

From Variable Importance to Economic Scenarios: Interpreting Machine Learning Models for Firm Failure Risk Management

Vlad TEODORESCU

Bucharest University of Economic Studies, Bucharest, Romania

**Corresponding author, vlad.teodorescu@fin.ase.ro*

Catalina-Ioana TOADER

Bucharest University of Economic Studies, Bucharest, Romania

Center of Financial and Monetary Research (CEFIMO)

catalina.toader@fin.ase.ro

Abstract. *Firm failure risk is an important topic for managers, credit institutions, investors and policymakers. In recent years, machine learning models have been increasingly used in this field. Although they often achieve high predictive performance, they typically lack interpretability. Existing literature on machine learning model interpretability focuses on individual variable importance with limited practical guidance for firm management. We use a large firm-level dataset of Romanian companies (360.249 observations) to train various XGBoost models and to examine their interpretability. We use a wide range of explanatory variables: firm size, industry classification, financial statement items scaled by total assets and 44 financial ratios. We use AUROC as performance measure, while firm failure is proxied by the transition from positive to negative equity. We calculate variable importance using the permutation feature importance method. Our results show that variable importance in high-dimensional settings can be unstable, as there are considerable redundancies and compensation effects between explanatory variables. Furthermore, individual variable changes have limited effects on predicted probabilities of failure, hence having little practical relevance. We illustrate that using coherent scenario-based modifications to financial statements has substantially larger and economically meaningful effects on predicted failure risk. Compared to variable importance ranking and score assignment, scenario-based analysis provides clearer interpretations and more relevant insights for management. Our contribution aims to shift the focus from individual variable importance to economically meaningful insights derived from machine learning models.*

Keywords: firm failure prediction, machine learning, model interpretability, variable importance, scenario analysis, xgboost, financial ratios.

Introduction

Firm failure, financial distress, credit default, bankruptcy are all closely related topics relevant for credit institutions, investors and government policy makers.

In recent years, increasingly sophisticated models have been used, leading to significant gains in predictive accuracy. Machine learning models are particularly suited for firm failure prediction tasks as they are able to learn non-linearities and complex patterns in financial data. However, on the downside, such models are increasingly harder to interpret. In many cases, they provide accurate probabilities of failure, default or bankruptcy, but offer limited insight into why a firm is deemed riskier than another, or what management can do to improve its financial position.

An increasing number of studies aims to provide techniques for interpreting black-box machine learning models. Many such studies, however, stop at identifying important explanatory variables and describing the relationship between individual variables and model predictions.

Machine learning models can handle a multitude of explanatory variables with little or no constraints regarding their inclusion. It is therefore common practice to include as many variables as possible, in the hope that each may shed a slightly different light on the target variable. Indeed, we show that including redundant, highly correlated, duplicate or irrelevant variable does not hurt performance, while it may marginally improve it.

While this approach is potentially beneficial for predictive accuracy, it poses serious challenges for model interpretation and variable importance analysis. If multiple variables convey similar and overlapping information, variable importance becomes unstable and highly dependent on which other variables are present. We show that eliminating even the most important variables from a model configuration can be compensated for by the remaining ones with only slightly lower model performance. As a result, an important question arises: what is, then, the practical value of calculating variable importance for a high dimensionality machine learning model?

To address this issue, we illustrate how a black box machine learning model can be used in practice to derive meaningful insights for firm management. We construct a simple, but consistent scenario and show that its results can give managers a better understanding of a firm's failure risk.

Overall, we argue that model interpretation should be viewed less as a tool to identify individual important variables and more as a tool to shed light on why a firm is predicted to have a higher or lower probability of failure.

This paper makes three main contributions. First, we provide empirical evidence that variable importance may have limitations in settings where explanatory variables have information overlap and redundancy. Second, we show that isolated single feature adjustments have limited effects on model predicted probabilities of failure. Third, we demonstrate that an economically coherent scenario produces a much larger shift in predicted probabilities of default, offering more informative insights to practitioners. Taken together, these contributions suggest that the a scenario-based analysis approach may provide more economically meaningful interpretations of black-box machine learning models.

Firm failure, default and bankruptcy are distinct, but related concepts. Bankruptcy typically refers to a formal legal procedure, while default refers to a company's failure to repay some of its due amounts in time. In a context where bankruptcy or default indicators are not available, accounting-based proxies are frequently employed. In this paper, we define firm failure as the transition from positive to negative equity. This measure does not coincide with legal bankruptcy or payment default, but it typically signals firm distress and an erosion of creditor protection.

Literature review

Accounting-based and structural approaches to firm failure

The link between a company's debt structure and its risk of default is a key issue in corporate finance. This idea comes from early studies on structural credit risk models and predicting financial distress. The Merton (1974) framework formalizes the role of leverage and debt servicing capacity in determining default probability; however, contributions by Beaver (1966) and Altman (1968) demonstrate that accounting-based indicators possess substantial explanatory power regarding corporate default. In this area of study, the interest coverage ratio (ICR) has become an important sign of how well a company can pay interest and a common way to measure financial vulnerability.

In financial literature, the probability of default has been predominantly addressed by analyzing a firm's ability to meet its financial obligations from the cash flows generated by its assets. However, a theoretical turning point is represented by the structural model proposed by

Merton (1974). The transition from positive to negative equity is not just an accounting fluctuation, but signals the complete exhaustion of the safety zone for creditors. Campbell et al. (2008) emphasize that firms operating with negative equity exhibit extreme volatility and a low probability of survival, this indicator being a much more robust proxy for financial distress than isolated liquidity indicators. This perspective motivates the use of negative equity as an accounting-based proxy for firm failure in empirical applications where direct default indicators are unavailable.

Traditionally, logistic regression (Logit) has been the benchmark tool in credit risk modeling, valued for its structural simplicity and ability to provide directly interpretable coefficients in the form of odds ratios. However, modern literature criticizes the rigidity of this model, as Logit imposes a linear relationship between the log-odds of bankruptcy and financial indicators, while suffering from high sensitivity to multicollinearity — an endemic problem in financial data, where most solvency and liquidity ratios are correlated with each other.

Traditional statistical models and their limitations

Traditionally, logistic regression (Logit) has been the benchmark tool in credit risk modeling, valued for its structural simplicity and ability to provide directly interpretable coefficients in the form of odds ratios. However, modern literature criticizes the rigidity of this model, as Logit imposes a linear relationship between the log-odds of bankruptcy and financial indicators, while suffering from high sensitivity to multicollinearity — an endemic problem in financial data, where most solvency and liquidity ratios are correlated with each other. These limitations become increasingly relevant as the number of explanatory variables grows, with multiple ratios conveying overlapping financial information.

Financial ratios, firm characteristics and default risk

The central conclusion of traditional literature, which extends Logit-type models, confirms that profitability (ROA) and leverage remain the pillars of financial stability, but their impact is now analyzed through the lens of nonlinear interactions (Barboza et al., 2017). Recent studies emphasize that low profitability combined with a fragile capital structure acts as a dominant predictor of insolvency, but Alonso and Carbó (2021) demonstrate that these factors gain superior predictive power only when integrated into models that capture the dynamics of cash flows, not only balance sheet stocks. The importance of current liquidity, established by traditional research as an indicator of short-term survival, is reaffirmed by Lessmann et al. (2015), who show that working capital deficiency is a critical signal of technical insolvency, especially for small and medium-sized enterprises. This "size effect" is nuanced in contemporary literature - while small companies are much more sensitive to equity erosion, Jones et al. (2017) note that large firms tend to mask default risk by accessing refinancing until they reach a critical threshold of negative equity. Although Logit models offer direct interpretations, research such as Tian et al., 2015 argues that their view remains static, failing to capture the speed with which financial indicators precipitate default, a limitation that machine learning algorithms overcome by capturing early warning thresholds and high data variability.

Machine learning models in firm failure prediction

In the last decade, literature significantly shifted to Machine Learning (ML) research methods. Several studies like those of Barboza et al. (2017) demonstrate that algorithms like Random Forest or XGBoost (Chen and Guestrin, 2016) consistently outperform traditional econometric models in terms of prediction accuracy. The advantage of machine learning algorithms lies in their ability to

capture complex and non-linear interactions between financial indicators and firm failure risk without imposing rigid restrictions on the data set. In this context, Lin (2024) compared the performance of four algorithms (logistic regression, Random Forest, XGBoost and AdaBoost) and concluded that gradient boosting ensemble algorithms in general, and XGBoost in particular, offer superior performance in predicting default.

Variable importance

Alanis et al. (2022) also benchmark the performance of various machine learning models in financial distress prediction and find that gradient boosting algorithms outperform other models. They also perform permutation feature importance and find that “excess stock returns, idiosyncratic risk, and relative size are the more important variables for pre-dictions”.

Teodorescu and Toader (2024) found that variable importance, using permutation feature importance, differs across model types (XGBoost, random forest and logistic regression with regularization). This result highlights the model-dependent nature of such interpretations.

Recent work on interpretability stress that counterfactual explanations should be grounded in the causal structure of the data-generating process (Guidotti, 2022). This perspective suggests potential limitations of interpretations based on isolated changes in individual financial variables (without simultaneously and consistently changing other variables as well), particularly in high-dimensional settings characterized by strong interdependencies.

Methodology

Data

Firm-level financial data are obtained from the Orbis database provided by Bureau van Dijk. We construct a firm-level dataset for Romanian companies, applying the following selection criteria: (i) at least one employee reported in the most recent available financial statements, and (ii) positive equity in 2021, with equity information available for 2022. The financial statement items retrieved from Obis are presented in the Appendix.

As Orbis does not provide a direct indicator of default or bankruptcy for the firms in our sample, we use negative equity as a proxy for firm failure. This approach is common in empirical research, as negative equity is an indicator of financial distress and of a reduced creditor protection buffer. Nevertheless, this proxy has limitations, as companies with negative equity can continue to operate for extended periods of time. This may be true more often for smaller firms, especially if they are family owned or with a single shareholder. As a result, this approach could introduce a size-related bias. Although this measure remains informative, as it is indicative of a financial situation deterioration, these limitations should be considered when interpreting the results.

We define the dependent variable as $Y=1$ if a firm’s equity is positive in 2021 and turns negative in 2022, and $Y=0$ if equity remains positive over the same period.

Explanatory variables consist of balance sheet, income statement, and memorandum items from 2021. For out-of-time validation (OOT), we use corresponding financial statement data for 2022 and 2023, applying the same selection rules.

Since we use XGBoost, no ex-ante transformations or imputations are applied to the data. Missing and extreme values are left unchanged, as XGBoost has in-built mechanisms to handle them by optimal split assignment.

Table 1 shows the sample composition of train and OOT data, after all filters are applied:

Table 1. Sample composition for training and test sets

	Observations	Failure (Y=1)	Failure rate
2021/2022 training set	360,249	16,463	4.57%
2022/2023 OOT test set	386,353	24,054	6.23%

Source: Authors' own research.

k-fold cross-validation

We use k-fold cross-validation for model selection, both for hyperparameter tuning and for exploring different sets of explanatory variables.

The k-fold cross-validation technique divides the training data into k subsets. Iteratively, one set is set aside for testing/validation, while the remaining k-1 sets are used to train a model. The model is evaluated out-of-sample on the kth test/validation fold. In the end, after all k iterations are performed, out-of-sample predictions are obtained for the whole training data set.

The model's cross-validation performance is calculated as the average of the performance of all k models.

k-fold cross-validation was shown to be an effective model selection technique (Teodorescu and Brasoveanu, 2025).

We set k=4 following Teodorescu and Brasoveanu (2025), which show that large values for k do not significantly improve a model's capacity to generalize, while risking overfitting and requiring significantly more computational effort.

AUROC

We evaluate model performance using the area under the receiver operating characteristic curve (AUROC), which is widely used in the literature on binary classification and firm failure modeling. AUROC assesses the model's ability to discriminate between firms that experience financial failure and those that do not, independently of the classification threshold. AUROC measures the proportion of correctly ranked pairs of firms consisting of one failed and one non-failed firm.

Explanatory variables

We compute two sets of possible explanatory variables (44 financial ratios and financial statement items scaled by total assets). We use all computed explanatory variables to train a main model. We use various subsets of these features to assess how model performance and variable importance relates to the choice of the explanatory variable set.

Adding total assets in levels (raw values, as reported by companies) allows the model to capture firm size information, which is not present in the scaled financial statement items or in the financial ratios.

Additionally, all conditions included the 4-digit NACE code.

XGBoost

XGBoost is a Machine Learning (ML) algorithm with notable performances in structured data (Chen and Guestrin, 2016).

Deep Learning (DL) techniques may offer superior performance with unstructured data or when massive amounts of data are available. In settings such as credit risk, bankruptcy or firm failure, where data is structured and the number of observations runs between hundreds and millions, traditional machine learning models, such as Random Forest or XGBoost, remain superior (Grinsztajn et al. 2022).

XGBoost is a decision tree ensemble method. It is an iterative algorithm, where successive trees are built to minimize the previous forest's errors.

We use the *xgboost* library in R.

We performed a guided grid search in order to tune the model's most important hyperparameters: `eta`, `min.child.weight`, `subsample`, `max_depth`, `colsample_bytree`, `colsample_bylevel`, `colsample_bynode`, `scale_pos_weight`, `max_delta_step`, and `gamma`. We settled on a configuration with high relative k-fold cross-validation performance (using AUROC) and ran the rest of the experiments with that configuration alone to compare relative performance between various sets of explanatory variables.

Permutation feature Importance

XGBoost, like most ML models, is a black-box type of model. While exhibiting high predictive performance, it cannot offer an explanation as to why a certain score was derived, unlike, for example, a regression, which has coefficients that could be used to interpret results.

Permutation Feature Importance (PFI) is a model agnostic technique for measuring feature importance. It does not assume any functional form of the model and can work with layered models as well.

PFI for an explanatory variable X_i requires the permutation of all its values between test set observations and recomputing the model predictions. The deterioration in model performance is considered a measure of that feature's global importance. Given the way it is calculated, a feature's importance as assessed by the PFI technique depends both on how the model uses the respective feature and also on the distribution of its values among observations in the test set.

Given the stochastic nature of PFI calculations, we compute 10 random permutations for each model and average the results.

We also analyze the effects on model performance of the removal of both the least and most important features.

We also evaluate the impact on model predictions of changes in important variables, answering questions such as “what is the impact on predicted failure probabilities of a +/-10% change in each important variable?”.

Scenario analysis

Our results with PFI show that variable importance may be difficult to interpret in a context where many alternative features with overlapping information are included. Machine Learning models can benefit from including similar information presented in slightly different ways, for example, from the inclusion of both ROE and ROCE, as each one of them may prove slightly more informative in some contexts. However, the removal of any of them might not significantly degrade model performance, while the removal of both may. Hence, the assigned variable importance might be misleading and not very informative for firm management.

As a result, we illustrate how economically coherent scenarios can be used to provide meaningful insights for companies. To demonstrate this approach, we use a stylized pricing scenario where a company is able to increase prices by 10% without losing market share. The aim is not to derive firm-specific or policy recommendations, but rather to quantify the effects of coherent financial statement adjustments compared with single feature adjustments.

While we rely on a single illustrative scenario, the method presented can be employed with a wide range of scenarios, such as increasing or decreasing leverage, working capital changes or cost adjustments. In practical applications, scenario construction depends on firm specific context and managerial priorities.

Our choice for scenario-based interpretations is motivated by Guidotti's assertions that counterfactual explanations should be consistent with the causal structure of the data (Guidotti 2022). Explanatory variables typically used in firm failure prediction models (i.e. financial ratios) have two layers of causality: first, changes in accounting ratios are the result of changes in financial statement items (for example, ROE changes if net income, equity or both change). As a result, an isolated change in one financial ratio violates accounting identities. Instead, it should be accompanied by changes in related ratios. At a deeper level, changes in financial statement items stem from changes in the firm's economic activity in the real world. These two levels of causality justify our approach in using coherent scenarios rather than isolated variable changes.

Results and discussions

Descriptive statistics

According to our empirical results, the most important variables included in the main model are total assets in levels (a measure of company size), NACE 4-digit industry code, return on equity, debt to equity, other operating items (in P&L) and current assets scaled by total assets

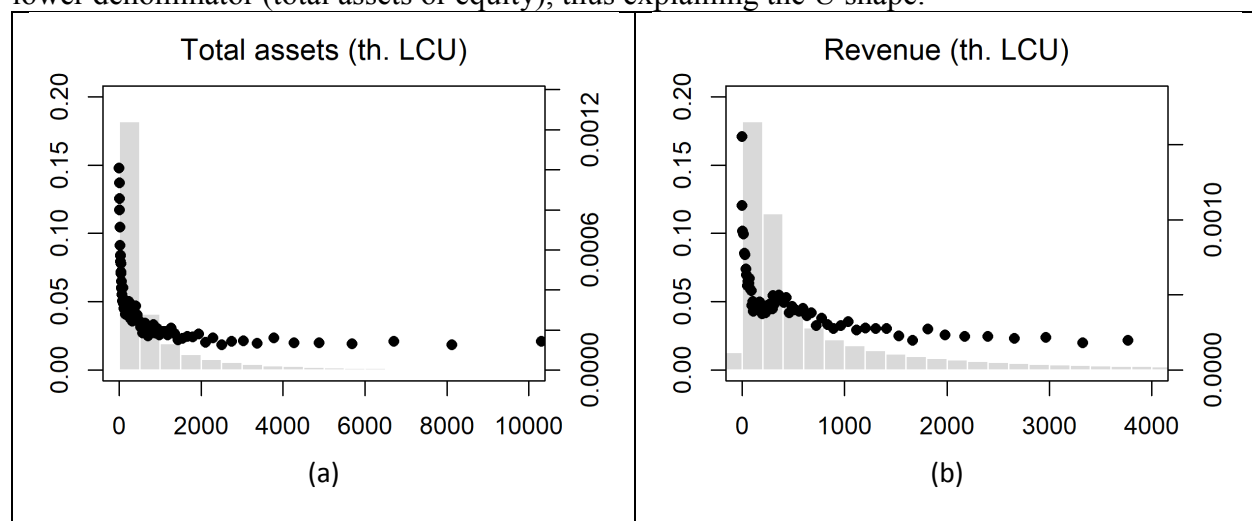
The relationship between failure rates and financial indicators is predominantly non-linear, with some variables exhibiting U-shaped patterns. Figure 1 presents the relationship between selected explanatory variables and observed failure rates, together with the distribution of firms across the values of these variables.

Measures of company size (total assets or revenue in levels – Figure 1 (a) and (b)) exhibit clear negative correlations with failure rates. For small companies, failure rates decline abruptly with company size, flattening out for medium and large firms.

Solvency-related measures (debt to equity, equity ratios – Figure 1 (c) and (d)) exhibit the expected relationship with default, higher indebtedness being associated with higher failure rates.

Current liabilities scaled by total assets (Figure 1 (e)) positively correlate with failure rates. Firms with low proportion of current liabilities exhibit slowly increasing failure rates. Failure rates rise sharply once current liabilities share in total assets exceeds 50 or 60 percent. This pattern suggests that short term liquidity pressures may correlate with firm failure.

Measures linked to profitability and revenue (ROE, revenues scaled by total assets, other operating items scaled by total assets – Figure 1 (f), (g) and (h)) exhibit a more pronounced U shape. Most likely, lower levels indicate poor performance, while higher levels are a result of a lower denominator (total assets or equity), thus explaining the U shape.



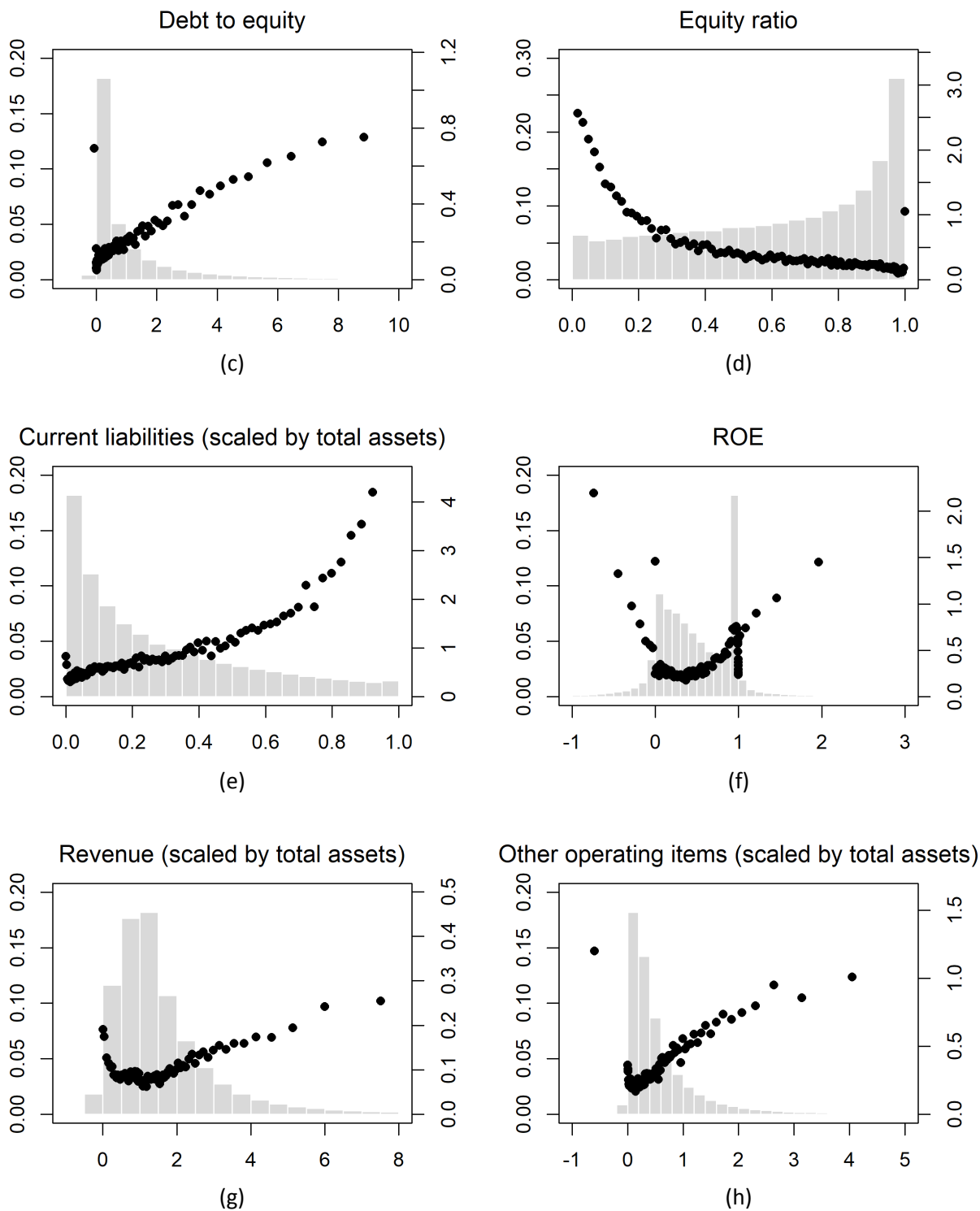


Figure 1. Observed failure rates (left axis) and density of firm distribution (right axis)

Source: Authors' own research.

There are considerable differences between economic sectors, as classified by NACE rev. 2 section. The sectors with highest failure rates are “Other service activities”, “Mining and quarrying” and “Education” with 8.5%, 6.8% and 6.5% failure rate, while the sectors with the lowest failure rates are “Information and communication” and “Financial and insurance activities” with 2.8% and 2.2%.

Overall, these descriptive statistics show the non-linear relationships between firm failure and firm size, industry, profitability, solvency and liquidity, thus motivating the use of machine learning models capable of capturing them.

Hyperparameter tuning

We train models with 70 different hyperparameter configurations. We perform guided grid search, where we iteratively set various hyperparameters as fixed to best performing values and modify others.

Since the XGBoost algorithm is stochastic and since data splitting into k folds is also stochastic, we run each model configuration on five different data splits, resulting in five different models for each configuration. For each configuration, the model’s performance is the average over the five trials.

After tuning hyperparameters, in the subsequent runs we use the following configuration: $\eta = 0.1$, $\text{min.child.weight} = 1$, $\text{subsample} = 0.6$, $\text{max_depth} = 5$, $\text{colsample_bytree} = 0.6$, $\text{colsample_bylevel} = 1$, $\text{colsample_bynode} = 1$, $\text{scale_pos_weight} = 1$, $\text{max_delta_step} = 0$ and $\gamma = 0$.

Tuning hyperparameters early on ensures that subsequent analyses and comparisons are based on a stable and well-performing model.

Model results

The model that includes both financial ratios and financial statements scaled by total assets has an average cross validation performance of 0.8445 and an out-of-time (OOT) performance of 0.8408.

The model is well calibrated for the 2021 training year (k-fold cross-validation predictions used), but to a lesser extent for the OOT 2022 test set (Figure 2). The model fails to capture the increased failure rate observed in 2022 and its computed probabilities underestimate the real values. The predicted failure rate for 2022 is very close to the actual failure rate in the training sample (4.77% compared with 4.56%), significantly below the actual 6.23% rate registered in 2022.

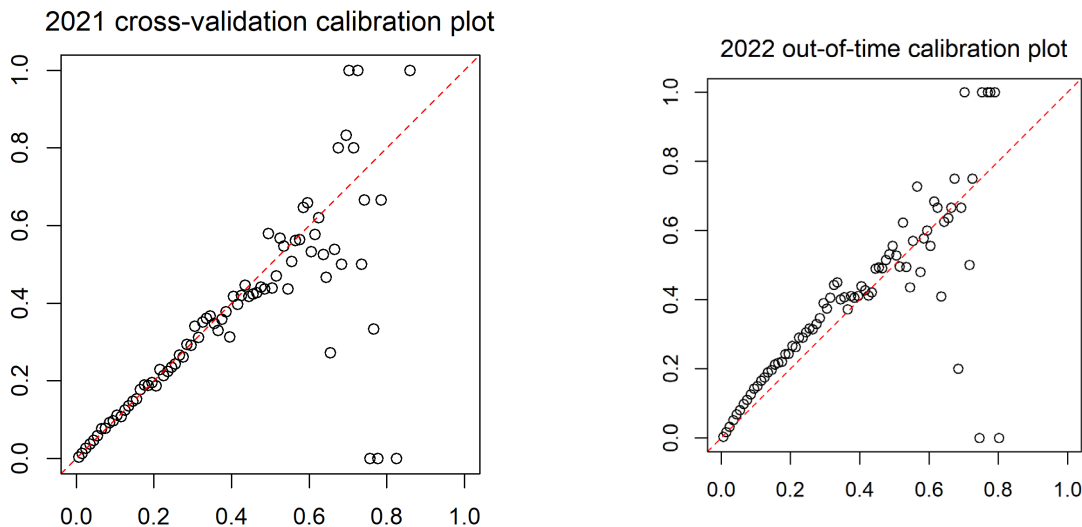


Figure 2. Calibration plots, observed failure rates (Y axis) vs. predicted failure probabilities (X axis)
Source: Authors' own research.

Binned scatter plots for explanatory variables and predicted probabilities of failure (Figure 3) closely mirror the binned scatter plots based on actual failure rates, indicating that the model accurately learns the underlying non-linear relationships between financial information and firm failure.

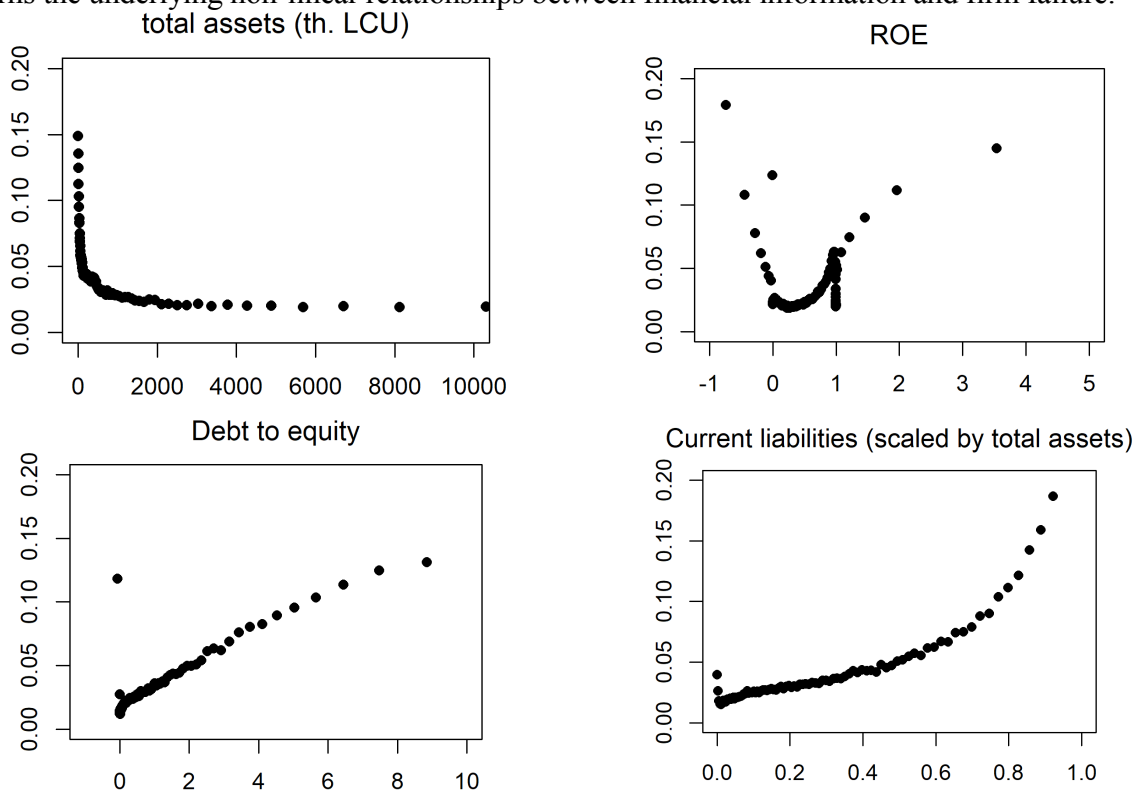
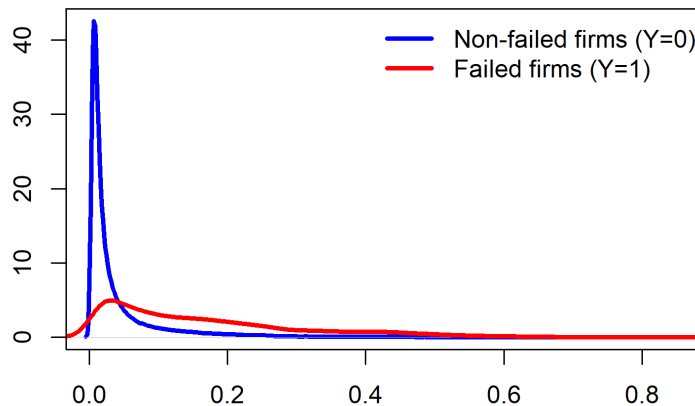


Figure 3. Binned scatter plots showing the relationship between selected explanatory variables (X axis) and predicted probabilities of failure (Y axis)
Source: Authors' own research.

These patterns support the use of machine learning models capable of learning complex, non-linear relationships, as linear models might fail to accurately reproduce them. The result is non-trivial, as the model is not explicitly constructed or trained to match specific patterns for individual explanatory variables.

The distribution of predicted probabilities of failure by outcome (Figure 4) shows that the predicted probabilities for failed firms are shifted to the right and also exhibit a higher variance. However, significant overlap remains between the two distributions, evidence of the inherent uncertainties in firm failure prediction. The average predicted probability for non-failed firms is 4%, below the observed 4.67% observed rate, while the average predicted probability for failed firms is significantly higher, at 16.35%. The standard deviation for predicted probabilities of failed firms is 14.15%, while for non-failed firms it is 6.36%.



**Figure 4. Distribution of predicted probabilities of failure by outcome
(cross-validation sample, 2021)**

Source: Authors' own research.

Explanatory variable importance

Using PFI, we calculate variable importance for the main model (with financial ratios, financial statement items scaled by total assets, NACE and total assets in level – 84 explanatory variables in total). Feature importance quickly decreases, with a few variables dominating, while the remaining have marginal contributions (Figure 5).

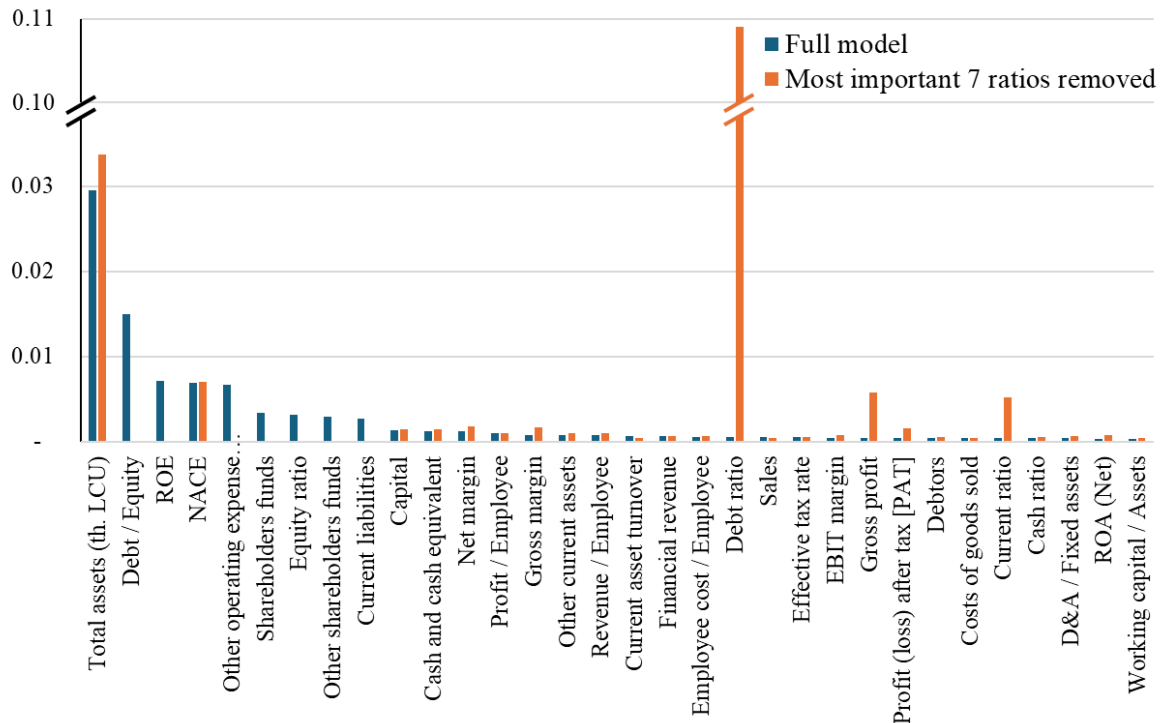


Figure 5. permutation feature importance

Source: Authors' own research.

Note: The y-axis includes a break due to extreme values.

We train each model configuration on 5 different random data splits (for the purpose of k-fold cross-validation) and average the AUROC performance results to mitigate model training stochasticity.

Removing 7 most important features (except Total assets in level and NACE) and retraining the model does not lead to a significant deterioration in performance, as measured by AUROC. The likely explanation stems from the fact that the information in the removed features is present in the remaining features and the model is able to extract it. Indeed, the debt ratio becomes extremely important in the second model, as it probably captures information from the solvency related removed features – debt to equity, shareholder funds scaled to total assets, equity ratio and other shareholder funds scaled by total assets. Similarly, gross profit scaled by total assets is replacing ROE and the current ratio is probably replacing current liabilities scaled by total assets.

We perform a number of experiments by removing variables from the main model (Table 2). Our results show that removing a considerable number of the least important variables does not improve the model's ability to extract information from the remaining, more important, variables. Additionally, our results show that removing such variables may lead to a marginal decrease in performance, however, the magnitude of this decrease is sufficiently small that it may be attributable to stochastic variation in model training. This observation supports the conclusion that the model is not hindered and possibly benefits by the inclusion of more features, even if some may later turn out to be irrelevant or duplicated.

Table 2. Model performance across explanatory variable sets

	Average performance across 5 random trials (on 5 data partitions)	
	2021 cross validation	2022 OOT test
Main model	0.8445	0.8408
Main model minus 7 most important features (other than total assets in level and NACE)	0.8429	0.8402
Main model minus 44 least important features	0.8441	0.8397
Main model minus 27 least important features	0.8442	0.8403
Main model - only 7 most important features + total assets in level and NACE	0.8372	0.8320

Source: Authors' own research.

However, removing the most important 7 features does degrade performance, but only slightly: a 0.0016 deterioration in average AUROC is above random fluctuations due to model training stochasticity, but of modest impact.

On the other hand, keeping only 7 most important variables plus total assets in levels and NACE does degrade performance to an important degree, although not severely.

These empirical observations are important in supporting the idea that variable importance is a diffuse concept: financial ratios have a lot of redundancy and informational overlap and are able to compensate one for another. As a result, it is difficult to ascertain that a certain variable is truly important for failure prediction. As a result, a better way of informing firms on which financial factors are important in failure prediction is needed.

The importance of classes of explanatory variables

A relevant analysis involves grouping explanatory variables into categories, adding their PFI scores and analyzing each group's total importance (Figure 6).

We analyze two groupings. The first one divides financial statement items scaled by total assets according to their general classification from the financial statements (assets, equity, liabilities, revenues, expenses and profits), while the financial ratios are grouped into categories (liquidity, solvency, turnover, profitability, interest coverage and productivity). The second grouping reclassifies most financial statement items scaled with total assets into categories of ratios (profits are reclassified as profitability ratios, equity and non-current debt positions as solvency, current assets and liabilities as liquidity ratios, revenues as turnover ratios, non-current assets as other and expenses are left separately).

Our results show that company size (proxied by total assets in level), followed by solvency and profitability indicators have the highest importance. Grouping variables into economically meaningful categories provides more stable results, although the findings remain general in nature, highlighting the need to complement this model interpretation approach with others capable of providing more in-depth insights.

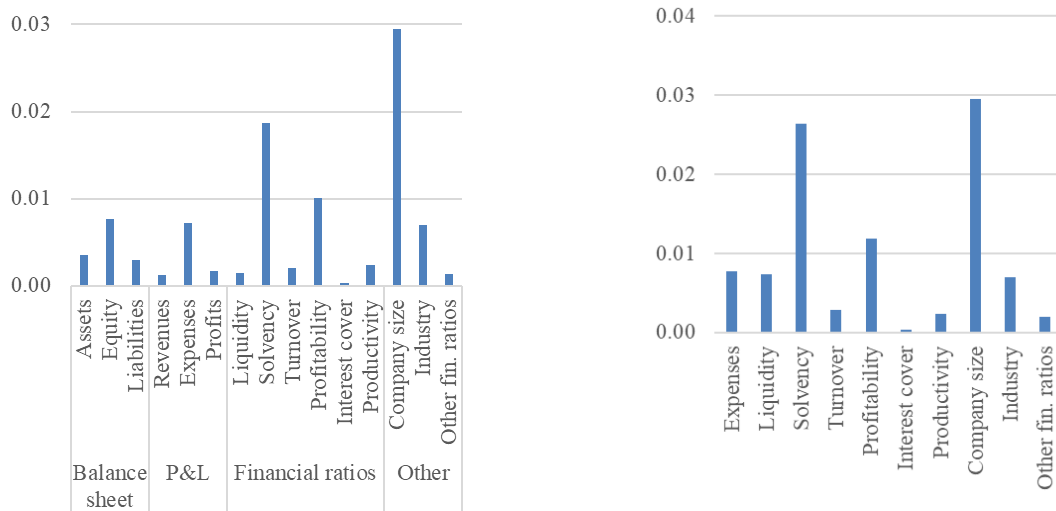


Figure 6. Permutation feature importance by groups of variables

Source: Authors' own research.

Probability of failure sensitivity to explanatory variables changes

A practitioner, having access to such a model and seeing that as little as 7 to 9 variables are important in predicting firm failure might conclude that taking managerial decisions leading to changes in those variables might decrease the firm's probability of failure.

In some cases, such as with solvency ratios, improving one of them (such as debt to equity) would lead to an improvement in most (equity ratio, debt ratio, etc.). With others, such as current liabilities share in total assets and current ratio, the situation is more nuanced: although decreasing current liabilities would also improve the current ratio, the latter can as well be improved by increasing current assets, with no effect on current liabilities as a share of total assets.

We analyze the impact on predicted probabilities of a +/-10% change in important variables and find that it is modest (Table 3). We conclude that a single variable modification is not enough to significantly shift probabilities of failure since the model relies on other variables as well. Moreover, a change in any of the accounting variables would, in reality, change many other financial statement items or ratios.

Table 3. Effects of a +/-10% change in individual explanatory variables on predicted probability of failure

	-10% change	+10% change
Total assets (th. LCU)	4.63%	4.48%
Debt / Equity	4.51%	4.60%
ROE	4.46%	4.70%
Other operating expense (income)	4.50%	4.61%
Shareholders funds	4.61%	4.54%
Equity ratio	4.61%	4.52%
Other shareholders funds	4.62%	4.56%
Current liabilities	4.55%	4.58%

Source: Authors' own research.

The modest impact of individual feature changes suggests that predicted failure probabilities are the result of complex interactions between multiple explanatory variables, limiting the value of single-variable interpretations.

Using scenarios to derive meaningful insights for firms

We construct a relatively simple, but coherent, scenario to illustrate how such a black box model could be used to derive meaningful insights for firm management.

We assume a 10% increase in prices with no corresponding decrease in quantities sold. Of course, such a scenario is not realistic because if a company were able to do so, it would have done it already. Our intention is not to present the scenario itself and derive specific recommendations, but rather to illustrate how such a scenario might be used and with what results.

An increase in sale price with no corresponding decrease in quantities sold or increase in costs other than cost of goods sold would have an impact on many items in the financial statements – from a considerable number of positions in the profit and loss statement, to positions in the balance sheet like inventory, debtors, cash, equity and total assets. All modifications are presented in Appendix.

After adjusting the financial statements, we compute the financial ratios and run the model to obtain predictions for all companies.

Our results show that the average probability of failure decreases significantly from 4.57% to 3.3%, significantly lower than in any 10% modification of individual important variables (Table 3). It should be noted that this scenario does not equate with an inflation rate of 10% because inflation would increase costs as well.

Not all firms are affected equally, and not all firms are affected in the assumed direction (i.e. some companies see their probability of failure increase) – as Table 4 shows. On the one hand, this is a counter-intuitive result, because, at least theoretically, a 10% increase in profitability should always improve a company's position. However, on the other hand, given the underlying non-linearities and complex relationships between variables, a model is expected to exhibit such behavior in a minority of cases. One explanation may be the observed U-shaped relationship between profitability and failure rates.

Table 4 shows that a significant number of firms experience a decrease in their predicted probabilities of failure, especially in the lower part of the distribution, where each percentage change impact a credit institution's decision to grant or deny a loan.

Table 4. Scenario effects: transition matrix of predicted failure probabilities before and after the scenario modification. Rows correspond to base predicted probability bins, while columns correspond to predicted probability bins after the scenario.

Values on the diagonal (in red) indicate unchanged probability bins

	0-1%	1-2%	2-3%	3-4%	4-5%	5-6%	6-7%	7-8%	8-9%	9-10%
0-1%	108,358	8,595	75	8	3	0	0	0	0	0
1-2%	22,547	55,046	4,423	298	59	14	9	1	0	1
2-3%	1,946	18,903	14,435	2,442	366	82	44	17	8	3
3-4%	513	4,558	9,938	5,575	1,432	332	98	47	24	10
4-5%	186	1,621	4,634	5,084	2,712	897	272	94	52	30
5-6%	98	764	1,904	3,259	2,815	1,533	560	216	86	44
6-7%	58	438	1,049	1,870	2,235	1,659	882	360	181	64
7-8%	33	280	557	1,088	1,445	1,593	1,098	555	251	160
8-9%	25	157	375	621	953	1,155	1,050	724	415	223
9-10%	26	102	266	436	633	850	876	790	520	303

Source: Authors' own research.

These results demonstrate that economically coherent scenarios can produce substantial shifts in predicted probabilities of failure, positioning scenario analysis as a tool that offers meaningful interpretations for machine learning models.

Conclusion

Machine Learning models achieve strong predictive performance for firm failure using financial information. Often, explanatory variables include a variety of financial ratios, many with overlapping information and high redundancy. Removing the least important features does not improve model predictive performance and may even marginally reduce it. In this context, variable importance is unstable and may be difficult to interpret, as even the most important variables do not convey unique or irreplaceable information about the drivers of failure. As such, interpretations based solely on variable importance ranking may be misleading.

Interpretations based on cumulative importance for variable groups offer more stable conclusions, although of limited depth. Indeed, we show that firm size, solvency and profitability indicators drive predicted probabilities of failure.

Adjusting individual explanatory variables has a limited impact on predicted probabilities of failure, as they are driven by multiple financial statement items, ratios and interactions. As a result, managers should not focus solely on isolated financial indicators, but rather on consistent business strategies that will have a broad and coherent impact on multiple financial statement items and ratios.

We show that using scenarios and modifying all affected financial statement items according to economic and accounting rules does change predicted probabilities of failure in a meaningful way. Scenario analysis outperforms individual variable-by-variable analysis in terms of interpretability and business insights.

One limitation of our analysis stems from the definition of failure – in the absence of a direct measure in the Orbis database, we use an accounting proxy, a common approach in the literature. Nevertheless, this approach may oversimplify the link between financial statements and firm failure. Indeed, if the model is trained to predict negative equity transition, it may assign a disproportionately high weight to previous year's equity level (and to other solvency and capital structure measures).

As an area of further research, more scenarios with increasing complexities could be analyzed. Additionally, more explanatory variable sets could be analyzed to shed more light on the interplay and tradeoffs between model complexity, performance and interpretability.

Our analysis shows that variable importance rankings provide limited insights for decision makers, while scenario analysis may offer a richer framework for interpreting machine learning models trained on high dimensional financial data.

References

- Alanis, E., Chava, S., & Shah, A. (2023). Benchmarking machine learning models to predict corporate bankruptcy. *Journal of Credit Risk*, 19(2), 77–110.
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589–609.
- Alonso, A., & Carbó, J. M. (2021). Understanding the predictors of corporate bankruptcy: A machine learning approach. *European Central Bank Working Paper Series*

- Barboza, F., Kimura, H., & Altman, E. (2017). "Machine learning models and bankruptcy prediction". Expert Systems with Applications.
- Beaver, W.H. (1966) Financial Ratios as Predictors of Failure, Empirical Research in accounting, Selected Studies. *Journal of Accounting Research*, 4, 71-111. <https://doi.org/10.2307/2490171>
- Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). "In Search of Distress Risk". *Journal of Finance*.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems* (Vol. 35). Curran Associates, Inc.
- Jones, S., Johnstone, D., & Wilson, R. (2017). Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting*, 44(1-2), 3-34.
- Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124-136.
- Lin, J. (2024). Research on loan default prediction based on logistic regression, random forest, XGBoost, and AdaBoost. *SHS Web of Conferences*, 181, 02008.
- Merton, R. C. (1974). On the pricing of corporate debt. *Journal of Finance*, 29(2), 449–470.
- Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109-131.
- Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), 101-124.
- Teodorescu, V., & Obreja Braşoveanu, L. (2025). Assessing the Validity of k-Fold Cross-Validation for Model Selection: Evidence from Bankruptcy Prediction Using Random Forest and XGBoost. *Computation*, 13(5), 127.
- Teodorescu, V., & Toader, C. I. (2024). Using machine learning to model bankruptcy risk in listed companies. In *Proceedings of the International Conference on Economics and Social Sciences* (pp. 610–619). Bucharest University of Economic Studies.
- Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts: Evidence from China. *International Review of Financial Analysis*, 42, 439-446.
- Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 59-82.

Note: This paper was co-financed by The Bucharest University of Economic Studies during the PhD program.

Appendix

Financial statement items retrieved from Orbis and changes to the financial statements for the scenario assuming 10% increase in sales price, no loss of market share and no increase in other costs

			Financial statement items retrieved from Orbis	Scenario changes to the financial statement items
Balance sheet	Assets	1	Non-current assets (2+3+4)	no change
		2	Intangible assets	no change
		3	Tangible fixed assets	no change
		4	Other non-current assets	no change
		5	Current assets (6+7+8)	increase in sales
		6	Inventory	no change
		7	Debtors	+10% of debtors, but no more than the increase in sales
		8	Other current assets	increase in sales - increase in debtors
		9	Of which cash and cash equivalent	increase in sales - increase in debtors
	Equity	10	Shareholders funds (11+12)	+ increase in sales
		11	Capital	no change
		12	Other shareholders funds	+ increase in sales
	Liabilities	13	Non-current liabilities (14+15)	no change
		14	Long term debt	no change
		15	Other non-current liabilities	no change
		16	Of which provisions	no change
		17	Current liabilities (18+19+20)	no change
		18	Loans & short-term debt	no change
		19	Creditors	no change
		20	Other current liabilities	no change
		21	Total shareholders' funds and liabilities (10+13+17)	+ increase in sales
Profit & Loss	22	Operating revenue (Turnover)	+ increase in sales	
	23	Of which sales	+ increase in sales	
	24	Costs of goods sold	no change	
	25	Gross profit (22-14)	+ increase in sales	
	26	Other operating expense (income)	no change	
	27	Operating profit (loss) [EBIT] (25-26)	+ increase in sales	
	28	Financial profit (loss) (29-30)	no change	
	29	Financial revenue	no change	
	30	Financial expenses	no change	
	31	Profit (loss) before tax [PBT] (27+28)	+ increase in sales	
	32	Income tax expenses (benefit)	no change	
	33	Profit (loss) after tax [PAT] (31-32)	+ increase in sales	
	34	Net extraordinary revenues (expenses) (35-36)	no change	
	35	Extraordinary revenues	no change	
	36	Extraordinary expenses	no change	
	37	Profit (loss) for the period [Net income] (33-34)	+ increase in sales	
Memorandum	38	Number of employees	no change	
	39	Working capital	+ increase in sales	

		Financial statement items retrieved from Orbis	Scenario changes to the financial statement items
	40	Enterprise value	no change
	41	Material costs	no change
	42	Costs of employees	no change
	43	Depreciation & Amortization	no change
	44	Other operating items	no change
	45	Interest paid	no change
	46	Research & Development expenses	no change
	47	Cash flow [Net Income before D&A]	+ increase in sales
	48	Added value	+ increase in sales
	49	EBITDA	+ increase in sales
	50	Exports	no change
	51	NACE	no change