EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Luca Nocciola, Samuele Scaglioni

Learning probability of default and stress testing

No 3277

Abstract

We analyze the Probability of Default (PD) of non-financial corporations in Europe using Random Forests (RF) and assess implications for stress testing the banking sector. To this end, we exploit data on firms' financial statements (Orbis) and banks' credit registry (Anacredit). We show that RF displays stronger risk sensitivity than logistic regression in stress testing, shedding new light on the non-linear effect of scenario severity on PD. Moreover, we show how RF-based PD can be used in a network of banks and firms to stress test the banking sector through loan exposures as a key transmission channel of adverse scenarios. A granular inspection of banks' riskiness indices derived from this network sheds light also on RF's superior ability in capturing non-linearity thanks to its capability in identifying "tail banks". Our work is relevant for central banks and banking supervisors alike.

JEL classification: C53, C55, C58, G17, G21

Keywords: Credit risk; Financial stability; Banking supervision; Random forests; Corporate exposures

Non-technical summary

This paper uses Machine Learning (ML), a subfield of Artificial Intelligence (AI), to build a granular stress-testing framework for banks' corporate credit risk. We use Random Forests (RF) and show that, beyond improving the prediction of firms' Probability of Default (PD) relative to logistic regression, RF is better able to capture the non-linear and heterogeneous transmission of stress to the banking sector.

In particular, we use RF to link non-financial firms' default risk to firm-level balance-sheet and profit-and-loss data, together with macroeconomic and sectoral conditions. We first exploit RF's variable-importance measures to identify the main drivers of firms' PD and guide how we shock firms' financial statements, ensuring that scenarios affect the most important determinants of default risk. We then use scenarios from an EU-wide stress test and map them into heterogeneous, firm-specific shocks: firms within the same country–sector are affected according to their financial structure and exposure to aggregate shocks.

Second, we use RF to project PDs under these shocked financials. RF is substantially more sensitive than logit to adverse macro and sectoral developments. This higher scenario sensitivity reveals non-linearity: increases in interest rates and drops in GVA growth have different non-linear effects on firms' PD once severity crosses certain thresholds, which can also inform monetary policy and its credit-risk effects.

Finally, we use the stressed PDs in a bipartite network of firms and banks, where loan exposures form the links. This network translates firm-level default risk into bank-level riskiness and allows us to assess the scenario impact on the whole banking system. The RF-based stress test produces risk metrics that are severe yet plausible and responsive to adverse scenarios, making it a suitable tool for stress testing. A granular inspection of banks' riskiness indices derived from this network also highlights RF's superior ability to capture non-linearity through its ability to identify "tail banks".

This paper thus fills a gap in current central-bank and supervisory practice, where ML is largely absent from stress-testing toolkits. By connecting ML, stress testing and network analysis in a single framework, we show that RF can be a valuable tool for stress tests, not just for improving forecasts.

1 Introduction

The Probability of Default (PD) is a key credit risk parameter used by banks to calculate expected credit losses for provisions and impairments and, indirectly, to compute unexpected credit losses.[1] In this paper, we leverage Machine Learning (ML), a subfield of Artificial Intelligence (AI), to predict and stress test PD of non-financial firms and, in turn, to stress test banks exposed to them. We use Random Forests (RF), a supervised ML algorithm capable of handling high-dimensionality and non-linearity thanks to its non-parametric nature. We first assess RF's in-sample and out-of-sample performance at the aggregate and micro level by comparing it against a benchmark used in the literature, i.e., logistic regression (logit). Second, we use RF to identify the main drivers of PD via variable importance measures, which is critical to stress testing. Third, we run a granular stress test by shocking firms' financial statements according to their variable importance ranking. We exploit the macro and sectoral scenarios used in the 2023 EU-wide banking sector stress test (EBA (2023)) and spread the shocks heterogeneously across companies belonging to the same country-sectors through their books. We also evaluate sensitivity of our results to scenario severity, thus unveiling potentially interesting non-linearity. Finally, we show how RF-based PD can be used in a bipartite network of firms and banks to stress-test banks, where loan exposures (EAD) act as a key transmission channel, and show also how RF-based PD can be used to derive granular and aggregate banks' risk indices.

This paper builds on several key studies. Barbaglia *et al.* (2023) use 12 million European mortgages to show that tree-based models outperform traditional methods in forecasting defaults across regions. Khandani *et al.* (2010) demonstrate the usefulness of tree ensembles and kernel machines for individual delinquency prediction using combined credit-bureau and transaction-level data.[2] Similarly, Petropoulos *et al.* (2020), Ekinci & Sen (2024) and Bellotti *et al.* (2021) report that RF is a superior method relative to other ML algorithms and logistic regression in predicting banks' insolvencies and recovery rates. Likewise, Jones *et al.* (2015) find that RF is a strong performer in classifying credit ratings changes of public companies, although boosting seems the strongest contender. Also, Galindo & Tamayo (2000) report that CART decision-tree models perform best in classifying mortgage-loan defaults. Kaposty *et al.* (2020) show that RF works well in predicting LGD, while Sadhwani *et al.* (2021) demonstrate the highly non-linear

[1] Together with the Loss Given Default (LGD) and Exposure At Default (EAD), it constitutes the backbone of credit risk modeling. Under the IFRS9 accounting standard, the PD is equivalent to the probability of transitioning from stage 1 or stage 2 loans to stage 3 loans, i.e., is a function of transition rates involving these stages.

[2] More generally, Butaru *et al.* (2016) also show the accuracy of ML algorithms to predict delinquency.

influence on mortgage borrowers' behaviour of loan-specific, meso and macro variables. Peng *et al.* (2025) find that RF yields the best performance in forecasting failure of credit unions, while Uddin *et al.* (2022) document the value of hybrid ML classifiers, including, e.g., a mix of RF and boosting, against base classifiers alone. By contrast, Beutel *et al.* (2019) show that, while ML algorithms have good in-sample fit, ML algorithms struggle in out-of-sample forecasting of banking crises against a benchmark logistic regression[3] and, similarly to du Jardin (2025), Kim *et al.* (2022) and Dendramis *et al.* (2025), find that neural networks[4] are superior (to RF and logit, respectively) in forecasting corporate bankruptcy, while Nazemi *et al.* (2022) show that deep learning works best in predicting recovery rates of consumer credit.

This literature has not made it into the central-banking stress testing toolbox yet. Budnik *et al.* (2024) summarize the ECB toolkit used for top-down stress testing purposes and quality assurance of banks' projections during the regular EU-wide stress tests, including a quantile model (QPD) to predict sectoral PD of non-financial corporations (Konietschke *et al.* (2026)), but no ML algorithm is used there. Chavleishvili & Manganelli (2024) introduce a quantile vector autoregression for stress testing, but, again, without using ML. Chan-Lau (2017) propose LASSO regression for forecasting and stress testing sectoral PD, but LASSO, contrary to RF, is not able to capture non-linearity.[5] Not explicitly linked to stress testing, Errico *et al.* (2025) document, via a generalized RF, the heterogeneous and non-linear response of firms to macro fluctuations, highlighting the importance of accounting for these two dimensions for understanding the transmission of aggregate shocks. The academic literature linking ML to stress testing is also scant. Gogas *et al.* (2018) show that Support Vector Machines (SVM) work well in forecasting bank failures and stress testing banks, but SVM perform less well than RF in feature selection, which is crucial in stress testing. Although not applied to stress testing, Conti & Morelli (2025) introduce a Bayesian logit enhanced by a Stochastic Gradient Langevin Dynamics algorithm to reduce overfitting often related to ML algorithms, but RF, averaging over many trees trained on different samples and features, reduces this issue, too.[6] This paper fills this gap, i.e. the rare application of ML algorithms to stress testing. Moreover, by using RF,

[3] Recently, some literature improved logistic regression in various directions, e.g., Audrino *et al.* (2019).

[4] Liu *et al.* (2024) find that Bayesian networks are superior in terms of forecasting performance and causal analysis to ML algorithms.

[5] Sarlin & Holopainen (2017) and Bräuning *et al.* (2019) suggest ML as an early-warning tool to predict financial distress.

[6] Skavysh *et al.* (2023) use quantum computing to improve the runtime of stress testing models and macro models with ML algorithms within them.

this paper enhances the current literature which uses, e.g., LASSO and SVM, which are not ideal to capture non-linearity and perform feature selection, both key aspects of stress testing. Finally, this paper also connects the ML, stress testing and network theory literature in one framework.[7]

Our study shows the usefulness of RF for stress-testing exercises, in addition to merely improving forecasting performance relative to a logit benchmark. On top of finding a superior prediction ability of RF compared to logit both at the aggregate and micro-level[8], when stressing firms' financial statements through scenarios, we find that RF is much more sensitive than logit to adverse macro and sectoral shocks. Scenario sensitivity analyses highlight interesting non-linear effects of an increase in interest rates and a depression in GVA growth on PD of non-financial companies, which can also inform monetary policy and its credit-risk effects. Also, by using the bipartite network of exposures among banks and firms and stress testing it using scenario-dependent PD estimated by RF, we showcase RF's ability in delivering severe-yet-plausible risk metrics for the whole banking sector. A granular inspection of banks' riskiness indices derived from this network also highlights RF's superior ability to capture non-linearity through its ability to identify "tail banks".

This paper is structured as follows. Section 2 describes the data and methodology. Section 3 reports model performance and interpretability. Section 4 introduces the stress testing framework. Finally, section 5 concludes.

2 Data and methodology

We use Orbis (Bureau van Dijk, a Moody's Analytics company) as the primary source for firm-level financial statements.[9] We draw a representative sample of 10,000 non-financial firms (i.e., without NACE sector K) from the Orbis universe. Sampling is quota-based: first by country,

[7] Another relevant paper investigating corporate PD using standard econometric models is by Castrén *et al.* (2010). Credit risk modelling is not the only domain of application of ML algorithms. Economists have been injecting ML into the broader econom(etr)ics' literature in recent years. In macroeconomics and finance, one of the main topics of application has been forecasting, see, e.g., Bluwstein *et al.* (2023), Lenza *et al.* (2025), Medeiros *et al.* (2021), Coulombe (2024), Coulombe *et al.* (2022), Buse *et al.* (2025), Aldasoro *et al.* (2025), Beck & Wolf (2025). In addition, Bie *et al.* (2024) use ML to deal with structural breaks, a known problem that under certain conditions may negatively affect forecasting performance (Nocciola (2022)). Another literature strand employs ML, in particular RF, for inference like hypothesis testing for unit roots in macro-financial time series, see, e.g., Nocciola *et al.* (2023).

[8] These findings are in line with Barbaglia *et al.* (2023), Khandani *et al.* (2010), Petropoulos *et al.* (2020), Ekinci & Sen (2024), Bellotti *et al.* (2021), Jones *et al.* (2015) and Peng *et al.* (2025).

[9] Orbis harmonizes accounting definitions across jurisdictions, enabling comparability over time and across countries.

proportionally to GDP weights, and then by sector within each country.[10] Within each country-sector cell, we draw firms uniformly at random.[11] To ensure sufficient data quality, we restrict attention to firms with at least five years of observations, at least one non-missing failure flag and non-missing core accounting items needed to compute the main ratios and dynamics mentioned below. Further details on exact quota allocation are provided in Appendix A.

From this sample, we extract firms' financial statements (balance sheet, income statement, cash flow) and demographics (industry, country, incorporation year). Starting from annual filings, we construct a quarterly firm-time panel. Each statement is time-stamped by its accounting reference date and aligned to calendar quarters.[12] Financial statements' items enter primarily through ratios or $\log(1+x)$-transforms to reduce scale and currency effects.[13] Categorical firm demographics (country, sector, legal status, consolidation code) are retained for one-hot encoding in the modeling stage. We then merge quarterly macro controls by calendar quarter and country of the firm. Additional details on processing are also reported in Appendix A.

Response variable: default status. Our default concept is cash-flow-based and equivalent to Budnik *et al.* (2024): a firm is said to default in quarter t if its cash inflows from operations are insufficient to cover its financial expenses in that quarter. Let $\mathrm{CI}_{i,t}$ denote cash on hand plus cash inflows from operating activities (from the cash-flow statement) and $\mathrm{FE}_{i,t}$ denote financial expenses (interest and other charges).[14] We define the firm-quarter default indicator

$$D_{i,t} \;=\; \mathbb{1}\{\,\mathrm{CI}_{i,t} < \mathrm{FE}_{i,t}\,\} \tag{1}$$

i.e., default occurs when cash inflows are lower than financial expenses.[15] When aggregating, the default rate or prevalence in a cell s (e.g., country, industry, or size bin) at time t is the share of firms that default, i.e., $DR_{s,t} = \frac{1}{N_{s,t}} \sum_{i \in \mathcal{S}_{s,t}} D_{i,t}$ where $\mathcal{S}_{s,t}$ is the set of borrower-firm observations in cell s at time t and $N_{s,t} = |\mathcal{S}_{s,t}|$. As pointed out in figure 1, panel (a), defaults

[10] Quotas are adjusted for availability constraints using standard remainder methods.

[11] We seed the random number generator to guarantee reproducibility.

[12] When within-year quarterly accounts are not available, the most recent filing is treated as the information set until the next filing becomes available. Where duplicate or synonymous fields exist (e.g., multiple liabilities totals or country codings), we keep a single canonical series for estimation.

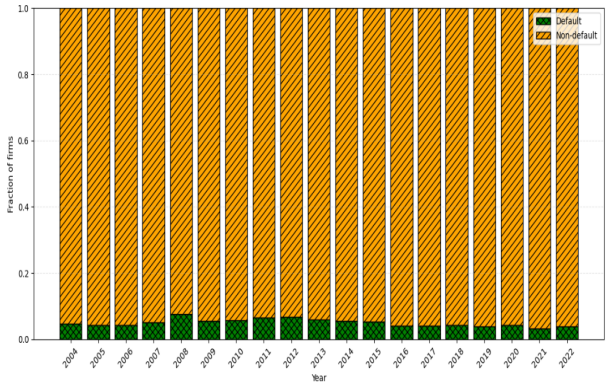
[13] Variables are winsorized at 1% to limit the influence of outliers. For dynamics (lags, differences, percentage changes), all computations are performed in firm-time order using only historical observations. Computations are available upon request.

[14] Here, cash inflows are defined as a "liquidity buffer" proxy that sums a stock (cash and cash equivalents, end-of-period) with a flow (cash flow over the period) following Budnik *et al.* (2024) and Konietschke *et al.* (2026).

[15] As an alternative, see definition of Gourinchas *et al.* (2020).

are rare, a phenomenon called "class imbalance".[16] When $DR_{s,t} \ll 1$, a trivial classifier that always predicts non-default achieves accuracy $1 - DR_{s,t}$, despite having no ability to identify defaults. Heterogeneity in $DR_{s,t}$ across geographies (figure 1, panel (b)) underscores that some are more prone to class imbalance than others. Appendix B further describes pitfalls under class imbalances and how we address these issues.

(a) Default rate over time



(b) Default rate by NUTS2 region

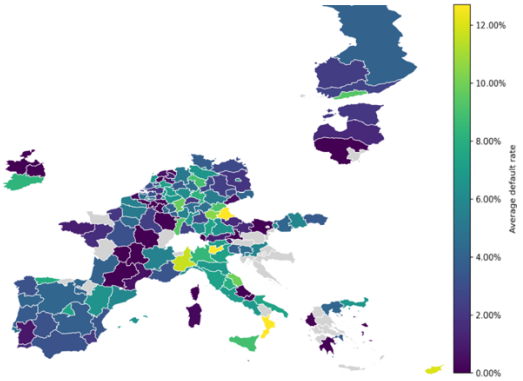


Figure 1: Class imbalance between defaulted and non-defaulted firms over time (panel (a)); and geolocalization of default rates by NUTS2 region (panel (b)), respectively. Source: authors' elaboration based on Orbis.

Firm-level characteristics. To predict PD, we use firms' financial statements (and macroeconomic variables) as features. These financials can be grouped in balance-sheet items (stocks) and income-statement variables (flows). In particular, among the balance-sheet items, we have asset-like controls, like total assets, tangible and intangible assets and liability-like controls, like total liabilities, current liabilities and debt. Among income-statement items, we have sales, financial revenues, and expenses. We also use additional variables, which are summarized and described in Appendix A.2, table A3.

2.1 Methodology

Breiman (2001) introduces RF, an ML algorithm which tries to approximate a mapping between the response variable and the features non-parametrically by growing a decision forest. Randomness kicks in the combinatorial choice of a subset of features to predict the response and in

[16] In our (aggregated) sample, class imbalance amounts to ca. 5% of defaults against ca. 95% of non-defaults. Class imbalance is a recognized problem in the literature, e.g., see Shahee & Patel (2025), Atif (2025), and remedies have been proposed, e.g., Leong (2016), Shahee & Patel (2025) and Chi *et al.* (2025).

bootstrapping the sample on which each tree is trained. This randomization allows each tree to produce a quasi-independent classification. RF then takes an aggregate decision by applying, e.g., majority voting among a possibly vast number of trees. Formally, an RF tries to learn the following relation

$$D_{i,t} \;=\; f(\mathbf{x}_{i,t-1}, \mathbf{z}_{s,t-1}) \tag{2}$$

where $D_{i,t}$ is the default status as defined in equation 1, $f()$ is a function (possibly highly nonlinear and misbehaved), $\mathbf{x}_{i,t}$ are, in our paper, NFC balance-sheet and income-statement items and $\mathbf{z}_{s,t}$ are sectoral, mesoeconomic and macroeconomic variables, which are all lagged to avoid any mechanical effect of the definition of default and preserve accounting-consistency. Appendix B provides further details and an example of RF's functioning. An advantage of RF is its ability to handle high-dimensional inputs and nonlinear interactions, as RF does not impose any distributional assumption a priori, contrary to econometric models like logistic regression.[17]

3 Model performance and interpretability

3.1 Model performance

In this subsection, we compare the performance of RF and logistic regression both in-sample and out-of-sample, at the aggregate and micro levels.

Micro-to-macro aggregation. We train the classifiers at the micro level—firm i observed in quarter t—to estimate conditional PDs $\hat{p}_{i,t}^{(H)} = \Pr(D_{i,t+H} = 1 \mid \mathbf{x}_{i,t}, \mathbf{z}_{s,t})$ for horizon H (≥ 1). Training and testing are strictly on observations with time-aware, group-purged splits (firms in a test fold are excluded from the train fold and chronology is respected), so that discrimination and calibration are judged where the models actually operate. Afterwards, to obtain a macro PD time series, we aggregate the probabilities (not hard labels) across firms active in quarter t as $\overline{p}_{t+H} = \frac{1}{N_t} \sum_{i \in \mathcal{A}_t} \hat{p}_{i,t}^{(H)}$. This two-layer design makes the training/testing logic transparent at the micro level while producing a macro series directly comparable to realized default rates. Importantly, any operating threshold (e.g., chosen to maximize F_2) is used only for observation-

[17] The superiority of ML algorithms compared to traditional methods has been argued to be linked to their ability to better capture non-linearities, which are especially acute in crisis times (see, e.g., Gambacorta *et al.* (2024)).

level summaries, while quarterly series are built from $\hat{p}_{i,t}^{(H)}$ themselves.

In- and out-of-sample model performance at the aggregate level. We assess model performance at the aggregate level (both in-sample and out-of-sample) using standard metrics, like MAE, RMSE and Correlation difference. Figure 2, panel (a) contrasts realized default rates with in-sample predicted PDs from RF and logit at the aggregate level. Numerically, RF outperforms logit over all metrics. RF attains lower errors (MAE = 0.006, RMSE = 0.005) and higher correlation of quarterly changes (Correlation difference = 0.677), indicating better adherence to observed defaults and their temporal variation. Figure 2, panel (b), shows boxplots of these metrics when calculated out of sample across different forecast origins (2021Q1, 2021Q2, 2021Q3 and 2021Q4).[18] Visually, the RMSE, MAE and Correlation difference highlight superior performance of RF relative to logit across multiple cutoffs. The median MAE and RMSE for RF are 0.0025 and 0.0031 against a median MAE and RMSE for logit of 0.0031 and 0.0034, respectively. Also, the correlation between quarterly changes of PD and the default rate is greater for RF (0.514) than logit (0.359).[19]

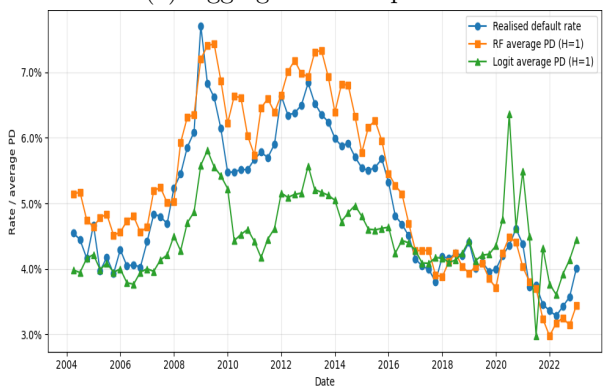
In-sample model performance at the micro level. Next, we evaluate the in-sample model performance at the micro level of RF against logit in the training sample. Figure 3, panel (a), plots the precision and recall (or sensitivity) of the two models in one diagram.[20] Panel (a) shows that RF is uniformly superior to logit: for any given precision (or recall), recall (or precision) of RF is strictly larger (orange solid curve) than those of logit (green dashed curve). Figure 3, panel (b), plots the score of precision and recall of RF and logit against the calibrated threshold. Again, precision and recall of RF (orange solid and orange dotted curves) are uniformly greater than those of logit (green dashed and green dotted curves) for all thresholds, including the calibrated ones. Figure 3, panel (c), shows the calibration curve of RF and logit against the

[18] All series are projected for 4 quarters and anchored to the last realized point at the evaluation cutoff. We choose a 4-quarter horizon trading-off between a sufficient number of quarters to be evaluated and a period short enough for an OOS exercise to be meaningful with RF. Contrary to aggregate time series models, like the autoregressive model, in which forecasts can be dynamically updated thanks to the autoregressive lag(s), an OOS exercise for RF requires firms' financial statements to be frozen at the forecast origin. Therefore, longer-term forecasts become less meaningful as the features are not updated over time.

[19] For the aggregate analysis, as an alternative to using logit at the micro level and then aggregating bottom-up, we could use an aggregate model directly, e.g., a linear model, an AR model, or a random walk.

[20] Precision and recall are defined in appendix C. These two measures trade off, as false positives (predicted defaults that are actual non-defaults) and false negatives (predicted non-defaults that are actual defaults) are inversely related and, therefore, the two measures trade off as well, i.e., a higher precision (recall) implies a lower recall (precision) and vice versa.

(a) Aggregate in-sample fit



(b) Aggregate out-of-sample performance

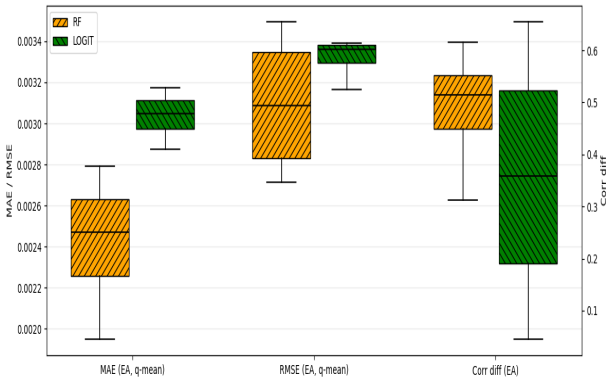


Figure 2: Panel (a): realized default rate vs predicted PDs (quarterly, 2001–2022) by model (RF in orange, logit in green). RF outperforms logit over all metrics: MAE, RMSE and Correlation difference (0.006 vs 0.009, 0.005 vs 0.007 and 0.677 vs 0.257, respectively). Panel (b): boxplots of out-of-sample metrics (MAE, RMSE and Correlation difference) comparing realized default rates with model forecasts (RF in orange, logit in green) across different forecast origins ($\tau^* \in \{2021Q1, 2021Q2, 2021Q3, 2021Q4\}$). Likewise, RF outperforms logit over all metrics: MAE, RMSE and Correlation difference (0.0025 vs 0.0031, 0.0031 vs 0.0034 and 0.514 vs 0.359, respectively).

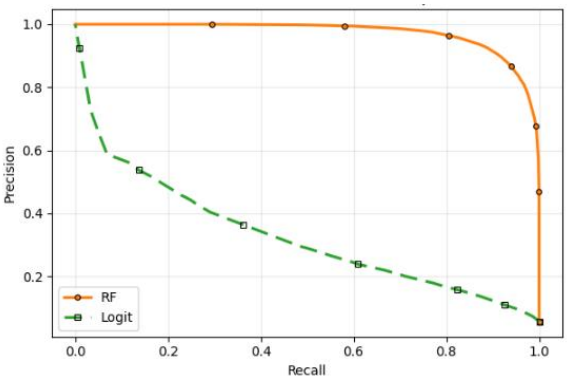
45-degree line, the perfectly calibrated model.[21] The RF calibration curve (orange solid curve) overall adheres much more closely to the 45-degree line than the logit calibration curve (green dashed curve). Finally, panel (d) plots the ROC curve in the true-positive-rate vs false-positive-rate plane.[22] A random classifier sits on the 45-degree line, for which the ROC-AUC is 0.5 (i.e., a true positive rate equal to false positive rate). A perfect classifier covers the upper-left borders of the quadrant, for which the ROC-AUC is 1 (i.e., a true positive rate equal to 1 and false positive rate equal to 0). The orange solid and green dashed ROC curves represent RF and logit, respectively, showing that RF performs better than logit because its ROC curve lies above logit's. These metrics indicate stronger recall-oriented discrimination, better probability calibration and overall evidence that RF is superior to logit across these metrics also at the micro level.[23]

Out-of-sample model performance at the micro level. All models are estimated using data up to a chosen cutoff quarter, which here is $\tau^* = 2019Q4$. For each forecast horizon

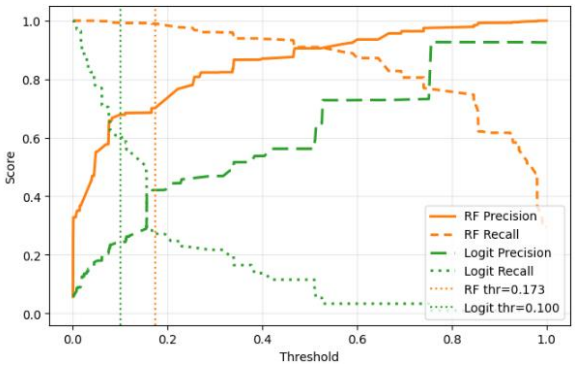
[21] The latter indicates that the fraction of positives (i.e., defaults) equals the mean predicted probability of positives, thus suggesting a perfectly calibrated model.

[22] These concepts are defined in appendix C. The true positive rate is equivalent to recall.

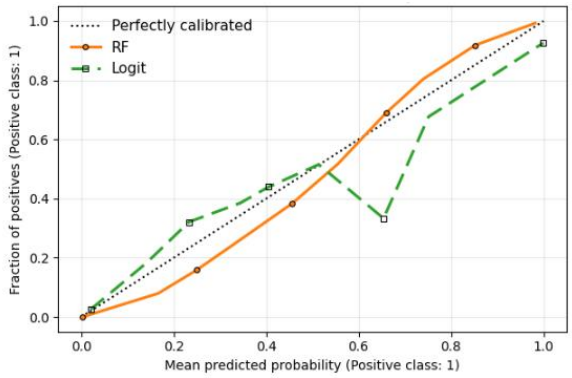
[23] Also, at their F_2-optimal operating points, RF attains a higher F_2 and a lower Brier score than logit (0.827 vs 0.408 and 0.015 vs 0.034, respectively).



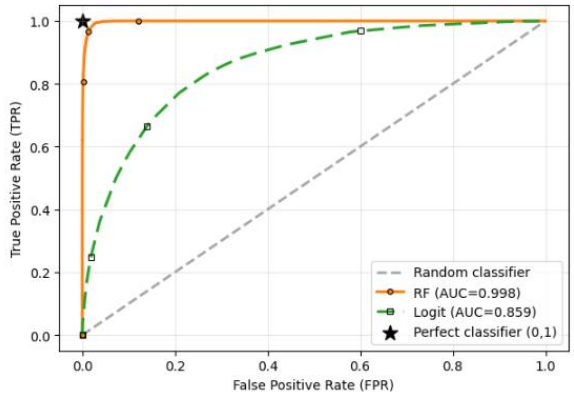
(a) Precision-recall (PR) curve.



(b) Precision and recall vs threshold (calibrated).



(c) Calibration curve.



(d) ROC curve.

Figure 3: In-sample model performance at the micro level: RF (orange) versus logit (green). Precision-recall (PR) curve (panel (a)), Precision and recall versus threshold (calibrated) (panel (b)), Calibration curve (panel (c)) and Receiving Operating Characteristic (ROC) curve (panel (d)) for both RF and logit. In Panel (a) the PR-AUC is the area under each PR curve, which is also known as Average Precision (AP), while in panel (d) the ROC-AUC is the area under each ROC curve. Calibration is performed via isotonic regression.

$H = 1, \ldots, 4$, RF predicts the average PD H quarters ahead, conditional on firms' financial statements observed at τ^*. These horizon-specific forecasts are then aggregated into a single projected time series of mean PDs from $\tau^* + 1Q$ to $\tau^* + 4Q$. Since RF is trained, calibrated, and thresholded exclusively on pre-τ^* data, all forecasts beyond τ^* constitute an OOS exercise. Across the four-quarter horizon, RF systematically achieves higher AP (mean 0.664 vs 0.577), higher ROC-AUC (mean 0.967 vs 0.917) and lower Brier (mean 0.020 vs 0.022) than logit. Short-term forecasts (e.g., H=1) show a clear edge for RF (AP 0.797 vs 0.756; ROC-AUC 0.985 vs 0.967, Brier 0.015 vs 0.016), and, although performance decays with the horizon for both models, RF maintains an advantage at longer H (see table 1). Likewise, lift statistics indicate

better ranking power relative to the baseline event rate (mean 8.705 for RF vs 7.840 for logit).[24] Appendix C reports the PR and ROC curves underlying table 1 at each H.

Table 1: Both models are trained on data up to $\tau^* = 2019Q4$ and evaluated OOS using isotonic calibration on an evaluation slice, with performance reported for each horizon H (quarters ahead). Historical feature pre-computation takes ∼22–23m (RF: 1341.29s; Logit: 1396.28s). Logit shows convergence warnings at some horizons, but still provides calibrated probabilities for evaluation. *Lift* is the ratio of AP to the event rate on the evaluation slice: higher is better. We report *Lift* at 10%, i.e. we look at the 10% of cases classified as most likely, e.g., a value of 9.505 indicates that the probability of finding a default in that group is 9.505 times the population average. Lower Brier indicates better probability calibration. Runtime per horizon: RF median $\tilde{t} \approx 304$s, logit median $\tilde{t} \approx 663$s. Total across H: RF ∼ 61m vs Logit ∼ 391m.

	RF			**Logit**			**Lift**†	
H	AP	ROC-AUC	Brier	AP	ROC-AUC	Brier	**RF**	**Logit**
1	**0.797**	**0.985**	**0.015**	0.756	0.967	0.016	**9.505**	9.144
2	**0.703**	**0.974**	**0.019**	0.607	0.927	0.021	**8.996**	7.978
3	**0.608**	**0.962**	**0.022**	0.502	0.896	0.025	**8.474**	7.345
4	**0.547**	**0.948**	**0.024**	0.443	0.877	0.027	**7.843**	6.894
Mean	**0.664**	**0.967**	**0.020**	0.577	0.917	0.022	**8.705**	7.840

3.2 Model interpretability: Shapley additive explanations

We assess variable importance in our RF through SHapley Additive Explanations (SHAP), introduced by Lundberg & Lee (2017) as a unified framework for explaining ML predictions and demystifying ML as a "black box".[25] SHAP are based on the concept of Shapley values in game theory (Shapley (1953)). A Shapley value can be loosely defined as the marginal contribution of a feature by averaging all relative prediction changes when such feature joins each possible combination of all other features. See appendix C for a recap on SHAP and Shapley values.[26] Importantly, SHAP values are not regression coefficients, but are heterogeneous local contributions, in this application to firm-level predicted PD. Firm-level SHAP values can be positive or negative for the same variable. Also, RF, as it is applied in this paper, is reduced-

[24] Operationally, RF also trains faster (median ∼5.1m per horizon; ∼1h total) than logit (median ∼11.1m; ∼6.5h total), which additionally exhibits convergence warnings despite extended `max_iter`. Overall, RF yields more accurate and more timely OOS PD projections over the full 4-quarter horizon.

[25] Previously, Štrumbelj & Kononenko (2011) and Štrumbelj & Kononenko (2014) also suggest to use Shapley values to explain ML predictions. Other papers which try to increase the explainability of ML algorithms are by Nazemi & Fabozzi (2024), Bussmann *et al.* (2021), Ozturkkal & Wahlstrøm (2025), Wang *et al.* (2024), Souza *et al.* (2024) and Buckmann & Potjagailo (2025), while other articles contribute to variable selection strategies like Zou *et al.* (2025) and Guo & Zhou (2022).

[26] SHAP are not the only variable importance measures that can be used.

form and predictive rather than causal or structural, therefore any sign of variable should not be interpreted causally.

Figure 4, panel (a), shows the average absolute impact of the top-ten items from firms' financial statements on PD prediction. As expected, cash-flow and cash-stock variables like financial expenses, total cash, cash & equivalents and cash flow rank as the most important features in classifying defaults followed by net income and balance-sheet items like liabilities, such as total debt, total liabilities, loans and non-current liabilities.[27] Interestingly, no macro variable shows up in the top ten, hinting that financial statements sufficiently capture the information needed to predict defaults.[28] Figure 4, panel (b), shows the distribution of SHAP and its impact on PD prediction. Higher expenses and lower cash inflows (red dots for these two features) are associated with a positive SHAP value, i.e. an increase in PD, while lower net income and larger liabilities (blue dots for these two features) also signal higher default risk.[29] Ideally, red dots (high value of the feature) and blue dots (low value of the feature) should be distinctly separated, as a clear segregation reflects the feature's effectiveness in predicting PD, resulting in a flat distribution of SHAP, as is the case for all top-four items related to cash flows and stocks. Also, the location of the dots, which reflects the sign of the impact on PD, is largely theory-consistent: blue dots (low value of the feature) for expenses, outflows and liabilities are positioned to the left of red dots (high value of the feature) mostly in the negative territory of SHAP values, i.e. a low value of those features is associated with a reduction in PD, while for revenues, inflows and assets they are positioned to their right and mostly in the positive territory of SHAP values, i.e. a low value of those features is associated with an increase in PD. For those firms where the sign is reversed, this finding is also intuitive. As an example, for financial expenses, a negative SHAP contribution may arise for firms whose higher financial expenses are associated with larger scale, better access to credit, or greater debt capacity, so SHAP can be negative because the predictive role of financial expenses depends on the interaction with other firm characteristics.

[27] To avoid a mechanical effect due to how the default indicator is defined, we consider lagged firms' financial-statement variables.

[28] This is intuitive as a macro shock can hide variation at the meso-micro scale, e.g., a negative regional GDP shock in Lombardy which may drive GDP of Italy down may not affect the default probability of a firm operating exclusively in Calabria.

[29] The importance of earnings-related variables for PD prediction has also been flagged in, e.g., Veganzones *et al.* (2023).

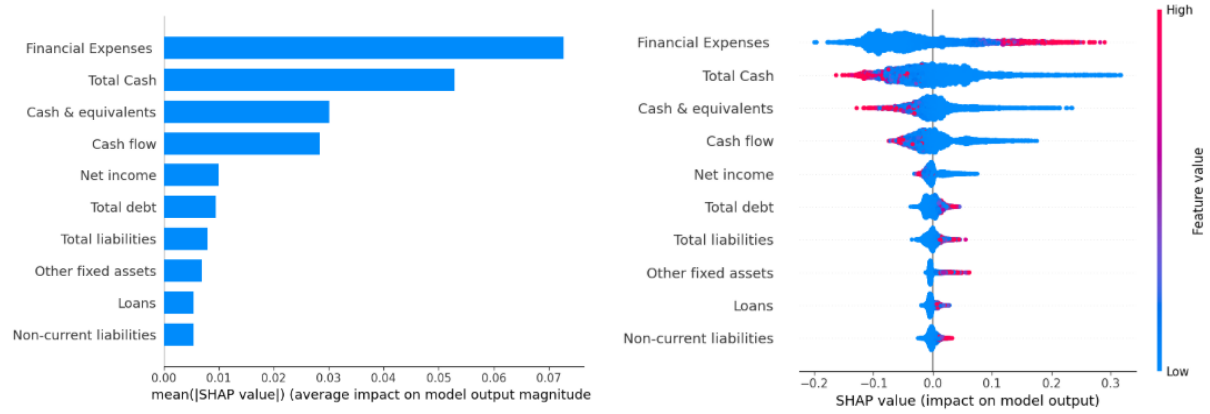


Figure 4: Model interpretability: variable importance with SHAP. Panel (a) depicts the mean of the absolute value of SHAP, i.e., the average impact of each feature on PD prediction. Panel (b) shows the distribution of SHAP and its impact on PD prediction: red dots are associated with high value of each feature, while blue dots are linked to low value of the feature.

4 Stress testing

We use RF to conduct stress testing by mapping macro scenarios to firms' financial statements. The importance of including financial statements has been under-recognized in the literature and in policy circles. Macro scenarios impact firms' default probability through various channels, which are missed if one stress-tests firms by linking defaults to macro variables only. By including firms' financial statements these channels can be explicitly accounted for. A shock to a macro variable can have different implications for different financial-statement items. For example, an increase in inflation hits financial expenses by boosting them, while a negative shock to GDP and, in turn, GVA can affect firms' revenue stream by depressing income.[30] In the next subsection we describe macro-meso-micro mappings.

4.1 Macro-meso scenarios

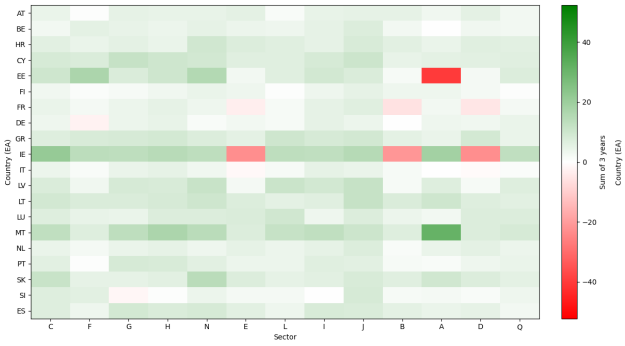
Starting from the macro, we focus on country-level GDP and take the GDP adverse and baseline scenarios from the EU-wide stress test 2023 (EBA (2023)).[31] Next, we exploit also the meso scenarios (at the country-sector level) provided therein, which are already made consistent accounting-wise with the macro scenarios. These meso scenarios include adverse and baseline scenarios for sectoral (real) GVA, which square with those designed for GDP. Figure 5 plots the

[30] When macro and borrower characteristics are jointly included, macro shocks affect both default probabilities and borrower characteristics.

[31] New scenarios for 2025 are also available (EBA (2025)), but as Orbis data are of sufficient quality only until end-2022, we take the 2023 edition.

matrix of country-sector pairs (cells) highlighting shock intensity to GVA within each cell both for the baseline and adverse scenario from EBA (2023). Panel (a) plots the baseline scenario, which consists of almost all greenish cells, pointing to benign dynamics for GVA. Panel (b) shows that, when moving to the adverse scenario, cells become more reddish, stressing that GVA is being depressed.

(a) Matrix of shocks to GVA - Baseline



(b) Matrix of shocks to GVA - Adverse

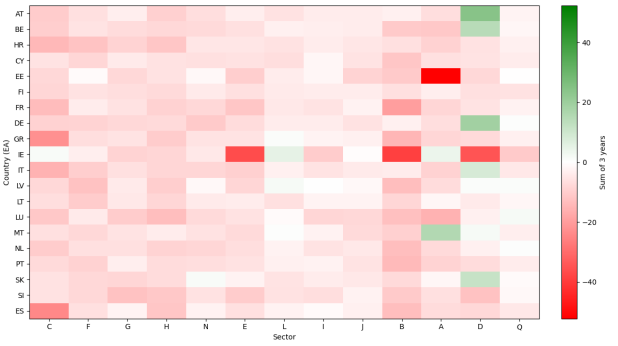


Figure 5: Country-sector shocks to GVA - Baseline scenario (panel (a)) and adverse scenario (panel (b)) of the EU-wide stress test of the banking sector 2023, respectively. Shocks are cumulated over the three-year horizon of the stress test and are expressed in percentage points.

We now translate these meso shocks into granular shocks. A way to do that is to spread shocks to all firms within each country-sector pair, noting that, although shocks are distributed homogeneously (as percentage points), the level impact on firms is heterogeneous, as the shock is multiplied by balance-sheet items which are heterogeneous across firms.[32] Before doing that, we use SHAP to reduce dimensionality from all features to the most important subset of features and stress only those in a way which is accounting-consistent.[33]

4.2 Mapping to granular shocks

We map meso shocks to micro shocks in three steps: we (i) link aggregate GVA growth to "driver" firm-level financial statements through elasticities; (ii) reconstruct the full balance sheet by recomputing the remaining items using firm-specific historical shares to preserve the balance sheet structure and ensure accounting consistency; and (iii) adjust interest expenses by applying a scenario-dependent interest rate which combines a GVA-driven component with

[32] In principle, each macro scenario can have a large number of meso and micro scenarios, i.e., many ways to heterogeneously spread shocks across sectors and firms, respectively, which makes the exercise interesting but challenging.

[33] For example, a GVA shock to the manufacturing sector in Germany is distributed equally within that sector in Germany, however, as firms are heterogeneous in, e.g., sales, the level shocks are heterogeneous.

calibrated spread add-ons that differ between the baseline and adverse scenarios.

Meso-micro mapping and anchoring. Let g_t^{bl} and g_t^{ad} denote the quarterly growth rates of GVA in the baseline and adverse scenarios.[34] For a generic firm-level quantity $x_{i,t}^{scn}$ we write:

$$\Delta \log x_{i,t}^{scn} \approx \beta_x^{(s)} \, g_t^{scn} \tag{3}$$

where $\beta_x^{(s)}$ is a sector-specific historical elasticity and g_t^{scn} is a scenario-dependent GVA growth rate.[35] We estimate elasticities for total assets ($TA_{i,t}$), total debt ($Debt_{i,t}$), operating profit ($OP_{i,t}$), cash flow ($CF_{i,t}$) and sales ($S_{i,t}$). We anchor each firm at its observed balance-sheet position in 2022Q4 and focus on a projection window from 2023Q1 to 2025Q4. Aggregating the quarterly log changes implied by the scenario and taking the exponential to obtain levels yields a cumulative path for a generic $x_{i,t}^{scn}$[36]:

$$x_{i,t}^{scn} = x_{i,0} \exp\left(\sum_{\tau=1}^{t} \Delta \log x_{i,\tau}^{scn} \right) \tag{4}$$

Reconstructing the full balance sheet. Given projected "driver" items (in particular, $TA_{i,t}$ and $Debt_{i,t}$), we reconstruct the remaining balance-sheet items, $\bar{x}_{i,t}^{scn}$, using firm-specific historical shares computed from pre-scenario observations. Let $Z_{i,t}^{x}$ be the parent, "driver", balance sheet item of item x. For each firm i, we compute the median share of each item over the last K quarters up to and including 2022Q4:

$$\bar{s}_i^x = \text{median}_{t \le 2022\text{Q4}} \left(\frac{x_{i,t}}{Z_{i,t}^{x}} \right) \tag{5}$$

for items like current assets, total liabilities, current and non-current liabilities, long-term debt, loans, tangible and intangible fixed assets, equity, etc., see appendix D.[37] We then allocate projected "driver" items, $Z_{i,t}^{scn,x}$, to the remaining balance sheet items, $\bar{x}_{i,t}^{scn}$, by scaling the

[34] These rates are expressed in decimal form (e.g. -0.01 for -1% quarter-on-quarter). For each firm i in sector s, we assume that (log) changes in key firm quantities respond linearly to the aggregate GVA shock, which, in turn, implies that changes in these quantities respond non-linearly to it.

[35] When sector-specific estimates are not reliable (e.g., due to limited data), we use global fallback values for the corresponding β.

[36] To avoid unrealistically large swings, we impose caps on the log changes: (i) per-quarter cap: $|\Delta \log x_{i,t}| \le 0.05$; (ii) cumulative cap: $\left|\sum_{\tau=1}^{t} \Delta \log x_{i,\tau}\right| \le 0.35$ for all the log-transformed quantities $x \in \{TA, Debt, OP, CF, S\}$. This keeps projected paths within economically plausible ranges, while still allowing differences between the baseline and adverse scenarios to emerge through the different GVA trajectories.

[37] Cash is calculated according to the same formula, but also simultaneously updated based on the cash flow.

former with the corresponding share:

$$\bar{x}_{i,t}^{scn} \approx \bar{s}_i^x \, Z_{i,t}^{scn,x} \tag{6}$$

This approach ensures that each firm retains its own typical balance-sheet structure, while scenarios primarily affect its size through the GVA-driven elasticities of the "driver" items.[38]

Financial expenses and spread add-ons. Financial expenses[39] are projected via an implied interest rate, which is derived through spread add-ons. These spread add-ons are, in turn, a function of a "positive wedge". For each quarter t in a calibration window[40], we calculate the positive wedge as:

$$\gamma_t = \max\big(r_t^{\mathrm{EA}} - r^*, \, 0\big) \tag{7}$$

where $r_t^{\mathrm{EA}} = \mathrm{median}_{i \in \mathrm{EA}} \, r_{i,t}$ is an aggregate benchmark like the euro-area (EA) median interest rate by quarter from Orbis and r^* is an anchor rate defined as the average EA median rate over the last $N_{\mathrm{anchor}} = 6$ quarters before and including 2022Q4. The distribution of γ_t summarizes how much EA funding conditions have historically deviated above the anchor level.

We use this distribution to calibrate constant spread add-ons for the baseline and adverse scenarios. Specifically, (i) the baseline spread add-on is set equal to the median wedge, $\mathrm{median}(\gamma_t)$; (ii) the adverse spread add-on is set equal to an upper quantile of the wedge distribution, e.g., the 90$^{\mathrm{th}}$ percentile, $Q_{0.90}(\gamma_t)$, corresponding to a severe but plausible funding environment.[41] For each firm and quarter in the projection window, the scenario-dependent annual interest rate is:

$$r_{i,t}^{\mathrm{scn}} = \big(r_{i,0} + \beta_r(-g_t^{\mathrm{scn}}) + \mathrm{SPREAD\_ADD}^{\mathrm{scn}}\big)\Big|_{[r_{\min}, \, r_{\max}]} \tag{8}$$

where $r_{i,0} = 4\,\mathrm{FE}_{i,0}/Debt_{i,0}$ is the interest rate in the anchor quarter, with $\mathrm{FE}_{i,0}$ denoting

[38] Our reconstruction is nested: items such as current assets and total liabilities are scaled off $TA_{i,t}^{scn}$, current versus non-current liabilities are split from total liabilities ($TL_{i,t}^{scn}$), debt components (e.g., loans and long-term debt) are split from $Debt_{i,t}^{scn}$, and tangible versus intangible components are split from fixed assets $FA_{i,t}^{scn}$. Cash is treated in a stock–flow consistent manner: it is initialized at the anchor date using the historical cash-to-assets share and then updated over the projection window by cumulating projected cash flows (with interest-expense effects reflected in the cash-flow projection).

[39] Here, financial expenses are loosely proxied with interest expenses. In general, interest expenses are a subcomponent of financial expenses.

[40] Here, the window corresponds to the last 30 quarters up to 2022Q4.

[41] In our calibration, the average EA median rate (the anchor) is $r^* \approx 1.12\%$, and the wedge distribution implies (i) $\mathrm{median}(\gamma_t) \approx 0.0025 (\approx 25$ bps); and (ii) $Q_{0.90}(\gamma_t) \approx 0.0064 (\approx 64$ bps). We therefore set (i) $\mathrm{SPREAD\_ADD}^{bl} = 0.0025$; and (ii) $\mathrm{SPREAD\_ADD}^{ad} = 0.0064$, which can be interpreted as constant annualized spread shifts over the projection horizon.

quarterly financial expenses and the factor 4 annualizing the rate[42], β_r captures the sensitivity of funding costs to GVA and $[r_{\min}, r_{\max}]$ is an admissible range[43]. We then translate this scenario-dependent interest rate into quarterly financial expenses:

$$\mathrm{FE}_{i,t}^{\mathrm{scn}} = \frac{1}{4}\, r_{i,t}^{\mathrm{scn}}\, Debt_{i,t}^{\mathrm{scn}}. \tag{9}$$

The baseline and adverse scenarios thus differ in two dimensions: the path of GVA growth g_t^{scn}, which drives the evolution of assets, debt and flows[44] through the elasticities, and the spread add-ons, $\mathrm{SPREAD\_ADD}^{\mathrm{scn}}$, which tilt funding costs upward more strongly under the adverse scenario.[45]

Visualizing the micro shocks. Based on the above mapping, we visualize the heterogeneous impact on firms' total assets of the country-sector-level shocks to GVA. Figure 6 shows the average percentage-point variation of total assets 2023-2025 against 2022Q4 for both the baseline and adverse scenarios. Panel (a) highlights the benign effect of shocks on total assets whose variation remains, to a great extent, within the positive territory. Instead, panel (b) shows a contraction of total assets for almost all the geo-localized firms, where geo-localization is achieved at zip-code level.

4.3 A stress test and scenario severity

In this subsection, we run a stress test and conduct sensitivity analyses with respect to scenario severity. Figure 7, panel (a), shows that RF is more sensitive to the adverse scenario than logit, as the projected aggregate PD jumps much more strongly starting from the second quarter of the horizon under the adverse scenario.[46] In the first four quarters, equivalent to the first year of the EU-wide stress test, PD roughly doubles under the adverse scenario. Interestingly, the spike and level of PD in the first year of the stress-test horizon are similar in magnitude to the historical spike and level observed in the default rate during the global financial crisis of

[42] The interest rate is computed for firms with positive debt only.

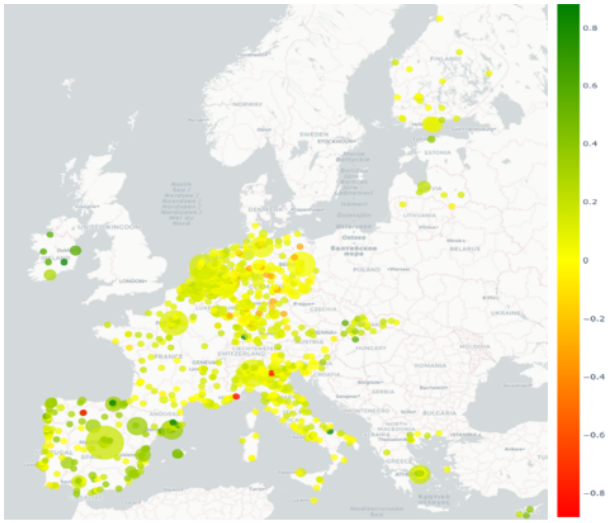
[43] Here, this range is $[0, 0.60]$. The spread add-on enters the interest rate equation with an implicit coefficient of one, as the spread add-on is an add-on on the interest rate itself, so it is the same variable with the same unit of measure.

[44] Cash flows and stocks, in turn, affect cash inflows as defined at the beginning of the paper and are used to define default.

[45] Shocks are homogeneous at the sector level (via common elasticities) while they affect firms heterogeneously through their idiosyncratic financial statements.

[46] In the first quarter PDs are still unaffected by the shock as shocks enter with a lag (H=1).

(a) Granular shocks to firms - Baseline



(b) Granular shocks to firms - Adverse

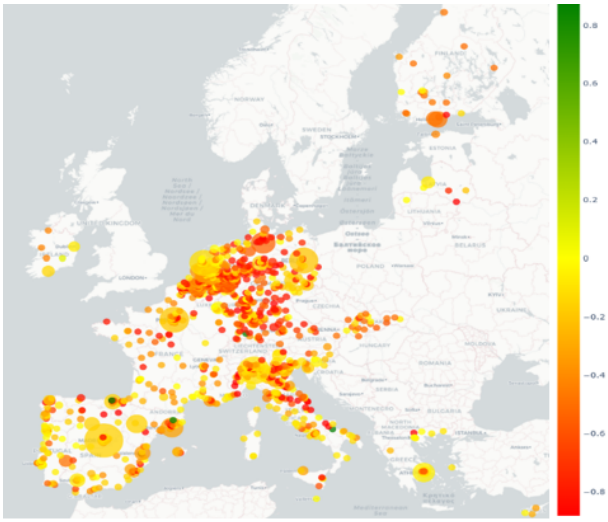


Figure 6: Heatmaps of granular shocks to firms' total assets - Baseline scenario (panel (a)) and adverse scenario (panel (b)) of the EU-wide stress test of the banking sector 2023, respectively. Shocks, stemming from GVA, are spread to firms, which are geo-localized at zip-code level, by hitting their total assets. Heatmaps report the average percentage-point variation of total assets over 2023-2025 against 2022Q4.

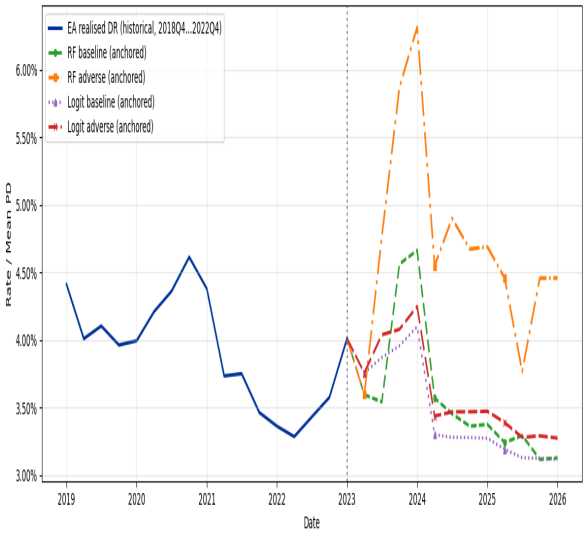
2007-2009 (see figure 2, panel (a)). In both cases, we observe a jump of circa 2.5 p.p., reaching circa 7% on aggregate.

A key question in stress testing is whether PD evolves non-linearly with scenario severity. This aspect is relevant in reverse stress testing and scenario calibration, so that stress tests deliver severe but plausible outcomes. A way to gauge any non-linear effect of scenarios on PD is to conduct a scenario sensitivity analysis and plot distributions of PD for different scenario severities. As a summary metric we take the average PD projected by RF over the stress test horizon. We plot this metric against a grid of scenario severities for the spread add-on and GVA growth on a 3D surface.[47] Figure 7, panel (b) shows the average PD plotted on the y-axis, the percentiles of the spread add-on distribution plotted on the x-axis and shift intensity of the GVA growth curve plotted on the z-axis. Scenario severity for the spread add-on and GVA growth are described in more detail in Appendix D (see figure 13).

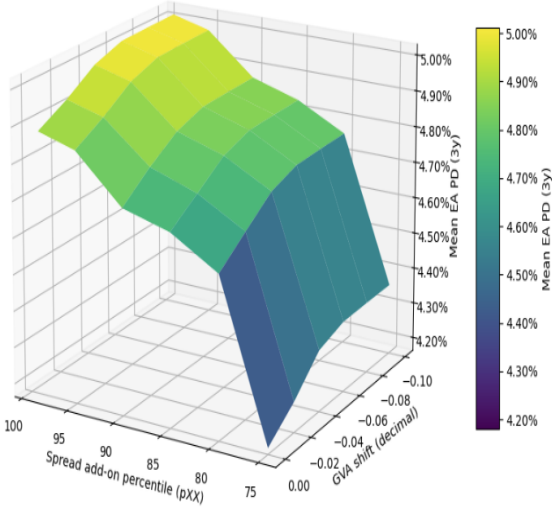
First, PD grows as scenario severity increases for both the spread add-on and GVA growth. However, the shape of this increase varies depending on which variable is shocked, although

[47] Figueres *et al.* (2025) design a framework for scenario severity distinguishing between backward-looking (historical) indicators and forward-looking (model-based) indicators. They show that scenarios are increasingly more severe over time, peaking in 2023 and stabilizing in 2025. Here, we focus on backward-looking severity, as forward-looking severity requires a model, which in itself is subject to uncertainty.

(a) Stress testing



(b) Scenario severity



(c) RF versus logit

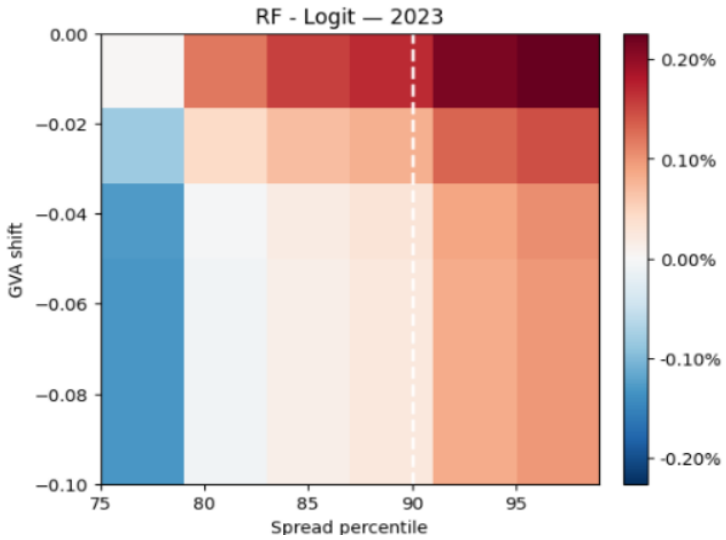


Figure 7: Stress testing and scenario severity. Panel (a): comparison of stressed (adverse) and baseline PD from RF and logit from 2023Q1 onward (the stress test three-year horizon). The vertical dotted line marks the evaluation cutoff (τ = 2023Q1). The historical default rate (solid blue) is extended by the PD under each scenario and model: Logit (baseline: violet, dotted; adverse: red, dashed) and RF (baseline: green, dashed; adverse: orange, dashed). All series are anchored to the first observation at the cutoff. Panel (b): 3D surface with the average PD projected by RF plotted on the y-axis, the percentiles of the spread add-on distribution plotted on the x-axis and the GVA growth curve scaling factor plotted on the z-axis. Panel (c): comparison of the sensitivity of RF and logit with respect to scenario severity in a 2D plane with the GVA shift on the vertical axis and spread add-on percentiles on the horizontal axis. Cells report the percentage-point variation in PD as projected by RF and logit in the first year of the horizon (when the shock hits the most).

both display a non-linear pattern. Greater severity of the spread add-on leads to a *staircase*-shaped evolution of PD. It initially grows rapidly, then suddenly plateaus, and subsequently re-accelerates after a threshold. Interestingly, the sharp jump in PD between the 75th and 80th percentile of the spread add-on distribution is associated with a mild change in interest rates (see figure 13, panel (a), in appendix D), hinting at strong non-linearity. Instead, greater severity of the GVA growth shock causes a logarithmic increase in PD, but the effect is weaker than that induced by the spread add-on. The stronger impact of the spread add-on may be due to the fact that the latter directly impacts financial expenses, which are the strongest predictor of default according to the SHAP ranking (see figure 4). The two-dimensional charts in which each variable is muted in turn are reported in Appendix D (see figure 14).

Finally, figure 7, panel (c), compares the sensitivity of RF and logit with respect to scenario severity in a 2D plane with the GVA shift on the vertical axis and spread add-on percentiles on the horizontal axis. We plot the percentage-point variation in PD as projected by RF and logit in the first year of the horizon (when the shock hits the most). When the scenario is more severe than the adverse scenario (corresponding to a spread add-on at the 90th percentile), RF delivers more severe PD growth compared to logit as shown by cells becoming redder, while when the scenario is less severe than the adverse scenario (as we approach the baseline), RF delivers equal or milder PD growth relative to logit as shown by cells becoming whiter and then blue. This acceleration is due to the fact that the RF surface gets steeper as a function of scenario severity.

4.4 Bank-firm network and systemic risk

In this subsection, we map firm-level data to banks' data by matching firm identifiers and time stamps available in Orbis to those available in a credit registry, Anacredit. Matching these two datasets generates a bipartite network of banks and firms, akin to Landaberry *et al.* (2021), whose edges are loan exposures (EAD), as plotted in figure 8, panel (a).[48] In the 2D donut chart, red nodes along the inner ring represent banks, blue nodes along the outer ring represent firms, while links between the two rings' nodes represent EAD. Notably, some banks (e.g., top-right region) are more interconnected to the real economy than others (e.g., bottom-

[48]For simplicity, contrary to Landaberry *et al.* (2021), we do not model the interfirm and interbank networks. Interfirm networks arise through supply chains, which, as COVID-19 has shown, are important for understanding the propagation of defaults in the real economy. Similarly, interbank networks are key to understanding default contagion within the banking sector, as the global financial crisis pointed out. Nevertheless, for the purpose of this article, we mute these two networks and focus on the bipartite bank-firm network.

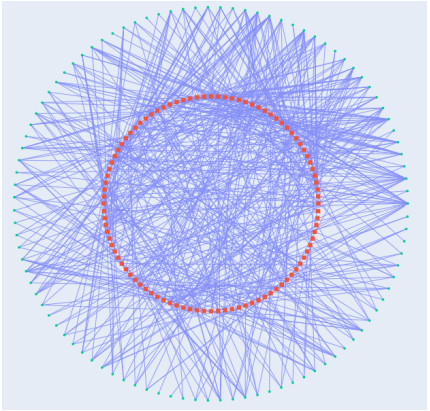
left region). Figure 8, panel (b), sketches the distribution of EAD shares of total exposures of the top-20 banks. This distribution is highly skewed with the top-5 banks already capturing a good chunk of the exposure pie. This bar chart could be interpreted as a systemic importance ranking, although with the caveat that the temporal information, i.e., how exposures evolve, is aggregated here and could alter the ranking (see Franch *et al.* (2024)). We use EAD to weight PD estimated via RF in order to construct a risk metric which can be used as a "stressometer" (see next paragraphs).

We enrich the descriptive panels (a) and (b) in figure 8 with two new elements: PD and stress testing. In figure 8, panel (c), we plot an index of systemic risk against an index of interconnectedness. The 'Riskiness' index is calculated as the product of the in-sample PD estimated via RF (orange) and logit (green) in the previous sections with EAD-weights from Anacredit across all banks, while the "Interconnectedness" index is computed as the EAD-weighted average of banks' percentile ranks in the cross-sectional distribution of the number of distinct borrowers they are exposed to. For the period 2019Q4 - 2022Q4, the two indices are inversely related, i.e., an increase in interconnectedness reduces riskiness and vice versa. When comparing RF with logit, this negative relation is captured more strongly by RF (fitted orange line) than logit (fitted green line), with RF delivering a statistically significant coefficient and an R^2 of 0.41, while logit fails in both dimensions. This negative relation could be explained by diversification: as banks diversify across borrowers, they insulate themselves from the default of any single borrower taken in isolation. However, it is known in the literature that this finding is not universally valid, e.g., Battiston *et al.* (2012) show that diversification reduces PD only up until a certain threshold, after which PD grows again. Nevertheless, their result is driven by financial acceleration, which is absent in our case.[49] One reason why riskiness decreases monotonically in our case may be the assumption of muted interfirm and interbank networks, whose modelling may trigger feedback loops currently absent in our paper (see, e.g., Wang *et al.* (2026) and Li *et al.* (2025)).

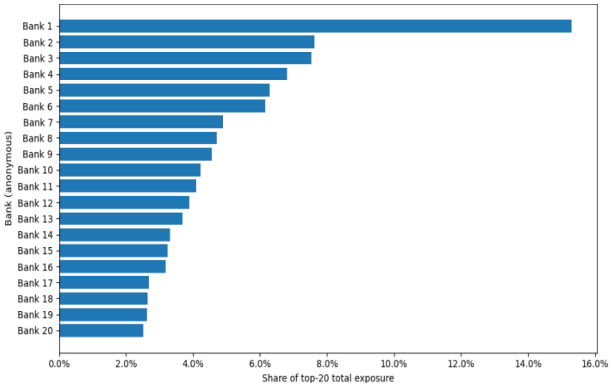
Next, figure 8, panel (d), shows how this 'Riskiness' index obtained from RF and logit varies under stress conditions in the adverse scenario relative to the baseline scenario used in EBA (2023). To this end, we employ RF and logit to project PD under the two scenarios and multiply them by EAD weights, i.e., the ratio of EAD over total EAD of each bank, which are

[49] When financial acceleration is assumed away, they also find a monotonic, though non-linear, reduction in PD.

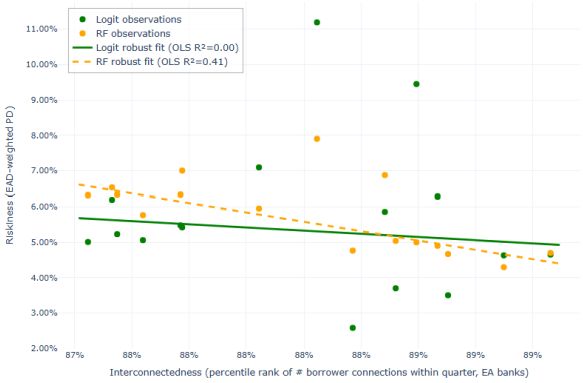
(a) Bipartite bank-firm network



(b) Distribution of loan exposures (EAD)



(c) Banks' riskiness – interconnectedness plane



(d) Banks' riskiness index and stress testing

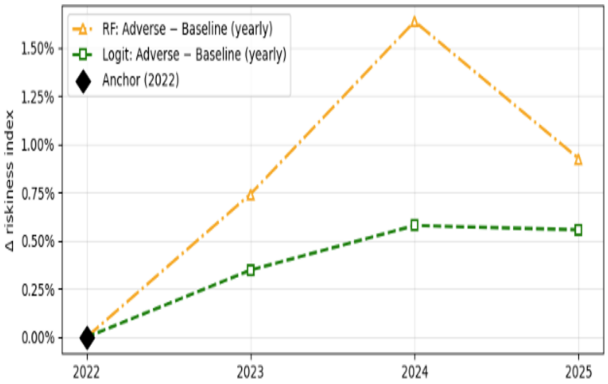


Figure 8: Panel (a): aggregate bipartite bank-firm network represented in a 2D donut chart. Red nodes along the inner ring represent banks, blue nodes along the outer ring represent firms, while links between the two rings' nodes (aggregated over time) represent loan exposures (EAD). For visualization purposes, only a subset of nodes and links are displayed based on an EAD threshold: the first 96 banks are selected based on a total exposure ranking. Then, for each bank we take the top 10 exposures and filter the firms with the highest exposure until we reach the cap of 100 firms. The actual network is composed of 6,215 firms and 2,039 banks over 2019Q1 - 2022Q4. Our sample of banks covers 88% of the universe of Anacredit banks as of 2022Q4 and 73% of their total assets available from COREP. Panel (b): Distribution of EAD across banks (top-20) as shares of total exposure, which are used to calculate the 'Riskiness' index (see bottom panels). Panel (c): 'Riskiness' index versus 'Interconnectedness' index. The 'Riskiness' index is calculated as the product of the in-sample PD estimated via RF (orange) and logit (green) in the previous sections with EAD-weights from Anacredit across all banks. The "Interconnectedness" index is computed as the EAD-weighted average of banks' percentile ranks in the cross-sectional distribution of the number of distinct borrowers they are exposed to. Fitted lines are obtained via Huber's robust M-estimator to reduce the influence of outliers. Panel (d): variation in the 'Riskiness' index obtained from RF (orange) and logit (green) under the adverse and baseline scenarios used in EBA (2023). Source: authors' elaboration based on Orbis and Anacredit.

kept frozen at the starting point.[50] In order to assess each model's sensitivity to the adverse scenario, for each model we then calculate the difference between the Riskiness index obtained under the adverse scenario and the Riskiness index obtained under the baseline scenario. The variation in the 'Riskiness' index is much more pronounced for RF (orange line) than for logit (green line), in particular during the first two years of the stress-test horizon during which the RF-based riskiness index increases from 5.5% to above 8.5%.

4.5 Granular inspection of banks' riskiness

The previous results highlight a stronger scenario sensitivity of RF relative to logit at the aggregate level both for firms and banks. However, from those findings, it is hard to understand why that is. In this subsection, we dig deeper and perform a granular analysis at the bank level, shedding light on these results.

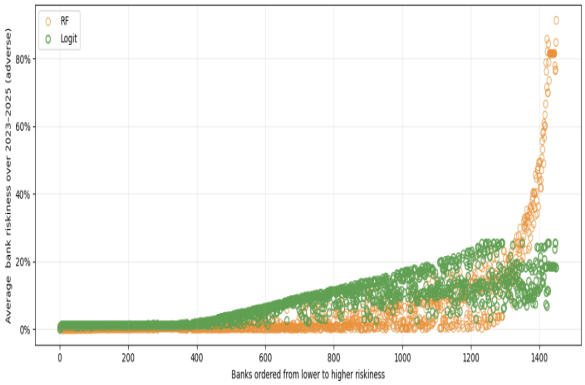
First, we use banks' riskiness indices computed in the previous subsection and plot them for each bank. Figure 9, panel (a), shows the average bank riskiness over 2023-2025 in the adverse scenario (y-axis), while the x-axis represents the banks' ranking from lower to higher riskiness. The orange circles, which are bank-level riskiness indices as estimated by RF, behave strongly non-linearly, showing that a few banks (the tail banks) are associated with substantially higher riskiness than the majority. The green circles, which are bank-level riskiness indices as estimated by logit, behave less sharply and do not capture the same non-linearity. Second, figure 9, panel (b), shows the share of banks above various thresholds of the riskiness index (y-axis) as depicted on its x-axis. The share of banks above a riskiness index of circa 25% remains positive for RF up to riskiness-index values above 80%, while for logit this share quickly falls to zero at a riskiness-index value of around 25%. Finally, figure 9, panel (c), plots the dynamics of banks' riskiness indices above the 95th quantile of the riskiness index's distribution (y-axis) over the stress test horizon (x-axis) for both RF (orange) and logit (green) and adverse (dashed) and baseline (continuous) scenarios. The chart shows that (i) RF is substantially more conservative than logit at this quantile, as its riskiness indices are higher; and (ii) the spread between RF's adverse- and baseline-scenario riskiness indices is substantially wider than logit's at the tail.

These different tail behaviors captured by the two models are at the root of the higher scenario sensitivity that RF is able to capture. The reason for this better ability is that RF is a

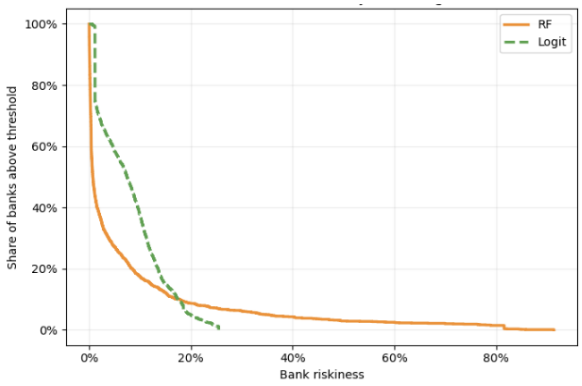
[50] This assumption is in line with the static balance sheet assumption used in the EU-wide stress test. Moreover, for simplicity, we assume away the collateralization status of the loan, implying an LGD equal to one (or a recovery rate equal to zero).

non-parametric algorithm which is able to fit the functional form relating defaults and scenarios without imposing a pre-specified form ex ante, as logit does.

(a) Tail banks



(b) Survival curve



(c) Tail banks – scenario sensitivity

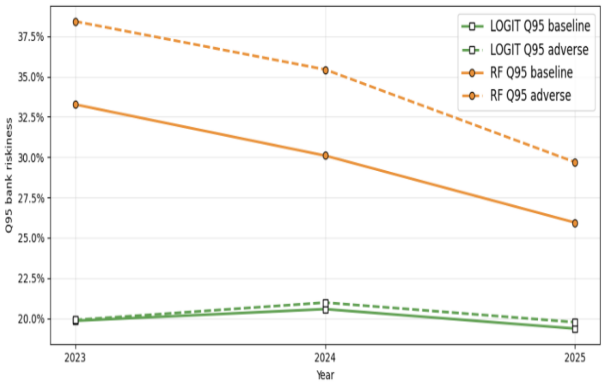


Figure 9: Panel (a): average bank riskiness over 2023-2025 in the adverse scenario. Orange circles are bank-level riskiness indices as estimated by RF, while green circles are bank-level riskiness indices as estimated by logit. The x-axis represents banks' ranking from lower to higher riskiness. Panel (b): share of banks above each threshold of 'Riskiness' index as depicted on x-axis. Orange line represents the share of banks as estimated by RF, while the green dashed line shows the share of banks as estimated by logit. Panel (c): dynamics of banks' riskiness indices above the 95th quantile of the riskiness index's distribution (y-axis) over the stress test horizon (x-axis) for both RF (orange) and logit (green) and adverse (dashed) and baseline (continuous) scenarios. Source: authors' elaboration based on Orbis and Anacredit.

5 Conclusion

Our paper offers a framework to conduct granular stress testing of the banking sector. In particular, we highlight the usefulness of RF for stress-testing exercises, in addition to merely improving forecasting performance relative to a logit benchmark. On top of finding a superior

prediction ability of RF compared to logit both at the aggregate and micro-level, when stressing firms' financial statements through scenarios, we find that RF is much more sensitive than logit to adverse macro and sectoral shocks. Scenario sensitivity analyses highlight interesting non-linear effects of an increase in interest rates and a depression in GVA growth on PD of non-financial corporations, which can also inform monetary policy and its credit-risk effects. Also, by using the bipartite network of exposures among banks and firms and stress testing it using scenario-dependent PD estimated by RF, we point out RF's ability to deliver severe-yet-plausible risk metrics for the whole banking sector. A granular inspection of banks' riskiness indices derived from this network also highlights RF's superior ability to capture non-linearity through its ability to identify "tail banks".

Our work can be extended in multiple directions. Our model can accommodate the analysis of different quantiles of the default rate distribution via a quantile RF (see, e.g., Lenza *et al.* (2025)).[51] Also, we could use clustering techniques to group firms or loans into homogeneous groups, thus allowing the development of cluster-specific RF. Next, our approach can be used to analyze default contagion not only via the bipartite network of banks and firms, but also within the interfirm and interbank networks, which could unveil potential feedback loops muted in the present paper. Finally, our article can be used for applications like climate stress testing (see Alogoskoufis *et al.* (2021), Emambakhsh *et al.* (2023), Abbondanza *et al.* (2025)), e.g., to simulate the impact of physical or transition risks on banks' credit risk. For example, our work can be used to translate natural disasters (like droughts, earthquakes, floods) in specific geographies into consequences for firms' balance sheets, and, in turn, banks' portfolios. Finally, our work can also be employed to gauge the effects of geopolitical risk by translating exposure to tariffs and/or wars into consequences for firms' balance sheets and, in turn, banks' portfolios. We leave these extensions and applications for future research.

References

Abbondanza, A., Albertazzi, U., Caccavaio, M., Djekic, D., Gattinoni, V., Georgescu, O. M., & Ponte Marques, A. 2025. Integrating climate risk into the 2025 EU-wide stress test: the effects of climate risks for firms. *ECB Macroprudential Bulletin*, **32**.

[51] Kellner *et al.* (2022) introduce a quantile neural network for LGD prediction, while Buse *et al.* (2025) use a generalised RF adapted to quantile prediction of cryptoasset prices.

Aldasoro, I., Hördahl, P., Schrimpf, A., & Zhu, S. 2025. *Predicting financial market stress with machine learning.* Tech. rept. 1250. Bank for International Settlements.

Alogoskoufis, S., Dunz, N., Emambakhsh, T., Hennig, T., Kaijser, M., Kouratzoglou, C., Muñoz, M.A., Parisi, L., & Salleo, C. 2021. *ECB's economy-wide climate stress test.* Tech. rept. 281. ECB Occasional Paper Series.

Atif, D. 2025. VAE-INN: Variational AutoEncoder with Integrated Neural Network classifier for imbalanced credit scoring, utilizing weighted loss for improved accuracy. *Computational Economics.*

Audrino, F., Kostrov, A., & Ortega, J.-P. 2019. Predicting U.S. bank failures with MIDAS logit models. *Journal of Financial and Quantitative Analysis*, **54**(6), 2575–2603.

Barbaglia, L., Manzan, S., & Tosetti, E. 2023. Forecasting loan default in Europe with machine learning. *Journal of Financial Econometrics*, **21**(2), 569–596.

Battiston, S., Delli Gatti, D., Gallegati, M., Greenwald, B. C., & Stiglitz, J. E. 2012. Liaisons dangereuses: Increasing connectivity, risk sharing, and systemic risk. *Journal of Economic Dynamics and Control*, **36**(8), 1121–1141.

Beck, E., & Wolf, M. 2025. *Forecasting inflation with the hedged random forest.* Tech. rept. 2025-07. SNB Working Paper Series.

Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. 2021. Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting*, **37**(1), 428–444.

Beutel, J., List, S., & von Schweinitz, G. 2019. Does machine learning help us predict banking crises? *Journal of Financial Stability*, **45**, 100693.

Bie, S., Diebold, F. X., He, J., & Li, J. 2024. *Machine learning and the yield curve: tree-based macroeconomic regime switching.* PIER Working Paper 24-028. Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

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**, 103773.

Breiman, L. 2001. Random forests. *Machine Learning*, **45**(1), 5–32.

Bräuning, M., Malikkidou, D., Scricco, G., & Scalone, S. 2019. A new approach to early warning systems for small European banks. *ECB Working Paper Series*.

Buckmann, M., & Potjagailo, G. 2025. *Infusing economically motivated structure into machine learning methods*. Tech. rept. Bank of England Staff Working Paper No. 1,144.

Budnik, K., Ponte Marques, A., Giglio, C., Grassi, A., Durrani, A., Figueres, J. M., Konietschke, P., Le Grand, C., Metzler, J., Población García, F. J., Shaw, F., Groß, J., Sydow, M., Franch, F., Georgescu, O. M., Ortl, A., Trachana, Z., & Chalf, Y. 2024. *Advancements in stress-testing methodologies for financial stability applications*. Tech. rept. ECB Occasional Paper Series No. 348.

Buse, R., Görgen, K., & Schienle, M. 2025. Predicting value at risk for cryptocurrencies with generalized random forests. *International Journal of Forecasting*, **41**(3), 1199–1222.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. 2021. Explainable machine learning in credit risk management. *Computational Economics*, **57**(1), 203–216.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. 2016. Risk and risk management in the credit card industry. *Journal of Banking & Finance*, **72**, 218–239.

Castrén, Olli, Dées, Stéphane, & Zaher, Fadi. 2010. Stress-testing euro area corporate default probabilities using a global macroeconomic model. *Journal of Financial Stability*, **6**(2), 64–78.

Chan-Lau, J. A. 2017. *Lasso regressions and forecasting models in applied stress testing*. Tech. rept. IMF Working Paper Series WP/17/108.

Chavleishvili, S., & Manganelli, S. 2024. Forecasting and stress testing with quantile vector autoregression. *Journal of Applied Econometrics*, **39**(1), 66–85.

Chi, G., Bai, F., Tan, H., & Zhou, Y. 2025. Default prediction framework with optimal feature set and matching ratio. *Journal of Forecasting*, **44**(7), 2067–2088.

Conti, A, & Morelli, G. 2025. Bayesian probability of default models with Langevin dynamics. *Quantitative Finance*.

Coulombe, P.G. 2024. The macroeconomy as a random forest. *Journal of Applied Econometrics*, **39**(3), 401–421.

Coulombe, P.G., Leroux, M., Stevanovic, D., & Surprenant, S. 2022. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, **37**(5), 920–964.

Dendramis, Y., Tzavalis, E., & Cheimarioti, A. 2025. Measuring the default risk of small business loans: improved credit risk prediction using deep learning. *Journal of Forecasting*, **44**(7), 2277–2297.

du Jardin, P. 2025. A quantification approach of changes in firms' financial situation using neural networks for predicting bankruptcy. *Journal of Forecasting*, **44**(2), 781–802.

EBA. 2023. *2023 EU-wide Stress Test*. Technical Report. European Banking Authority.

EBA. 2025. *2025 EU-wide Stress Test*. Technical Report. European Banking Authority.

Ekinci, A., & Sen, S. 2024. Forecasting bank failure in the U.S.: A cost-sensitive approach. *Computational Economics*, **64**(6), 3161–3179.

Emambakhsh, T., Fuchs, M., Kördel, S., Kouratzoglou, C., Lelli, C., Pizzeghello, R., Salleo, C., & Spaggiari, M. 2023. *The road to Paris: stress testing the transition towards a net-zero economy*. Tech. rept. 328. ECB Occasional Paper Series.

Errico, M., Pesce, S., & Pollio, L. 2025. Nonlinearities and heterogeneity in firms' response to aggregate fluctuations: what can we learn from machine learning? *ECB Working Paper Series*.

Figueres, J. M., Montero Prieto, B., Scalone, V., t'Hooft, J., Ter Steege, L., & Vallotto, C. 2025. A framework to assess the severity of adverse scenarios in EU-wide stress tests. *ECB Macroprudential Bulletin*, **32**.

Franch, F., Nocciola, L., & Vouldis, A. 2024. Temporal networks and financial contagion. *Journal of Financial Stability*, **71**, 101224.

Galindo, J., & Tamayo, P. 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics*, **15**(1-2), 107–143.

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. 2024. How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. *Journal of Financial Stability*, **73**, 101284.

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.

Gourinchas, P. O., Kalemli-Özcan, S., Penciakova, V., & Sander, N. 2020. *COVID-19 and SME failures*. NBER Working Paper 27877. National Bureau of Economic Research.

Guo, W., & Zhou, Z.Z. 2022. A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction. *Journal of Forecasting*, **41**(6), 1248–1313.

Jones, S., Johnstone, D., & Wilson, R. 2015. An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking & Finance*, **56**, 72–85.

Kaposty, F., Kriebel, J., & Löderbusch, M. 2020. Predicting loss given default in leasing: A closer look at models and variable selection. *International Journal of Forecasting*, **36**(2), 248–266.

Kellner, R., Nagl, M., & Rösch, D. 2022. Opening the black box – Quantile neural networks for loss given default prediction. *Journal of Banking & Finance*, **134**(C), 106334.

Khandani, A. E., Kim, A., & Lo, A. W. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, **34**(11), 2767–2787.

Kim, H., Cho, H., & Ryu, D. 2022. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Computational Economics*, **59**(3), 1231–1249.

Konietschke, P., Metzler, J., & Ponte Marques, A. 2026. *A quantile probability model for sectoral corporate defaults in Europe*. ECB Working Paper Series.

Landaberry, V., Caccioli, F., Rodriguez-Martinez, A., Baron, A., Martinez-Jaramillo, S., & Lluberas, R. 2021. The contribution of the intra-firm exposures network to systemic risk. *Latin American Journal of Central Banking*, **2**(2), 100032.

Lenza, M., Moutachaker, I., & Paredes, J. 2025. Density forecasts of inflation: A quantile regression forest approach. *European Economic Review*, **178**.

Leong, C. K. 2016. Credit risk scoring with bayesian network models. *Computational Economics*, **47**(3), 423–446.

Li, Z., Fu, D., & Li, H. 2025. The influence of bank-firm loan network structure on systemic risk: from the perspective of complex networks. *Frontiers in Physics*, **13**, 1548204.

Liu, J., Zhang, X., & Xiong, H. 2024. Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation. *Journal of Forecasting*, **43**(5), 1625–1660.

Lundberg, S. M., & Lee, Su-In. 2017. A unified approach to interpreting model predictions. *In: Advances in Neural Information Processing Systems*, vol. 30.

Lundberg, S. M., Erion, G. G., & Lee, Su-In. 2019. Consistent individualized feature attribution for tree ensembles. *arXiv*.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. 2021. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, **39**(1), 98–119.

Nazemi, A., & Fabozzi, F. J. 2024. Interpretable machine learning for creditor recovery rates. *Journal of Banking & Finance*, **164**, 107187.

Nazemi, A., Rezazadeh, H., Fabozzi, F. J., & Höchstötter, M. 2022. Deep learning for modeling the collection rate for third-party buyers. *International Journal of Forecasting*, **38**(1), 240–252.

Nocciola, L. 2022. Finite sample forecast properties and window length under breaks in cointegrated systems. *Pages 167–196 of:* Chudik, A., Hsiao, C., & Timmermann, A. (eds), *Essays in Honor of M. Hashem Pesaran: Prediction and Macro Modeling*. Advances in Econometrics.

Nocciola, L., Ollech, D., & Webel, K. 2023. Unit root test combination via random forests. *Pages 31–45 of:* Valenzuela, O., Rojas, F., Herrera, L. J., Pomares, H., & Rojas, I. (eds), *Theory and Applications of Time Series Analysis and Forecasting*. Contributions to Statistics.

Ozturkkal, B., & Wahlstrøm, R.Â.R. 2025. Explaining mortgage defaults using SHAP and LASSO. *Computational Economics*, **66**, 3291–3325.

Peng, Q., McKillop, D., Quinn, B., & Liu, K. 2025. Modeling and predicting failure in US credit unions. *International Journal of Forecasting*, **41**(3), 1237–1259.

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.

Sadhwani, A., Giesecke, K., & Sirignano, J. 2021. Deep learning for mortgage risk. *Journal of Financial Econometrics*, **19**(2), 313–368.

Sarlin, P., & Holopainen, M. 2017. Toward robust early-warning models: a horse race, ensembles and model uncertainty. *Quantitative Finance*, **17**(12), 1885–1902.

Shahee, S.A., & Patel, R. 2025. An explainable ADASYN-based focal loss approach for credit assessment. *Journal of Forecasting*, **44**(4), 1513–1530.

Shapley, L. S. 1953. A value for n-person games. *Pages 307–317 of:* Kuhn, H. W., & Tucker, A. W. (eds), *Contributions to the Theory of Games*. Annals of Mathematics Studies, vol. II, no. 28. Princeton University Press.

Skavysh, V., Priazhkina, S., Guala, D., & Bromley, T. R. 2023. Quantum Monte Carlo for economics: stress testing and macroeconomic deep learning. *Journal of Economic Dynamics and Control*, **153**, 104680.

Souza, J. G., Castro, D. T., Peng, Y., & Gartner, I. R. 2024. A machine learning-based analysis on the causality of financial stress in banking institutions. *Computational Economics*, **64**(3), 1857–1890.

Uddin, M.S., Chi, G., Al Janabi, M. A. M., Habib, T., & Yuan, K. 2022. Modeling credit risk with a multi-stage hybrid model: An alternative statistical approach. *Journal of Forecasting*, **41**(7), 1386–1415.

Veganzones, D., Séverin, E., & Chlibi, S. 2023. Influence of earnings management on forecasting corporate failure. *International Journal of Forecasting*, **39**(1), 123–143.

Wang, G., Zhao, H., Song, L., & Shu, L. 2026. Do multilayer networks amplify systemic risk? Evidence from the Chinese real estate industry. *Computational Economics*, **67**, 123–145.

Wang, L., Yu, Z., Ma, J., Chen, X., & Wu, C. 2024. A two-stage interpretable model to explain classifier in credit risk prediction. *Journal of Forecasting*.

Zou, S., Cai, Z., Chu, C., Zhao, Z., Cai, H., Shen, N., & Ren, J. 2025. Variable selection and fusion sampling of unbalanced data in credit risk assessment. *Computational Economics*.

Štrumbelj, E., & Kononenko, I. 2011. A general method for visualizing and explaining black-box regression models. *Pages 21–30 of:* Dobnikar, A., Lotrič, U., & Šter, B. (eds), *Adaptive and Natural Computing Algorithms*. Springer.

Štrumbelj, E., & Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, **41**(3), 647–665.

Appendix

A Data processing and description

The following subsections discuss the data processing steps to obtain the final sample used in our analysis, while the last subsection describes the variables used.

A.1 Sampling

1. **Firm selection:** we first identify firms that have at least five years of observations, at least one non-missing value of the binary failure variable and reside in euro-area countries. All other firms are excluded.

2. **Country quotas:** we fix a target sample size of $M = 10{,}000$ firms. Quotas by country are computed proportionally to their GDP weights. Since integer rounding may cause discrepancies, we use the largest remainder method to ensure that quotas sum exactly to M. When the number of available firms in a country is smaller than the quota, the quota is reduced accordingly and the remainder redistributed to other countries with spare capacity.

3. **Sector quotas:** within each country, firms are allocated proportionally to GVA across sectors. Again, the largest remainder method and capacity checks ensure that sector quotas match availability.

4. **Random sampling:** within each (country, sector) cell, firms are randomly selected using a uniform random number generator seeded for reproducibility. This yields a final set of sampled firm identifiers.

The resulting percentage of firms sampled by country and sector (i.e., sample representativeness) is highlighted in figure 10.

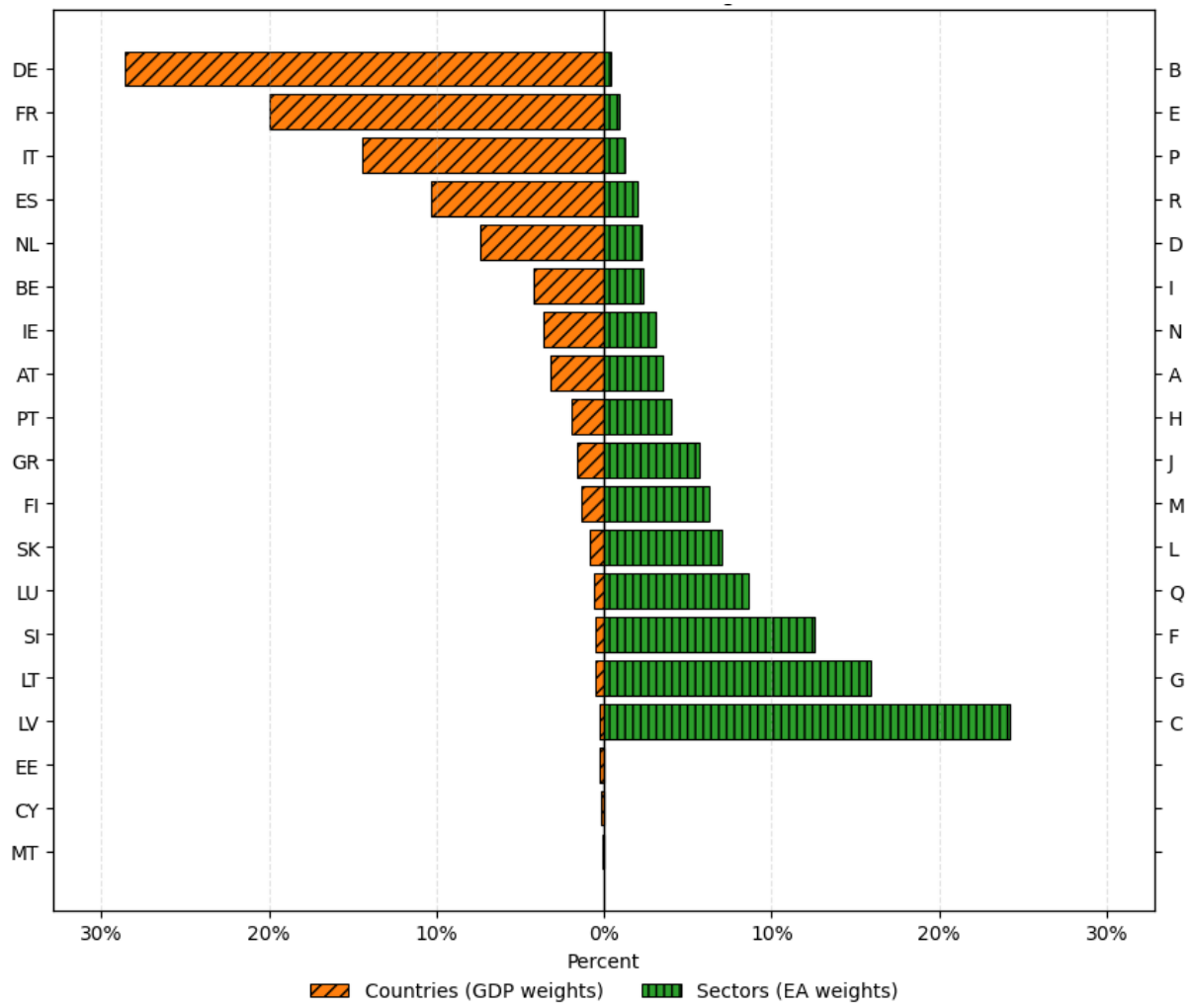


Figure 10: Percentage of firms sampled (i.e., sample representativeness) by country (left bar chart) and sector (right bar chart), respectively. The NACE section mapping to sector names is reported in table A2.

A.2 Imputation

To build a balanced panel of the sampled firms, we apply the following steps:

1. **Date construction:** for each record we construct a proper date (end of month), based on the year and month variables.

2. **Categorical variables:** missing categorical values are filled in two stages: first using the firm-specific mode, and then (if still missing) using the overall mode across all firms.

3. **Temporal alignment:** firm-level data are resampled at a quarterly frequency. For stock variables (balance sheet levels), the last observation of each quarter is retained; for flow variables (income statement items), quarterly means are used.

4. **Interpolation:** for flow variables, missing values are interpolated using a polynomial method of order two, with fallbacks to time interpolation and linear interpolation if necessary. Remaining gaps are filled using firm-specific medians.

5. **Non-negativity:** variables that cannot be negative (e.g., assets, liabilities, employment) are clipped at zero if imputation produced invalid values.

6. **Imputation:** variables that still have missing values are imputed via a random forest imputer (classifier for categorical, regression for continuous variables). This applies to all variables except the binary failure variable.

7. **Derived variables:** where possible, derived measures, like firm age, are recomputed. As an additional step, missing values of the binary variable that represents default status are imputed based on a cash flow test: if cash flow plus cash equivalents is less than financial expenses, the firm is classified as defaulted.

8. **Consistency check:** for consistency, firm identifiers are verified across all records after imputation. In case of conflicts, the most frequent identifier is retained.

This process produces a balanced and consistent panel dataset of sampled firms, ready for further econometric analysis.

A.3 Data description

Table A2 describes the sectors, while table A3 summarizes the raw fields used before feature engineering. Where synonyms exist (e.g., `country` vs. `cntry`; multiple liabilities totals), we keep one canonical column and document any recoding in the replication code.

Table A2: Mapping from sector labels to NACE section codes

Sector label (dataset)	NACE section
Manufacturing	C
Wholesale and retail trade	G
Administrative and support service activities	N
Real estate activities	L
Construction	F
Accommodation and food service activities	I
Transportation and storage	H
Professional scientific and technical activities	M
Information and communication	J
Arts entertainment and recreation	R
Water supply sewerage waste management and remediation Activities	E
Electricity gas steam and air conditioning supply	D
Human health and social work activities	Q
Mining and quarrying	B
Education	P
Agriculture forestry and fishing	A

Table A3: Variables by category (raw inputs prior to feature engineering)

Category	Fields (examples)	Notes
Identifiers & calendar	`bvd_id`, `date`, `ym`, `date_month_start`, `date_month_end`	Unique firm key and time stamps. We standardize to a quarterly `date` and sort within firm to build lags/differences.
Assets (stocks)	`totalassets`, `currentassets`, `fixedassets`, `tangiblefixedassets`, `intangiblefixedassets`, `otherfixedassets`, `cashcashequivalent`, `sum_cash`	Size, composition, and liquidity stock measures. Used to form log levels and ratios (e.g., tangibility).
Liabilities (stocks)	`currentliabilities`, `noncurrentliabilities`, `othercurrentliabilities`, `othernoncurrentliabilities`, `creditors`; totals `total_liabilities`, `longtermdebt`, `totaldebt`, `totaldebt_adj`, `loans`, `shareholdersfunds`, `capital`	Funding structure and maturity. We harmonize to a single liabilities total and prefer `totaldebt` (with `totaldebt_adj` as a sensitivity).
Income statement (flows)	`sales`, `operatingrevenueturnover`, `operatingplebit`, `ebitda`, `plforperiodnetincome`, `financialrevenue`, `financialexpenses`, `cashflow`, `addedvalue`	Profitability, operating strength, interest burden, and internal cash generation. Used to form margins and coverage ratios.
Employment	`numberofemployees`	Firm scale and labor intensity.
Categoricals (firm)	`country`; `sector`; NACE codes: `nace_code`, `nace_lev2`, `nacerev2corecode4digits`, `lev2`, `reclev2`, `lev2_win`; `status`, `consolidationcode`; `originalcurrency`	One-hot encoded for country/sector/status/consolidation. We reconcile duplicate encodings (prefer a single `country`, a single NACE resolution, and one `sector`). Currency retained for metadata; ratios/logs limit currency effects.

B Random forests: details

RF are nonparametric ensembles that approximate an unknown mapping $y = f(x)$ without specifying a functional form. An RF aggregates B randomized decision trees $\{h_b\}_{b=1}^{B}$, each trained on a bootstrap sample and, at every node, splitting on a random subset of features of size m_{try}. For classification with classes $c \in \{1, \ldots, C\}$, the forest's class probability is the average of leaf proportions,

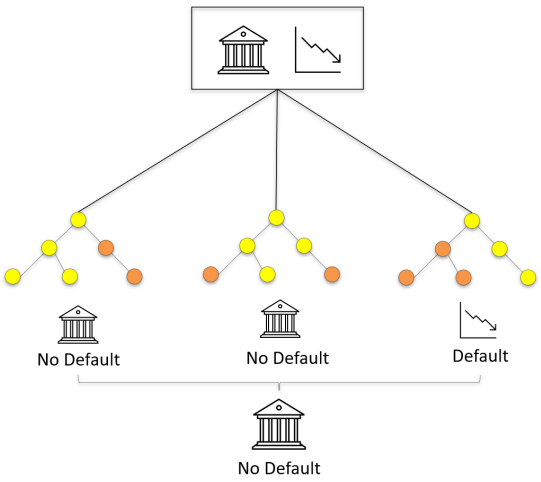
$$\widehat{\pi}_c(x) \;=\; \frac{1}{B}\sum_{b=1}^{B} \widehat{\pi}_c^{(b)}(x),$$

where $\widehat{\pi}_c^{(b)}(x)$ is the empirical class frequency in the terminal leaf reached by x in tree b. Majority voting $\hat{y}(x) = \arg\max_c \widehat{\pi}_c(x)$ yields a hard classifier.

Two sources of randomness—bootstrapping observations and subsampling features—decorrelate trees. Individual trees are grown deep (low bias, high variance), while averaging across weakly correlated trees reduces variance with modest bias increase, delivering strong out-of-sample performance. RF offers practical benefits: (i) out-of-bag (OOB) samples give near-unbiased error estimates and class probabilities; (ii) few hyperparameters drive performance (B, m_{try}, depth and leaf size, and class weights); and (iii) global interpretability via variable importance (permutation importance and impurity-based measures), supported by partial dependence or accumulated local effects for marginal relationships.

In this paper, we forecast firm default at horizons $H \in \{1, 2, 3, 4\}$ quarters. Let $\mathbf{x}_{i,t}$ denote features for firm i at time t, and $D_{i,t+H} \in \{0, 1\}$ indicates default at horizon H. Figure 11 provides an example of RF's functioning. Panel (a) shows a forest with three trees in which the first and second trees predict "No default", while the third tree predicts "Default". Under majority voting, the aggregate decision taken by RF is then "No default". Importantly, each tree is trained on a different bootstrap sample and learns with a different subset of features. Panel (b) shows a decision region when two features at random are used: EBITDA over total assets and profitability. Along these two dimensions, a tree in the forest is able to segregate defaulted firms (orange dots) from non-defaulted ones (blue dots). When EBITDA over total assets and profitability become negative, it is fairly easy for a tree to learn to classify a firm as a defaulted one. Likewise, when both are positive, a tree is able to recognize a firm as a non-defaulted one.

(a) Subset of a forest



(b) Decision regions

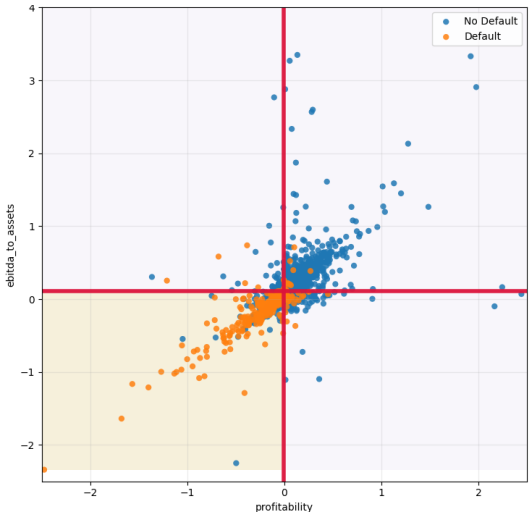


Figure 11: Subset of a forest (panel (a)) and decision region of forest's tree (panel (b)), respectively).

Problem setup and label construction. For each horizon H, labels are formed by a forward within-firm shift as follows

$$D_{i,t+H} \;=\; \text{shift}_{-H}\{D_{i,t}\},$$

so the training pair $(\mathbf{x}_{i,t}, D_{i,t+H})$ uses only information available at t. The last H observations per firm are unlabeled by construction and excluded from training, inducing right-censoring near the sample end. Right-censoring is handled by time-aware cross validation (see below).

Features and leakage control. We engineer point-in-time ratios and logs (liquidity, profitability, leverage, tangibility; log of capital/loans/assets), then add dynamics: one-quarter lags, first differences, and one-step percentage changes. Variables are winsorized at 1% to dampen outliers, categorical controls (country, sector, status, consolidation code) are one-hot encoded, numeric features with missingness larger than 40% are dropped, and remaining gaps are imputed via date-wise medians with a global fallback. All steps are computed in strict time order within each firm so that lags/differences use only history.

Time-aware cross validation, hyperparameters and search. Model selection and assessment use an expanding-window[52] split by date with group purging: at split k, the training

[52]Structural breaks are not always a concern when using an expanding window, see Nocciola (2022).

window runs through a cut t_k and the test window is the next date t_{k+1}; firms appearing in the test set are excluded from the training set. This blocks both temporal leakage and entity leakage (same firm in train/test at different dates). We tune {#trees, max depth, min samples split/leaf, max features} via Bayesian optimization. The objective is AP (PR-AUC) computed on the time-aware folds, which is informative under severe class imbalance (see figure 1, panel (a)). The final model is refit on the last expanding training fold.

Pitfalls under class imbalance and survivorship. Defaults are rare. Under severe class imbalance, off-the-shelf classifiers minimize a loss dominated by the majority class, tilting decisions toward non-default. A fixed threshold $\tau = 0.5$ is ill-suited when the base rate $\pi_{s,t}$ is small. Resampling during training can distort posterior probabilities unless followed by proper calibration and "accuracy" is largely uninformative because a trivial majority rule can score highly without catching any defaults. Compounding this, because defaults are rare and defaulting firms typically exit the panel, samples conditional on long reporting histories over-represent survivors (non-defaults), which pushes the base rate downward, shifts covariates toward larger, older and fitter firms, and inflates apparent accuracy for rules that favor the majority class.

How we address these issues: decision thresholding. We align training, decision, and evaluation with the rare-event objective by

- using cost-sensitive learning via balanced subsample (`class_weight='balanced_subsample'`) so that splits account for the minority class;
- selecting an operating classification threshold τ on an out-of-time (OOT) fold to maximize an asymmetric utility (here F_2, favoring recall - see definitions in appendix C) rather than fixing $\tau = 0.5$, yielding a recall-tilted decision rule suitable for early-warning settings;
- avoiding any resampling at evaluation time and applying probabilistic post-calibration (isotonic with sigmoid/Platt fallback) on the OOT fold that preserves the fold's base rate;
- optimizing hyperparameters and reporting minority-sensitive metrics — PR-AUC (AP) as the model-selection objective, along with OOT F_2 and Brier score; and
- enforcing time-aware cross validation with an expanding, group-purged split to prevent temporal and entity leakage.

Collectively, these steps yield calibrated probabilities, recall-aware decisions, and performance estimates that reflect the default environment.[53]

Out-of-sample projections. To generate projections at historical cutoffs $\tau \geq$ 2019Q4, we freeze features at τ by taking the last observed row per firm with $t \leq \tau$, recompute lags/differences using only data up to τ, and score with the horizon-specific calibrated RF to obtain $\widehat{p}_{i,\tau}^{(H)} \approx \mathbb{P}(D_{i,\tau+H} = 1 \mid \mathbf{x}_{i,\tau}, \mathbf{z}_{s,\tau})$. Because one-hot levels can change over time, we align the score-time design matrix to the training feature set (adding missing columns as zeros and dropping extras) to ensure identical columns and order.

Summary. Our RF (i) constructs horizon-consistent labels without leakage, (ii) engineers ratios and dynamics with robust imputation, (iii) uses expanding, group-purged validation, (iv) tunes hyperparameters for PR-AUC under class imbalance, (v) calibrates probabilities on an OOT fold, (vi) selects a recall-tilted threshold via F_2, and (vii) produces projection paths by re-scoring at each historical cutoff using frozen features.

C Model performance metrics and variable importance

This appendix reports some useful definitions of the metrics used for model evaluation in the main body of the paper.

C.1 Aggregate metrics

Let DR_t denote the realized default rate at time t (here we abstract from the subscript s) and $\overline{p}_t$ the corresponding model (RF and logit) prediction. Then, the evaluation metrics are defined as follows:

[53] However, we do not (yet) implement explicit survivorship-bias corrections. Future work could estimate observation/exit hazards to form inverse-probability-of-censoring weights (IPCW), include tenure and filing-gap controls, and report performance on panels that retain real-world attrition and class proportions.

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}\left|\overline{p}_t - DR_t\right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\overline{p}_t - DR_t\right)^2},$$

$$\text{Corr\_diff} = \text{corr}\left(\Delta\overline{p}_t,\, \Delta DR_t\right),$$

where Δ is the difference operator. These metrics are evaluated in-sample and out-of-sample, i.e., over horizon H.

C.2 Micro metrics

Let True Positives (TP) be realized defaults that are predicted as defaults, False Positives (FP) be realized non-defaults that are predicted as defaults, True Negatives (TN) be realized non-defaults that are predicted as such and False Negatives (FN) be realized defaults that are predicted as non-defaults, as summarized in the following table:

Actual \ Predicted	0	1
0	TN	FP
1	FN	TP

Table C4: Confusion matrix: actual non-defaulted (first row), actual defaulted (second row), predicted non-defaulted (first column) and predicted defaulted (second column).

Moreover, let P be Precision and R be Recall, which are defined as follows:

$$P = \frac{TP}{TP + FP},$$
$$R = \frac{TP}{TP + FN}$$

P sheds light on the ability of the model to identify how many defaults occurred out of the total defaults predicted. R[54] instead measures the model's ability to identify how many of the defaults that actually occurred were correctly predicted. As FP and FN are inversely

[54]Recall is also called power in statistics, sensitivity in medicine or simply TP Rate (TPR).

related, the two measures trade off as well, i.e. a higher precision (recall) implies a lower recall (precision) and vice versa.

Given the above definitions, the evaluation metric Average Precision (AP), which is the PR-AUC (the area under the PR curve), is given by the area of a trapezoid formula:

$$\text{AP} = \sum_{n=2}^{N} (R_n - R_{n-1}) \left(\frac{P_n + P_{n-1}}{2} \right)$$

AP tells how well, on average across all thresholds, the model performs in classifying actual defaults as such (P).

Another measure that is used in the paper is the ROC-AUC (the area under the Receiving Operating Characteristic curve). Let $FPR = FP/(FP + TN)$ be the FP Rate or type I error (as known in statistics) and $TPR = R$. The evaluation metric ROC-AUC is given again by the area of a trapezoid formula

$$\text{ROC-AUC} = \sum_{n=2}^{N} (FPR_n - FPR_{n-1}) \left(\frac{R_n + R_{n-1}}{2} \right)$$

Next, we can define F_β with $\beta = 2$ as

$$F_\beta \;=\; (1 + \beta^2) \frac{PR}{\beta^2 P + R},$$

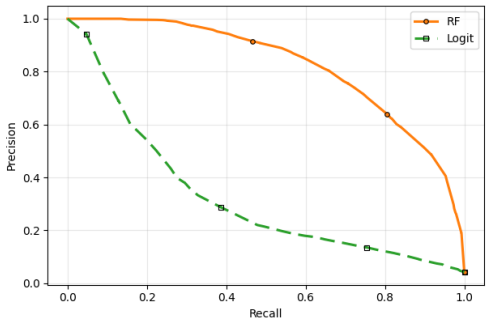
where P and R are precision and recall at τ.

Finally, the Brier score, or simply Brier, is an MSE applied to binary classification. Its formula is given by

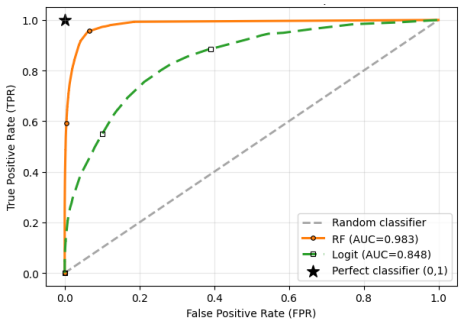
$$\text{Brier} = \frac{1}{N} \sum_{n=1}^{N} \left(\hat{p}_n^{(H)} - D_n \right)^2$$

where $\hat{p}_n^{(H)}$ is the estimated conditional PD for observation n (a firm-time pair) at horizon H and D_n is the corresponding observed default indicator.

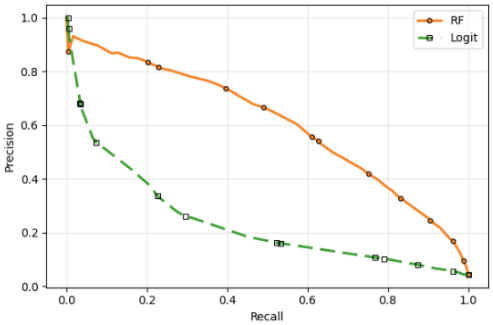
C.3 Out-of-sample performance at the micro level



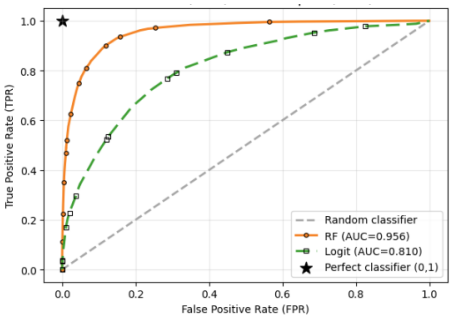
(a) PR curve. H=1.



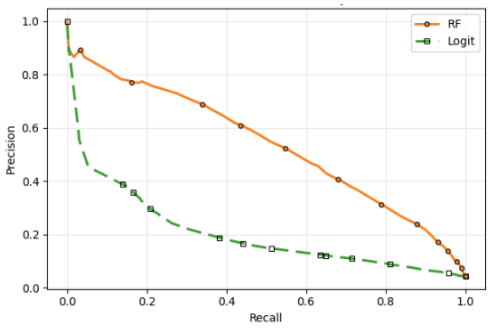
(b) ROC curve. H=1.



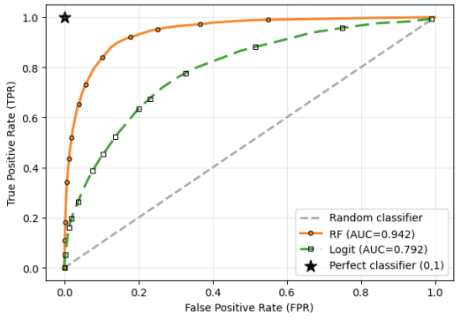
(c) PR curve. H=2.



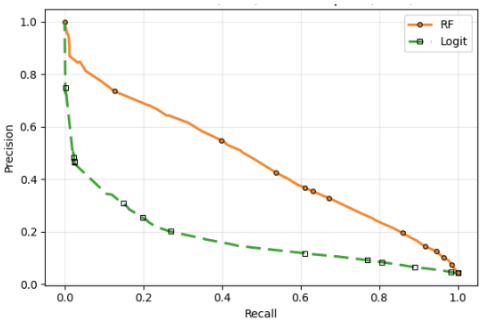
(d) ROC curve. H=2.



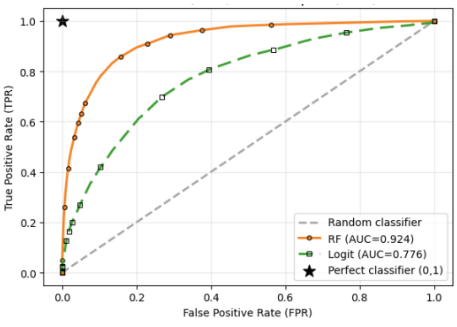
(e) PR curve. H=3.



(f) ROC curve. H=3.



(g) PR curve. H=4.



(h) ROC curve. H=4.

Figure 12: OOS model performance at the micro level for each H: PR curve (panel (a, c, e, g)) and ROC curve (panel (b, d, f, h)) for both RF (orange) and logit (green). See also notes to figure 3.

C.4 Shapley additive explanations

This part summarizes SHAP by starting from Shapley values (Shapley (1953)). Assume that the relative prediction is a payoff that each feature j (in this case, the fictitious players) gains in a cooperative game, with "relative prediction" defined as the difference between a prediction based on a feature and the average prediction based on the remaining features. A Shapley value ϕ_j provides a fair way[55] to distribute the payoffs among the features by computing each feature's contribution through the average of all relative prediction changes when each feature joins each possible combination (or "coalition") of other features.[56] Formally, Shapley values are defined via a payoff or value function v of features in C in a cooperative game as

$$\phi_j = \sum_{C \subseteq N \setminus \{j\}} \frac{|C|!(|N| - |C| - 1)!}{|N|!} \Big(v(C \cup \{j\}) - v(C) \Big)$$

where N is the set of all features, $|N|$ is the number of features and the sum is over all subsets C of N not containing feature j, empty set included.[57] SHAP (Lundberg & Lee (2017)) represents Shapley values as coefficients of the features in a linear model:

$$p(\mathbf{y}') = \phi_0 + \sum_{j=1}^{|N|} \phi_j \mathbf{y}'_j$$

where p is the predicted PD (hat and subscripts assumed away for simplicity), $\mathbf{y}' = (y_1, \ldots, y_M)' \in \{0,1\}^{|N|}$ is the "coalition" vector of features[58], $|N|$ is the maximum coalition size and $\phi_j \in \mathbb{R}$ are Shapley values for feature j, where ϕ_0 is the Shapley value of an empty coalition. Estimation of Shapley values in an RF is based on the TreeSHAP algorithm, which aggregates path-wise contributions across all trees in the forest (Lundberg *et al.* (2019)).[59]

D Stress test: scenario severity and additional stress test results

D.1 Mapping of shocks: details of balance sheet reconstruction

[55] Fairness is a byproduct of the properties of Shapley values.
[56] Computation time is exponential in the number of features.
[57] The first term of the Shapley value can be expressed as a binomial coefficient, too.
[58] The entries of this binary vector take value 1 if a feature is present and 0 if it is not.
[59] See also the Python package `shap`.

Item	Formula
currentassets	share of totalassets
fixedassets	= totalassets currentassets
cashcashequivalent	share of totalassets, then updated via cashflow cumulation
total_liabilities	share of totalassets
currentliabilities	share of total_liabilities
noncurrentliabilities	= total_liabilities currentliabilities
loans	share of total debt
longtermdebt	share of total debt
tangiblefixedassets	share of fixedassets
intangiblefixedassets	share of fixedassets
otherfixedassets	= fixedassets tangible intangible
operatingrevenueturnover	= sales_proj
ebitda	= sales_proj × EBITDA margin
operatingplebit	= sales_proj × EBIT margin
addedvalue	= sales_proj × (added value / sales share)
financialrevenue	= cash_stock × (financial revenue / cash share)
creditors	= current liabilities × creditors share
othercurrentliabilities	= current liabilities × share
othernoncurrentliabilities	= non-current liabilities × share
capital	= equity × (capital / equity share)

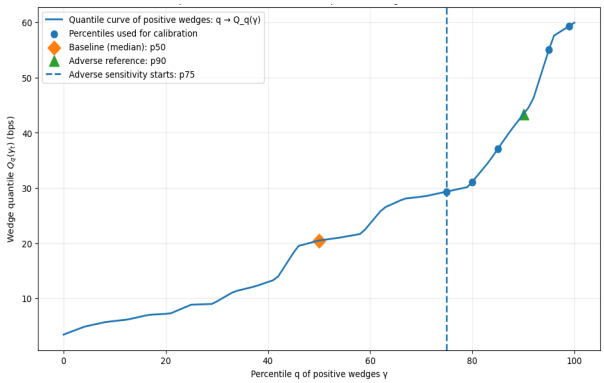
Table D5: Balance sheet (first block) and flow P&L (second block) items rebuilt from shares with reconstruction formulas. Cash is first "share-initialized", but then "flow-updated". Total assets, total debt, sales, operating profit and cash flow are obtained via the elasticities.

D.2 Scenario sensitivity and additional stress test results

The scenario severity exercise consists of varying the intensity of scenarios. We do so for the spread add-on to the interest rate used to project financial expenses (as proxied by interest expenses) and GVA growth. Starting from the spread add-on, we use the distribution of the positive wedge, as described in the paper, and select different percentiles corresponding to different severity levels. We start from the 75th percentile and increase it to test our model under different stress conditions including (and surpassing) the adverse scenario, noting that increases in interest rates follow a quasi-exponential curve (see figure 13, panel (a)). Second, we depress GVA growth across country-sectors beyond the adverse scenario, which, in turn, inflates the interest rate on financial expenses and boosts the total debt position through the estimated elasticity. We assess different shifts downward of the GVA growth curve across country-sectors matching various severity intensities (see figure 13, panel (b), for the most severe, full-freeze scenario).

Having described how we construct scenario severity, we then decompose the 3D surface

(a) Spread add-on severity curve



(b) Full-freeze GVA growth

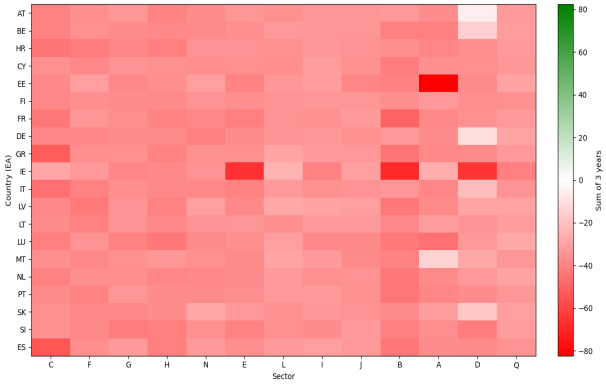
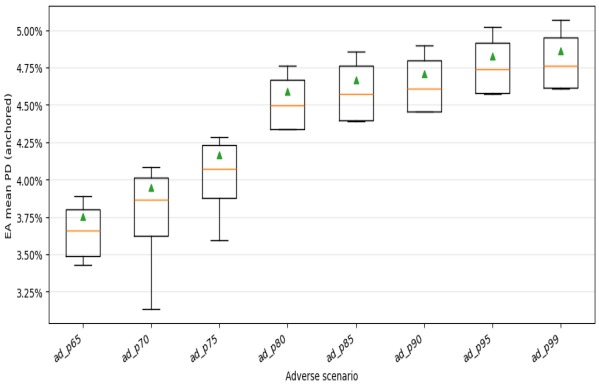


Figure 13: Scenario severity. Panel (a): Spread add-on severity curve. Percentile of positive wedges γ on the x-axis. Wedge quantile in bps on the y-axis. The vertical dashed blue line marks where the adverse-scenario sensitivity analysis starts, blue dots are the tested severities, the orange square is the baseline scenario and the green triangle is the adverse scenario. Panel (b): Full-freeze GVA growth scenario corresponding to the most adverse GVA growth curve shift downward tested. All other scenario severities lie between this full-freeze scenario (right pole) and the adverse scenario (left pole, see figure 5, panel (b)).

displayed in figure 7, panel (b), and analyze each horizontal axis independently: spread add-on and GVA growth. We look in 2D at scenario severity in each case by muting stress along the remaining axis. For example, we switch off stress on GVA growth and analyze the PD evolution when the spread add-on to the interest rate used to project financial expenses varies along the distribution of γ. The adverse scenario assumes a value of this parameter at the 90th percentile. We vary this percentile to simulate different scenario severities. Similarly, we switch off stress on the spread add-on and evaluate the PD evolution when GVA growth is stressed following a downward shift in its curve across country-sectors.

Interestingly, figure 14, panel (a), shows a non-linear relation of the S-shaped type between PD and scenario severity for the spread add-on. Up until the 75th percentile of the spread add-on distribution, PD grows linearly. In the interval 75th-80th percentile, PD suddenly jumps. After the 80th percentile, PD grows again linearly but, this time, more slowly. In sum, after the 75th percentile we observe a *staircase*-shaped behavior of PD. Instead, figure 14, panel (b), suggests a non-linear relation of the logarithmic type between PD and scenario severity for GVA growth.

(a) Scenario severity - Spread add-on



(b) Scenario severity - GVA growth

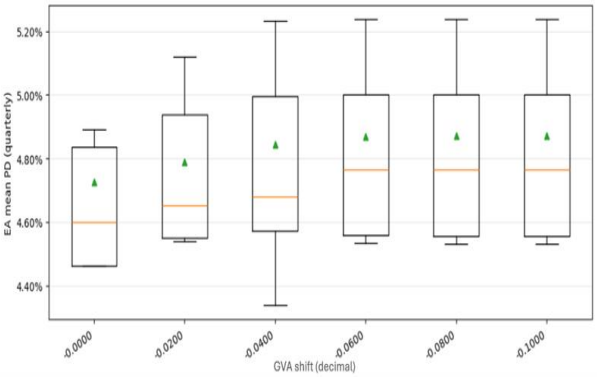


Figure 14: Scenario severity. Panel (a): PD evolution (in boxplot format) when the spread add-on to the interest rate used to project financial expenses varies along the distribution of γ. Panel (b): PD evolution (in boxplot format) when the GVA growth curve across country-sectors is shifted downward. For both panels, the green triangle represents the mean PD, the orange line represents the median PD and the black lines of the box represent the 95% confidence interval.

Acknowledgements

We thank Katrin Assenmacher, Ugo Albertazzi, Garbrand Wiersema, Aurea Ponte Marques, Costanza d'Acri Rodriguez, Matthias Sydow, Julian Metzler, Duc Thi Luu, Sylvain Barde, Jelena Zivanovic and the Editorial Boardof the European Central Bank (ECB) Working Paper Series and participants in the EcoMod2026 International Conference on Economic Modeling and Data Science, the Computing in Economics and Finance Conference 2026 and the Directorate General Macroprudential Policy and Financial Stability Seminar Series at the ECB for insightful comments and suggestions. The datasets employed in this Working Paper contain confidential statistical information. Its use for the purpose of the analysis described in the text has been approved by the relevant ECB decision making bodies. All the necessary measures have been taken during the preparation of the analysis to ensure the physical and logical protection of the information. In the write-up of this paper Generative AI and AI-assisted technologies were only used to improve the readability and language of the text as well as research assistance to improve coding. We carefully reviewed the results from these tools and are ultimately responsible for the content of this article.
The views expressed in this paper are our own and do not necessarily represent the views of the ECB or the Eurosystem.

Luca Nocciola
European Central Bank (affiliation during the development of this paper), Frankfurt am Main, Germany;
email: luca.nocciola@ecb.europa.eu

Samuele Scaglioni (corresponding author)
European Central Bank, Frankfurt am Main, Germany; email: samuele.scaglioni@ecb.europa.eu

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

PDF ISBN 978-92-899-7989-4 ISSN 1725-2806 doi:10.2866/8508602 QB-01-26-205-EN-N