) |

Third, we opt for estimating models that incorporate time and industry fixed effects, which are meant to control for business cycle effects and industry differences, respectively. We tabulate all related results in online Supporting Information: Appendix Tables E7 and E8. We confirm that controlling for time and industry effects can improve the predictive accuracy of models related to the DH model but adding time effects cannot significantly influence the predictive ability of simple classifiers in the whole sample. To sum up, the RF model generates more accurate predictions than other models, which confirms our main findings.

Fourth, we create an early warning distress indicator to account for the fact that firms in financial distress do not necessarily go out of business due to legislation on exit and restructuring barriers (Farinha et al., 2019). To this end, we employ FAME's rich information about firm credit ratings, which measure the likelihood of company failure in the 12 months following the date of calculation. A rating score ranges from 0 to 100, with higher values indicating improved financial health and lower risk. We create a dummy variable that equals 1 if the firms attain a score of 40 or below and 0 otherwise. The results in online Supporting Information: Appendix Tables E9 and E10 corroborate our main findings. In the out-of-sample exercises, the RF model provides more accurate forecasts than its DH benchmark. To sum up, even when employing an early warning indicator, the out-of-sample predictions show that the RF models outperform the benchmark model.

Fifth, we check the sensitivity of our results given that failed firms are under-sampled relative to surviving firms. We create a balanced sample by year in the following way. For each year, we get a random sample of $X$ non-failed firms, where $X$ is the number of failed firms in that year to construct a balanced data set. The results, reported in Tables E12 and E13, corroborate our main findings.

Finally, our results may be sensitive to the inclusion of listed firms in the sample.[17] We address this issue by removing the listed firms and re-assessing our models. In Tables E14 and E15, we continue to observe that RF displays the highest predictive ability compared to the other competing models. Therefore, our results are not affected by the inclusion of listed firms.

## 6 | CONCLUSION

A corporate failure should be carefully assessed because it can bring significant wealth losses for market participants and potentially lead to economic depression. Thus, a reasonable margin of accuracy in failure predictions can bring many benefits for market participants, firm managers and policymakers. In this study, we model the prediction of failure using the DH, the BMA-DH models, the LASSO-DH, the NB, the k-NN and the RF ML classifiers. We rely on annual data of firm-specific factors and macroeconomic variables for a period of about 20 years (1994–2019) as the input in the benchmark model (the DH model) and all competing models. This model selection not only follows the literature of binary-dependent variable models but also compares the predictive performance of the reduced-form models and some of the more promising ML models.

Our results show that ML classifiers offer significant gains in predictive ability. When comparing all candidate models, the RF model performs better in out-of-sample prediction than DH models, mostly adopted in previous studies. In addition, we find that the two major economic episodes and firm and industry heterogeneity can influence the predictive power of each candidate model. These results suggest that RF is widely applicable for predicting failure because it does not require a priori knowledge of this method and does not need to satisfy assumptions carefully. Over time, the reasons for failure will change, which implies that the best model would also change. To solve the parameter and model uncertainty, the BMA version of the DH model is a reasonable model selection to improve predictive accuracy in further research.

This study has implications for managers, particularly during periods of extreme economic events. We conclude that managers should rely on rich accounting and macroeconomic data to improve their assessment of business failures. In addition, our results can help policymakers and managers better assess credit risk and financing needs for younger firms. However, our

[17] As noted in the data description, our sample contains only a small number of listed firms (approx. 1% of the total sample). We check whether the distance to default (or Merton DD) is a valuable predictor for the sample of listed firms. Our findings, despite the limited number of observations, confirm the superiority of RF to predict firm failures. The results of this exercise are available upon request.

findings confirm to some extent the scepticism of several academics and practitioners in machine learning. Our study does not argue that all ML models have value in failure prediction. We argue that specific models (such as the RF or LASSO and BMA combined with DH in some settings) can offer a comparative advantage to financial managers. ML should not be treated as a panacea but as a set of promising specifications that require good technical knowledge for successful application. In particular, ML models can support the development of risk-management strategies using both operational and financial hedging when dealing with companies in financial distress. For the correct choice of ML predictor, managers should be aware of the different properties that these models possess.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the Bureau Van Dijk Electronic Publishing in the FAME database. Restrictions apply to the availability of these data, which were used under license for this study.

## REFERENCES

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking and Finance*, *32*, 1541–1551.

Akyildirim, E., Nguyen, D. K., Sensoy, A., & Šikić, M. (2021). Forecasting high-frequency excess stock returns via data analytics and machine learning. *European Financial Management*, 1–54.10.1111/eufm.12345

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

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management and Accounting*, *28*, 131–171.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalised random forests. *The Annals of Statistics*, *47*, 1148–1178.

Audretsch, D. B. (1991). New firm survival and the technological regime. *Review of Economics and Statistics*, *73*, 441–450.

Audretsch, D. B. (1995). *Innovation and industry evolution*. The MIT Press.

Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2021). Machine learning in finance: A topic modeling approach. *European Financial Management*, 1–27. 10.1111/eufm.12326

Baker, S., Bloom, N., & Davies, S. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, *131*, 1593–1636.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems With Applications*, *83*, 405–417.

Beaver, W. H., McNichols, M. F., & Rhie, J.-W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies*, *10*, 93–122.

Bellovary, J., Giacomino, D., & Akers, M. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, *33*, 1–42.

Bergstein, D. A., Ozkumur, E., Wu, A. C., Yalçin, A., Colson, J. R., Needham, J. W., Irani, R. J., Gershoni, J. M., Goldberg, B. B., Delisi, C., Ruane, M. F., & Unlü, M. S. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*, 1–37.

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *The Review of Financial Studies*, *21*, 1339–1369.

Bloom, N., Bond, S., & van Reenen, J. (2007). Uncertainty and investment dynamics. *Review of Economic Studies, 74*, 391–415.

Bloom, N., Bunn, P., Chen, S., Mizen, P., Smietanka, P., & Thwaites, G. (2019). *The impact of Brexit on UK firms* (Working paper no. 26218). NBER.

Bonfim, D. (2009). Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking and Finance*, *33*, 281–299.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *Proceedings of the fifth annual workshop on Computational learning theory*. (pp. 144–152). ACM Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Bunn, P., & Redwood, V. (2003). Company accounts-based modelling of business failures and the implications for financial stability (Working paper no. 210). Bank of England.

Byrne, J. P., Spaliara, M. E., & Tsoukas, S. (2016). Firm survival, uncertainty and financial frictions: Is there a financial uncertainty accelerator? *Economic Inquiry*, *54*, 375–390.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63*, 2899–2939.

Chang, Y., Hsieh, Y. T., Liu, W., & Miu, P. (2020). Intra-industry bankruptcy contagion: Evidence from the pricing of industry recovery rates. *European Financial Management*, *26*, 503–534.

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *European Finance Review*, *8*, 537–569.

Chen, S., Härdle, W., & Moro, R. (2011). Modeling default risk with support vector machines. *Quantitative Finance*, *11*(1), 135–154.

Davies, R., & Studnicka, Z. (2018). The heterogeneous impact of Brexit: Early indications from the FTSE. *European Economic Review*, *110*, 1–17.

DeLong, E., DeLong, D., & Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

Dimitras, A., Zanakis, S., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, *90*, 487–513.

Ding, A., Tian, S., Yu, Y., & Guo, H. (2012). A class of discrete transformation survival models with application to default probability prediction. *Journal of the American Statistical Association*, *107*, 990–1003.

Doumpos, M., Andriosopoulos, K., Galariotis, E., Makridou, G., & Zopounidis, C. (2017). Corporate failure prediction in the European energy sector: A multicriteria approach and the effect of country characteristics. *European Journal of Operational Research*, *262*, 347–360.

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

Esteve-Pérez, S., & Mañez-Castillejo, J. A. (2008). The resource-based theory of the firm and firm survival. *Small Bussiness Economics*, *30*, 231–249.

Farinha, L., Spaliara, M.-E., & Tsoukas, S. (2019). Bank shocks and firm performance: New evidence from the sovereign debt crisis. *Journal of Financial Intermediation*, *40*, 100818.

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

Görg, H., & Spaliara, M.-E. (2018). Export market exit and financial health in crises periods. *Journal of Banking and Finance*, *87*, 150–163.

Guariglia, A., Spaliara, M. E., & Tsoukas, S. (2016). To what extent does the interest burden affect firm survival? Evidence from a panel of UK firms during the recent financial crisis. *Oxford Bulletin of Economics and Statistics*, *78*, 576–594.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and techniques, 3rd ed., Burlington: Elsevier Science.

Härdle, W., & Simar, L. (2012). Applied multivariate statistical analysis (3rd ed.). Springer Verlag.

Härdle, W., Lee, Y. J., Schäfer, D., & Yeh, Y. -R. (2009). Variable selection and oversampling in the use of smooth support vector machine for predicting the default risk of companies. *Journal of Forecasting*, *28*, 512–534.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). Random forests, *The elements of statistical learning* (pp. 587–604). Springer.

Im, K. S., Pesaran, M. H., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, *115*, 53–74.

Kim, H., Cho, H., & Ryu, D. (2020). Corporate default predictions using machine learning: Literature review. *Sustainability*, *12*, 6325.

Knaus, M., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte Carlo evidence. *Econometrics Journal*, *24*, 134–161.

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

Lin, W. Y., Hu, Y. H., & Tsai, C. F. (2011). Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, *42*, 421–436.

Medeiros, M., Vasconcelos, G., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business and Economic Statistics*, *39*, 98–119.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, *29*, 449–470.

de Moor, L., Luitel, P., Sercu, P., & Vanpee, R. (2018). Subjectivity in sovereign credit ratings. *Journal of Banking and Finance*, *88*, 366–392.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective'*. MIT Press.

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

Olson, D., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*, 464–473.

Psillaki, M., Tsolas, I., & Margaritis, D. (2010). Evaluation of credit risk based on firm performance. *European Journal of Operational Research*, *201*, 873–881.

van Reenen, J. (2016). Brexit's long-run effects on the U.K. economy. *Brookings Papers on Economic Activity*, *47*, 367–383.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46.

Robb, A. (2002). Small business financing: Differences between young and old firms. *Journal of Enterprenerial Finance*, *7*, 45–64.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 569–575.

Sarkar, S., & Sriram, R. S. (2001). Bayesian models for early warning of bank failures. *Management Science*, *47*, 1457–1475.

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

Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking and Finance*, *52*, 89–100.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.

Traczynski, J. (2017). Firm default prediction: A Bayesian model-averaging approach. *Journal of Financial and Quantitative Analysis*, *52*, 1211–1245.

Varian, H. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*, 3–28.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Sermpinis, G., Tsoukas, S., & Zhang, Y. (2023). Modeling failure rates with machine-learning models: Evidence from a panel of UK firms. *European Financial Management*, 29, 734–763. https://doi.org/10.1111/eufm.12369