Malikkidou, Despo; Strohbach, Wolfgang

Working Paper

Predicting bank distress in Europe: Using machine learning and a novel definition of distress

EBA Staff Paper Series, No. 21

Provided in Cooperation with:
European Banking Authority (EBA), Paris La Défense

Suggested Citation: Malikkidou, Despo; Strohbach, Wolfgang (2025) : Predicting bank distress in Europe: Using machine learning and a novel definition of distress, EBA Staff Paper Series, No. 21, ISBN 978-92-9245-978-9, European Banking Authority (EBA), Paris La Défense, https://doi.org/10.2853/8659719

This Version is available at:
https://hdl.handle.net/10419/311181

European Banking Authority

EBA STAFF PAPER SERIES: N. 21 – JANUARY 2025

PREDICTING BANK DISTRESS IN EUROPE

USING MACHINE LEARNING AND A NOVEL DEFINITION OF DISTRESS

by Despo Malikkidou and Wolfgang Strohbach

ABSTRACT

This paper develops an early warning system for predicting distress for large European banks. Using a novel definition of distress derived from banks' headroom above regulatory requirements, we investigate the performance of three machine learning techniques against the traditional logistic model. We find that the random forest model shows superior performance both out-of-sample and out-of-time. Unlike previous studies, we also employ a series of sampling techniques showing that they significantly improve the ability to identify distress events irrespective of the model used. Moreover, we show that ensemble techniques can help improve performance relative to the single best performing model. Finally, using the latest machine learning interpretability tools, we show that the variables closely tied to bank profitability and solvency are important drivers for predicting bank distress. Overall, our paper has important practical implications for bank supervisors and macroprudential authorities who can utilise our findings to identify bank weaknesses ahead of time and adopt pre-emptive measures to safeguard financial stability.

KEYWORDS

Bank distress; early warning system; machine learning; neural networks; decision tree; random forest; risk assessment; banking supervision

JEL CODES

C14; C33; C38; C45; C52; C53; G21

1. Introduction

The primary objective of banking supervision is to safeguard the stability of the banking system and protect depositors from bank failure. This pivotal role requires a comprehensive evaluation of banks' capacity to withstand future adverse economic developments and shocks. This evaluation extends beyond a current point-in-time assessment, including a forward-looking perspective on banks' future sustainability.

To accomplish this task, supervisory authorities often deploy early warning systems designed to identify in advance banks that are prone to face financial difficulties in the future. The results of this analysis can serve as the foundation for implementing targeted measures and corrective actions to strengthen banks' resilience to financial distress. These proactive interventions can prevent banks of facing vulnerabilities in the first place, enhancing the resilience of the financial sector overall.

In recent years, early warning systems have also proved to be useful tools beyond micro-prudential supervision. Detken et al. (2014) use early warning models to identify leading indicators that can be helpful in guiding macroprudential authorities on the activation and release of the countercyclical capital buffer (CCyB). In addition, Lo Duca & Peltonen (2013), Aldasoro, Borio, & Drehmann (2018) and Tölö (2020) rely on early warning indicators to predict systemic banking crisis or recessions, while Lang, Peltonen, & Sarlin (2018) illustrate how the traditional bank-level early warning systems can support both micro- and macro-prudential policy analysis.

Traditionally, logistic regression models have been used in designing early warning systems. Various academic papers found that machine learning approaches are superior to traditional techniques, which are unable to deal with the complexities and non-linearities that exist in the underlying relationships. Suss & Treitel (2019) for example, compared the performance of six different approaches for predicting distress of UK banks: two logistic regression models and four machine learning approaches – K-Nearest Neighbours (KNN), Random Forest, Boosting, and Support Vector Machines (SVM). The results show that machine learning approaches, in particular the Random Forest algorithm significantly and substantively outperform traditional techniques. They also demonstrate the benefits of ensemble techniques and find that a stacked ensemble using a linear regression as the second-level model provides better results than the Random Forest alone.

Similarly, Le & Viviani (2018) assessed various traditional statistical techniques and machine learning techniques to predict bank failures in the US. They found that machine-learning techniques are more accurate than traditional techniques, and in particular Artificial Neural Networks and K-Nearest Neighbour methods. They do not, however, assess the Random Forest algorithm. Petropoulos, Siakoulis, Stavroulakis, & Vlachogiannakis (2020) also applied machine learning modelling techniques to predict bank insolvencies of US-based financial institutions, with their results confirming that the method of Random Forests has a superior performance when compared with other techniques like logistic regression, linear discriminant analysis, Support Vector Machines or Artificial Neural Networks.

Building on the existing academic literature, our paper develops an early warning system to predict distress for large banks in the EU using machine learning techniques. Our work contributes to the literature in several ways. First, we make use of a unique and comprehensive supervisory dataset for a sample of large EU banks between 2017 - 2023. Previous studies have mainly focused on US or UK banks or small EU banks. Part of the reason is that outright failures of large EU banks do not happen often, making the estimation of early warning systems particularly challenging. To overcome this problem, we use a novel approach in defining bank distress events, which is directly aligned with the supervisory risk assessment framework. In addition, we utilise sampling techniques to account for the imbalance between distress and non-distress events, improving the model's performance, particularly for predicting distress events.

We compare four distinct early warning systems employing Random Forest, Logistic Regression, Neural Network, Decision Tree models to forecast bank distress. Our analysis reveals that the Random Forest model consistently

outperforms the other approaches both out of sample and out of time, particularly in its capacity to identify distress events. Unlike previous studies, we also employ a series of sampling techniques showing that they significantly improve the ability to identify distress events irrespective of the model used. The most influential variables, as determined by Shapley values, are closely tied to bank profitability, including average interest expense of deposits and asset-deposit spread for non-financial corporations, as well as solvency indicators such as equity to total liabilities and equity. We also show that ensemble techniques can sometimes perform better than the single best performing model. Our findings are robust across various prediction horizons and alternative definitions of distress, providing valuable insights for the development of effective early warning systems in the banking sector.

The rest of the paper is organised as follows. Section 2 describes the data used for the construction of the bank distress events and the explanatory variables. Section 3 describes the different machine learning models employed in the paper. Section 4 presents the results and robustness tests. Section 5 concludes the paper.

2. Data

2.1. Sample

Our sample covers 176 banks from 27 EEA countries which reported supervisory data at the highest level of EU/EEA consolidation from 2017 Q1 to 2023 Q3 (Table 10 in the Annex).[1] The data period includes the COVID pandemic, which triggered a sharp economic shock in 2020 followed by a rebound in 2021 and 2022. For the banking sector, the impact was less severe due to the introduction of various public support measures and European banks have reported solid profitability and strong solvency positions throughout that period.[2] We exclude data for public banks, United Kingdom's (UK) banks and banks with data of insufficient quality.[3]

While the EBA started to collect supervisory data for all banks in the EU/EEA from 2020 Q4 as part of EUCLID, it still does not collect balance-sheet and income-statement data for most of the medium-sized and small banks.[4] Therefore, we restrict our sample to the largest banks in the EU/EEA, given that we are not able to construct most of the explanatory variables used in our models for the remaining banks (see section 3.3 for details).

2.2. Distress events

Traditionally, early warning systems use outright failures (e.g. insolvencies, bankruptcies, liquidations and defaults) to represent distress. However, these distress events are rare among large banks, making the estimation of early warning systems in the banking sector particularly challenging. To overcome this problem, several authors have relaxed the traditional notion of distress and instead used an 'extended' definition, which encompasses a wider range of distress episodes beyond outright failures.

Betz, Oprică, Peltonen, & Sarlin (2014) take into account state interventions and forced mergers to capture bank distress for large European banks. Suss & Treitel (2019) make use of confidential supervisory assessments on the riskiness of UK banks and building societies between 2006 and 2012. They classify banks in distress as those that have received a 'high risk' score by bank supervisors. Moreover, Ferriani et al. (2019) identify distress events on the basis of the Italian regulatory framework using information on compulsory administrative liquidation, extraordinary administration, temporary administration voluntary liquidation, merger in distress, disposal of assets, resolution, intervention of the depositors' guarantee funds, and notification of financial deterioration to the ECB. Similarly, Bräuning, Malikkidou, Scalone, & Scricco (2020) use a series of early warning events prescribed in the European Bank Recovery and Resolution Directive, in short BRRD (EU, 2014b). These include failing or likely to fail triggers (Article 32 of the BRRD), early intervention triggers (Article 27 of the BRRD), special or temporary administration and notifications of financial deterioration to the ECB.[5]

[1] While supervisory reporting started in 2014 Q1, we restrict our sample to the period after 2017 Q1 as some of the main variables used to construct the distress events were only available after this date.

[2] As a robustness check, we carry out the analysis by excluding the Covid period between 2020 Q1-Q3 and the results are qualitatively similar. The results are not presented in the paper for the sake of brevity.

[3] Following the UK's departure from the EU, supervisory reporting data for banks domiciled in the UK are no longer collected by the EBA as of 2020 Q2.

[4] https://www.eba.europa.eu/sites/default/documents/files/document_library/News%20and%20Press/Communication%20materials/Factsheets/1025098/Factsheet%20on%20EUCLID.pdf

[5] For more details on the failing or likely to fail triggers and early intervention triggers see the EBA guidelines on failing or likely to fail (EBA/GL/2015/07) and EBA guidelines on early intervention triggers (EBA/GL/2015/03)

In this paper we adopt a new definition of distress, which aims to capture "weak" banks that merit higher supervisory attention. By doing so, we can build an early warning system which identifies weak banks at an early stage and provides enough time for supervisors to intervene. This is particularly useful for supervisors who are often interested in early signs of distress, rather than outright bank failures.

The Basel Committee on Banking Supervision defines a weak bank as one "whose liquidity or solvency is impaired or will soon be impaired unless there is a major improvement in its financial resources, risk profile, business model, risk management systems and controls, and/or quality of governance and management in a timely manner" (BCBS, 2015).6 Drawing upon this definition, we consider a bank to be "weak" if any of the following conditions is met:

- Common Equity Tier 1 (CET1) ratio breaches an early warning threshold.
- Leverage ratio (LR) breaches an early warning threshold.
- Liquidity Coverage ratio (LCR) breaches an early warning threshold.

Supervisors often use these metrics to impose intervention measures on banks when they drop below a minimum level. As such, modelling distress based on these metrics will provide supervisors with a practical tool that is directly aligned with their risk assessment framework.

To construct the distress events, we rely on quarterly supervisory data available to the EBA. We consider a distress event to start when an early warning threshold is breached and to end when the threshold is no longer breached (or an actual bank failure occurs, and the bank exits the sample). In our analysis, we assess different early warning threshold levels motivated by the solvency and liquidity requirements set in the European banking regulation, as presented in the next section.

CET1 ratio

The CET1 ratio was introduced after the financial crisis that started in 2008 as one of the main measures of a bank's capacity to absorb unexpected losses. It measures a bank's CET1 capital against its risk weighted assets. CET1 capital mostly consists of a bank's share capital, reserves and retained earnings and as such is the highest quality of capital for loss absorption purposes. In the EU, the CRD IV package − comprising the Capital Requirements Regulation (CRR) and the Capital Requirements Directive (CRD) − and competent authorities via the Supervisory Review and Evaluation Process (SREP) determine the level of capital (for each layer separately) banks are required to hold (EBA, 2022; EU, 2013a; EU, 2013b). This is called the 'own funds requirement' and is usually expressed as a percentage of risk weighted assets.

Figure 1 presents the stacking order of own funds requirements and Pillar 2 Guidance (P2G). Pillar 1 requirements are the minimum capital requirements applicable to all banks. They ensure that banks hold enough capital to cover for unexpected losses related to credit, market and operational risks. Their level, as set in the CRR, stands at 4.5%, 6% and 8% for the CET1, Tier 1 and Total capital ratio. Pillar 2 requirements (P2R) are additional bank-specific capital requirements, applied on top of the minimum Pillar 1 requirements, which aim to cover risks that are underestimated or not covered by the minimum Pillar 1 requirements. The level of the P2R is set by the competent authority as part of the SREP and is bank-specific.[7] Pillar 1 requirements and P2R, together known as the Total SREP Capital Requirement (TSCR), are legally binding and should be met at all times. Any breach of the TSCR can have direct legal consequences for banks, including a potential withdrawal of authorisation (Article 48 of CRD).

[6] BCBS (2015) Guidelines for identifying and dealing with weak banks

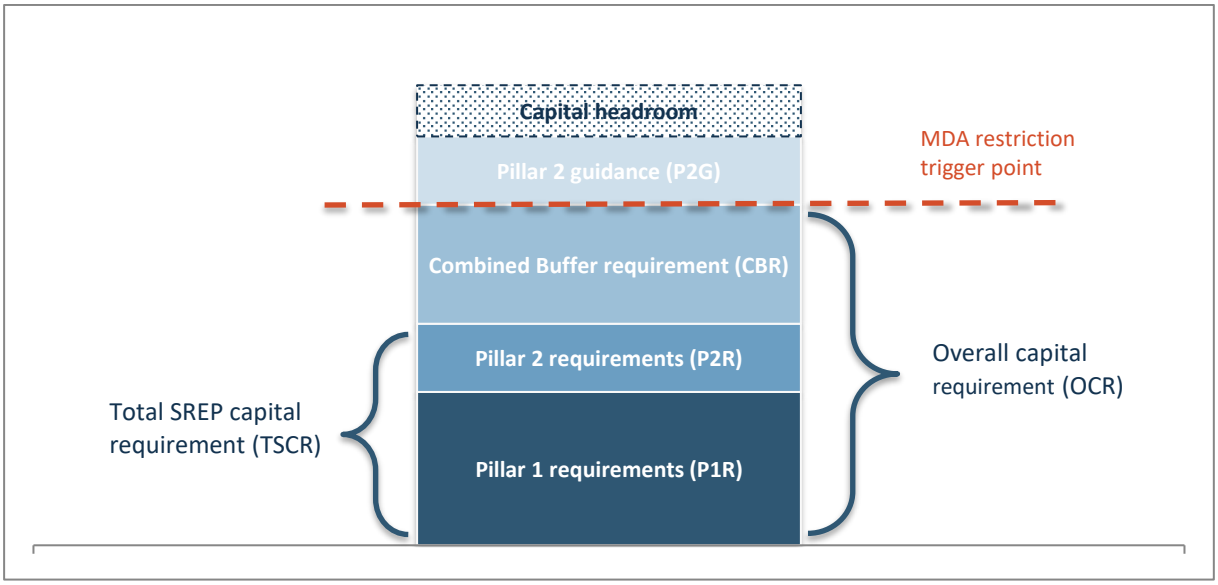
[7] Under the new Capital Requirements Directive V (CRDV), which came into effect in January 2021, banks can fulfil P2R with a minimum 56.25% CET1 capital (EU, 2019a) (EU, 2019b). The remaining P2R can be filled with Additional Tier 1 and Tier 2 capital.

In addition to the TSCR, banks are subject to several capital buffers of a macroprudential nature. These include the capital conservation buffer (CCB), the countercyclical buffer (CCyB), the systemic risk buffer (SyRB) and buffers for global (G-SII) and other systematically important institution (O-SII), together constituting the combined capital buffer requirement (CBR). The level of these buffers is either directly determined in the regulatory framework (e.g. 2.5% for the capital conservation buffer) or set individually by national macroprudential authorities and should be met fully in CET1 capital. The buffer framework is designed in such a way so that banks can operate below the CBR when needed (e.g. in a period of stress) subject to automatic restrictions on distributions.[8,9] In practise, banks must calculate the Maximum Distributable Amount (MDA) as soon as they fail to meet the Overall Capital Requirement (OCR) – the sum of Pillar 1 requirements, P2R and CBR.

Finally, as part of the SREP, competent authorities can set a Pillar 2 Guidance that sits on top of the CBR and acts as additional buffer of protection against losses from adverse scenarios. This is not a legally binding requirement but rather a supervisory expectation/recommendation of the adequate levels of capital a bank must have to be sufficiently protected during stressed conditions. Therefore, if a bank's capital falls below the P2G level, it would not trigger any automatic supervisory action neither would lead to any restrictions on the distributable amount. P2G is expected to be fully met with CET1 capital.

While banks should meet different capital requirements for each layer of capital, in this paper we make use only of the CET1 ratio requirements to determine our distress events. CET1 capital is not just the highest quality of capital but also the most expensive form of capital from a bank's point of view. Banks usually focus on managing their CET1 capital levels and set a CET1 capital target to maintain an additional capital buffer above capital requirements. At the same time, they are reluctant to hold excessive CET1 capital above capital requirements since this is too costly.

Figure 1: Stacking order of own funds requirements and P2G



Note: The scale is not meaningful/indicative only.

Figure 2 shows the distribution of the CET1 ratio (fully loaded) over time and the number of related distress events using different early warning threshold levels. Over the last years, European banks have continuously

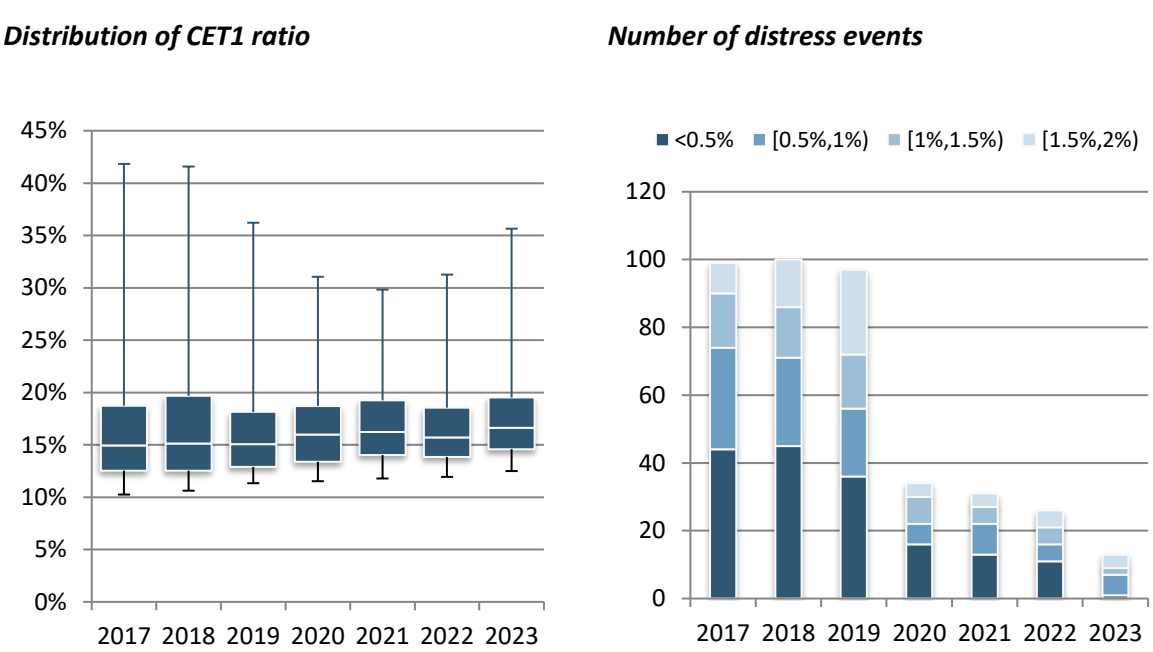
[8] Banks are required to rebuild their capital levels in a timely manner when operating within the buffer range.

[9] The restrictions cover dividend payments, payments on variable remuneration and payments on Additional Tier 1 (AT1) instruments.

increased their CET1 ratios (fully loaded), starting from 14.9% in 2017 and reaching 16.6% in 2023. This significant improvement was primarily driven by an increase in capital sources, most notably retained earnings, but was also supported by banks' de-leveraging and de-risking.

The early warning threshold levels are measured in terms of percentage points (p.p.) in excess of the OCR and P2G. For example, an early warning threshold level of 0.5 p.p. counts the number of banks that have a CET1 ratio, which is not higher than 0.5 p.p. above their OCR and P2G level. Our choice of the early warning threshold relative to the OCR and P2G was based on the fact that supervisors and market participants expect banks to keep a certain level of excess capital to manage fluctuations while complying with requirements and supervisory expectations at all times.[10] Therefore, banks who are closer to breaching this level are likely to receive higher supervisory attention and can be considered weak(er) relative to the rest of the banks. We chose the threshold level at 0.5 p.p. above OCR and P2G based on the 5th percentile of banks' actual CET1 ratio buffer in the period 2017 Q1 to 2023 Q3 (0.49 p.p.), which leads to 166 distress events during that period.

Figure 2: CET1 ratio (fully loaded) - Distribution (left) and number of distress events using alternative early warning threshold levels (right)



Source: EBA supervisory data and EBA calculations.

Notes: For the quarters between March 2017 and September 2018, the capital requirements as of December 2018 has been used as a proxy due to unavailability of data. Data for 2023 covers only the first three quarters of the year.

Leverage ratio

The leverage ratio captures the relation between a bank's capital and its assets, irrespective of how risky those are. It has been introduced in EU banking regulation to act as a backstop to risk-based capital requirements by constraining the building up of excessive leverage during economic upturns.

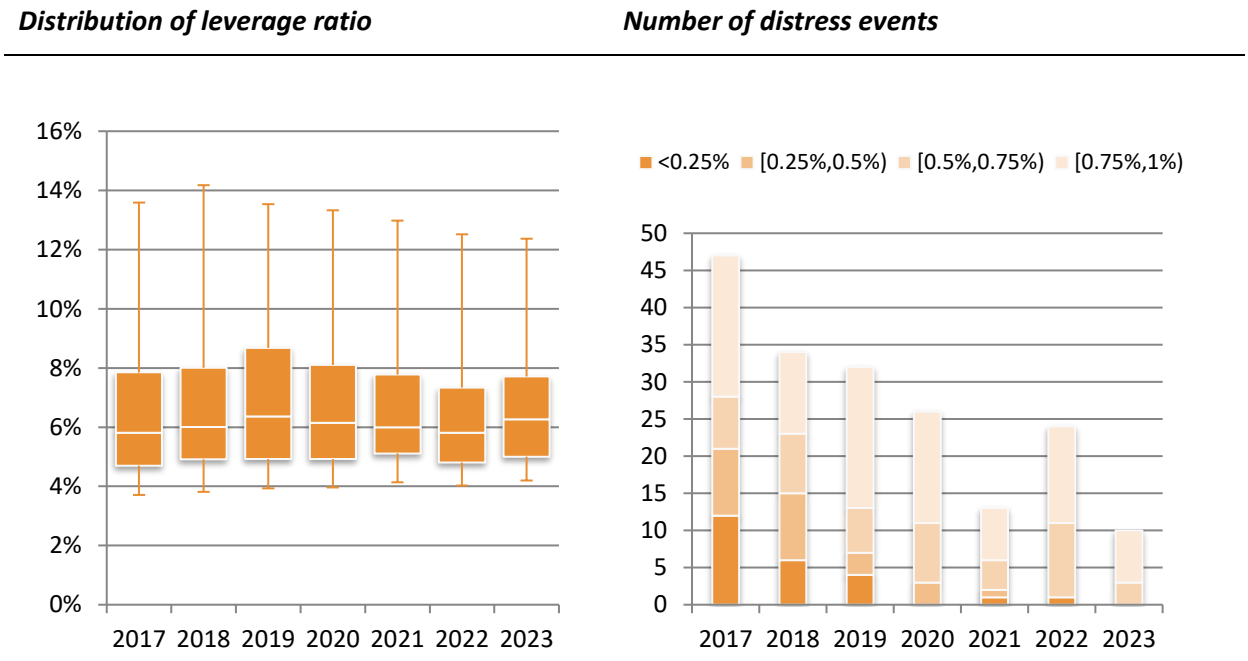
[10] On 12 March 2020, the ECB announced that it would allow banks to operate temporarily below the level of capital defined by P2G as part of the relief measures in reaction to coronavirus: https://www.bankingsupervision.europa.eu/press/pr/date/2020/html/ssm.pr200312~43351ac3ac.en.html

A bank's leverage ratio is calculated as a bank's Tier 1 capital divided by its total leverage ratio exposure measure, which includes its assets and off-balance-sheet items. The CRR introduced a uniform definition for leverage ratio, which was later amended to align it with the revised international standards on the leverage ratio, published in December 2017 by the BCBS (BCBS, 2017).

While the minimum leverage ratio requirement of 3% has been applicable since June 2021, we treat it as binding from 2017 for the purposes of our analysis. We consider this assumption reasonable, given that banks had to report and publicly disclose the ratio from as early as 2016; hence it is likely that they had frontloaded the leverage requirement before its actual application date. Figure 3 shows the distribution of the leverage ratio over time and the number of related distress events using different early warning threshold levels. As can be seen, banks reported ratios well above the minimum requirement since 2017 confirming that banks frontloaded the leverage requirement. As of September 2023, the average leverage ratio stood at 6.3%. With ratios generally on the rise between 2017 and 2021, the number of distress events declined over the same period and increased in 2022.

Similar to the rationale behind the CET1 threshold, our choice reflects banks' need to operate above minimum requirements at all times while providing a signal to supervisors regarding banks that might operate too close to those requirements. Our chosen threshold level of 1 p.p. is close to the 5th percentile of banks' actual leverage ratio buffer in the period 2017 Q1 to 2023 Q3 (0.95 p.p.) and leads to 186 distress events during that period.

Figure 3: Leverage ratio (fully loaded) – Distribution (left) and number of distress events using alternative early warning threshold levels (right)



Source: EBA supervisory data and EBA calculations.

Notes: Data for 2023 covers only the first three quarters of the year.

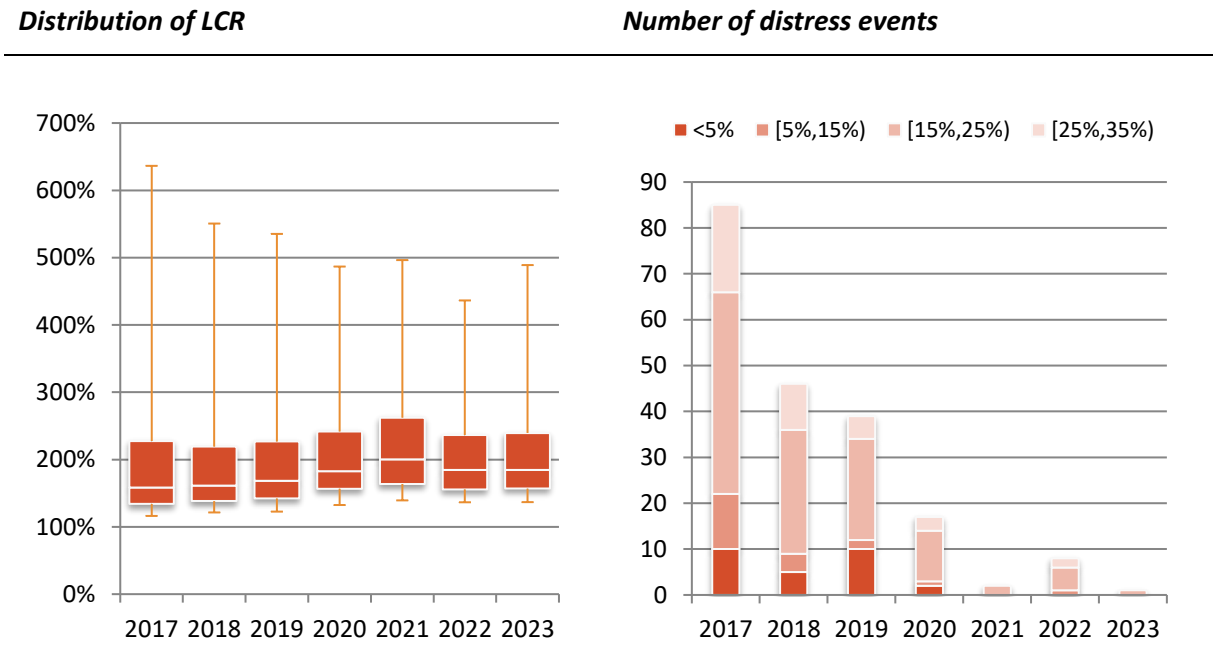
Liquidity coverage ratio

The CRR has also set out a general liquidity coverage requirement for credit institutions under Article 412(1). The objective of this requirement is to ensure that banks hold liquid assets to meet withdrawals demands under

gravely stressed conditions over a period of thirty days. In October 2014, the European Commission adopted a Delegated Act on the Liquidity Coverage Ratio (LCR), specifying in detail how to apply the liquidity coverage requirement (EU, 2014a). The LCR is defined as the stock of high-quality liquid assets (HQLAs) over the net liquidity outflows arising during a 30-calendar-day stress period. It was introduced on 1 October 2015, with a minimum requirement set at 60%, which was gradually phased-in to reach 100% on 1 January 2018. Given the long monitoring period that preceded the phase-in, banks generally complied with the final requirements already in 2017. We therefore used the fully-loaded LCR requirement for the entire period studied.

Figure 3 shows the number of distress events based on the LCR using different early warning threshold levels. Supervisory data on the LCR (based on the Commission's Delegated Act definition) became available from 2016 Q3 and stood at 158.6% in 2017 for the median bank. Driven by public measures, including central bank liquidity support, the LCR increased in the years leading to 2021. In 2023, and due to the phasing out of support measures, the LCR declined to reach an average of 184.5% in September 2023. The 5th percentile LCR buffer of 25 p.p. above requirements matches with our chosen threshold of 25 p.p. With this choice, we observe 159 distress events during the period 2017 Q1 to 2023 Q3.

Figure 4: Liquidity coverage ratio - Distribution (left) and number of distress events using alternative early warning thresholds (right)



Source: EBA supervisory data and EBA calculations.

Notes: Data for 2023 covers only the first three quarters of the year.

Total distress events

Table 1 illustrates the number of distress events per category using the aforementioned early warning threshold levels. In total, there are 455 distress events, accounting for around 14% of the total observations. The distress events are spread almost equally across the categories. The occurrence of distress events is not mutually exclusive across the categories (i.e. a bank may breach multiple early warning thresholds in the same quarter). Hence, the categories do not sum up to the total. In addition, the number of distress events exceeds the number of banks, as a distress event is identified at the bank-quarter level and the same bank can be in distress in multiple quarters.

Table 1: Number of distress events by category

Distress category	Frequency	Probability (%)
CET1 ratio (0.5% headroom)	166	5.0
Leverage ratio (1% headroom)	186	5.7
Liquidity Coverage Ratio (25% headroom)	159	4.9
Total	455	13.8

Source: EBA supervisory data and EBA calculations.

Notes: The statistics are derived from a sample of 176 banks with 3298 observations over the period of 2017Q1 to 2023Q3. Probability is calculated as the ratio of the number of distress events by the total number of observations. The total number of distress events does not sum up to the individual categories because they are not mutually exclusive.

2.3. Explanatory variables

The existing literature mainly relies on CAMELS indicators to capture bank's vulnerability to distress. These include measures of capital adequacy, asset quality, management, earnings, liquidity and sensitivity to market risk. However, there is inconclusive evidence and an ongoing debate in the current literature regarding the level of significance of the explanatory variables used in predicting bank failures. Depending on the models and time horizons used, different variables tend to be more significant in predicting bank failures or distress events.

Petropoulos, Siakoulis, Stavroulakis, & Vlachogiannakis (2020) assessed the importance of each variable used as input for each model in their study. They found that for most models, profitability and capital indicators were the most important drivers across all models. Cost of Funding Earnings Assets (CFEA) and leverage ratio (LEV) were identified to be leading indicators in bank failure forecasting. In addition to CFEA, earnings related indicators such as Return on Equity were also identified as important determinants. On the other hand, Loan loss allowance to non-performing loans and non-performing loans to loans appear to be the ones with the lower importance across all models. Furthermore, Liquidity risk as measured by the Net Loans to Core Deposits and Asset Quality as measured by the distance from the sector of Loss allowance to loans were found to have increased significance in the Support Vector Machine (SVM) and Neural Network (NN) models.

Le & Viviani (2018) observed that three groups of variables play a more important role in predicting bank failures, namely operation efficiency, profitability and liquidity. Variables that were found to be more relevant than others were Impaired Loans to Gross Loans, Tier 1 capital ratio, Capital funds to Total assets, Other Operation Income to Average Assets, Net interest revenue to Average Assets, Non Operation Items and taxes to Average Assets, Return on Average Assets, Cost to income ratio, Net Loans to Total Asset, Net loans to Deposit and Short Term funding and Net Loans to Total Deposit and Borrowing.

Gogas, Papadimitriou, & Agrapetidou (2018) found two variables that provided the highest forecasting accuracy for the model they studied with a 1-year forecasting horizon. These are Tier 1 capital over total assets and total interest expense over total interest income. When adapting the model to a two-year forecasting horizon, the following variables turned out to be most predictive: Tier 1 to Total assets, Provision for loan losses to Total

interest income, Loan loss allowance to Total assets and Volatile liabilities to Total assets. When applying a three-year forecasting horizon, the Tier 1 to Total assets, Loan loss allowance to Total assets and Total interest expense to Total interest income ratios proofed best suited.

Suss & Treitel (2019) found that lagged macroeconomic variables were very important for predicting distress of UK banks. For the random forest model, a measure of average real UK earnings was the single most important variable. As regards bank-specific financial ratios, they identified the ratio of trading book to total assets, capital buffer, and net interest margin to be the most predictive variables.

Betz et al. (2014) complemented bank-specific vulnerabilities with indicators for macro-financial imbalances and banking sector vulnerabilities. In their view, this improves model performance and yields useful out-of-sample predictions of bank distress.

In our model, and in line with previous research, we use two types of indicators to capture early signals of a bank's vulnerability to distress. These include:

- bank-specific indicators;
- country-specific macro-financial indicators.

Bank-specific indicators

Building on the existing literature, we use a comprehensive list of indicators that reflect a bank's main risks and balance sheet structure. The indicators are computed from the list of EBA Risk Indicators using quarterly supervisory data. The EBA Risk Indicators capture a range of risk dimensions: liquidity, funding, asset quality, profitability, concentration risk, solvency, operational risk, market risk, SME, sovereign risk. The indicators resemble traditional proxies for the CAMELS rating system (i.e. Capital Adequacy, Asset Quality, Management, Earnings, Liquidity, and Sensitivity to market risk). [11]

Country-specific macro-financial indicators

In addition to the above bank-specific variables, we consider a range of country-specific macro-financial indicators retrieved from Eurostat, ECB Statistical Data Warehouse and Bloomberg. These include GDP growth, inflation rate, unemployment rate, Government debt-to-GDP, Residential Real estate index, cost of borrowing for households and corporations, long term interest rates and sovereign yields.

Changes in variables over time

In addition to point-in-time readings for both bank-specific and macro-financial indicators, we also use the quarter-over-quarter change and year-over-year change of all our indicators. This allows us to capture changes or shifts in an indicator over time in predicting distress. For example, a significant shift in a bank's reliance on client deposits could signal an increase in liquidity risk.

Prediction horizon

For each of the above bank-specific and macro-financial indicators (including the transformation of the indicators over time), we create lagged variables on a quarterly basis starting from 1 to 8 quarters (i.e. 2 years). We treat all the variables of the same indicator but different lag structure as separate predictors in our model.

[11] For a full list of the EBA Risk indicators and their description, see https://www.eba.europa.eu/risk-analysis-and-data/guides-on-data

For our model specification, we only include predictors that are lagged by 4 up to 8 quarters relative to the dependent variable, i.e. we apply a 1-year prediction horizon as a minimum for our model. We do so, given that the purpose of our model is to detect possible distress events well in advance. In this way, supervisors will have enough time to take action aimed at preventing or mitigating negative consequences of distress or potential bank failure. Depending on the supervisory approach chosen by the relevant authorities that wish to use our model, the prediction horizon can be adjusted. We test our model's performance under various prediction horizons as a robustness check (see section 4.3).

3. Methodology

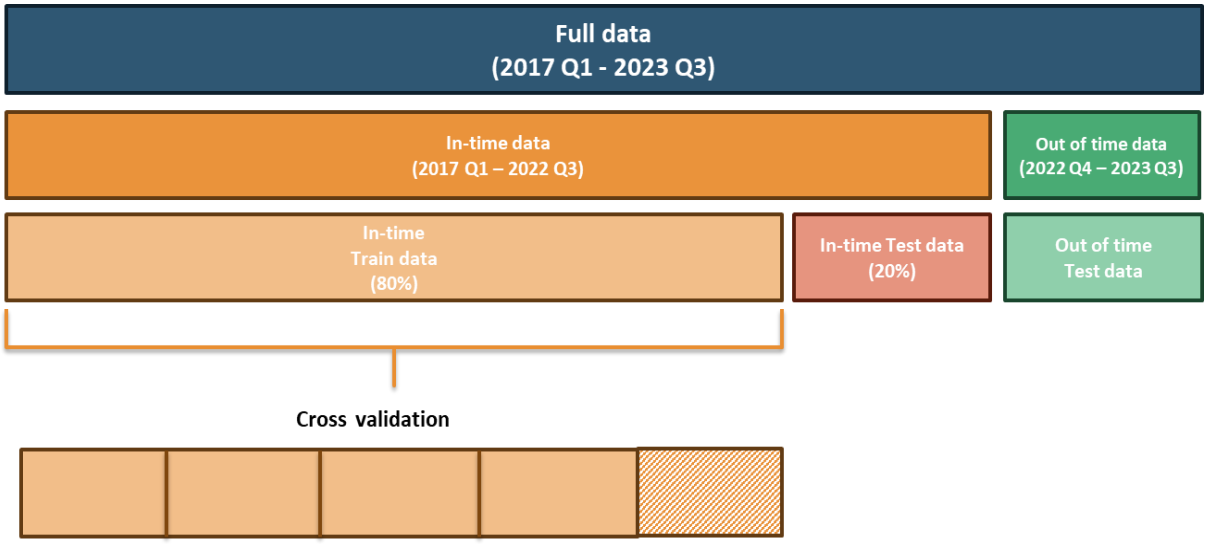
3.1. Pre-processing

Data splitting

Before building our models, we split the full dataset into three parts (Figure 5):

- the full in-time dataset that covers the period between 2017 Q1 – 2022 Q2 and is divided into two sub-parts using a randomly stratified sampling technique[12] :
  - a) the in-time training dataset consisting of 80% of the total observations; and
  - b) the in-time test dataset consisting of the remaining 20% of the observations;
- the out-of-time dataset that comprises data between 2022 Q3 – 2023 Q3.

Figure 5 Training, validation and test datasets



The in-time training dataset is used to train and develop the model. During the development of our model, the training dataset is further divided in k subsets called "folds". The model is then trained on a combination of k-1 folds and tested on the remaining fold, often referred to as the validation set. This technique is commonly used in machine learning to fine-tune and select the optimal model parameters and hyperparameters, such as the number of layers in neural network models (for more information on parameters and hyperparameters see section 3.2). This helps optimizing the model's performance, while preventing overfitting. [13]

[12] The technique ensures that proportion of distress and non-distress events are preserved in our training and test datasets.

[13] Overfitting is a common problem in machine learning, where the model "memorizes" the training set and their corresponding labels instead of learning the true underlying relationships in the data. When this occurs, the model usually gives very accurate predictions for the training data but not for new data.

We select k=5 for the number of folds, in other words we use a 5-fold cross-validation. This is a common choice for k among practitioners and suits better to the size of our dataset which is relatively small.[14] To reduce the impact of the random partitioning of the data when dividing into folds, we use repeated k-fold cross validation, where the entire k-fold cross validation process is repeated multiple times, with different random partitions each time. The model is then assessed using averages of the performance metrics across all partitions and repeats. This technique is particularly useful when working with smaller datasets as it provides a more robust evaluation of the model's performance, independent of any specific partitioning.

The in-time test data is used to evaluate the performance of the trained model. By evaluating our model on unseen data, we can assess its ability to generalize and make accurate predictions. In this way, we mitigate the problem overfitting where the model performs very well on our existing data, but fails to generalize on new, unseen data.

The out-of-time test dataset assesses the performance of the trained model during a different (future) time period. Using an out-of-time dataset can help identify if the model is robust to changes and variation that might occur over time. It also helps assess the model under a real-world scenario, where supervisors will use the trained model to predict future distress events.

Feature selection procedure

Before training our models, we employ a feature selection procedure to reduce the number of explanatory variables – often called features in machine learning – that will go into our models (Figure 6). The technique has multiple benefits as outlined in the past literature by Kuhn & Jonson (2019), Li, et al. (2017), Murphy (2012) and Sarkar, Bali, & Sharma (2017). First, it ensures that only the most relevant and informative features are selected for training the model. Second, it reduces model complexity and enhances interpretability. Finally, it makes the training process faster and more efficient. To avoid any leakages between our training and testing datasets, we carry out the feature selection procedure only on the training dataset rather than the full dataset.

We start with an initial set of 7,058 potential predictors. As a first step in our feature selection procedure, we exclude 5,864 variables with poor coverage, which have more than 15% of the values missing. Secondly, we exclude 12 variables which have near zero variation.[15] Thirdly, we exclude variables that are considered irrelevant or redundant. In particular, we employ a range of filter techniques and rank the variables based on their importance.[16] We then exclude 1,017 variables that are not ranked among the top 100 important variables in any of these techniques. As filter techniques, we use the correlation with the dependent variable, the information value with respect to the dependent variable, the information gain with respect to the dependent variable and the area under the ROC curve for each feature. Finally, we exclude 159 variables that exhibit pair-wise correlation of more than 0.7.[17] We end up with a final set of 33 number of features to train our model.

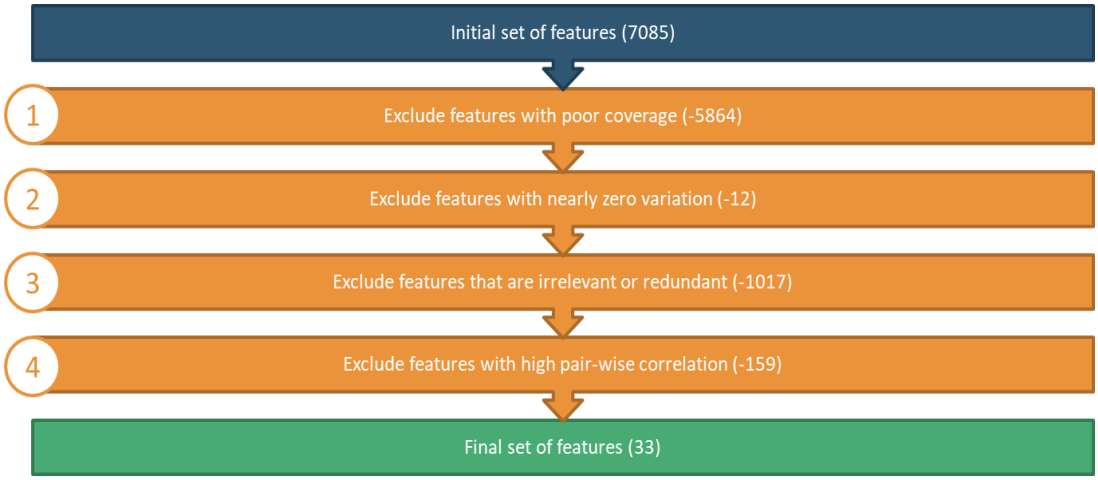
[14] Other common choices are k=10.

[15] We use the "nearZeroVar" function in R. The function diagnoses predictors that have one unique value (i.e. are zero variance predictors) or predictors that are have both of the following characteristics: they have very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large.

[16] There are generally 3 classes of feature selection methods for eliminating irrelevant and redundant features: filter methods, wrapper methods and embedded methods. We use filter methods because they are simple, computationally efficient and independent of a specific machine learning algorithm. These methods rely on a statistical measures or heuristic metrics, such as the correlation with the target variable, to rank and select the most relevant features.

[17] We use the function "findCorrelation" in R. The function searches through a correlation matrix and returns a vector of integers corresponding to columns to remove to reduce pair-wise correlations. If two variables have a high correlation, the function looks at the mean absolute correlation of each variable and removes the variable with the largest mean absolute correlation.

Figure 6 Feature selection procedure



Note: The number of features excluded in each step is indicated in brackets.

For modelling purposes, we only consider observations which have all the 33 predictors available (i.e. not missing), resulting in a total of N = 2,371 observations over 108 banks and 341 distress events.

Selected features

The final set of selected features covers a range of areas including asset quality, liquidity and funding, profitability, solvency (capital) as well as macroeconomic indicators (Table 11 in the Annex). Some indicators appear more than once due to multiple transformations of the same indicator we consider in our model. For example, the asset encumbrance ratio features in our model twice, both in its point-in-time version as well as in its year-over-year change. Table 12 and Figure 9 in the Annex show their summary statistics and variation across distress / non-distress events.

Many of our selected features are commonly found as important drivers of bank distress in many studies in the field. Among the common indicators, those associated with solvency (capital), profitability, liquidity and funding appear in most studies. For example, variables such as return on equity/assets and the ratio of Tier 1 capital over total assets consistently rank among the most influential predictors of distress in many studies such as in Gogas, Papadimitriou, & Agrapetidou (2018), Cole & Wu (2018), Suss & Treitel (2019) and Petropoulos, Siakoulis, Stavroulakis, & Vlachogiannakis (2020).

As regards liquidity and funding, we find that changes over time transformations of various indicators have a higher predictive power compared to their point-in-time readings. In addition, banks' reliance on customer deposits and interest rates paid on customer deposits emerge as important features for our model. This likely reflects that many banks in our sample operate a business model with a focus on funding via deposits.

In contrast to other studies, we also find that indicators linked to government indebtedness (as share of GDP) and banks' share of exposures to governments over total assets as important variables for our model. This might reflect the sovereign-bank nexus that came to the fore during and after the sovereign debt crisis that took place in several European countries in the 2010s.

3.2. Model development

In this paper, we apply selected machine learning techniques and compare their performance with traditional logistic regression models. We consider the following models which are described in more detail in the next sections: logistic regression, boosted decision tree (C5.0), random forest and neural networks (see Bishop, 2006; Murphy, 2012 for more information).

Logistic regression

Logistic regression is one of the most conventional methods for predicting distress. It relates the log-odds of distress to a linear combination of predictor variables. The logistic function takes the form:

$$\log\left(\frac{P(Y = Distress)}{1 - P(Y = Distress)}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N$$

where α is the intercept term, β_i is the coefficient or parameter associated with the predictor variable X_i and X_i is a set of continuous or categorical variables.

The coefficients are estimated through maximum likelihood estimation.[18] The predicted probability of distress is estimated by the inverse logistic function:

$$P(Y = Distress) = \frac{\exp(\alpha + \beta_1 X_1 + \cdots + \beta_N X_N)}{1 + \exp(\alpha + \beta_1 X_1 + \cdots + \beta_N X_N)}$$

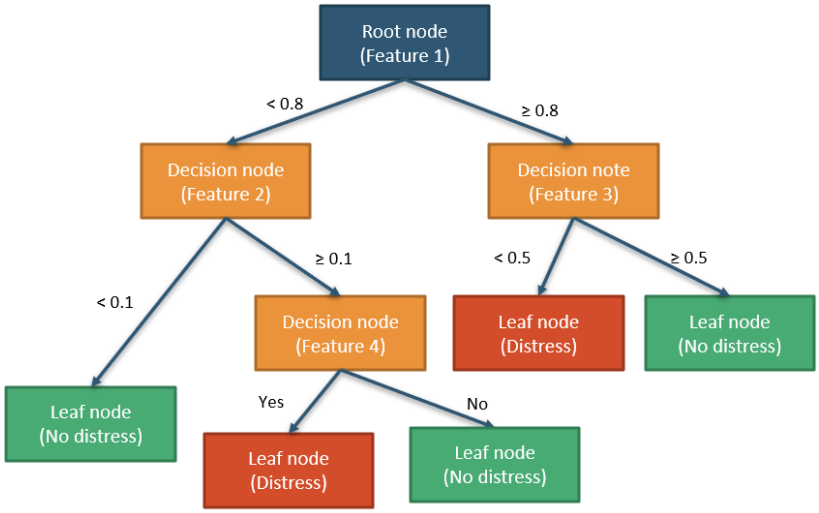
Following the existing literature (e.g. Duca & Peltonen, 2013; Lang, Peltonen, & Sarlin, 2018), we employ a pooled logistic model. Pooled logistic models have been shown outperform on out-of-sample data (even though in-sample performance is lower), which is the intention of our model, i.e. to predict future distress events (Fuertes & Kalotychou, 2017).

Boosted decision trees

A decision tree is a supervised machine learning algorithm that is widely used for both classification and regression problems. It creates a tree-like model that acts as a decision support tool (Figure 7). The tree consists of internal nodes (representing features or attributes), branches (representing decision rules), and leaf nodes (representing the predicted outcome or class label). Moving from the root to a leaf, one can visually understand how decisions are taken and what are their possible outcomes.

[18] We implemented logistic regression in R by using the 'caret' package.

Figure 7: Decision tree schematic



Decision trees have several advantages, including their interpretability, their ability to visualise classification rules and their ability to handle missing values. However, they are prone to overfitting, they are generally sensitive to changes in the underlying training dataset, and they are less able to deal with complex relationships in the underlying data. Also, the tree can quickly become deep and complex making interpretability harder.

To overcome these issues, we implement decision trees using the Quinlan C5.0 algorithm in R.[19] The algorithm uses a boosting method to create a series of decision trees forming an ensemble.[20] All trees (trials) in the ensemble are then combined to produce a final prediction. In this way, the model can make better and more robust predictions, reducing overfitting.

As part of the modelling phase, we use grid search to tune three hyperparameters: the number of trees (trials) grown, the model type (model) and the winnowing parameter which controls whether a winnowing algorithm is applied to reduce the number of variables in the model (winnow). For trials we consider the values: 5, 10, 15. For model we consider the tree, the rules and their combination. For winnow, we consider both cases where the winnowing algorithm is applied and not applied.

Random forest

Random Forest is a widely used machine learning algorithm for modelling classification and regression problems (Breiman, 2001). It belongs to the family of ensemble algorithms that combine the predictions of multiple individual models to improve model performance and prediction accuracy. It consists of a collection of decision trees, with each tree trained on a random subset of the training data and a random subset of the features. When making predictions, each tree in the random forest produces its own prediction independently and the final prediction is determined by combining these individual predictions. For classification problems, the final prediction is determined by majority voting, i.e., the predicted class is the one that received the most votes

[19] For a literature review of Data Mining Algorithms see Wu et al. (2008). The relative R environment used in this paper refers to Kuhn et al. (2015). The algorithm uses an entropy-based approach to construct decision trees. It recursively splits the data based on features that maximize the information gain or minimize the entropy at each step.

[20] Boosting is a technique for generating and combining multiple classifiers to improve the predictive accuracy of the model. Instead of using a single tree, n separate decision trees (trials) are grown and combined to make predictions. The error rate of the boosted classifier is often substantially lower than that of single trees.

across all the trees. For regression problems, the prediction is usually the average or median prediction across all the trees.

Random forests have several advantages. By combining the predictions of multiple trees, they can often achieve better performance compared to a single tree. They are robust to overfitting because of the randomness introduced when constructing of the trees. They can also handle large datasets with many features. While Random Forest and the C5.0 algorithm discussed above are both ensemble learning models that uses decision trees they differ in certain aspects. These include a) the way the trees are build, with random forest randomisation process ensuring a greater diversity among the trees; b) the variable selection process, with C5.0 considering the importance of the variables based on their predictive performance while Random Forest selects the variables randomly; c) the way prediction are made, with the Random Forest aggregating predictions through majority voting, while C5.0 combines the predictions of the best individual trees (weighted voting).

We tune two hyperparameters using grid search: the number randomly selected features used in each tree (mtry) and the number of trees to be grown (ntree). We consider the following possible values for mtry: $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$, $2\sqrt{m}$ where m is the number of the predictors fed into the model (Breiman, 2001). For ntree we consider the following possible values: 100, 250, 500.

Neural networks

Neural networks are advanced machine learning models inspired by the structure and functioning of the human brain. They are comprised of multiple nodes, or neurons, which are connected to one another forming a network. The nodes are organised into node layers, consisting of an unput layer, one or mode hidden layers and an output layer. The input layer receives the input data and passes them into the next layers. The hidden layers carry out complex calculations and pass them to the output layer which produces the final output (e.g. in the form of predictions). The nodes are the basic processing units of the network and each of them can be thought of as an individual model, composed of input data, weights, a bias (or threshold), and an output. The output of each node is passed through an activation function and if it exceeds a given threshold, it activates the node, passing data to the next layer in the network. The output of one node then becomes the input of the next node until the final output is obtained.

Neural networks are highly flexible and can be applied to a wide range of tasks. They can capture complex non-linear relationships in the data and are good in handling large datasets. They can also automatically learn which are the most useful features; hence data preprocessing and feature engineering is not as necessary. On the down side, neural networks are often seen as opaque and black box models.

We construct a single layer perceptron neural network. We tune two hyperparameters using grid search: the number of hidden units (neurons) in the hidden layer of the neural network (size) and the regularisation parameter which helps avoid overfitting (decay). We consider the following possible values for 'size': 2, 5, 7, 10. For 'decay' we consider the following possible values: 0.001, 0.01, 0.1, with smaller values resulting in stronger regularization.

Sampling techniques

To deal with the fact that our dataset is highly imbalanced, i.e. the distress class is represented by significantly lower proportion relative to the non-distress class, we use a variety of sampling techniques before estimating the models:

- Under sampling: randomly remove observations from the majority class (non-distress) from the training dataset
- Over sampling: randomly replicate observations from the minority class (distress) and add to the training dataset.

- Synthetic Minority Over-sampling Technique (SMOTE): synthesize new observations from the minority instances and add to the training dataset.

The above techniques aim to balance the class distribution, by either reducing the number of instances in the majority class (undersampling) or increasing the number of instances in the minority class (oversampling, SMOTE) of the training set. In this way, we avoid building a biased model that performs poorly on the minority class (distress), which is the primary focus of the supervisors.

4. Results

4.1. Model evaluation

Performance metrics

In this section, we compare the performance of the different models described above. To do so, we make use of several common evaluation metrics for classification problems. Most of them rely on the elements of the confusion matrix, a table that summarizes the performance of a classification model. It presents the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. In our study, the positive class corresponds to a distress event, while the negative class to a non-distress event.

The matrix is constructed as follows:

Table 2: Confusion matrix

		Prediction	
		Positive (distress)	**Negative (non-distress)**
Actual	**Positive (distress)**	True positive (TP)	False negative (FN)
	Negative (non-distress)	False positive (FP)	True negative (TN)

The components of a confusion matrix are the following:

- True Positives (TP): The number of instances that are correctly predicted as positive by the model. These are the cases where the model correctly identified distress events.
- False Positives (FP): The number of instances that are incorrectly predicted as positive by the model. These are the cases where the model predicted a distress event, but the true class was actually non-distress. Also known as a Type I error.
- True Negatives (TN): The number of instances that are correctly predicted as negative by the model. These are the cases where the model correctly identified non-distress events.
- False Negatives (FN): The number of instances that are incorrectly predicted as negative by the model. These are the cases where the model predicted a non-distress event, but the true class was actually distress. Also known as a Type II error.

Table 3 summarises the performance metrics used in this paper to assess the performance of the above models. When building an early warning system, the supervisor is often faced with a trade-off between two types of errors: Type I errors (missing distress events) or Type II errors (issuing false alarms). We consider Type I errors to be more costly than Type II errors, based on the assumption that the supervisor has a stronger preference in correctly identifying as many actual distress events as possible rather than issuing a false alarm. While the latter has the risk of damaging the supervisor's credibility, we consider this to be limited, because the early-warning

signal would act only as a trigger for a more in-depth analysis of the bank, which will give the supervisor the chance to assess if the signal is false.

Therefore, we pay particular attention to how the model performs in terms of Recall / Sensitivity / True positive rate. Also, given that our dataset is highly imbalanced we interpret with caution the overall Accuracy of the model. Instead, we focus on Balanced Accuracy and Weighted Balanced Accuracy, particularly WBA2, weighted balance accuracy 2 (WBA2), which implicitly assigns a higher weight to Type I error (75%) than Type I error (25%).[21]

Table 3 Summary of performance metrics

Performance metric	Description	Formula
Accuracy	Measures the overall correctness of the model's predictions	$\frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$
Sensitivity / Recall / True positive rate	Measures the proportion of correctly predicted positive instances out of all actual positive instances. Equals to 1 – Type I error	$\frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$
Specificity / True negative rate / Negative prediction rate	Measures proportion of correctly predicted negative instances out of all actual negative instances. Equals to 1 – Type II error	$\frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})}$
Balanced Accuracy	Measures the overall correctness of the model's predictions, taking into account the imbalance in the data	$0.5 * Sensitivity + 0.5 * Specificity$
Weighted Balanced Accuracy (1)	Measures the weighted average balance accuracy that weights specificity more than sensitivity (75%/25%)	$0.25 * Sensitivity + 0.75 * Specificity$
Weighted Balanced Accuracy (2)	Measures the weighted average balance accuracy that weights specificity less than sensitivity (25%/75%)	$0.75 * Sensitivity + 0.25 * Specificity$
AUC-ROC	Measures the area under the Receiver Operating Characteristics (ROC) curve. The ROC curve is a graphical representation of the trade-off between true positive rate and false	Area under Receiver Operating Characteristics (ROC) curve

[21] Sensitivity = 1 – Type I error and Specificity = 1 – Type II error

Performance metric	Description	Formula
	positive rate for different classification thresholds	
Brier score	Measures the mean squared difference between the predicted probability and the actual outcome across all instances	$Brier\ score = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2$ where N is the total number of observations; P_i is the predicted probability of distress for instance I; O_i is the actual outcome for the i-th instance, where 1 indicates a distress and 0 indicates a non-distress

Effect of sampling techniques

Table 4 and Table 5 presents the performance metrics for the various methods under different sampling techniques.

For the logistic regression, the baseline model performs well in terms of accuracy but does a poor job in predicting distress events as showcased by the very low sensitivity. All sampling techniques perform similarly, leading to an improved sensitivity at the cost of lower specificity and overall accuracy. This is translated to lower values for WBA1 and higher values for WBA2 under all sampling techniques compared to the baseline. AUC ROC values are consistently high across all sampling techniques and close to the baseline. On the other hand, Brier Score is lowest for the baseline model, indicating better calibration of the predicted probabilities.

For random forest, the baseline and oversampling lead to the highest accuracy, with SMOTE closely follows. However, all three methods show a very low Sensitivity. Undersampling improves Sensitivity significantly, while maintaining a good balance in Specificity, yielding the highest WBA2. Similarly with logistic regression, AUC ROC values are consistently high across all sampling techniques. Brier score is low and very similar for the baseline, oversampling and SMOTE, while is higher for undersampling technique.

For decision tree, accuracy is the highest with oversampling. Sensitivity is relatively low for the baseline, oversampling and SMOTE and improves considerably with the undersampling technique. Specificity is consistently high across sampling methods, while slightly lower for the undersampling technique. WBA2 is the significantly higher for undersampling compared to other sampling techniques, while WBA1 is consistently high for all sampling techniques. AUC ROC does not change significantly across sampling techniques, while Brier score deteriorates with undersampling.

For the neural networks, the baseline exhibits the highest accuracy followed closely by other sampling techniques. Sensitivity varies across sampling techniques, with undersampling and oversampling yielding a significant improvement relative to baseline although at the expense of lower specificity. WBA2 is the highest for oversampling followed closely by the undersampling method. As in other methods, AUC ROC remains high across all sampling techniques and Brier is lowest for the baseline model.

Overall, the baseline models demonstrate good accuracy but have low sensitivity, indicating a potential issue in identifying distressed banks. Employing sampling techniques leads to a significant increase in sensitivity, suggesting an improved identification of distress banks, although at the cost of lower specificity. AUC ROC values

for all sampling techniques are close to the baseline, indicating good discrimination ability. Brier Score is usually lower for the baseline model compared to others, suggesting better calibration.

As we are particularly interested in correctly predicting distress banks, we rely on Sensitivity as well as WBA2 to choose the preferred sampling technique. For logistic regression, the oversampling and SMOTE techniques produce the highest Sensitivity and WBA2, while for random forest and decision tree the undersampling technique. For the neural network, WBA2 and Sensitivity is highest for oversampling, closely followed by undersampling.

For consistency, we rely on a single sampling technique when comparing model performance across methods. We choose the undersampling technique, which appears to work well across models. In the remainder of the paper, we present the results based on the undersampling technique for all methods.

Table 4 Validation results for logistic and random forest models across sampling techniques based on in-time test dataset

Logistic					Random forest				
Performance metric	**Baseline**	**Under-sampling**	**Over-sampling**	**SMOTE**	**Performance metric**	**Baseline**	**Under-sampling**	**Over-sampling**	**SMOTE**
Accuracy	0.8632	0.7447	0.7579	0.7579	Accuracy	0.8737	0.8132	0.8895	0.8921
Sensitivity	0.3881	0.806	0.8209	0.8209	Sensitivity	0.403	0.8806	0.5522	0.6716
Specificity	0.9649	0.7316	0.7444	0.7444	Specificity	0.9744	0.7987	0.9617	0.9393
Balanced Accuracy	0.6765	0.7688	0.7827	0.7827	Balanced Accuracy	0.6887	0.8397	0.757	0.8055
WBA1	0.8207	0.7502	0.7635	0.7635	WBA1	0.8316	0.8192	0.8593	0.8724
WBA2	0.5323	0.7874	0.8018	0.8018	WBA2	0.5458	0.8601	0.6546	0.7386
AUC ROC	0.83	0.8299	0.8361	0.8331	AUC ROC	0.9225	0.9046	0.9264	0.9273
Brier score	0.1073	0.1784	0.1697	0.1684	Brier score	0.0802	0.143	0.08	0.0848

Table 5 Validation results for decision tree and neural network models across sampling techniques based on in-time test dataset

	Decision tree (C5.0)				Neural network				
Performance metric	**Baseline**	**Under-sampling**	**Over-sampling**	**SMOTE**	**Performance metric**	**Baseline**	**Under-sampling**	**Over-sampling**	**SMOTE**
Accuracy	0.8632	0.7974	0.8737	0.8605	Accuracy	0.8421	0.7342	0.7553	0.7974
Sensitivity	0.5224	0.8358	0.4627	0.4478	Sensitivity	0.4776	0.8209	0.8507	0.6119
Specificity	0.9361	0.7891	0.9617	0.9489	Specificity	0.9201	0.7157	0.7348	0.8371
Balanced Accuracy	0.7292	0.8125	0.7122	0.6983	Balanced Accuracy	0.6989	0.7683	0.7928	0.7245
WBA1	0.8327	0.8008	0.8369	0.8236	WBA1	0.8095	0.742	0.7638	0.7808
WBA2	0.6258	0.8242	0.5874	0.573	WBA2	0.5882	0.7946	0.8218	0.6682
AUC ROC	0.902	0.878	0.8834	0.9015	AUC ROC	0.8513	0.8494	0.8554	0.806
Brier score	0.0953	0.147	0.0918	0.0888	Brier score	0.1213	0.198	0.1571	0.1712

Model comparison

We now evaluate the performance across our four methods —Logistic Regression, Random Forest, Decision Tree (C5.0 algorithm), and Neural Networks based on the under-sampling technique.

We first focus on the predictive performance of each model on the in-sample test data (Table 6). Random Forest is the best performing model, demonstrating the highest AUC ROC (0.9064), Sensitivity (0.8806), Specificity (0.7987), leading to superior (weighted) balanced accuracies (WBA1, WBA2). It also has the best overall calibration of the fitted probabilities, having the lowest Brier score (0.143). Decision Trees follow closely, showcasing comparable results across various performance metrics. Logisitc regression and Neural Networks demonstrate satisfactory performance, although they exhibit slightly lower weighted balanced accuracies and AUC ROC.

Turning to the out-of-time performance, presented in Table 7, Random forest provides again the best fit across most of the performance metrics. Logistic regression and Decision tress follow closely in terms of AUC ROC and Brier Score, but Neural networks are ranked as the second best method in terms of Sensitivity (0.7857) and WBA2 (0.7864).

Overall, random forest appears to be the best performing model in both the in-time test data and out-of-time test data. The remaining methods also demonstrate adequate performance. Decision tree and neural network appear to follow closely in terms of Sensitivity and WBA2, while logistic regression underperforms in these metrics.

Table 6 Validation results for final models based on in-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.7447	0.8132	0.7974	0.7342
Sensitivity	0.806	0.8806	0.8358	0.8209
Specificity	0.7316	0.7987	0.7891	0.7157
Balanced Accuracy	0.7688	0.8397	0.8125	0.7683
WBA1	0.7502	0.8192	0.8008	0.742
WBA2	0.7874	0.8601	0.8242	0.7946
AUC ROC	0.8299	0.9046	0.878	0.8494
Brier score	0.1784	0.143	0.147	0.198

Table 7 Validation results for final models based on out-of-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.8395	0.8884	0.8837	0.7884
Sensitivity	0.7143	0.8571	0.7143	0.7857
Specificity	0.8438	0.8894	0.8894	0.7885
Balanced Accuracy	0.779	0.8733	0.8019	0.7871
WBA1	0.8114	0.8814	0.8456	0.7878
WBA2	0.7467	0.8652	0.7581	0.7864
AUC ROC	0.9069	0.9317	0.8776	0.8226
Brier score	0.1223	0.1052	0.0894	0.1522

4.2. Model explainability

While machine learning algorithms often perform better than traditional algorithm, they come at the cost of lower interpretability and explainability. In the recent years, several tools have been developed to help interpret and explain machine learning predictions, for example local interpretation tools (e.g. LIME, Shapley values, etc.), global interpretable tools (feature importance, partial dependence plot, etc.) and sensitivity analysis (EBA, Follow-up report on the use of machine learning for internal ratings-based models, 2023). In this paper, we use Shapley values to identify the drivers behind our predictions.

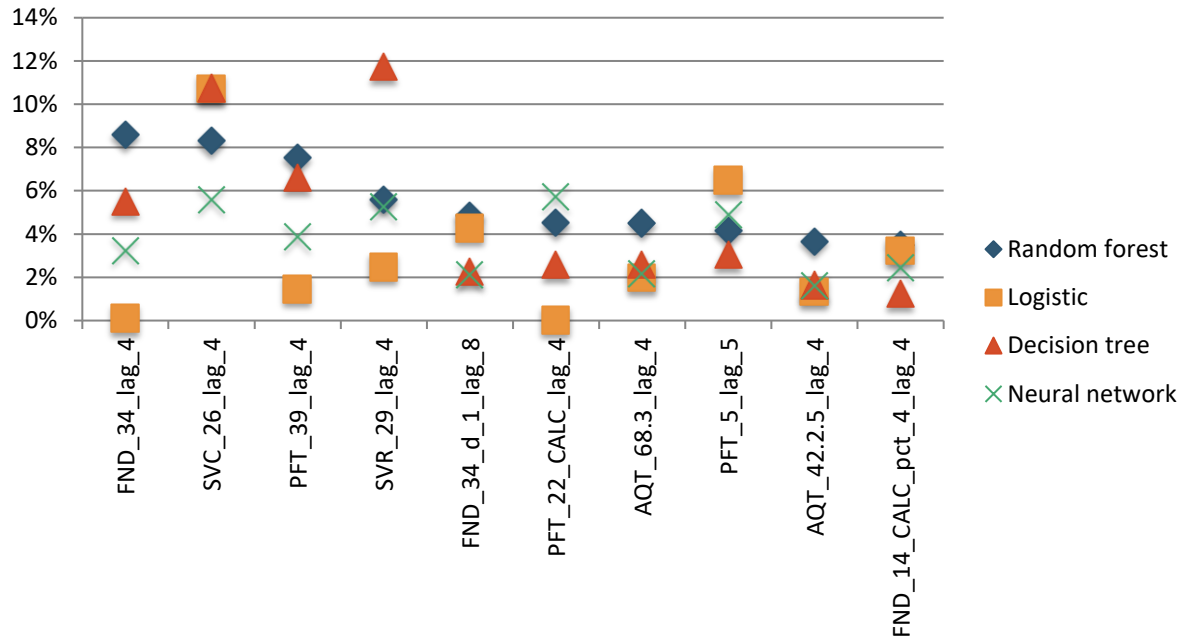
Shapley values

Shapley values provide the average marginal contribution of a feature to the prediction across all possible coalitions of features, where a coalition represents a subset of features. They are computed per observation (i.e. bank/quarter) and help explain individual predictions. We follow Bluwstein, Buckmann, Joseph, Kapadia & Şimşek (2023) and calculate the average absolute Shapley value per predictor across all test observations to help us understand the average behaviour of the model.

For the Random forest, our best performing model, we find the most predictive indictors to be related to bank profitability (average interest expense of deposits, asset-deposit spread for non-financial corporations) and to solvency (equity to total liabilities and equity). The top 3 indicators each have a Shapley value of close to 8% or above. Other indicators with Shapley values above 4% relate to profitability (Interest income from households, Return on regulatory capital requirements), the sovereign-bank nexus (the Share of Sovereign Exposures of Total Assets) and the share of financial instruments measured at (amortised) cost in total financial instruments.

Several of the most predictive indicators are found across models. Two of the top 10 indicators for the Random forest are also found in the top 10 for all the four other models. These two are the ratio of equity to total assets and the interest income from households. Another three indicators from the top 10 for the Random forest (Share of Sovereign Exposures of Total Assets, Average interest expense of deposits, Asset-deposit spread for non-financial corporations) are among the top 10 for three of the four models.

Figure 8: Shapley values TOP 10 indicators



4.3. Ensemble models

Previous studies have shown that ensemble models often outperform individual models on their own (Suss & Treitel, 2019). We apply two ensemble techniques to assess the performance is improved. We use a simple average of all four models and a stacked procedure with Gradient Boosting model (GBM) and Generalised Linear model (GLM) as the meta-model. The stacked model basically combines the prediction of the individual models using GLM or GBM.

Table 8 and Table 9 present the performance metrics for the ensemble models for the in-time and out-of-time test datasets. Regarding in-time performance, the Ensemble (GBM) is the best performing model in terms of AUC ROC and Brier score. On the other hand, the Ensemble (GLM) and Ensemble (Simple average) demonstrate the highest Sensitivity and WBA2. Random forest follows closely, showcasing comparable results across various performance metrics. Turning to the out-of-time performance, Random forest provides the best fit across most of the performance metrics. The ability of Ensemble (GBM) and Ensemble (GLM) to correctly identify distress events in the out-of-time dataset is considerably reduced as showcased by the drop in Sensitivity and WBA2 metrics.

Overall, ensemble techniques appear to improve the predictive performance compared to the single random forest model in the in-time test data. However, the random forest remains the best performing model in the out-of-time test data.

Table 8: Validation results for ensemble models based on in-time test data

Performance metric	Random Forest	Ensemble (simple average)	Ensemble (GLM)	Ensemble (GBM)
Accuracy	0.8132	0.7763	0.8079	0.8342

Performance metric	Random Forest	Ensemble (simple average)	Ensemble (GLM)	Ensemble (GBM)
Sensitivity	0.8806	0.9104	0.9104	0.8806
Specificity	0.7987	0.7476	0.7859	0.8243
Balanced Accuracy	0.8397	0.829	0.8482	0.8524
WBA1	0.8192	0.7883	0.8171	0.8384
WBA2	0.8601	0.8697	0.8793	0.8665
AUC ROC	0.9046	0.9012	0.9119	0.9133
Brier score	0.143	0.1442	0.1291	0.122

Table 9 Validation results for ensemble models based on out-of-time test dataset

Performance metric	Random Forest	Ensemble (simple average)	Ensemble (GLM)	Ensemble (GBM)
Accuracy	0.8884	0.8651	0.8930	0.9070
Sensitivity	0.8571	0.8571	0.7857	0.7143
Specificity	0.8894	0.8654	0.8966	0.9135
Balanced Accuracy	0.8733	0.8613	0.8412	0.8139
WBA1	0.8814	0.8633	0.8689	0.8637
WBA2	0.8652	0.8592	0.8134	0.7641
AUC ROC	0.9317	0.9219	0.9174	0.9031
Brier score	0.1052	0.2193	0.0791	0.0741

4.4. Robustness checks

We carry a series of robustness checks to assess if our results are sensitive to the definition of distress and prediction horizon chosen.

Definition of distress

First, we examine how sensitive are the results to the early warning threshold used to define the distress events. Instead of the 5th percentile, we re-run our analysis using the 10th percentile as the early warning threshold level.

The list of selected features increases from 33 to 42 with most of the added indicators relating to asset quality and some new entries relating to liquidity, market and operational risk. 50% of the indicators selected in our default model (5th percentile) also feature in the model when using the 10th percentile early warning threshold. Most of the other indicators selected in our default model appear with a different expression, either with some variation in the change over time or with a different prediction horizon. The two macroeconomic indicators selected in our default model are no longer selected. The Shapley values for the TOP 10 indicators range within 3.4% and 4.9% with no single or group of indicators standing out.

In terms of in-time test performance (Table 13 in the Annex), random forest remains the best performing method in most of the performance metrics (AUC ROC, Brier score, Sensitivity, WBA2). Decision tree and neural network follow closely, with logistic regression again showing the lowest performance. When looking at the out-of-time test dataset (Table 14 in the Annex), random forest remains an adequate tool for predicting bank distress, with overall good performance. However, the performance of the remaining methods, particularly in terms predicting distress events as showcased by Sensitivity, is substantively reduced.

Prediction horizon

Second, we examine if our results are sensitive to the choice of a 1-year prediction horizon. We re-run our analysis using different prediction horizons: two quarters, three quarters and six quarters.

The list of selected features increases with a longer prediction horizon. While a prediction horizon of two quarters results in 24 selected features, a prediction horizon of six quarters considers 43 features. The individual indicators selected for each prediction horizon stay broadly the same. In general, we observe that indicators selected for a shorter prediction horizon also appear in models with longer time horizons. Indicators that are added in models with longer prediction horizons are either already selected indicators with different lags or indicators that related to the same risk. Starting from a prediction horizon of 4 quarters, macroeconomic indicators appear, and their importance increases with longer horizons.

Table 15 to Table 20 in the Annex show the models' performance across the three different horizons. Overall, the random forest performs best in both the in-time test data and out-of-time test data, albeit in shorter horizons (2, 3) decision tree also gain some prominence. Logistic regression and neural network maintain adequate levels of performance, although with reduced performance in predicting distress (Sensitivity) for longer horizons.

5. Conclusions

Machine learning techniques can be powerful ingredients of early warning systems. We found Random Forest to be the technique with the best predictive performance, demonstrating the highest AUC ROC, Sensitivity and Specificity scores. It also has the best overall calibration of the fitted probabilities, having the lowest Brier score. Decision Trees follow closely, showcasing comparable results across various performance metrics. Logistic regression and Neural Networks demonstrate satisfactory performance albeit with some shortcomings.

To overcome the lack of a sufficient number of bank failures in our dataset, we adopt a new definition of distress to build an early warning system for predicting distress of large EU banks. The definition establishes specific thresholds above the regulatory requirements for capital, liquidity and allows to identify banks that fall below these set thresholds. This novel approach is directly aligned with the supervisory risk assessment framework and captures "weak" banks that merit higher supervisory attention. To adjust the model to the local approach to banking supervision, a multi-class model could be employed, which can distinguish between different severity levels (using multiple early warning distress thresholds) and types (liquidity, capital, leverage) of distress.

We also use a holistic feature selection approach to identify the most significant indicators of bank distress – prior to the model development phase – from a comprehensive list of bank-specific and country-specific macro-financial indicators. This improves the efficiency of the modelling process and ensures that only the most relevant variables enter our model. We find that the most informative indictors are associated with bank profitability, solvency and the sovereign-bank nexus.

Unlike previous studies, we employ a series of sampling techniques and show that these can significantly improve the model's ability to identify distress events, irrespective of the method used. Our study provides an important contribution to the literature of early warning systems, which has rarely used these techniques in the past to improve model performance. Our results suggest that sampling techniques are highly relevant when building early warning systems for such rare events as bank distress.

Using the latest machine learning interpretability tools, we find that the most influential variables, as determined by Shapley values, are closely tied to bank profitability, including average interest expense of deposits and asset-deposit spread for non-financial corporations, as well as solvency indicators such as equity to total liabilities and equity.

Finally, we test the performance of three ensemble techniques and find that sometimes they can outperform the single best performing model. Our findings are robust across various prediction horizons and alternative definitions of distress.

Overall, our study provides valuable insights for an effective implementation of early warning systems for the banking sector. Banking supervision and macroprudential authorities can utilise our findings to identify bank weaknesses ahead of time and adopt pre-emptive measures to safeguard financial stability.

6. References

Aldasoro, I., Borio, C. E., & Drehmann, M. (2018). Early Warning Indicators of Banking Crises: Expanding the Family. *BIS Quarterly Review*.

BCBS. (2015). *Guidelines for identifying and dealing with weak banks.* Basel Committee on Banking Supervision.

BCBS. (2017). *Basel III: Finalising post-crisis reforms.* Retrieved from https://www.bis.org/bcbs/publ/d424.htm

Betz, F., Oprică, S., Peltonen, T. A., & Sarlin, P. (2014). Predicting distress in European banks. *Journal of Banking & Finance, 45*, 225-241. doi:https://doi.org/10.1016/j.jbankfin.2013.11.041

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, Berlin, Heidelberg.

Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics, 145*.

Bräuning, M., Malikkidou, D., Scalone, S., & Scricco, G. (2020). A New Approach to Early Warning Systems for Small European Banks. *LOD 2020: International Conference on Machine Learning, Optimization, and Data Science* (pp. 551-562). Siena: Springer.

Breiman, L. (2001). Random Forest. *Machine Learning, 45*, 5-32.

Cole, R. A., & Wu, Q. (2018). *Hazard versus probit in predicting U.S. bank failures: A regulatory perspective over two crises.*

Detken, C., Weeken, O., Alessi, L., Bonfim, D., Boucinha, M., Castro, C., . . . Kakes, J. (2014). Operationalising the countercyclical capital buffer: indicator selection, threshold identification and calibration options. *ESRB Occasional paper Series*.

Duca, M. L., & Peltonen, T. A. (2013). Assessing systemic risks and predicting systemic events. *Journal of Banking & Finance, 37*(17), 2183-2195.

EBA. (2022). *Guidelines on common procedures and methodologies for the supervisory review and evaluation process (SREP) and supervisory stress testing under Directive 2013/36/EU.* Retrieved from https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Guidelines/2022/EBA-GL-2022-03%20Revised%20SREP%20Guidelines/1028500/Final%20Report%20on%20Guidelines%20on%20common%20procedures%20and%20methodologies%20for%20SREP%20and

EBA. (2023). *Follow-up report on the use of machine learning for internal ratings-based models*. Retrieved from https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf

EU. (2013a). *Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32013R0575

EU. (2013b). *Directive (EU) 2013/36 of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32013L0036

EU. (2014a). *Commission Delegated Regulation (EU) 2015/61 of 10 October 2014 to supplement Regulation (EU) No 575/2013 of the European Parliament and the Council with regard to liquidity coverage requirement for Credit Institution.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015R0061

EU. (2014b). *Directive (EU) 2014/59 of the European Parliament and of the Council of 15 May 2014 establishing a framework for the recovery and resolution of credit institutions and investment firms*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0059

Ferriani, F., Cornacchia, W., Farroni, P., Ferrara, E., Guarino, F., & Pisanti, F. (2019). *An early warning system for less significant Italian banks.* Bank of Italy, Economic Research and International Relations Area.

Fuertes, A.-M., & Kalotychou, E. (2017). Optimal design of early warning systems for sovereign debt crises. *International Journal of Forecasting, 23*(1), 85-100.

Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2018). Forecasting bank failures and stress testing: A machine learning approach. *International Journal of Forecasting, 34*(3), 440-455.

Kuhn, M. &. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models.* Chapman and Hall/CRC.

Lang, J., Peltonen, T., & Sarlin, P. (2018). A framework for early-warning modeling with an application to banks. *ECB Working paper*.

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

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection: A Data Perspective. *ACM Computing Surveys (CSUR), 50*(6), 1 - 45.

Lo Duca, M., & Peltonen, .. (2013). Assessing Systemic Risks and Predicting Systemic Events. *Journal of Banking & Finance, 37*(7), 2183-2195.

Murphy, K. (2012). *Probabilistic machine learning : an introduction.* MIT press.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting, 36*(3), 1092-1113.

Sarkar, D., Bali, R., & Sharma, T. (2017). *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems.* Apress Berkeley, CA.

Suss, J., & Treitel, H. (2019). Predicting bank distress in the UK with machine learning. *Bank of England Staff Working Paper* , No. 831.

Tölö, E. (2020). Predicting systemic financial crises with recurrent neural networks. *Journal of Financial Stability, 49*.

7. Annex

7.1. Sample

Table 10: Sample composition

Country	Number of banks
AT	8
BE	7
BG	1
CY	5
DE	29
DK	5
EE	2
ES	15
FI	5
FR	12
GR	4
HU	2
IE	9
IS	3
IT	18
LI	3
LT	3
LU	8
LV	2

Country	Number of banks
MT	5
NL	6
NO	3
PL	2
PT	7
RO	1
SE	6
SI	5
Total	176

7.2. Selected features

Table 11 Final set of features, definitions and transformations, sources

Category	Variable	Definition & transformation	Source
Asset quality	AQT_14	Post-CRM exposure to original exposure	EBA risk indicators
Asset quality	AQT_42.2.5	Forbearance ratio (gross amount) for loans and advances- Non-financial corporations	EBA risk indicators
Asset quality	AQT_68.1a_d_1	Share of financial instruments measured at FV through P&L in total IFRS 9 assets, q-o-q change	EBA risk indicators
Asset quality	AQT_68.3	Share of financial instruments measured at (amortised) cost in total financial instruments	EBA risk indicators
Funding	FND_14_CALC_pct_1	Total assets, q-o-q percentage change	EBA risk indicators
Funding	FND_14_CALC_pct_4	Total assets, y-o-y percentage change	EBA risk indicators
Funding	FND_18_d_4	Customer deposits to total liabilities, y-o-y change	EBA risk indicators
Funding	FND_33_d_4	Asset encumbrance ratio, y-o-y change	EBA risk indicators
Funding	FND_33	Asset encumbrance ratio	EBA risk indicators
Funding	FND_34_d_1	Average interest expense of deposits, q-o-q change	EBA risk indicators
Funding	FND_34_d_4	Average interest expense of deposits, y-o-y change	EBA risk indicators
Funding	FND_34	Average interest expense of deposits	EBA risk indicators
Macroeconomic and sectoral statistics	Govt_Debt_to_GDP	Government debt (consolidated) (as % of GDP)	SDW
Macroeconomic and sectoral statistics	HICP_YoY_growth	HICP overall index, annual rate of change	SDW

Category	Variable	Definition & transformation	Source
Liquidity	LIQ_5	Withdrawable funding (% of total liabilities)	EBA risk indicators
Profitability	PFT_22_CALC	Return on regulatory capital requirements	EBA risk indicators
Profitability	PFT_26	Net fee and commission income to total net operating income	EBA risk indicators
Profitability	PFT_32	Net income to total net operating income	EBA risk indicators
Profitability	PFT_39	Asset-deposit spread for non-financial corporations	EBA risk indicators
Profitability	PFT_45	Impairment and provisioning on financial asset to Net Ordinary Operating Income	EBA risk indicators
Profitability	PFT_5	Interest income from households	EBA risk indicators
RDB	RDB_3	Debt securities on Total Assets	EBA risk indicators
RDB	RDB_6	Other assets on Total Assets	EBA risk indicators
SME	SME_14	Post-CRM SME exposure to original SME exposure	EBA risk indicators
SME	SME_16	Increase in CET1 capital ratio with the application of SME supporting factor	EBA risk indicators
Solvency	SVC_23	Retained earnings and reserves to total equity	EBA risk indicators
Solvency	SVC_26	Equity to total liabilities and equity	EBA risk indicators
Solvency	SVC_28_CALC _log_pct_1	Total RWA, q-o-q percentage change	EBA risk indicators
Solvency	SVC_28_CALC _pct_4	Total RWA, y-o-y percentage change	EBA risk indicators

Category	Variable	Definition & transformation	Source
Sovereign	SVR_29	Share of Sovereign Exposures of Total Assets	EBA risk indicators

Notes: Only 30 variables are presented in this table instead of the total of 33 variables that enter in the model, since 3 variables enter twice with different lags.

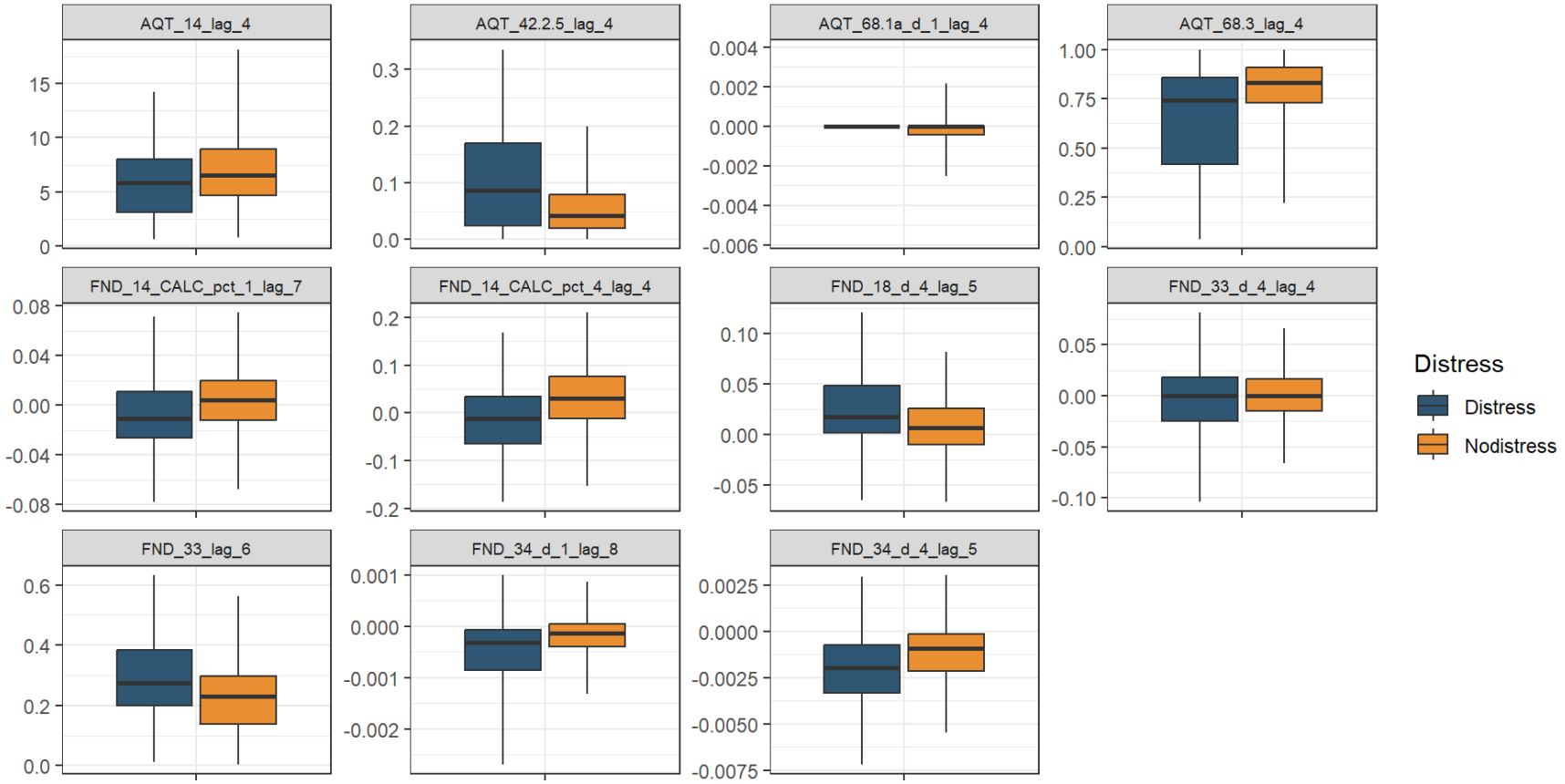
Table 12 Summary statistics of selected features

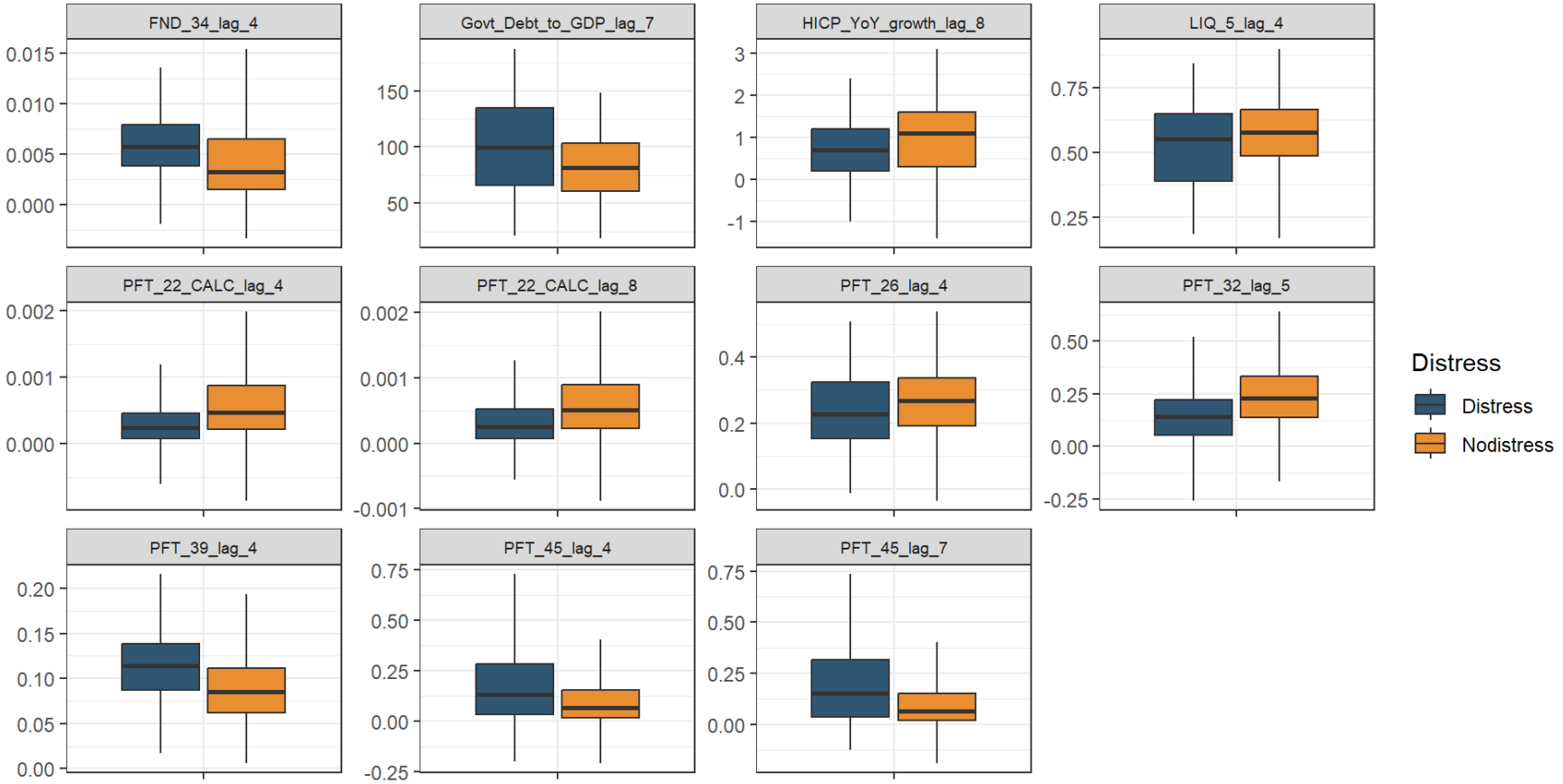
Variable	N	Mean	St. dev.	25th Pct	Median	75th Pct
AQT_14_lag_4	2371	10	15	4.8	7	10
AQT_42.2.5_lag_4	2371	0.073	0.075	0.021	0.046	0.096
AQT_68.1a_d_1_lag_4	2371	-0.033	0.17	-0.0019	0	0.00064
AQT_68.3_lag_4	2371	0.73	0.25	0.63	0.82	0.91
FND_14_CALC_pct_1_lag_7	2371	0.0062	0.042	-0.015	0.0045	0.025
FND_14_CALC_pct_4_lag_4	2371	0.036	0.11	-0.019	0.033	0.078
FND_18_d_4_lag_5	2371	0.0085	0.04	-0.01	0.0069	0.027
FND_33_d_4_lag_4	2371	0.00086	0.039	-0.016	0.00023	0.019
FND_33_lag_6	2371	0.25	0.15	0.15	0.25	0.33
FND_34_d_1_lag_8	2371	-0.00041	0.001	-0.00063	-0.0002	0.0000061
FND_34_d_4_lag_5	2371	-0.0014	0.0025	-0.0024	-0.0012	-0.00023
FND_34_lag_4	2371	0.0048	0.0055	0.0011	0.0035	0.0069
Govt_Debt_to_GDP_lag_7	2371	90	39	62	85	114
HICP_YoY_growth_lag_8	2371	0.97	1.1	0.2	1	1.7

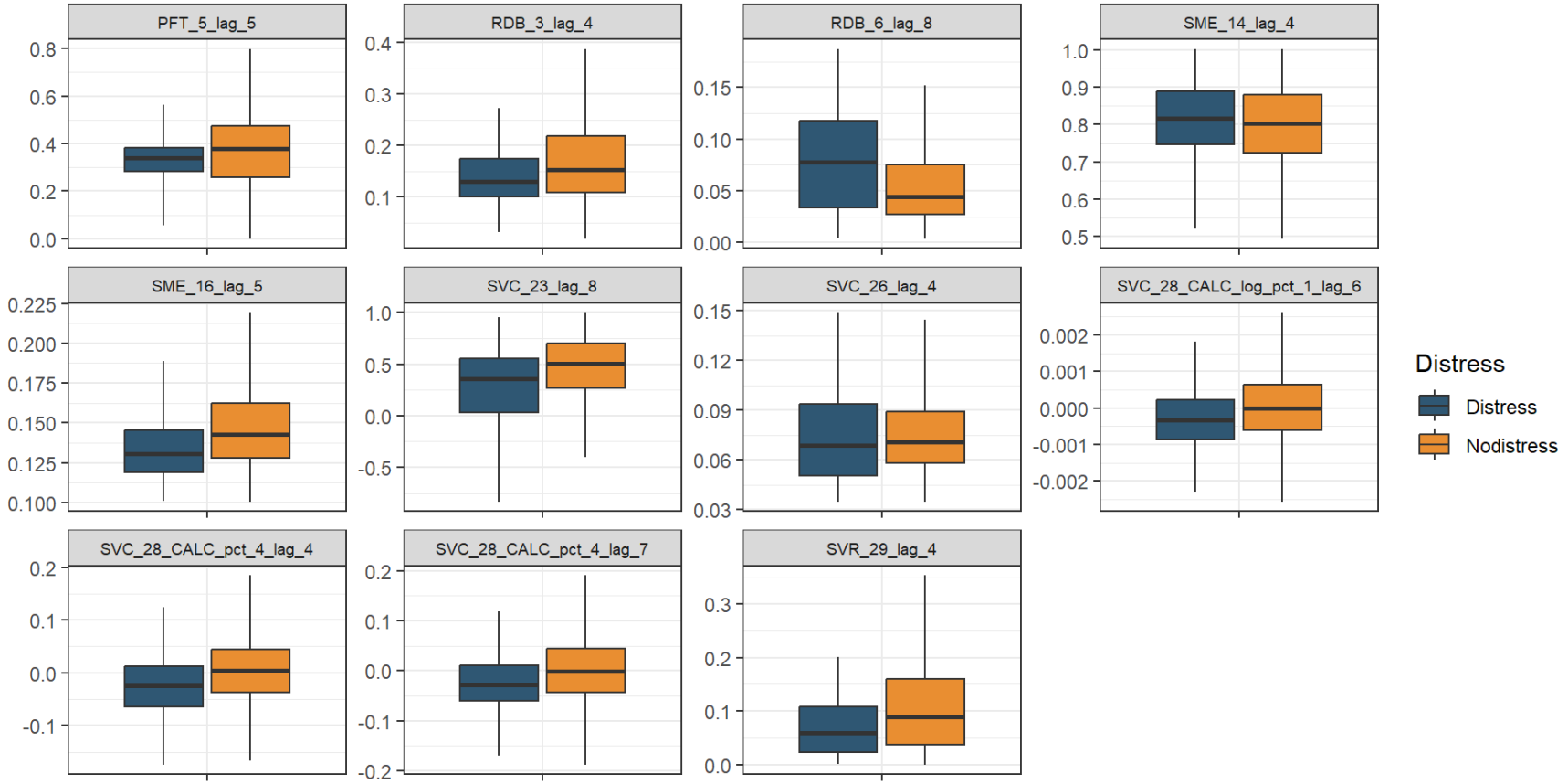
Variable	N	Mean	St. dev.	25th Pct	Median	75th Pct
LIQ_5_lag_4	2371	0.79	0.58	0.49	0.61	0.76
PFT_22_CALC_lag_4	2371	0.00056	0.00084	0.00019	0.00047	0.00092
PFT_22_CALC_lag_8	2371	0.00054	0.00087	0.00018	0.00046	0.00089
PFT_26_lag_4	2371	0.27	0.14	0.18	0.27	0.34
PFT_32_lag_5	2371	0.2	0.36	0.11	0.22	0.33
PFT_39_lag_4	2371	0.096	0.051	0.064	0.088	0.12
PFT_45_lag_4	2371	0.12	0.2	0.017	0.069	0.17
PFT_45_lag_7	2371	0.15	0.25	0.021	0.076	0.19
PFT_5_lag_5	2371	0.35	0.18	0.25	0.36	0.47
RDB_3_lag_4	2371	0.16	0.08	0.11	0.15	0.21
RDB_6_lag_8	2371	0.058	0.042	0.026	0.046	0.079
SME_14_lag_4	2371	0.79	0.12	0.72	0.8	0.88
SME_16_lag_5	2371	0.15	0.046	0.13	0.14	0.17
SVC_23_lag_8	2371	0.42	0.41	0.25	0.5	0.69
SVC_26_lag_4	2371	0.077	0.028	0.056	0.07	0.09
SVC_28_CALC_log_pct_1_lag_6	2371	0.000049	0.0017	-0.00066	-0.0000059	0.00064

Variable	N	Mean	St. dev.	25th Pct	Median	75th Pct
SVC_28_CALC_pct_4_lag_4	2371	0.015	0.12	-0.041	0.0049	0.052
SVC_28_CALC_pct_4_lag_7	2371	0.0092	0.12	-0.046	0.0006	0.046
SVR_29_lag_4	2371	0.11	0.089	0.039	0.087	0.16

Figure 9 Boxplot of selected features, by distress/non-distress events


AQT_14_lag_4
AQT_42.2.5_lag_4
AQT_68.1a_d_1_lag_4
AQT_68.3_lag_4
FND_14_CALC_pct_1_lag_7
FND_14_CALC_pct_4_lag_4
FND_18_d_4_lag_5
FND_33_d_4_lag_4
FND_33_lag_6
FND_34_d_1_lag_8
FND_34_d_4_lag_5
Distress
Distress
Nodistress


FND_34_lag_4
Govt_Debt_to_GDP_lag_7
HICP_YoY_growth_lag_8
LIQ_5_lag_4
PFT_22_CALC_lag_4
PFT_22_CALC_lag_8
PFT_26_lag_4
PFT_32_lag_5
PFT_39_lag_4
PFT_45_lag_4
PFT_45_lag_7
Distress
Distress
Nodistress


PFT_5_lag_5
RDB_3_lag_4
RDB_6_lag_8
SME_14_lag_4
SME_16_lag_5
SVC_23_lag_8
SVC_26_lag_4
SVC_28_CALC_log_pct_1_lag_6
SVC_28_CALC_pct_4_lag_4
SVC_28_CALC_pct_4_lag_7
SVR_29_lag_4
Distress
Distress
Nodistress

7.3. Robustness checks

Table 13 Validation results for models based on the 10th percentile distress definition, in-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.7447	0.8132	0.7974	0.7342
Sensitivity	0.806	0.8806	0.8358	0.8209
Specificity	0.7316	0.7987	0.7891	0.7157
Balanced Accuracy	0.7688	0.8397	0.8125	0.7683
WBA1	0.7502	0.8192	0.8008	0.742
WBA2	0.7874	0.8601	0.8242	0.7946
AUC ROC	0.8299	0.9046	0.878	0.8494
Brier score	0.1784	0.143	0.147	0.198

Table 14 Validation results for models based on the 10th percentile distress definition, out-of-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.8395	0.8884	0.8837	0.7884
Sensitivity	0.7143	0.8571	0.7143	0.7857
Specificity	0.8438	0.8894	0.8894	0.7885
Balanced Accuracy	0.779	0.8733	0.8019	0.7871
WBA1	0.8114	0.8814	0.8456	0.7878
WBA2	0.7467	0.8652	0.7581	0.7864
AUC ROC	0.9069	0.9317	0.8776	0.8226
Brier score	0.1223	0.1052	0.0894	0.1522

Table 15 Validation results for models based on 2 quarter prediction horizon, in-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.6966	0.8301	0.8131	0.7379
Sensitivity	0.6912	0.8382	0.8529	0.8235
Specificity	0.6977	0.8285	0.8052	0.7209
Balanced Accuracy	0.6944	0.8334	0.8291	0.7722
WBA1	0.696	0.8309	0.8172	0.7466
WBA2	0.6928	0.8358	0.841	0.7979
AUC ROC	0.7859	0.8953	0.8979	0.831
Brier score	0.1856	0.1326	0.1371	0.2245

Table 16 Validation results for models based on the 2 quarter prediction horizon, out-of-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.7727	0.8593	0.8074	0.7208
Sensitivity	0.6667	0.7333	0.7333	0.7333
Specificity	0.7763	0.8635	0.8098	0.7204
Balanced Accuracy	0.7215	0.7984	0.7716	0.7268
WBA1	0.7489	0.831	0.7907	0.7236
WBA2	0.6941	0.7659	0.7525	0.7301
AUC ROC	0.8594	0.8758	0.8888	0.7745
Brier score	0.1429	0.1212	0.1484	0.249

Table 17 Validation results for models based on the 3 quarter prediction horizon, in-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.7321	0.8061	0.8367	0.7168
Sensitivity	0.7727	0.8182	0.803	0.7576
Specificity	0.7239	0.8037	0.8436	0.7086
Balanced Accuracy	0.7483	0.8109	0.8233	0.7331
WBA1	0.7361	0.8073	0.8334	0.7208
WBA2	0.7605	0.8146	0.8132	0.7453
AUC ROC	0.8114	0.8958	0.896	0.8233
Brier score	0.1861	0.1392	0.1291	0.2204

Table 18 Validation results for models based on the 3 quarter prediction horizon, out-of-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.8018	0.9099	0.8896	0.6486
Sensitivity	0.6667	0.7333	0.8667	0.7333
Specificity	0.8065	0.9161	0.8904	0.6457
Balanced Accuracy	0.7366	0.8247	0.8786	0.6895
WBA1	0.7716	0.8704	0.8845	0.6676
WBA2	0.7016	0.779	0.8726	0.7114
AUC ROC	0.8758	0.9345	0.9556	0.7392
Brier score	0.1281	0.0922	0.0929	0.2818

Table 19 Validation results for models based on the 6 quarter prediction horizon, in-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.7288	0.8274	0.7973	0.7726
Sensitivity	0.7288	0.9492	0.8644	0.8136

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Specificity	0.7288	0.8039	0.7843	0.7647
Balanced Accuracy	0.7288	0.8765	0.8244	0.7891
WBA1	0.7288	0.8402	0.8043	0.7769
WBA2	0.7288	0.9128	0.8444	0.8013
AUC ROC	0.8171	0.9219	0.8976	0.85
Brier score	0.1771	0.1335	0.1422	0.185

Table 20 Validation results for models based on the 6 quarter prediction horizon, out-of-time test dataset

Performance metric	Logit	Random forest	Decision tree (C5.0)	Neural network
Accuracy	0.6721	0.8776	0.9007	0.6697
Sensitivity	0.7692	0.8462	0.6923	0.3846
Specificity	0.669	0.8786	0.9071	0.6786
Balanced Accuracy	0.7191	0.8624	0.7997	0.5316
WBA1	0.6941	0.8705	0.8534	0.6051
WBA2	0.7442	0.8543	0.746	0.4581
AUC ROC	0.8141	0.8857	0.9207	0.6126
Brier score	0.2075	0.1194	0.0903	0.2561

ACKNOWLEDGEMENTS

We would like to thank an anonymous referee for his/her comments. Despo Malikkidou:
Team Leader, European Banking Authority

Wolfgang Strohbach:
Senior Bank Expert, European Banking Authority

EUROPEAN BANKING AUTHORITY

20 avenue André Prothin CS 30154
92927 Paris La Défense CEDEX, France

Tel. +33 1 86 52 70 00
E-mail: info@eba.europa.eu

https://eba.europa.eu/

ISBN 978-92-9245-978-9
ISSN 2599-7831

doi: 10.2853/8659719
DZ-01-25-000-EN-N