

DRIVERS OF DEFAULT RISK IN SWEDISH SMEs

A MACHINE LEARNING APPROACH WITH EXPLAINABLE AI

VILHELM KARLIN

MINNA OLSSON

Bachelor Thesis

Stockholm School of Economics

2025



Drivers of Default Risk in Swedish SMEs: A Machine Learning Approach with Explainable AI

Abstract:

This paper investigates whether gradient-boosted decision trees can offer both predictive performance and economic interpretability in default prediction for Swedish SMEs. Using firm-level accounting data from the Serrano database, we compare LightGBM with logistic regression and apply SHAP and ALE, two explainable AI methods. LightGBM achieves stronger predictive accuracy and calibration than logistic regression. ALE analysis shows that both models generally agree on the direction of effects, but also reveals meaningful non-linearities such as threshold effects, diminishing marginal returns, and non-monotonic patterns that linear specifications struggle to capture. Interaction effects are modest overall but concentrate along the industry dimension, indicating that moderation is predominantly sector-driven rather than arising from complex nonlinear interactions among financial variables. The findings indicate that machine-learning models, when paired with interpretability tools, can uncover economically meaningful non-linearities and interaction patterns in addition to achieving superior predictive performance.

Keywords:

Credit Risk, Default Prediction, Explainable Machine Learning, LightGBM, SHAP, ALE

Authors:

Vilhelm Karlin (26012)
Minna Olsson (26027)

Tutor:

Viktor Thell, PhD, Department of Finance

Examiner:

Ramin Baghai, Professor, Department of Finance

Bachelor Thesis

Bachelor Program in Business & Economics

Stockholm School of Economics

© Vilhelm Karlin and Minna Olsson, 2025

I. Introduction

Predicting corporate default is a central concern in modern finance. Accurate default-prediction models underpin capital requirements under Basel III, inform loan pricing and portfolio allocation, and enable early warning systems for distressed firms. Small improvements in model discrimination translate into meaningful financial gains: loan pricing and regulatory capital are direct functions of estimated probability of default, meaning that prediction errors compound into mispriced risk and inefficient capital allocation. The choice of modelling approach thus carries consequences well beyond academic interest.

This paper investigates whether machine-learning methods combined with explainable AI techniques can advance both the predictive accuracy and economic understanding of corporate default among Swedish SMEs. We construct a LightGBM gradient-boosted model using publicly filed financial statements, corporate registry information, and macroeconomic data benchmarking it against logistic regression and applying SHAP and Accumulated Local Effects methods to interpret the learned relationships. This approach simulates the information set available to external stakeholders, suppliers, customers, and alternative lenders who lack the granular loan-level data possessed by primary banks. Our analysis addresses three research questions: (1) Can we construct a competitive default-prediction model for SMEs using only publicly filed accounting information? (2) What can a gradient-boosted model reveal that is difficult or impossible to learn from logistic regression? (3) Do the model’s predictive features vary across time or different macro-financial regimes? Beyond these empirical questions, the analysis also illustrates how gradient-boosted models paired with explainable AI methods can function as tools for exploratory empirical research, surfacing patterns that inform the specification of traditional regression models.

Our empirical setting comprises approximately 300,000 firm-year observations of Swedish SMEs from the Serrano database spanning 1998–2023. We restrict attention to independent limited-liability companies satisfying European Commission size criteria, excluding financial firms and group entities. The feature set draws on the foundational variables from Altman (1968) and Ohlson (1980), augmented with growth indicators, volatility measures and macroeconomic controls.

The results yield several findings of both practical and methodological interest. First, we demonstrate that publicly available accounting data contain substantial predictive signal for SME default. As reported in Table I, our LightGBM model achieves an AUC of 0.897 compared to 0.869 for logistic regression, a gain of approximately 2.8 percentage points in discriminatory power. The 7% reduction in Log Loss indicates that LightGBM also produces more accurate probability estimates. Crucially, both models are well-calibrated, with expected calibration errors below 0.5%, meaning that predicted probabilities correspond reliably to observed default frequencies (Figure 1).

Second, we use SHAP to examine what each model relies upon. Both models agree on the primacy of *Dividend Payer* and *Cash Ratio* as top predictors, but the distribution of importance differs structurally (Figure 2). Logistic regression concentrates predictive power in the top two features, while LightGBM distributes

importance more evenly across the feature set, extracting signal from variables whose effects are conditional or non-linear. Notably, we find that liquidity buffers and dividend signalling, rather than profitability metrics, emerge as the dominant predictors for Swedish SMEs (Figure 3).

Third, we analyse the functional forms through which features influence default risk. The logistic regression coefficients (Table II) broadly align directionally with the classical specifications of Altman (1968) and Ohlson (1980): working capital, retained earnings, and firm size reduce risk, while leverage increases it. We find *Total Asset Turnover* to be positively correlated with default, unlike Altman (1968), though this is likely a consequence of dataset composition. The ALE analysis reveals that LightGBM captures important non-linearities that linear specifications miss (Figure 9). *Interest Coverage* exhibits diminishing marginal returns beyond approximately three times coverage. *Net Profit Margin* operates almost as a binary signal, with the critical transition occurring between -5% and -18% ; positive margins provide negligible additional protection. *Company Age* displays a striking non-monotonic pattern: newly founded firms carry the *lowest* predicted default risk, risk peaks between one and ten years, then declines thereafter. *Equity Ratio* and *Return on Equity* both exhibit U-shaped relationships, patterns that cause logistic regression coefficients to collapse toward zero as opposing effects cancel.

Fourth, the SHAP interaction analysis reveals that the model’s learned structure is predominantly additive (Figure 5). Pairwise interactions account for approximately 30% of total model output on average (Figure 6), but this interaction budget concentrates heavily along the industry dimension: the predictive signal of financial ratios is sector-dependent, while ratio-on-ratio interactions are sparse. This finding suggests that much of the performance gap between linear and ensemble methods may be recoverable by introducing industry-conditional slopes into linear specifications, a tractable extension that would retain interpretability.

Fifth, we examine whether predictor importance varies across economic regimes. Contrary to prior work documenting pronounced temporal shifts (Cascarino, Moscatelli, and Parlapiano, 2022), we find remarkable stability in feature importance ranks across six periods spanning 2004–2023 (Figure 7). The top-ranked features, *Dividend Payer*, *Cash Ratio*, and *Industry (SNI)* maintain their positions from the pre-crisis environment through the post-COVID rate-tightening period. This stability may reflect either genuine constancy in the fundamental relationships between financial health and default, or methodological dampening from our cumulative training approach.

The remainder of this paper proceeds as follows. Section II reviews related work on default prediction and positions our contribution. Section III describes the data and methodological framework. Section IV presents the empirical findings. Section V concludes.

II. Related Works

The literature on corporate default prediction has evolved through several methodological traditions. We situate our contribution relative to three strands: traditional accounting-based models, market-based structural approaches and recent machine-learning applications.

The foundational work in the accounting tradition is Altman (1968), whose Z-score combined financial ratios into a linear discriminant function for bankruptcy classification. Ohlson (1980) subsequently introduced logistic regression to estimate default probabilities, offering calibrated outputs rather than binary classifications. These approaches remain widely used because coefficients directly quantify each variable’s contribution to risk, facilitating the transparent assessments that regulators require. The core insight, that standard accounting ratios capturing profitability, leverage and liquidity contain substantial predictive signal motivates our feature construction.

An alternative tradition originates with Merton (1974), who recognised that equity can be valued as a call option on firm assets. Jackson and Wood (2013) compare market-based models against accounting-based approaches and find that the former outperform when market prices are observable, though accounting-based models still achieve meaningful discriminatory power. The critical limitation is data availability: structural models require publicly traded equity, which SMEs lack.

The limitations of linear specifications have driven interest in flexible methods that capture non-linear relationships without manual feature engineering. Gradient-boosted decision trees demonstrate strong performance in credit-risk settings (Carmona, Climent, and Momparler, 2019), though opacity conflicts with regulatory requirements for explainable assessments. Methods such as SHAP (Lundberg and Lee, 2017) and Accumulated Local Effects (Apley and Zhu, 2020) address this tension by decomposing predictions into interpretable components. Another related application is Brunlid (2025), whose bachelor thesis applies CatBoost, a gradient boosting model together with SHAP to predict active S&P credit ratings.

The most relevant prior work is Cascarino et al. (2022), who apply XAI techniques to default prediction among Italian non-financial firms over 2009–2020. Combining balance sheet data with credit behavioural indicators from the Italian Credit Register, including credit utilization ratios, overdraft status, and non-performing loan flags they demonstrate that gradient boosting outperforms logistic regression while revealing how predictor importance shifts across the economic cycle. Our work takes a similar methodological approach but differs in setting and data. We focus on Swedish SMEs over a more recent period (2004–2023), covering the post-COVID environment and subsequent monetary tightening that existing studies have not examined. Where Cascarino et al. (2022) supplement balance sheet ratios with credit behavioural indicators from the Italian Credit Register, including utilization ratios and non-performing loan status, our feature set instead emphasizes growth dynamics, revenue trends, drawdowns and earnings volatility alongside a dividend payer indicator that emerges as highly predictive. We extend their analysis in two directions: a more detailed examination of functional forms through ALE analysis, and a systematic decomposition of the interaction effects captured by the gradient boosting model.

III. Data and Methodology

A. Data

Firm-level data are obtained from the Serrano database (Swedish House of Finance, 2025). The Serrano database includes information on firms' income statement and balance sheet, as well as data from Bolagsverket, such as the firms' sector, date of registration and dates of other notable events such as mergers, reconstructions and bankruptcies. To aid comparability, especially over time, Serrano adjusts, corrects, and handles things like broken accounting periods, short and long accounting periods, etc. From this raw accounting material, financial ratios are constructed to capture liquidity, profitability, solvency, efficiency and growth characteristics. To account for macroeconomic conditions, the dataset is augmented with data on lending rates from SCB (Statistics Sweden, 2025). Other macro-economic data from SCB and the Riksbank (Sweden's central bank) on GDP growth, unemployment, and inflation were experimented with, but eventually excluded from the final analysis due to low predictive power.

The initial dataset contains around 12 million firm-year observations (1998–2023). It is then restricted to active Swedish SMEs as defined by the European Commission (European Commission, 2003). Small enterprises have fewer than 50 employees and turnover or total assets below €10 million, while medium-sized enterprises have fewer than 250 employees and turnover below €50 million or assets below €43 million. Micro enterprises have fewer than 10 employees and turnover or total assets below €2 million and are not included in the sample. To ensure cross-firm comparability and avoid intra-group distortions, only independent firms are retained, and financial and real estate firms are also excluded, leaving roughly 300,000 firm-year observations. The final dataset focuses exclusively on active Swedish SME limited-liability companies (AB) and excludes all other legal forms, inactive firms and group entities. In short, only standalone active Swedish SME ABs remain.

Restricting the analysis to a single country is partially down to convenience, but also has benefits in increasing comparability, thereby possibly allowing cleaner inference. Similarly, analysing the full SME population rather than only firms large enough to obtain external credit ratings ensures that the model captures the broad spectrum of financial conditions among Swedish firms. Excluding non-independent firms is also necessary for comparability and generalisation, as group effects otherwise end up dominating the dataset.

The dependent variable, *Distress Event*, equals 1 if a firm enters bankruptcy or reorganisation in a given year, and 0 otherwise. Following standard practice in the credit-risk literature, the model predicts distress one year ahead, using financial and macroeconomic data from year $t - 1$ to forecast outcomes in year t . The dataset is split into an 80% training set and a 20% test set for model evaluation.

After training some initial models, a significant data leakage issue was identified in the data. Firms about to go bankrupt had repeated financial statement information, resulting in year-over-year metrics all being zero. This likely stems from the Serrano dataset's handling of, and adjustment for, fiscal versus calendar years. The issue was rectified by removing the duplicated year, and transferring the target variable to the previous year, thereby restoring the integrity of the predictive framework.

B. Feature Selection Strategy

Extensive feature engineering was conducted to capture the dynamic nature of firm distress. This process began with a core set of financial ratios established in the foundational literature by Altman (1968) and Ohlson (1980), including standard measures of liquidity, profitability and leverage. To augment these static snapshots, we generated features capturing temporal dynamics, such as year-over-year changes, multi-year compound growth rates and volatility measures.

Over 100 variables were considered initially. This extensive search was necessitated by the limitations of the available credit usage data. Cascarino et al. (2022), for instance, utilise data from the Italian Credit Register to construct behavioural variables like drawn-to-granted ratios, which often rank among the most important predictors in banking models. Although the raw source material for the Serrano database contains fields for granted and utilised bank overdrafts, these variables are not retained in the final, processed Serrano dataset. We exclude them here due to significant underlying data quality issues, combined with the practical infeasibility of accurately replicating the complex standardisation procedures required to merge them into the processed sample. To compensate for the exclusion of these granular credit-line metrics, our strategy relied on extracting maximum signal from public financial statements through feature engineering.

The final feature set was constructed through a holistic process combining domain knowledge, multicollinearity diagnostics, stability selection and out-of-sample performance evaluation. While Gradient Boosted Decision Trees (GBDTs) are theoretically robust to uninformative features and capable of handling correlated predictors, the inclusion of irrelevant variables can degrade generalisation performance and obscure interpretation. Furthermore, logistic regressions struggle significantly with multicollinearity. Statistical diagnostics addressed these concerns: we iteratively removed features with extreme Variance Inflation Factors (VIF) and utilised the Boruta algorithm (Kursa and Rudnicki, 2010) to establish a formal significance benchmark. Conflicts between statistical selection and domain theory were resolved by prioritising features with robust economic definitions, for example, ratios with stable denominators (e.g., Total Assets) were preferred over those prone to explosion due to the "small denominator problem" common in SME data.

C. Feature Composition

The final model comprises 22 features, fully detailed in Table IV in Appendix A.

C.1. Traditional Accounting Ratios.

From Altman (1968)'s Z-score, we retain *Working Capital / Total Assets* (X_1), *Retained Earnings / Total Assets* (X_2), and *Total Asset Turnover* (X_5). From Ohlson (1980)'s O-score, we include *Log Total Assets* (SIZE) and *Debt Ratio* (TL/TA). These are augmented with standard ratios: *Cash Ratio* and *Quick Ratio* for liquidity; *Net Profit Margin*, *Gross Margin*, and *Return on Equity* for profitability; and *Equity Ratio* and *Interest Coverage* for leverage. In total, these account for 13 of the 22 features.

C.2. Efficiency Metrics.

Beyond *Total Asset Turnover*, we include *Days Sales Outstanding* and *Days Payables Outstanding* to capture the working capital cycle and cash conversion speed.

C.3. Growth and Volatility Indicators.

Year-over-year revenue growth is augmented by *Revenue CAGR (3Y)* to measure medium-term trajectory. *EBITDA Volatility (3Y)*, defined as the standard deviation of EBITDA scaled by total assets, captures earnings instability. *Revenue Drawdown (5Y)* measures the maximum peak-to-trough decline in net sales over five years, identifying firms that may appear stable currently but carry latent risk from recent shocks.

C.4. Structural and Contextual Controls.

Company Age addresses the “liability of newness” (Stinchcombe, 1965). *Dividend Payer* is a binary indicator serving as a summary statistic for cash-flow capacity. *Industry (SNI)* controls for sector-specific baseline risk. The *Term Spread* captures macroeconomic conditions via the slope of the yield curve.

D. Methodological Framework

A fundamental distinction exists between the objectives of classical econometrics focused on causal modelling and machine learning focused on prediction. This dichotomy was articulated by Shmueli (2010) and is also partially related to the “two cultures” of statistical modelling identified by Breiman (2001). In traditional empirical finance, the primary objective is causal explanation: the isolation of unbiased parameter estimates ($\hat{\beta}$) with the objective of testing hypotheses about causal or structural relationship between covariates and a dependent variable. This paradigm relies very strongly on the underlying assumption that the data are generated from a stochastic process with some known functional form, typically linear plus some error term. Using this framework, model selection is driven by in-sample goodness-of-fit as well as the statistical significance of individual coefficients under these assumptions.

The machine learning paradigm, however, prioritises prediction, i.e. the generalisation error of the model on unseen data ($\hat{y}$). This shift allows for the functional form of the data-generating process to be treated as unknown rather than imposing a more rigid structure on the data. The ML algorithms then approximate the underlying function $f(x)$ by learning directly from the data. This shift in objective from explaining the past to forecasting the future is subtle but very significant, and therefore necessitates a different set of tools and validation protocols.

In the context of corporate default prediction, this shift is not merely academic. The primary objective of financial institutions is to obtain an accurate model for estimating the probability of default of future firms. A model that yields statistically significant coefficients but fails to discriminate between solvent and insolvent firms therefore offers very limited practical value for the actual operator of the model. Superior predictive accuracy directly translates to more efficient capital allocation,

as real-world loan pricing and regulatory capital requirements (e.g., under Basel III (Basel Committee on Banking Supervision, 2011)) are functions of the estimated probability of default (PD). Consequently, a model that reduces prediction error improves the pricing of risk and enhances the stability of the financial system.

Crucially, however, prediction and causal explanation need not be mutually exclusive. As Shmueli (2010) emphasizes, predictive models can inform and refine theory by revealing patterns that simpler specifications miss. The main challenge then becomes methodological, more expressive models are almost by definition harder to interpret, and standard machine learning tools are effectively "black boxes" whose internal mechanisms are near impossible to interpret. This opacity is not only unfortunate from an academic perspective, but also has real world consequences, with both regulators and the operators of requiring (European Banking Authority, 2021) model transparency to be confident in their use. Explainable AI (XAI) techniques address this tension by providing post-hoc methods which decompose predictions into interpretable components. By applying such methods, namely SHAP values and Accumulated Local Effects to a high-performing GBDT model, this paper aims to recover economic insights from a high-performing model without sacrificing predictive accuracy.

E. Model Specifications

E.1. Logistic Regression

To provide a benchmark for performance and interpretability, we employ a standard Logistic Regression model estimated via a Generalized Linear Model (GLM) framework.

To address the statistical heterogeneity typical of firm-level financial data, we apply a preprocessing pipeline specifically for the logistic regression model, following established practice in the default prediction literature (Nyitrai and Virág, 2019). *Interest Coverage*, which tends toward extreme values when interest expense approaches zero, is first clipped to the range $[-5, 20]$ based on domain knowledge. Beyond these thresholds the ratio loses economic meaning, and leaving such values untreated would disproportionately influence coefficient estimates given the linear specification. Remaining predictors exhibiting extreme outliers or heavy tails are then Winsorised at the 1st and 99th percentiles to reduce the influence of extreme observations while retaining sample information. Numeric variables are scaled using a robust scaler based on the median and interquartile range. The *Industry (SNI)* is encoded using Weight of Evidence (WoE) transformation with Laplace smoothing ($\alpha = 0.5$), where each category is replaced by the log-ratio of the within-category event distribution to the non-event distribution (Siddiqi, 2006).

The target value is transformed such that the model estimates the log-odds of default as a linear combination of input features (logistic regression). To control for potential heteroscedasticity, we report HC3 (Heteroscedasticity-Consistent) robust standard errors. The implementation is done in Python using `scikit-learn` for feature transformation and `statsmodels` for GLM estimation.

In industry practice, credit risk modelling with logistic regression is done using Weight of Evidence (WOE) binning. This enables logistic regression to capture

non-linear patterns (Siddiqi, 2006). For this analysis, we avoid doing this deliberately, applying minimal preprocessing to isolate the architectural differences between linear and tree-based models.

E.2. Gradient Boosted Decision Trees

To capture non-linearities and interaction effects without manual specification, we employ Gradient Boosted Decision Trees (GBDT). These models are, unlike logistic regression, invariant to monotonic transformations of input features, meaning that predictors can be left in their raw form to preserve interpretability.

Decision Trees

The fundamental building block of this approach is the classification and regression tree (CART) (Breiman, Friedman, Stone, and Olshen, 1984). CARTs are able to approximate arbitrarily complex relationships by recursively partitioning the feature space into J distinct, non-overlapping regions (leaves) R_j . For an input vector x , the CART T predicts a constant value c_j corresponding to the region into which x falls:

$$T(x; \Theta) = \sum_{j=1}^J c_j \mathbf{1}(x \in R_j)$$

where $\Theta = \{R_j, c_j\}_1^J$ represents the model parameters and $\mathbf{1}$ is the indicator function. These regions are able to be effectively constructed via a greedy algorithm which, at each step, splits the feature set on the single variable x_k at a point s which partitions the data into two half-planes that maximise the reduction in impurity.

Gradient Boosting

While single CARTs are able to approximate arbitrarily complex functions and remain very understandable while doing so, single CARTs are extremely prone to over-fitting, and are therefore not used in practice. Ensemble methods which combine many weak learners (shallow trees) into a single predictive model $F_m(x)$ have, however, been shown to be extremely performant, especially on tabular data. Gradient Boosting (Friedman, 2001), is one of the most popular and powerful ensemble methods, and it works by building the tree sequentially. At each iteration m , a new tree $h_m(x)$ is fitted to the negative gradient of the loss function with respect to the previous ensemble’s prediction $F_{m-1}(x)$. The model updates in a stagewise fashion:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

where η is the learning rate and scales the contribution of each tree. Our implementation minimizes the Logloss (1). This objective can be shown to provide unbiased estimates of the real probabilities. Because the negative gradient of the LogLoss is the residual $(y - p)$, the boosting algorithm effectively fits trees to the errors of the probability estimates at each step. This formulation allows the GBDT to be interpreted as an additive logistic regression model, where the log-odds are constructed as a sum of weak learners rather than a single linear equation.

We choose to implement this algorithm using LightGBM (Ke, Meng, Finley, Wang, Chen, Ma, Ye, and Liu, 2017) in Python. This choice is primarily motivated by the great computational efficiency of this implementation. LightGBM utilizes histogram-based algorithms and a leaf-wise tree growth strategy, which allows for faster training speeds and lower memory consumption on large tabular datasets. Furthermore, LightGBM can natively handle both missing and categorical values.

Hyperparameters are selected using the Optuna framework (Akiba, Sano, Yanase, Ohta, and Koyama, 2019), which employs Bayesian optimisation to efficiently search the parameter space. Because gradient-boosted trees can approximate arbitrarily complex functions, regularisation is essential to ensure the model generalises to unseen data rather than memorising the training sample. Our specification incorporates three forms of regularisation: L1 and L2 penalties on leaf weights, which shrink coefficient magnitudes and encourage sparsity; a maximum tree depth constraint, which limits the complexity of individual trees; and row and column subsampling, which introduces stochasticity by training each tree on a random subset of observations and features, thereby reducing variance.

F. Evaluation Criteria

The distinction between prediction and causal explanation, as well as the weaker assumptions on the underlying data, shapes how models should be evaluated. Standard econometric practice emphasizes in-sample fit and coefficient significance. These metrics are designed to assess whether estimated relationships are statistically robust, and less about evaluating overall model power. Predictive modelling, by contrast, requires metrics that assess overall model performance on unseen data, and which relate to the operational objectives of credit risk management. We therefore evaluate models along three complementary dimensions: discrimination, probabilistic accuracy, and calibration.

ROC-AUC

The Area Under the Receiver Operating Characteristic Curve is designed to measure discriminative ability, or the ability of the model to distinguish high PD and low PD firms. Concretely, it measures the probability that the model assigns a higher risk score to a randomly chosen defaulting firm than to a randomly chosen solvent firm. A value of 0.5 indicates no discriminatory power, while 1.0 represents perfect separation. Credit-scoring models typically achieve values between around 0.75 and around 0.90 (Lessmann, Baesens, Seow, and Thomas, 2015). The downside of ROC-AUC, especially when considering the objective of estimating PD is that it only evaluates rank-ordering, not the accuracy of probability levels. This is a critical limitation, as a model which perfectly separates defaulters from survivors but assigns arbitrary probability values would receive a ROC-AUC of 1.0 while being useless for pricing or capital allocation.

Probabilistic Accuracy: LogLoss and Brier Score

For this model, perhaps the most important metrics are ones which directly evaluate the assigned probabilities of default for the model. Since loan pricing and regulatory capital requirements under Basel III are direct functions of the estimated probability

of default, the accuracy of these estimates are what determine economic outcomes. Both LogLoss, which is the negative log-likelihood, defined as

$$L(y, p) = -[y \log(p) + (1 - y) \log(1 - p)], \quad (1)$$

and the Brier Score (Brier, 1950), which is the mean squared error of probability forecasts and is given by

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

penalize confident but incorrect predictions more heavily than tentative ones. In the above equations, $y_i \in \{0, 1\}$ is the observed outcome and p_i is the predicted probability. LogLoss operates on a logarithmic scale, imposing steeper penalties for overconfident errors. These metrics effectively reward models that are "honest" about uncertainty: a prediction of $p = 0.50$ for genuinely ambiguous cases outperforms arbitrary assignments of $p = 0.10$ or $p = 0.90$, even if both rank firms identically.

Calibration: Expected Calibration Error (ECE)

A well-calibrated model is one where predicted probabilities correspond to empirical default frequencies; among firms assigned a 5% default probability, approximately 5% should default in reality. This property is therefore a great sanity check, and ensures that the model is not systematically incorrect about a certain type of prediction. We assess calibration using the Expected Calibration Error (ECE) (Naeini, Cooper, and Hauskrecht, 2015) with quantile binning. Predictions are partitioned into $M = 100$ equally sized buckets, and ECE is calculated as

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$$

where $\text{acc}(B_m)$ is the observed default rate and $\text{conf}(B_m)$ is the average predicted probability within bucket B_m .

G. Explainable AI Methods

While it is effectively impossible to directly examine the contents of ML models to derive insights about the functionality of the models, post-hoc methods for recovering interpretable insights have recently been developed and gained prominence. We employ two of these methods which together end up being complementary, namely SHAP which provides local explanations by decomposing individual predictions into feature contributions, and ALE which reveals the average functional relationship between each predictor and the target across the entire dataset.

G.1. SHapley Additive exPlanations (SHAP)

A central challenge in interpreting predictive models is attribution: when a model predicts a 15% probability of default, which features drove this prediction, and by how much? For linear models, coefficients provide a direct answer; for non-linear models

with interactions, attribution becomes ambiguous since a feature’s contribution depends on the values of other features.

SHAP resolves this ambiguity using a solution from cooperative game theory (Lundberg and Lee, 2017). Each feature is treated as a ”player” in a coalition game where the payout is the model’s prediction; Shapley values provide the unique fair allocation of this payout among features. Concretely, the Shapley value for a feature is computed by considering all possible subsets of features and measuring how much the prediction changes when that feature is added to each subset. This marginal contribution is then averaged across all subsets, weighted by the number of ways each subset can form. For feature j , the Shapley value is:

$$\phi_j(v) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{j\}) - f_x(S)] \quad (2)$$

where F is the set of all features, S is some subset excluding feature j , and $f_x(S)$ denotes the model’s prediction using only features in S . The SHAP value ϕ_j can thus be interpreted as the average marginal contribution of feature j to the prediction, controlling for all possible interactions with other features.

In econometric terms, SHAP values represent the local marginal effect of each variable on log-odds for the specific observation. To construct global importance measures, SHAP values can be aggregated; the mean absolute SHAP value indicates a feature’s average impact across the dataset.

We implement SHAP using the `shap` package in Python. For our LightGBM model, the TreeSHAP algorithm (Lundberg, Erion, Chen, DeGrave, Prutkin, Nair, Katz, Himmelfarb, Bansal, and Lee, 2020) exploits the hierarchical structure of decision trees to compute exact Shapley values efficiently. For our logistic regression, the additive model structure yields a closed-form solution:

$$\phi_j = \beta_j(x_j - \mathbb{E}[x_j])$$

where each feature’s contribution is simply its coefficient-weighted deviation from its mean. Both approaches circumvent the exponential $2^{|F|}$ subset enumeration that directly evaluating (2) would otherwise require.

SHAP Interaction Values

Machine learning models often also rely on complex interaction effects between variables. SHAP is also a powerful tool for computing the importance of pairwise interaction terms. For any two features j and k , the SHAP interaction value $\phi_{j,k}$ captures the portion of the prediction attributable to their joint effect, that is, the effect which cannot be separated into the sum of their individual contributions. This value is computed by distributing the Shapley value of feature j across its interactions with other features. A non-zero value of $\phi_{j,k}$ indicates that the effect of feature j depends on the value of feature k .

G.2. Accumulated Local Effects (ALE)

While SHAP explains features’ contributions to individual predictions, it is often of use to obtain a global view of how the model’s output changes, on average, as a given input feature varies across its range. Partial Dependence Plots (PDP) are a common tool for this purpose, but they suffer from a critical flaw when predictors are correlated (as financial ratios typically are). The PDP for variable j is computed as the average prediction across all the observations while varying variable j across its entire range. This can easily result in unrealistic combinations of features (e.g., a firm with very low assets but very high revenue), meaning that the model will not have been trained on such inputs. This extrapolation produces misleading estimates of feature effects.

Accumulated Local Effects (Apley and Zhu, 2020) is a method developed specifically to solve this problem. Instead of averaging across the entire dataset, ALE considers only observations that naturally occur near each feature value. The method works by (1) partitioning the feature’s range into small intervals; (2) for observations in a given interval, calculate the local effect by taking the difference between the prediction at the interval’s upper and lower boundaries, and then averaging these differences across all observations within the interval; (3) accumulate these local effects from the lowest feature value upward. Formally, the ALE function is defined as

$$\hat{f}_{ALE}(x) = \int_{z_{0,j}}^x \mathbb{E} \left[\frac{\partial f}{\partial x_j}(X_j, X_{-j}) \mid X_j = z_j \right] dz_j - C$$

where the expectation over (X_j, X_{-j}) is taken conditioned on $X_j = z_j$, meaning we average over values of the other features that are consistent with $X_j = z_j$ according to the joint distribution of the data. C is a constant which centres the plot such that the average effect is zero. This conditional averaging ensures that the ALE curve reflects the true local effect of a variation of the feature across the actual data distribution, rather than in extrapolated regions.

ALE is implemented using the `alibi` package in Python.

IV. Results

In this section, we present the empirical findings from running our models and applying explainable AI. The analysis proceeds in five stages. First, we compare the actual predictive performance of our LightGBM model to our logistic regression model to quantify how much additional performance the increased expressiveness of GBDT’s yield. Second, we examine global feature importance for both models via SHAP to understand what each model relies upon to make its predictions. Third, we analyse the functional forms of the features in the models to reveal how features influence default risk. Fourth, we examine the feature interactions in the LightGBM model to try to understand what portion of its advantage can be attributed to capturing feature dependencies which have not been added to the logistic regression model. Lastly, we look at how feature SHAP importance varies over time for LightGBM to see if the model relies on different features during different regimes.

A. Model Performance Comparison

In Table I, out-of-sample performance metrics for both models are presented. LightGBM achieves an AUC of 0.897 compared to 0.869 for logistic regression, representing an improvement of approximately 2.8 percentage points. The 7% reduction in Log Loss likewise shows that LightGBM produces more accurate probability estimates. Importantly, both models are well calibrated with ECEs below 0.5%, indicating that the gains in discrimination and probabilistic accuracy do not come at the expense of reliability.

These improvements are meaningful but not groundbreaking, both models perform very well. The magnitude of gains is consistent with benchmarking studies comparing gradient boosting methods to logistic regression (Lessmann et al., 2015). Figure 1 visualizes these results: Panel A shows that LightGBM consistently achieves a higher true positive rate across false positive rate levels, while Panel B confirms that both models are well calibrated, with LightGBM exhibiting slightly smaller deviations from perfect calibration.

Table I
Model performance comparison

Out-of-sample performance on a 20% held-out test set ($N = 60,830$). ROC-AUC measures discriminatory power (higher is better); Log Loss and Brier Score measure probabilistic accuracy (lower is better); ECE (Expected Calibration Error) measures alignment between predicted and observed default frequencies using 100 quantile bins (lower is better). The final row reports model differences.

Model	ROC-AUC	Log Loss	Brier Score	ECE
LightGBM	0.8972	0.0660	0.0159	0.0032
Logistic Regression	0.8685	0.0711	0.0165	0.0041
Δ (LightGBM – Logistic)	+0.0287	−0.0051	−0.0006	−0.0008

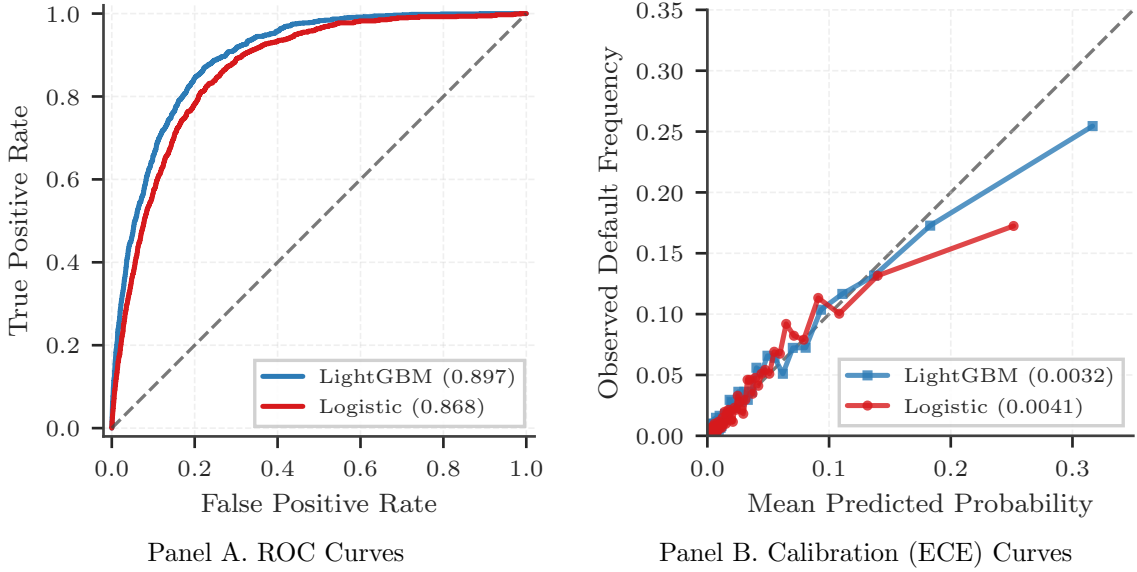


Figure 1: Model performance: discrimination and calibration. This figure compares out-of-sample performance for the LightGBM model and the Logistic Regression baseline. Panel A displays Receiver Operating Characteristic curves, plotting the true positive rate against the false positive rate across classification thresholds; the dashed diagonal represents a no-skill classifier, and values in parentheses report the Area Under the Curve. Panel B displays calibration curves, plotting observed default frequencies against mean predicted probabilities across 100 quantile bins; the dashed diagonal represents perfect calibration, and values in parentheses report Expected Calibration Error. Both panels are evaluated on the 20% held-out test set ($N = 60,830$).

B. Global Feature Importance

Figure 2 contrasts the mean absolute SHAP values of the LightGBM model against the logistic baseline. While both models agree on the primacy of the *Dividend Payer* and *Cash Ratio* variables, the distribution of importance reveals a structural divergence. The logistic model concentrates the vast majority of its predictive power in the top two features, whereas LightGBM exhibits a flatter importance profile, distributing weight more evenly across the feature set.

This difference is a consequence of model architecture. The logistic regression is constrained to additive, global coefficients, forcing it to rely exclusively on features with strong, monotonic linear relationships to the target. In contrast, LightGBM is not bound by this linearity constraint. It can extract value from features that exhibit complex, non-monotonic patterns which a linear model might discard as noise. Crucially, the tree-based structure allows the model to capture high-order interactions, assigning significant importance to variables that may possess little predictive power in isolation but become highly relevant when considered in concert with other features.

Comparing our findings to the existing literature highlights the importance of the specific economic context. While Jabeur, Gharib, Mefteh-Wali, and Ben Arfi (2021) employed CatBoost on French corporate data, they found profitability metrics, specifically return on shareholder funds and return on capital employed to be the dominant predictors. Conversely, our analysis of Swedish SMEs indicates that liquidity buffers and dividend signaling are the primary determinants of default risk.

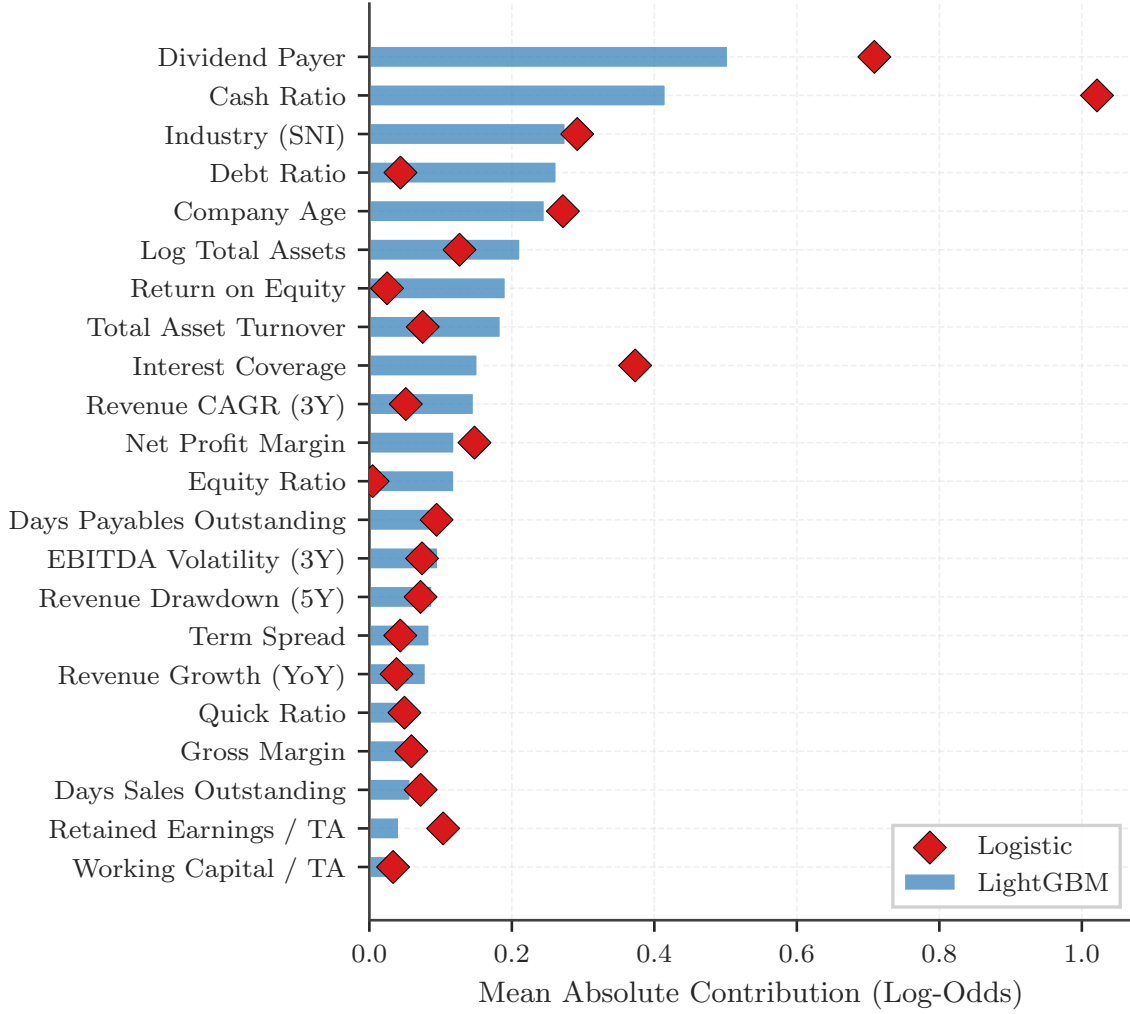
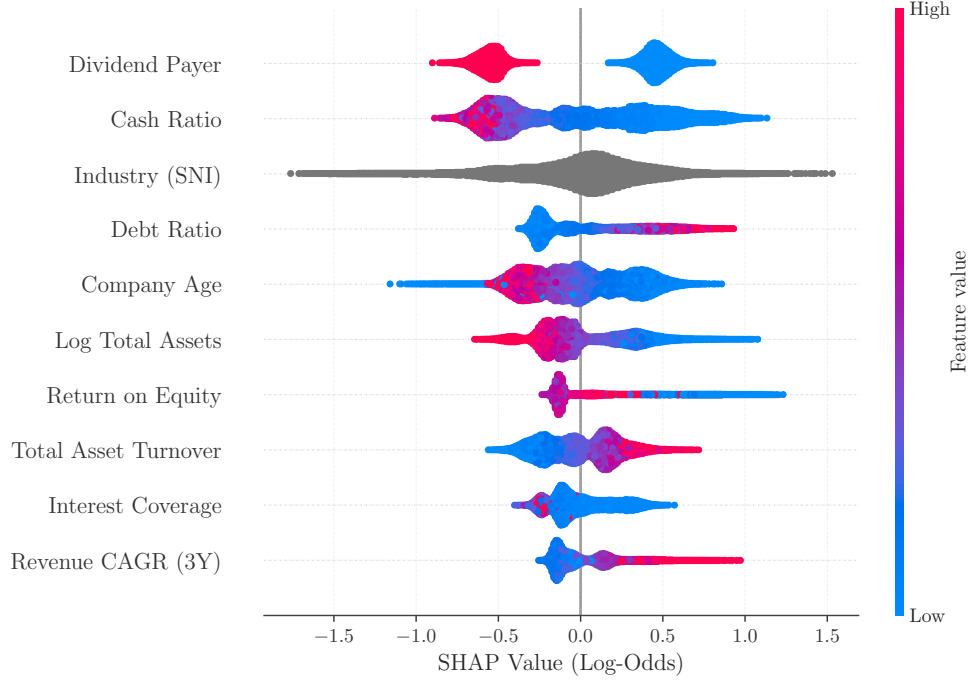


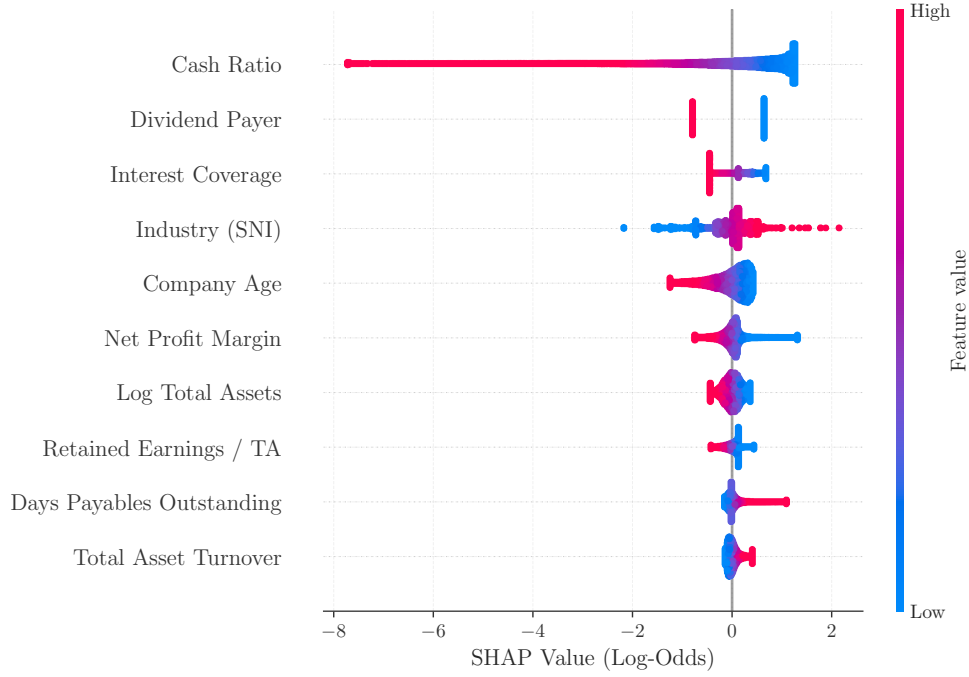
Figure 2: SHAP feature importance comparison. This figure displays mean absolute SHAP values for the LightGBM model and the Logistic Regression baseline. SHAP values decompose each prediction into additive feature contributions on the log-odds scale; the mean absolute value across observations measures a feature’s average impact on predictions, capturing importance magnitude regardless of direction. Features are ordered by their importance ranking in the LightGBM model. SHAP values are computed using TreeExplainer for LightGBM and LinearExplainer with correlation-dependent feature perturbation for Logistic Regression. Both models are evaluated on the held-out test set ($N = 60,830$).

This suggests that for Swedish SMEs, immediate solvency and cash flow resilience (liquidity) may outweigh profitability as warning signals.

To analyse the direction and heterogeneity of these effects, Figure 3 presents SHAP value distribution plots where each dot represents a single firm-year observation. Panel B displays the strict monotonicity of a linear model. *Cash Ratio* shows a perfect colour gradient, high values (red) consistently reduce risk and *Dividend Payer* forms two narrow vertical clusters. This is a consequence of the logistic regression’s fixed, global coefficients, where a feature’s marginal effect remains constant regardless of other firm characteristics. In contrast, Panel A (LightGBM) exhibits significant horizontal dispersion and overlapping feature values. The spread of the red *Dividend Payer* cluster implies that the protective effect of dividends is not fixed, but highly



Panel A. LightGBM: SHAP value distribution



Panel B. Logistic Regression: SHAP value distribution

Figure 3: How features contribute to default predictions. This figure displays the distribution of SHAP values for the top 10 features in each model. Each point represents a firm-year observation. Horizontal position indicates the feature's contribution to predicted log-odds of default (positive values push predictions toward default); colour indicates the feature's value (blue for low, red for high). For a feature with a consistent positive relationship with default risk, high values (red) appear on the right. Features are ordered by mean absolute SHAP value. SHAP values are computed using TreeExplainer for LightGBM and LinearExplainer with interventional feature perturbation for Logistic Regression. Evaluated on the held-out test set ($N = 60,830$).

context-dependent. Similarly, the mixed distribution of *Company Age* suggests that "youth" contributes to risk only under specific conditions. This absence of clear gradients signals that the model leverages either non-monotonic relationships or complex variable interactions, mechanisms we distinguish and explore in detail in the subsequent sections.

C. Functional Forms

While the primary focus of this paper is the LightGBM model, the logistic regression offers a unique advantage: its coefficients are directly comparable to the classical bankruptcy prediction literature. Table II reports coefficient estimates with heteroscedasticity-consistent standard errors, while Figure 4 presents Accumulated Local Effects (ALE) curves for four features exhibiting particularly instructive patterns. Full ALE curves for all features are provided in Figure 9.

C.1. Comparison with Classical Models

The signs on our core variables align with both Altman (1968) and Ohlson (1980). Consistent with Altman’s X_1 and X_2 , we find that *Working Capital/TA* and *Retained Earnings/TA* carry negative coefficients, indicating that liquidity and accumulated earnings reduce default risk. Similarly, *Log Total Assets* and *Debt Ratio* align with Ohlson’s SIZE and TL/TA variables.

However, coefficient magnitudes differ substantially from the classical specifications. *Debt Ratio* carries a coefficient of just 0.17 versus Ohlson’s 6.03, and *Working Capital/TA* is only marginally significant ($p = 0.08$) with a coefficient of -0.14 compared to Ohlson’s -1.43 . These attenuations likely reflect our richer feature set: *Interest Coverage* and *Cash Ratio* absorb solvency signals, while *Days Sales Outstanding* and *Days Payables Outstanding* decompose working capital into its constituent timing dynamics. In contrast, *Log Total Assets* (-0.53) aligns closely with Ohlson’s SIZE coefficient (-0.41), suggesting the firm size effect is robust and not absorbed by other variables.

Several coefficients carry signs that appear counterintuitive relative to classical specifications but reflect the distinct economics of SME populations. *Total Asset Turnover* is positive, contrary to Altman’s original model where higher turnover indicated operational efficiency and reduced default risk. This discrepancy likely reflects differences in sample composition: Altman’s model was estimated on large U.S. manufacturing firms, where high asset turnover signalled productive capital use, whereas among SMEs it may instead indicate asset-light operations with limited buffers or unsustainable sales generation relative to the asset base. Similarly, both *Revenue Growth (YoY)* and *Revenue CAGR (3Y)* associate with higher default probability. Though variables absent from the classical models but consistent with the “overtrading” phenomenon whereby rapidly growing firms exhaust working capital faster than they generate cash (Delmar, McKelvie, and Wennberg, 2013).

Table II
Logistic Regression Coefficient Estimates

This table reports coefficient estimates from the logistic regression model predicting one-year-ahead default. The dependent variable equals one if a firm enters bankruptcy or reorganisation in the following year. Coefficients represent the change in log-odds of default for a one-unit change in the standardised predictor. Interest coverage is clipped to $[-5, 20]$; remaining continuous features are Winsorised at the 1st and 99th percentiles and standardised using the median and interquartile range. Industry (SNI) is encoded using Weight of Evidence transformation with Laplace smoothing. Robust standard errors (HC3) are reported to account for heteroscedasticity. The model is estimated on the 80% training sample ($N = 243,319$). Features are ordered by statistical significance. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Feature	Coefficient	Robust SE	z-value	p-value	
Intercept	-4.4164	0.0353	-124.97	0.0000	***
Dividend Payer	-1.4325	0.0721	-19.88	0.0000	***
Industry (SNI)	0.7243	0.0373	19.40	0.0000	***
Interest Coverage	-1.1295	0.0803	-14.07	0.0000	***
Days Payables Outstanding	0.3929	0.0290	13.55	0.0000	***
Company Age	-1.1344	0.0880	-12.89	0.0000	***
Net Profit Margin	-0.6536	0.0519	-12.60	0.0000	***
Cash Ratio	-3.6419	0.2951	-12.34	0.0000	***
EBITDA Volatility (3Y)	0.3208	0.0359	8.93	0.0000	***
Revenue Drawdown (5Y)	-0.3139	0.0402	-7.81	0.0000	***
Retained Earnings / TA	-0.3850	0.0509	-7.56	0.0000	***
Log Total Assets	-0.5260	0.0727	-7.24	0.0000	***
Revenue CAGR (3Y)	0.1831	0.0256	7.15	0.0000	***
Revenue Growth (YoY)	0.1223	0.0182	6.73	0.0000	***
Days Sales Outstanding	-0.2811	0.0426	-6.59	0.0000	***
Debt Ratio	0.1726	0.0271	6.38	0.0000	***
Total Asset Turnover	0.3193	0.0589	5.42	0.0000	***
Term Spread	-0.2732	0.0670	-4.08	0.0000	***
Return on Equity	0.1061	0.0274	3.87	0.0001	***
Gross Margin	-0.2524	0.0667	-3.78	0.0002	***
Working Capital / TA	-0.1376	0.0792	-1.74	0.0823	*
Quick Ratio	0.1957	0.1537	1.27	0.2030	
Equity Ratio	-0.0182	0.1094	-0.17	0.8680	

C.2. Beyond Linear Specifications

The ALE curves reveal that LightGBM largely corroborates the directional relationships identified in the logistic regression, but captures important non-linearities that help explain several coefficient patterns. For *Debt Ratio*, LightGBM exhibits a substantially steeper slope in the low-to-moderate leverage range before plateauing around a ratio of 10 (Figure 9); this pattern explains the attenuated logistic coefficient, as the linear specification averages over a steep initial region and a flat tail, producing a coefficient that understates the risk gradient where most observations lie. *Total Asset Turnover* displays similar behaviour, with a steeper initial slope followed by a plateau around a ratio of 4. For *Working Capital/TA*, the models agree over the economically meaningful range, though LightGBM identifies a sharp discontinuity around -0.30 , though this may be an artifact of sparse data rather than a genuine threshold. *Retained Earnings/TA* presents a starker divergence: while the logistic coefficient (-0.39) aligns with Altman’s X_2 , LightGBM exhibits step-change behaviour around zero, with negative retained earnings actually associated with lower default risk than modestly positive values.

A distinct class of non-linearity appears in *Equity Ratio* and *Return on Equity*, both of which exhibit U-shaped relationships with elevated default risk at extreme values (Figure 9). For Equity Ratio, this non-monotonicity explains its statistical insignificance in the logistic regression ($\beta = -0.02$, $p = 0.87$): forced to fit a single slope through a U-shaped pattern, the linear model estimates a coefficient near zero as positive and negative effects cancel. Return on Equity displays a similar pattern, with elevated risk at both very negative values (indicating distress) and very high values (above approximately 0.7). The elevated risk at high ROE may reflect volatile or highly leveraged business models, though it could also stem from dataset composition: firms with very high ROE often have small equity bases, making them disproportionately young or thinly capitalised. These U-shaped relationships represent a class of non-linearity that tree-based methods can exploit but that require explicit polynomial or spline terms in parametric specifications.

Exhaustive treatment of every feature, while admirable in principle, exceeds our present ambitions; Figure 4 highlights four with particularly instructive non-linear patterns.

Interest Coverage.

Panel A shows that default probability decreases steeply as interest coverage rises from negative values toward approximately 3, then flattens considerably for higher values. This diminishing marginal effect broadly validates findings from Blume, Lim, and MacKinlay (1998), who, in their analysis of corporate credit ratings, documented that interest coverage improvements from 0 to 5 strongly improved ratings while increases beyond 10–20 provided negligible incremental benefit. Although their study examined credit ratings rather than default prediction directly, the convergence of findings across outcome measures and institutional contexts suggests this non-linearity reflects a genuine feature of how debt-servicing capacity relates to firm risk. The logistic regression, constrained to a linear specification, captures the average gradient but cannot represent this threshold structure.

Net Profit Margin.

Panel B reveals that LightGBM exhibits near-binary behaviour with respect to profitability. The transition occurs between approximately -5% and -18% margin, where default probability rises steeply; beyond -18% , further deterioration adds little incremental signal, while positive margins have essentially no marginal effect. The logistic regression, by contrast, naturally imposes a linear slope.

This threshold behaviour connects to a modelling distinction between Altman (1968) and Ohlson (1980). Altman treated profitability ratios as continuous predictors where every marginal decrease corresponds to a proportional increase in default risk. Ohlson, by contrast, augmented standard ratios with a binary indicator for negative net income two years in a row (INTWO), explicitly recognising that the transition into loss-making represents a qualitatively distinct regime. LightGBM appears to have organically rediscovered this insight, treating net profit margin as a regime indicator where the sign of profitability contains the primary signal.

Interestingly, *Gross Margin* exhibits a more linear relationship across most of its range (Figure 9), though LightGBM identifies a step change around 10%. This linearity may reflect that gross margin captures more fundamental firm economics: pricing power, operational efficiency and differentiated products, where each incremental improvement provides additional buffer against adverse shocks. Net margin, by contrast, determines ultimate viability; crossing zero represents a regime change that gross margin does not.

Days Payables Outstanding.

Panel C displays a non-linear relationship rarely examined in traditional default-prediction studies. LightGBM reveals a muted effect up to approximately 40 days, with even a slight uptick in default probability for firms paying very rapidly (< 20 days). Between 40 and 80 days, the effect rises sharply.

This pattern resonates with established findings on trade credit. The slight risk elevation for rapid payers, if genuine, aligns with Deloof (2003), who characterises accounts payable as flexible, inexpensive financing. Underutilising this credit window may indicate suboptimal liquidity management. The muted risk profile up to 40 days is consistent with Jacobson and von Schedvin (2015), who document average trade credit durations of approximately 31 days in Sweden, suggesting that payment behaviour within this window reflects normal operations rather than distress. The sharp increase beyond 40 days corroborates Deloof (2003)’s finding that less profitable firms systematically delay payments, making abnormal DPO a proxy for liquidity constraints.

Company Age.

Panel D shows that LightGBM identifies firms between approximately 1 and 10 years old as carrying elevated default risk, consistent with the liability of newness documented in organisational ecology (Stinchcombe, 1965). Risk then decreases with diminishing returns, stabilising for firms beyond roughly 30 years, reflecting survivorship bias, whereby persistent firms are systematically healthier (Brown, Goetzmann, Ibbotson, and Ross, 1992).

Strikingly, firms at year zero exhibit by far the lowest predicted default probability. This likely reflects that newly founded firms possess starting capital and have not yet

accumulated the operational difficulties or debt burdens that precipitate default. The logistic regression captures the general negative relationship but imposes a strictly linear decay, missing both the year-zero effect and the plateau at maturity. This pattern illustrates how insights from flexible models can inform simpler specifications: a logistic regression augmented with a binary indicator for newly founded firms (age < 1 year) could capture this non-linearity while retaining full interpretability.

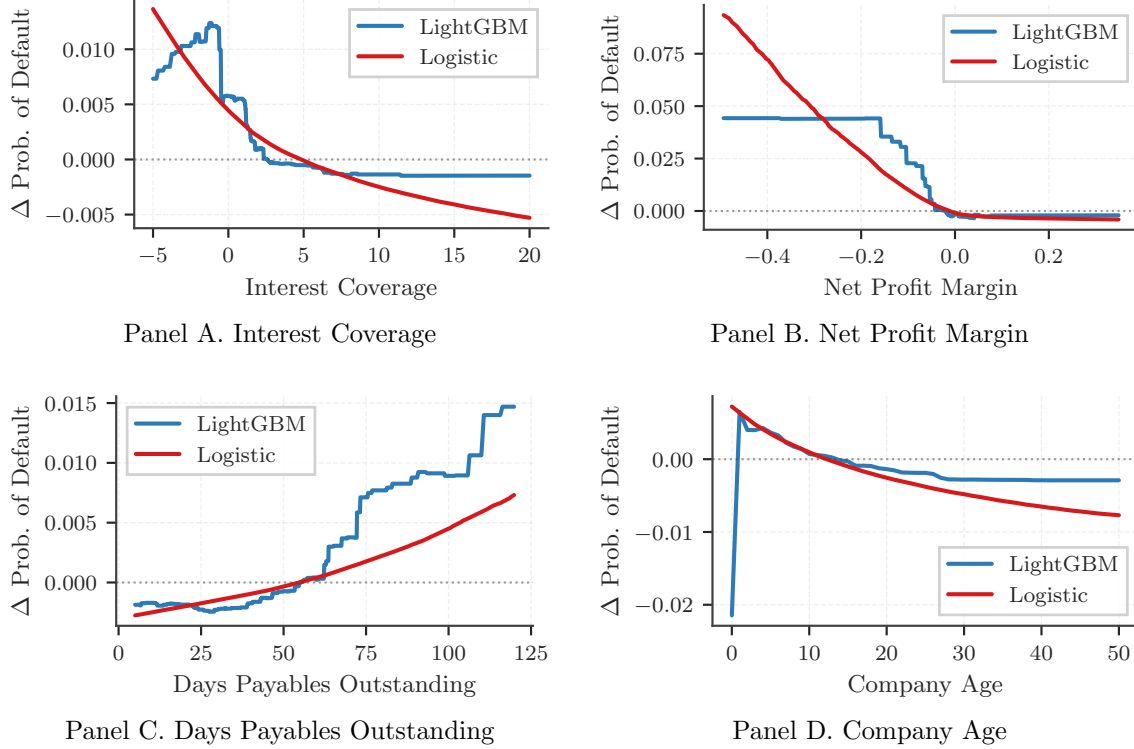


Figure 4: Non-linear effects of selected features on default probability. This figure plots the Accumulated Local Effects (ALE) of four key predictors on the predicted probability of default: (A) interest coverage, (B) net profit margin, (C) days payables outstanding, and (D) company age. Panels A and B display unitless ratios; Panel C is measured in days; Panel D in years. Results are estimated using a LightGBM gradient boosting model and a logistic regression baseline. The y-axis displays the ALE value, representing the change in predicted default probability relative to the sample average. The horizontal dotted line indicates zero effect. Evaluated on the held-out test set ($N = 60,830$).

D. Interaction Effects

While the preceding analysis demonstrated LightGBM’s ability to model non-linear main effects, a key theoretical advantage of gradient boosting is its capacity to capture feature interactions without explicit specification. Figure 5 reveals a striking sparsity of strong interactions in our model. Even the strongest pairwise effect, the interaction between *Industry SNI* and *Revenue Growth (YoY)* registers only 0.040 log-odds units in mean absolute SHAP value, representing at most a 1% change in the predicted probability, indicating that individual interactions exert relatively minor individual effects on average.

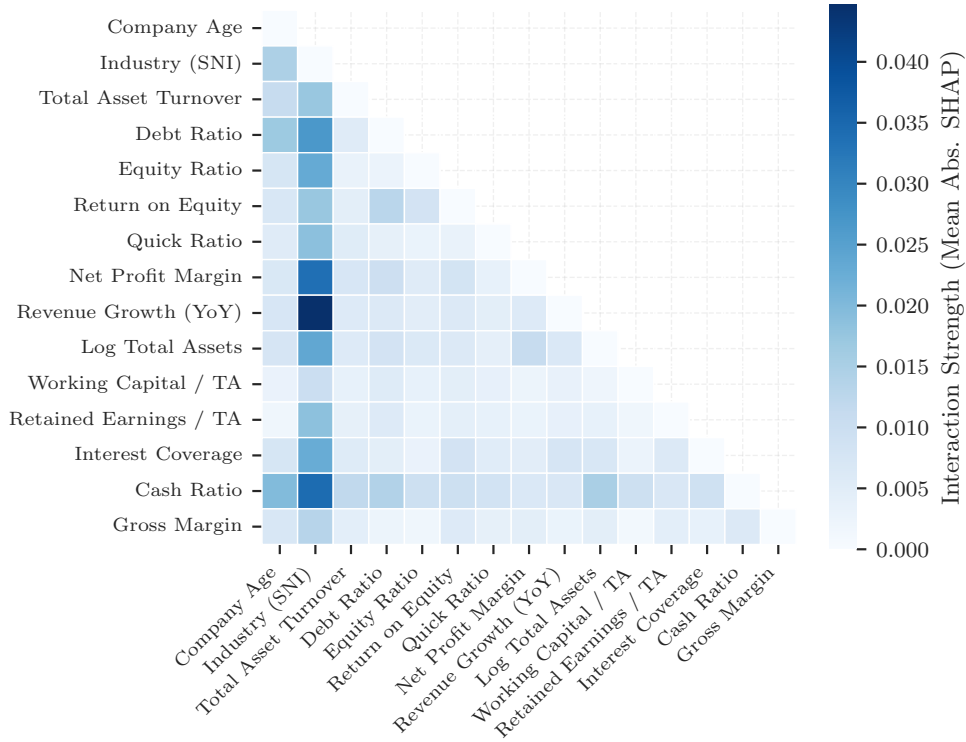


Figure 5: Pairwise feature interactions in default prediction. This figure displays mean absolute SHAP interaction values for the top 15 features in the LightGBM model. The matrix shows pairwise interaction effects (off-diagonal elements); a non-zero interaction value for features j and k indicates that the effect of feature j on predictions depends on the value of feature k . Cell colour indicates interaction strength, with darker shading representing stronger interactions. The matrix is symmetric; only the lower triangle is displayed. Evaluated on the held-out test set ($N=60,830$).

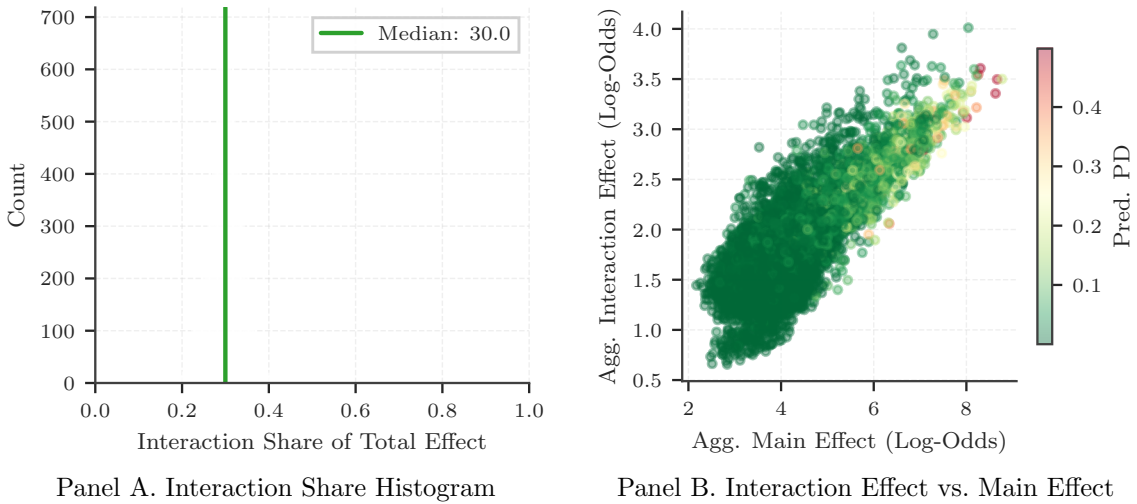


Figure 6: Collective importance of feature interactions. This figure examines the aggregate contribution of interaction effects relative to main effects in the LightGBM model. Panel A displays the distribution of interaction-to-total-effect ratios across firm-year observations, computed as the sum of absolute pairwise interaction SHAP values divided by the sum of absolute total SHAP values for each observation; the vertical line indicates the median. Panel B plots aggregate interaction effects against aggregate main effects for each observation, with point colour indicating the model's predicted PD. Evaluated on the held-out test set ($N = 60,830$).

Despite this general sparsity, *Industry SNI* emerges as a clear exception. The heatmap reveals a discernible band of elevated interaction values along the industry dimension, with the strongest effects involving *Revenue Growth (YoY)*, *Net Profit Margin* and *Cash Ratio*. This pattern indicates that the predictive signal of these accounting metrics is conditional on the firm’s operating sector. This aligns with evidence from bankruptcy hazard models, where Chava and Jarrow (2004) demonstrate that industry groupings affect not only the intercept (baseline hazard) but also the slope coefficients of financial covariates, a finding consistent with the accounting intuition that financial ratios should be interpreted relative to industry norms. For instance, liquidity constraints that signal distress in retail may reflect normal operating cycles in construction. LightGBM captures these sector-specific sensitivities automatically, effectively learning that the relationship between firm characteristics and default risk is moderated by industry context.

While individual interactions are small, their aggregate contribution is non-trivial as illustrated in Figure 6, interactions typically account for 20% to 40% of the total model output (median: 30.0%). Importantly, however, this interaction budget is not diffusely allocated. The concentration along the industry dimension indicates that the non-additive structure in our model is sparse and interpretable, primarily reflecting the sector-specific sensitivities discussed above rather than complex interdependencies among financial ratios.

Our results diverge notably from Ben Jabeur, Gharib, Mefteh-Wali, and Ben Arfi (2021), who document strong interactions between financial ratios, specifically between return on shareholder funds and return on capital employed in French corporate data. This divergence may reflect differences in feature specification. Ben Jabeur et al. (2021) rely exclusively on financial ratios, explicitly omitting industry indicators. Without industry features, ratio-on-ratio interactions may emerge as proxies for sector-specific capital structures. By explicitly modelling industry, our specification absorbs sectoral heterogeneity directly, reducing both the structural need for and the appearance of complex ratio-on-ratio interactions.

D.1. Implications for Model Choice.

The sparse, structured interaction pattern we observe has encouraging implications for the interpretability-accuracy trade-off in credit risk modelling for SMEs. We find LightGBM’s learned structure to be approximately additive: even a fully additive model without any interactions would perform reasonably well. Because the interactions that do occur concentrate along the industry dimension rather than densely among arbitrary feature pairs, predictions can be meaningfully explained as sums of individual feature contributions, with industry providing necessary moderation.

This finding connects to broader evidence on model complexity. Lou, Caruana, Gehrke, and Hooker (2013) demonstrate that models restricted to main effects and pairwise interactions achieve accuracy comparable to full-complexity random forests across a range of prediction tasks, establishing that higher-order interactions (involving three or more features simultaneously) contribute negligibly to ensemble performance. Our analysis suggests that credit risk prediction may require considerably less complexity even within the pairwise space. Moreover, the pairwise

structure that does exist is concentrated along the industry dimension, representing a particularly tractable form of non-additivity.

Importantly, the industry interactions we observe reveal that industry’s effect on default risk is more complex than what standard approaches such as fixed effects or weight-of-evidence encoding can capture. These methods allow industry to shift the baseline risk level, but our results show that industry also moderates the effects of other variables, the predictive signal of a given financial ratio differs across sectors. This moderation can be introduced into linear frameworks via frequentist random-slope models (Gelman and Hill, 2007) or Bayesian hierarchical specifications with group-level slope parameters (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin, 2013), both of which permit feature coefficients to vary by sector. If the pattern we document generalises, a substantial share of the performance gap between linear and ensemble methods attributable to interactions could be closed by this single, principled extension.

Gradient boosted decision trees nonetheless retain practical advantages beyond exploratory analysis. In prediction, GBDTs will most likely outperform even carefully specified linear models: they capture non-linear main effects automatically, and while ratio-on-ratio interactions appear weak on average, they may contribute meaningfully in tail cases where multiple risk factors compound. The contribution of the present analysis is not to argue against ensemble methods, but to characterise the structure of what they learn. Our results suggest that for practitioners operating under interpretability constraints, augmenting linear models with industry-conditional slopes offers an empirically motivated path to recovering much of ensemble performance.

E. Feature Importance Over Time

To examine whether the predictive drivers of default vary across economic regimes, we compute SHAP importance for models trained on cumulative historical data and evaluated on six distinct periods: Pre-Crisis (2004–2006), Financial Crisis (2008–2010), Sovereign Debt Crisis (2011–2013), Recovery (2014–2018), COVID-19 (2020–2021), and Post-COVID (2022–2023). Figure 7 presents a heatmap of feature importance ranks across these periods.

The most striking finding is the lack of meaningful temporal variation. The top-ranked features, *Dividend Payer*, *Cash Ratio*, and *Industry SNI* maintain their positions with remarkable consistency across all six periods. While modest shifts appear in the middle ranks, the overall hierarchy remains stable from the pre-crisis environment through the post-COVID period.

Two features warrant specific discussion. *Term Spread* exhibits the largest variation, jumping from rank 22 in the pre-crisis period to rank 7 during the 2008–2010 financial crisis. This shift likely reflects a mechanical effect: the yield curve exhibited substantially greater variation during financial stress, providing more discriminatory power. During stable periods when the term spread is compressed, the feature contains less information. *Total Asset Turnover* shows a gradual increase in importance over time, rising from rank 18 in the pre-crisis period to rank 8 in recent years, though the economic interpretation of this trend is unclear.

Several factors may explain the overall stability. First, our methodology trains each period’s model on all preceding data, meaning that crisis-period models are

dominated by pre-crisis observations. This choice is driven by data constraints. With approximately 10,000 firm-year observations per year and a default rate of 1.8%, each year contains fewer than 200 default events, insufficient for stable estimation of a gradient-boosted model with 22 features. Cumulative training ensures adequate sample size, particularly for earlier periods, and reflects realistic deployment scenarios where practitioners would train on all available history. However, this approach dampens regime-specific signals, as crisis-period models inherit relationships estimated predominantly from pre-crisis data. Second, the relatively low base rate of defaults ($\approx 1.8\%$) limits statistical power to detect subtle shifts in feature importance within shorter time windows. Third, the fundamental relationships between financial health and default may genuinely be stable: highly leveraged firms with poor liquidity face elevated default risk regardless of the macroeconomic environment, even if the *level* of risk varies cyclically.

These findings contrast with Cascarino et al. (2022), who document more pronounced shifts in predictor relevance across the Italian economic cycle. Using a one-year rolling window approach, whereby they train on year $t - 1$ and evaluating on year t , they find that liquidity and indebtedness indicators become particularly influential during financial stress. This methodological difference is consequential: their approach allows the model to adapt more rapidly to changing relationships, while our cumulative training smooths over regime-specific patterns. Whether our stability reflects genuine constancy in the underlying relationships or limitations in statistical power given our training approach remains an open question. The Italian context may also differ substantively, with potentially higher default rates and different firm characteristics enabling sharper identification of temporal variation.

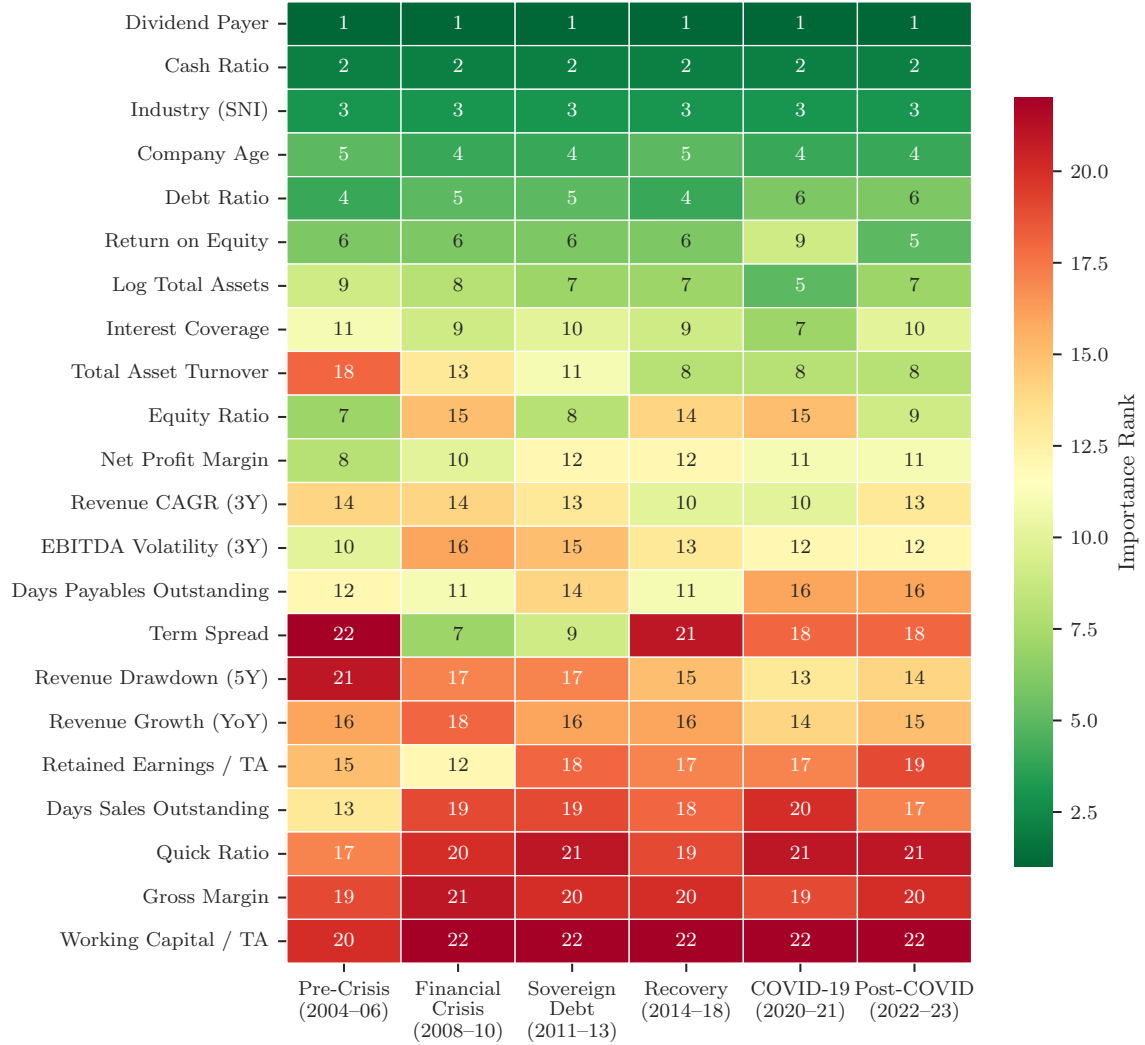


Figure 7: Feature importance stability across economic regimes. This figure displays SHAP feature importance ranks for the LightGBM model across six economic periods: Pre-Crisis (2004–2006), Financial Crisis (2008–2010), Sovereign Debt Crisis (2011–2013), Recovery (2014–2018), COVID-19 (2020–2021), and Post-COVID (2022–2023). For each period, a model is trained on all preceding data and evaluated on observations within that period; features are then ranked by mean absolute SHAP value, with rank 1 indicating highest importance. Cell colours indicate importance rank (green for higher importance, red for lower). Features are ordered by their average rank across all periods.

V. Conclusion

A. Summary of Findings

This paper investigated whether machine learning methods combined with explainable AI techniques can advance both predictive accuracy and economic understanding of corporate default among Swedish SMEs.

A competitive default-prediction model for SMEs can indeed be constructed using only publicly filed accounting information. Our LightGBM model achieves strong discriminatory power that exceeds a logistic regression baseline, with both models exhibiting reliable calibration. These results suggest that external stakeholders lacking granular credit-line data can nonetheless build effective default-prediction systems from financial statements alone.

Gradient-boosted models reveal several patterns that are difficult or impossible to learn from logistic regression. The SHAP importance analysis shows that logistic regression concentrates predictive power in its top features, while LightGBM distributes importance more evenly, extracting signal from variables whose effects are conditional or non-linear. The ALE analysis uncovers threshold effects in interest coverage and profitability, non-monotonic relationships in firm age and capital structure, and U-shaped patterns that cause certain logistic regression coefficients to collapse toward zero. The interaction analysis reveals a predominantly additive model structure, with pairwise effects concentrating along the industry dimension rather than diffusing across financial ratios, suggesting that much of the ensemble advantage attributable to interactions could be recovered by introducing industry-conditional slopes into linear specifications.

The model’s predictive features remain largely stable across time and macro-financial regimes. Contrary to prior work documenting pronounced temporal shifts in predictor relevance, feature importance ranks exhibit remarkable consistency across six economic periods spanning two decades. The top-ranked features maintain their positions from the pre-crisis environment through the post-COVID period, though our cumulative training methodology may dampen regime-specific signals.

B. Methodological Reflections

The findings carry implications for how machine learning might be positioned within empirical finance research. We identify two complementary roles: as an exploratory tool for informing traditional econometric specifications, and as a direct prediction engine in domains where forecasting accuracy is the operational objective. For the exploratory use case, it is worth noting that careful data analysis could uncover the same relationships. The advantage of the ML + XAI approach is that it surfaces relevant patterns systematically without requiring the researcher to know where to look in advance. Three observations develop these themes.

First, gradient-boosted models combined with XAI methods offer a form of automated specification search. Rather than manually testing quadratic terms, threshold dummies, and interaction effects, researchers can fit a flexible model and examine SHAP dependence plots and ALE curves to understand what patterns the data support. Our analysis provides three concrete illustrations. The binary

Dividend Payer variable, which emerged as one of the most important predictors, was itself identified through preliminary ALE analysis of ratios involving dividend payouts; examination revealed that the only relevant signal was whether a dividend was paid at all, not its magnitude, prompting the construction of a binary indicator that would not have been specified *ex ante*. The *Company Age* ALE plot reveals a year-zero effect that the linear specification missed; having identified this pattern, a researcher could construct a binary indicator for newly founded firms and incorporate it into a logistic regression, recovering the non-linearity within a fully interpretable framework. Similarly, the near-binary behaviour of *Net Profit Margin* suggests that Ohlson (1980)’s approach of including an indicator for consecutive losses captures the essential signal, with continuous margin adding limited incremental information. More generally, observing diminishing returns, plateaus, or U-shaped relationships in ALE plots can motivate experimenting with feature transformations such as logarithms, polynomials, or piecewise splines. The machine learning model thus serves as an exploratory tool that informs subsequent confirmatory analysis.

Second, SHAP interaction values provide a principled method for identifying which variable pairs exhibit meaningful dependencies, and how those dependencies operate. Where traditional approaches require researchers to specify interactions *ex ante* based on theory or intuition, interaction heatmaps reveal empirically which combinations matter and which do not. Once identified, these interactions can be encoded into linear specifications through interaction terms, hierarchical random slopes, or stratified estimation preserving interpretability while capturing the relevant structure. This workflow addresses a long-standing challenge in applied econometrics: the combinatorial explosion of possible interactions makes exhaustive testing infeasible, yet omitting important interactions biases coefficient estimates and degrades predictions.

Third, these methods warrant attention because prediction is often the operationally relevant objective across many domains in finance. Credit-risk modelling is one example, but the same logic applies to fraud detection, where identifying suspicious transactions matters more than understanding why fraud occurs; to algorithmic trading, where forecasting short-term price movements is the objective; and to insurance pricing, where accurate loss predictions determine premium-setting. In such contexts, using tools optimised for prediction and being able to understand what those tools learn is directly valuable. The growing availability of XAI methods makes this combination increasingly practical across applied finance research.

C. Limitations

Several limitations warrant acknowledgment. Most fundamentally, gradient-boosted decision trees and their associated XAI methods do not provide *p*-values or confidence intervals in the traditional econometric sense. While SHAP values quantify feature contributions and ALE plots reveal functional forms, these tools do not directly address statistical significance. We cannot say, in the formal sense, whether the threshold effect observed in *Interest Coverage* is “significant” or whether it might arise from sampling variation. The machine learning paradigm optimises for predictive accuracy, not inferential validity. The appropriate question shifts from “is this effect

statistically significant?” to “does incorporating this pattern improve predictions on unseen data?”, both questions are valuable, but they are not equivalent.

Our reliance on publicly filed accounting data imposes several constraints. Market-based structural models, which have shown strong performance for listed firms (Jackson and Wood, 2013), require observable equity prices unavailable for privately held SMEs. Primary lenders possess granular loan-level data and transaction histories which they can use for internal models that we also lack access to. Additionally, although the raw Serrano database contains fields for granted and utilised overdrafts, we exclude these due to merging difficulty and data quality concerns, variables that proved important in Cascarino et al. (2022).

The temporal analysis is constrained by our cumulative training methodology and by sample size. Alternative approaches such as rolling windows or period-specific models might reveal more pronounced temporal variation, though at the cost of smaller training samples. The relatively short post-COVID period (2022–2023) also limits conclusions about the current high-interest-rate environment.

Certain modelling choices warrant scrutiny. *Industry SNI* is encoded at the 3-digit SNI level, a granularity selected via analysis of variance but which may introduce overfitting relative to coarser 1, or 2-digit classifications. The *Term Spread* variable, while intended to capture macroeconomic conditions, takes a single value per calendar year, effectively encoding time into the model. Its ALE plot exhibits a counterintuitive slope, suggesting that the variable may be proxying for unobserved temporal trends rather than capturing the intended yield-curve signal.

Finally, our analysis is restricted to Swedish SMEs. While this focus enables clean inference by holding institutional and accounting frameworks constant, it limits the generalisability of specific findings. The importance rankings, threshold locations, and interaction patterns we document may differ in other national contexts with different bankruptcy regimes, accounting standards or firm characteristics.

D. Directions for Future Research

Several extensions appear promising. First, the concentration of interactions along the industry dimension suggests value in developing sector-specific default models or hierarchical frameworks that allow coefficients to vary across sectors. Such an approach would clarify whether the non-linear patterns identified here are consistent across sectoral contexts or reflect industry-specific dynamics.

Second, the temporal dimension warrants further investigation. Our cumulative training approach, while necessitated by sample size constraints, may dampen regime-specific signals. Alternative methodologies such as time-weighted estimation that downweights older observations, or rolling-window approaches applied to larger datasets could assess whether the stability we observe reflects genuine constancy in the underlying relationships or is an artifact of our estimation strategy. More granular examination of the post-COVID period as additional data accumulate might be particularly valuable.

Third, while this paper estimates one-year default probabilities, many credit decisions require longer-term assessments. Comparing one-, three- and five-year prediction horizons would reveal how predictor importance and functional forms evolve with the forecast horizon.

Fourth, replication on other national datasets would establish which findings generalise and which are specific to the Swedish institutional context.

E. Concluding Remarks

This paper demonstrates that explainable AI methods can render gradient-boosted decision trees interpretable in economically meaningful ways. The combination of strong predictive performance with transparent explanations, revealing thresholds, functional forms and interaction structures addresses the tension between accuracy and interpretability that has limited machine learning adoption in regulated financial contexts.

The broader implication is that machine learning and econometrics need not be viewed as competing paradigms. When the objective is prediction, flexible models combined with post-hoc explanation methods offer capabilities that linear specifications cannot match, and crucially, XAI tools allow practitioners to understand what these models learn, not merely to trust their outputs. Beyond prediction, such methods can also serve an exploratory function, surfacing non-linearities and interaction structures that inform the specification of traditional econometric models. When the objective is causal identification and hypothesis testing, traditional regression with careful identification strategies remains essential. The growing maturity of XAI tools makes this integration increasingly practical.

References

- Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama, 2019, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 2623–2631.
- Altman, Edward I., 1968, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance* 23, 589–609.
- Apley, Daniel W, and Jingyu Zhu, 2020, Visualizing the effects of predictor variables in black box supervised learning models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, 1059–1086.
- Basel Committee on Banking Supervision, 2011, Basel iii: A global regulatory framework for more resilient banks and banking systems, Discussion paper Bank for International Settlements.
- Ben Jabeur, Sami, Cheima Gharib, Salma Mefteh-Wali, and Wissal Ben Arfi, 2021, Catboost model and artificial intelligence techniques for corporate failure prediction, *Technological Forecasting and Social Change* 166, 120658.
- Blume, Marshall E., Felix Lim, and A. Craig MacKinlay, 1998, The declining credit quality of u.s. corporate debt: Myth or reality?, *The Journal of Finance* 53, 1389–1413.
- Breiman, Leo, 2001, Statistical modeling: The two cultures, *Statistical science* 16, 199–231.
- Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen, 1984, *Classification and regression trees* (CRC press).
- Brier, Glenn W, 1950, Verification of forecasts expressed in terms of probability, *Monthly weather review* 78, 1–3.
- Brown, Stephen J, William Goetzmann, Roger G Ibbotson, and Stephen A Ross, 1992, Survivorship bias in performance studies, *Review of financial studies* 5, 553–580.
- Brunlid, Axel, 2025, Predicting corporate credit ratings using catboost, .
- Carmona, Pedro, Francisco Climent, and Alexandre Momparler, 2019, Predicting failure in the u.s. banking sector: An extreme gradient boosting approach, *International Review of Economics & Finance* 61, 304–323.
- Cascarino, Giulia, Marco Moscatelli, and Fabio Parlapiano, 2022, Explainable artificial intelligence: Interpreting default forecasting models based on machine learning, *Questioni di Economia e Finanza* (Occasional Papers) 724 Bank of Italy.
- Chava, Sudheer, and Robert A Jarrow, 2004, Bankruptcy prediction with industry effects, *Review of Finance* 8, 537–569.
- Delmar, Frédéric, Alexander McKelvie, and Karl Wennberg, 2013, Untangling the relationships among growth, profitability and survival in new firms, *Technovation* 33, 276–291.
- Deloof, Marc, 2003, Does working capital management affect profitability of belgian firms?, *Journal of Business Finance & Accounting* 30, 573–588.
- European Banking Authority, 2021, Discussion paper on machine learning for irb models, Discussion Paper EBA/DP/2021/04 European Banking Authority.
- European Commission, 2003, Commission recommendation of 6 may 2003 concerning the definition of micro, small and medium-sized enterprises, Discussion Paper L 124/36 Official Journal of the European Union.
- Friedman, Jerome H., 2001, Greedy function approximation: A gradient boosting machine, *The Annals of Statistics* 29, 1189–1232.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, 2013, *Bayesian Data Analysis* (Chapman and Hall/CRC: Boca Raton, Florida) third edn.
- Gelman, Andrew, and Jennifer Hill, 2007, *Data analysis using regression and multilevel/hierarchical models* (Cambridge university press).
- Jabeur, S. B., C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, 2021, Catboost model and artificial intelligence techniques for corporate failure prediction, *Technological Forecasting and Social Change* 166, 120658.
- Jackson, Richard H.G., and Anthony Wood, 2013, The performance of insolvency prediction and credit risk models in the UK: A comparative study, *The British Accounting Review* 45, 183–202.

- Jacobson, Tor, and Erik von Schedvin, 2015, Trade credit and the propagation of corporate failure: An empirical analysis, *Econometrica* 83, 1315–1371.
- Ke, Guolin, Thomas Meng, Qi andberg, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, 2017, Lightgbm: A highly efficient gradient boosting decision tree, in *Advances in neural information processing systems* , vol. 30.
- Kursa, Miron B, and Witold R Rudnicki, 2010, Feature selection with the boruta package, *Journal of statistical software* 36, 1–13.
- Lessmann, Stefan, Bart Baesens, H-V Seow, and Lyn C Thomas, 2015, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research* 247, 124–136.
- Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker, 2013, Accurate intelligible models with pairwise interactions, in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* pp. 623–631.
- Lundberg, Scott M, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Shion Bansal, and Su-In Lee, 2020, From local explanations to global understanding with explainable ai for trees, *Nature machine intelligence* 2, 56–67.
- Lundberg, Scott M., and Su-In Lee, 2017, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems* , vol. 30.
- Merton, Robert C., 1974, On the pricing of corporate debt: The risk structure of interest rates, *The Journal of Finance* 29, 449–470.
- Naeini, Mahdi Pakdaman, Gregory Cooper, and Milos Hauskrecht, 2015, Obtaining well calibrated probabilities using bayesian binning, in *Proceedings of the AAAI conference on artificial intelligence* , vol. 29.
- Nyitrai, Tamás, and Miklós Virág, 2019, The effects of handling outliers on the performance of bankruptcy prediction models, *Socio-Economic Planning Sciences* 67, 34–42.
- Ohlson, James A., 1980, Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research* 18, 109–131.
- Shmueli, Galit, 2010, To explain or to predict?, *Statistical science* 25, 289–310.
- Siddiqi, Naeem, 2006, *Credit risk scorecards: developing and implementing intelligent credit scoring* (John Wiley & Sons).
- Statistics Sweden, 2025, Lending rates to households and non-financial corporations, breakdown by fixation periods, Statistical Database Accessed: September 2025.
- Stinchcombe, Arthur L, 1965, Social structure and organizations, in James G. March, ed.: *Handbook of Organizations* . pp. 142–193 (Rand McNally: Chicago).
- Swedish House of Finance, 2025, Serrano database, Swedish House of Finance, Stockholm School of Economics Accessed: September 2025.

A. Appendix

This appendix provides additional detail on the dataset, sample construction, feature definitions, descriptive statistics, and the full ALE plots.

Data source

The primary data source is the Serrano database (Swedish House of Finance, 2025), a comprehensive panel dataset of Swedish firms maintained by the Swedish House of Finance. Serrano consolidates information from multiple administrative registers, including annual financial statements filed with Bolagsverket (the Swedish Companies Registration Office), corporate registry data, and records of significant corporate events such as bankruptcies and reorganisations. The database covers 1998–2023 and includes both balance sheet and income statement items, pre-computed financial ratios ("nyckeltal"), and firm-level metadata such as industry classification, geographic location, and legal form. Macroeconomic data on term spread is obtained from Statistics Sweden (2025).

Sample construction and filters

The raw Serrano dataset contains approximately 12.4 million firm-year observations. To construct an analytically coherent sample of active Swedish SMEs, four sequential filters are applied. Table III summarises the impact of each filter on sample size and default composition. The final sample contains 326,658 firm-year observations with an overall default rate of 1.76%.

Figure 8 displays the number of firms and the annual default rate over time. Default rates exhibit cyclical spikes around the early 2000s recession, the 2007–2009 financial crisis, and the 2021–2022 period.

Model features

Table IV lists the 22 features used in the predictive models, organised by category. Table V summarises the distribution of all features for defaulting and non-defaulting firms.

Full ALE plots

Figure 9 visualises the Accumulated Local Effects (ALE) curves for all continuous variables included in the model.

AI disclaimer

During the preparation of this paper, we made use of AI tools including ChatGPT, Claude and Google Gemini to assist with Python coding and grammatical refinement. However, all reasoning, empirical analysis, interpretations and conclusions are entirely the work of the authors.

Table III
Sample construction and filter analysis

This table summarises how sequential sample filters affect the size and default composition of the Serrano dataset. Columns report the number of firm-year observations (N), their share of the total sample (% of Total), and the default rate, defined as the percentage of firms entering bankruptcy or reorganisation in the following year. Default is measured before applying subsequent filters.

Filter	Category	N	% of Total	Default Rate (%)
Active Status	Inactive (0)	2,523,321	20.3	2.98
	Active (1)	9,892,403	79.7	0.72
Size Category	Micro	11,441,062	92.1	1.19
	Small	792,238	6.4	1.19
	Medium	141,884	1.1	0.70
	Large	40,540	0.3	0.29
Credit Reporting	Independent (1)	9,064,718	73.0	1.39
	Swedish subsidiary (2)	963,247	7.8	0.55
	Foreign subsidiary (3)	2,030,491	16.4	0.65
	Parent company (4)	357,265	2.9	0.55
Industry	Non-Financial	10,847,679	87.4	1.29
	Financial (40.0)	1,568,045	12.6	0.44

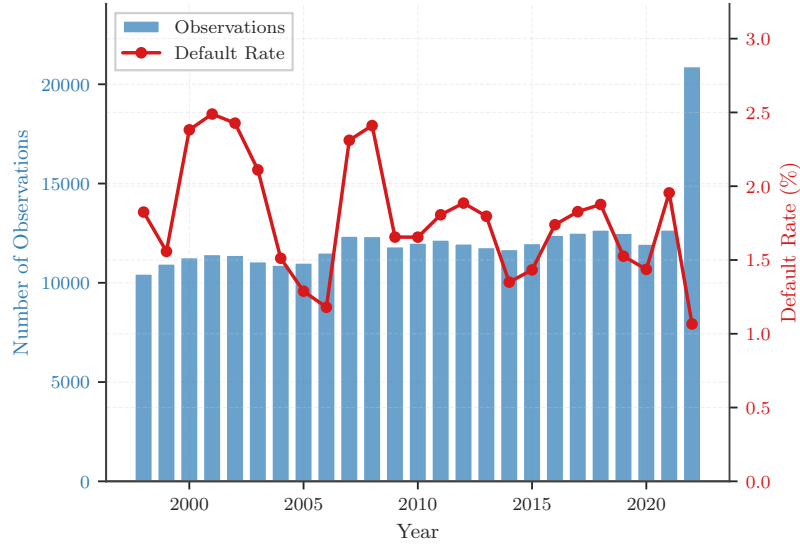


Figure 8: Sample composition and default rates over time. This figure displays the number of firm-year observations (bars, left axis) and the annual default rate (line, right axis) for the final sample across 1998–2022. Default rate is defined as the percentage of firms entering bankruptcy or reorganisation in the following year. The sample comprises active, independent, non-financial Swedish SMEs from the Serrano database.

Table IV
Model feature definitions

This table provides definitions for all 22 features used in the final model, organised by category. Parenthetical references indicate correspondence with classical bankruptcy prediction models (Altman X_1 – X_5 from the Z-score; Ohlson variables from the O-score). All monetary values in the underlying Serrano data are reported in thousands of Swedish kronor (kSEK). Log transformations use $\log(1+x)$ to handle zero values. YOY changes and multi-year features are computed at the firm level using lagged observations. Missing values are handled natively by LightGBM and imputed using training-set medians for Logistic Regression.

Feature	Definition
<i>Size & Age</i>	
Log Total Assets	$\log(\text{Total Assets})$ (Ohlson SIZE)
Company Age	Years since registration with Bolagsverket
<i>Liquidity</i>	
Working Capital / TA	$(\text{Current assets} - \text{Current liabilities}) / \text{Total assets}$ (Altman X_1 ; Ohlson WC/TA)
Cash Ratio	$\text{Cash and equivalents} / \text{Current liabilities}$
Quick Ratio	$(\text{Current assets} - \text{Inventories}) / \text{Current liabilities}$
<i>Profitability</i>	
Retained Earnings / TA	$\text{Retained earnings} / \text{Total assets}$ (Altman X_2)
Net Profit Margin	$\text{Net profit} / \text{Net sales}$
Gross Margin	$(\text{Net sales} - \text{Cost of goods sold}) / \text{Net sales}$
Return on Equity	$\text{Net profit} / \text{Shareholders' equity}$
<i>Leverage & Solvency</i>	
Debt Ratio	$\text{Total liabilities} / \text{Total assets}$ (Ohlson TL/TA)
Equity Ratio	$\text{Shareholders' equity} / \text{Total assets}$
Interest Coverage	$\text{EBIT} / \text{Interest expenses}$ (capped at $[-5, 20]$)
<i>Efficiency</i>	
Total Asset Turnover	$\text{Net sales} / \text{Total assets}$ (Altman X_5)
Days Sales Outstanding	$(\text{Accounts receivable} / \text{Net sales}) \times 365$
Days Payables Outstanding	$(\text{Accounts payable} / \text{Cost of goods sold}) \times 365$
<i>Growth & Volatility</i>	
Revenue Growth (YoY)	Year-over-year percentage change in net sales
Revenue CAGR (3Y)	Compound annual growth rate of revenue over 3 years
Revenue Drawdown (5Y)	Maximum peak-to-trough revenue decline over 5 years
EBITDA Volatility (3Y)	$\sigma(\text{EBITDA}_{t,t-2}) / \text{Total Assets}$
<i>Structural & Contextual</i>	
Dividend Payer	Binary indicator: 1 if firm paid dividends, 0 otherwise
Industry (SNI)	3-digit SNI07 industry classification code
Term Spread	Long-term bond rate – short-term rate

Table V
Descriptive statistics by default status

This table reports summary statistics for all model features, separately for non-defaulting firms ($N = 298,795$) and defaulting firms ($N = 5,354$). Default is defined as entering bankruptcy or reorganisation in the following year. Statistics reported are mean, standard deviation, skewness, and kurtosis. The high skewness and kurtosis values for several financial ratios reflect the heavy-tailed distributions typical of SME accounting data, motivating the use of robust scaling for Logistic Regression and the robustness of tree-based methods to outliers.

Variable	Non-Defaulting				Defaulting			
	Mean	Std	Skew	Kurt	Mean	Std	Skew	Kurt
Company Age	16.28	14.11	1.78	5.22	10.00	10.41	2.57	10.4
Total Asset Turnover	2.91	5.90	347	159,204	7.18	242.6	73.0	5,337
Debt Ratio	6.71	95.82	105	18,610	21.57	173.2	52.8	3,025
Equity Ratio	0.355	0.212	0.53	1.68	0.163	0.158	1.94	4.91
Return on Equity	0.029	9.46	-65.9	9,811	-1.59	11.36	-8.74	335
Quick Ratio	1.96	134.0	503	264,438	0.839	1.51	44.8	2,639
Net Profit Margin	-0.928	133.9	-304	118,424	-4.50	217.6	-60.1	3,816
Revenue Growth (YoY)	4.49	935.9	436	208,241	6.17	195.8	52.3	2,927
Log Total Assets	9.02	1.14	0.38	1.50	8.42	1.13	0.07	1.64
Working Capital / TA	0.225	1.41	-454	229,917	-0.757	46.35	-73.0	5,337
Retained Earnings / TA	0.149	1.34	-243	79,711	-0.328	22.95	-72.9	5,320
Interest Coverage	143.0	1,812	42.8	12,086	0.746	574.7	59.2	3,996
Cash Ratio	0.912	30.53	383	173,790	0.153	0.511	17.0	436
Gross Margin	-0.130	83.26	-320	127,998	-0.720	67.20	-71.7	5,163
Dividend Payer	0.466	0.499	0.14	-1.98	0.055	0.228	3.90	13.2
Days Sales Outstanding	138.4	10,694	330	133,062	108.8	2,091	67.9	4,795
Days Payables Outstanding	63.08	3,194	422	198,067	77.29	395.9	46.7	2,492
Revenue CAGR (3Y)	0.210	0.780	21.5	1,002	0.465	1.51	9.69	129
Revenue Drawdown (5Y)	-0.113	0.166	-2.36	6.88	-0.199	0.235	-1.57	2.26
EBITDA Volatility (3Y)	0.092	0.467	161	34,684	0.475	16.17	65.9	4,361
Term Spread	0.423	0.575	0.27	-0.07	0.392	0.574	0.41	-0.04

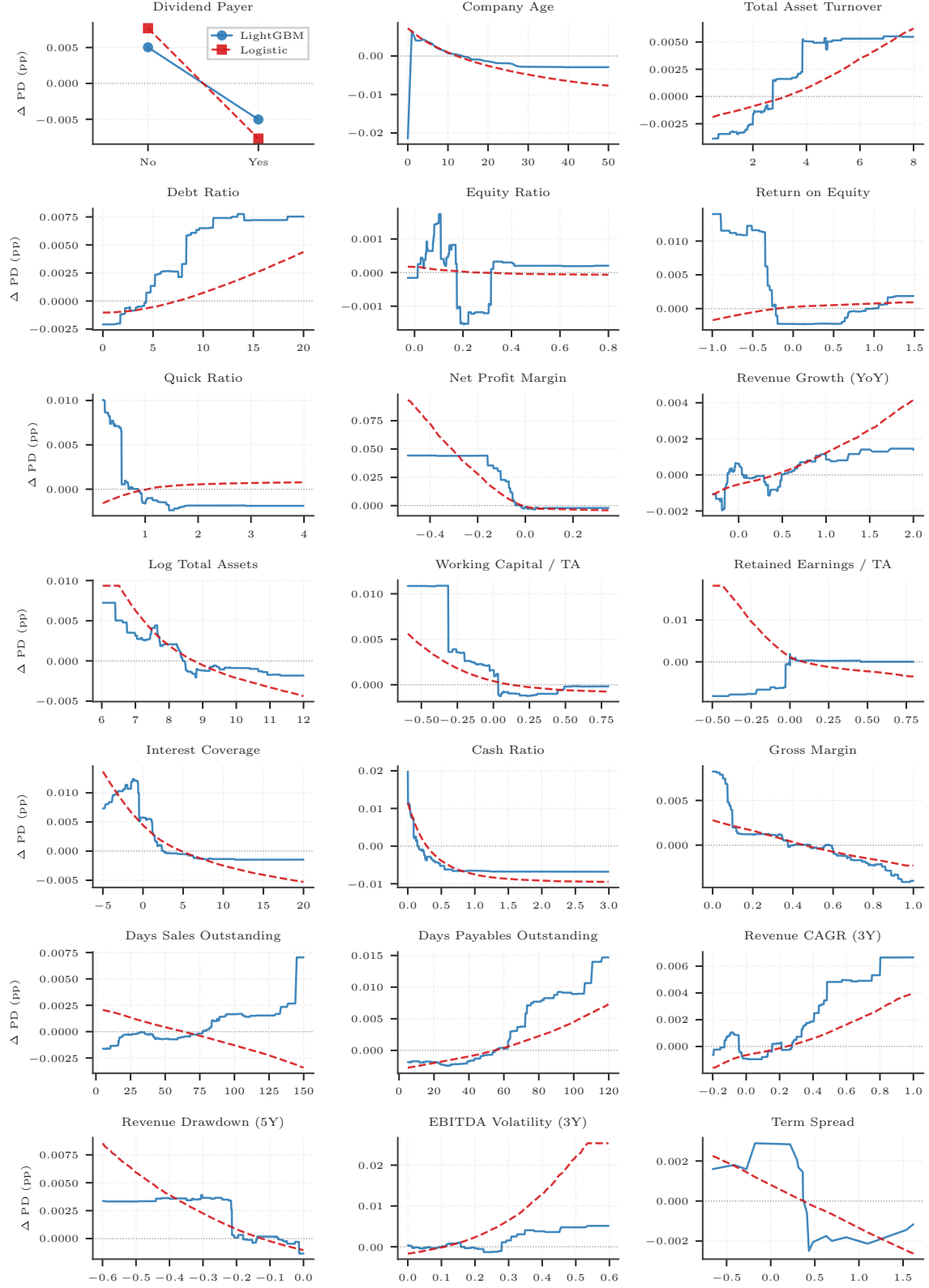


Figure 9: Accumulated Local Effects (ALE) for all continuous features. Each panel displays the ALE curve for one feature, estimated using the LightGBM gradient boosting model and the Logistic Regression baseline. The y -axis displays the ALE value, representing the change in predicted default probability (in percentage points) relative to the average prediction, accumulated across the distribution of the feature. Horizontal dotted lines indicate an effect of zero, corresponding to the average predicted probability. Feature definitions are provided in Table IV.