1 Data and Methodology

1.1 Data

1.1.1 Sample Construction

Firm-level data are obtained from the Serrano database. 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 variables from Statistics Sweden (SCB) and the Riksbank (Sweden's central bank) including GDP growth, the Consumer Price Index (CPI), the policy rate and the term spread. The modelling framework follows an iterative data-processing procedure, where variables are refined and transformed step by step to ensure comparability and consistency.

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. Small enterprises have fewer than 50 employees and either 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. 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 intentional; prior research shows that institutional and macroeconomic differences across countries constitute a major source of variation in credit-risk models (Brunelid, 2025). Concentrating solely on Sweden therefore enhances comparability and supports 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. Although excluding group-owned firms slightly reduces predictive performance, it safeguards independence across observations and avoids implicit support effects, resulting in a behaviourally consistent dataset suitable for machine-learning analysis.

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.

1.1.2 Data Quality and Leakage Correction

During the exploratory phase, a significant data leakage issue was identified, likely stemming from the Serrano dataset's handling of fiscal versus calendar years. We observed that for certain defaulting firms, financial records were duplicated across consecutive time steps. This artifact was detected via the engineered year-over-year features, which revealed zero-variance changes or perfect correlations inconsistent with genuine financial reporting. These anomalies inadvertently encoded future information (specifically, the cessation of new reporting due to default) into the training data. 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.

1.1.3 Feature Selection Strategy

To capture the dynamic nature of firm distress, we engaged in extensive feature engineering guided by domain expertise and prior literature. Beyond the standard liquidity and leverage ratios provided by Serrano, we constructed year-over-year changes, multi-year trend indicators, and volatility measures. This process initially yielded well in excess of 100 candidate features—an effort conducted partially in response to data limitations. Unlike recent studies in the credit risk literature, we lack granular loan-level data from credit registers. Cascarino et al. (2022), for instance, utilise data from the Italian Credit Register that includes variables such as drawn-to-granted ratios for different loan types, measuring the proportion of available credit a firm is actually using across credit lines. These credit behavioural variables proved highly predictive in their analysis, with drawn-to-granted ratios ranking among the most important features. The absence of access to equivalent Swedish credit register data represents a significant limitation.

While gradient boosted trees are theoretically robust to uninformative features and capable of handling correlated predictors natively, including irrelevant variables can degrade generalization performance and obscure interpretation. We therefore pursued a deliberate reduction from over 100 candidates to a parsimonious set, guided by a single objective: maximize predictive performance—as measured by LogLoss, ROC-AUC, and Brier Score on held-out data—while minimizing model complexity. Our approach was holistic rather than mechanical; no single algorithm determined inclusion, and we triangulated across multiple methods capturing different dimensions of feature relevance.

Domain knowledge provided the first filter. Financial ratios with established theoretical links to default—such as interest coverage, leverage, and liquidity measures—were given priority, while variables lacking economic interpretation were scrutinized more heavily regardless of their statistical performance. This grounding in financial theory ensured that the final model would yield interpretable insights rather than merely exploiting statistical artifacts.

Statistical diagnostics addressed multicollinearity and robustness. We iteratively removed features with a Variance Inflation Factor exceeding 10, ensuring that retained variables contributed unique information rather than duplicating signals captured elsewhere. Stability selection provided a complementary check: models were trained across bootstrap iterations, and only features ranking consistently among the most important were retained. This filtered out variables that appeared predictive in specific data slices but lacked robustness across samples.

The Boruta algorithm offered a formal significance benchmark by comparing feature importance against randomized shadow versions, retaining only those performing significantly better than noise. However, we observed that Boruta's strict rejection criteria—amplified by the multicollinearity inherent in accounting data—sometimes eliminated variables that, while individually weak, contributed meaningfully to ensemble accuracy. We therefore treated Boruta as informative rather than dispositive.

The binding criterion throughout was out-of-sample predictive performance evaluated via cross-validation. Features were retained if their inclusion improved LogLoss, ROC-AUC, and Brier Score; those degrading performance were removed regardless of theoretical appeal or standing on any individual diagnostic. SHAP-based importance provided a final model-specific check, quantifying each feature's marginal contribution to actual predictions and confirming alignment between statistical selection and the fitted model's behaviour.

These methods do not always agree—a variable may pass collinearity thresholds but fail stability selection, or survive Boruta but contribute little according to SHAP. We resolved such conflicts pragmatically, always deferring to generalization performance. This approach aligns with the study's predictive focus: a feature contains valuable signal if and only if it reduces error on unseen data. The final model utilizes 29 features, representing a balance between predictive power, interpretability, and theoretical grounding. These variables are defined in Appendix A.

1.2 Methodological Framework

1.2.1 Prediction versus Inference in Financial Modelling

A fundamental distinction exists between the objectives of classical econometric inference and machine learning prediction, a dichotomy articulated by Shmueli (2010) and often referred to as the ”two cultures” of statistical modelling (Breiman, 2001). In traditional empirical finance, the primary objective is inference: the isolation of unbiased parameter estimates ($\hat{\beta}$) to test hypotheses about the causal or structural relationship between covariates and a dependent variable. This paradigm relies on the assumption that the data are generated by a stochastic process of a known functional form—typically linear—plus an error term. Under this framework, model selection is driven by goodness-of-fit within the sample and the statistical significance of individual coefficients.

In contrast, the machine learning paradigm prioritizes prediction: minimizing the generalization error on unseen data ($\hat{y}$). Here, the functional form of the data-generating process is treated as unknown and potentially complex. Rather than imposing a rigid linear structure, machine learning algorithms approximate the underlying function $f(x)$ by learning patterns directly from the data. This shift in objective from explaining the past to forecasting the future necessitates a different set of tools and validation protocols.

In the context of corporate default prediction, this shift is not merely academic but operationally critical. For financial institutions and regulators, the utility of a risk model is defined by its ability to accurately quantify the probability of future distress. A model that yields statistically significant coefficients but fails to discriminate between solvent and insolvent firms offers limited practical value. Superior predictive accuracy directly translates to more efficient capital allocation, as loan pricing and regulatory capital requirements (e.g., under Basel III) 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 inference need not be mutually exclusive. As Shmueli (2010) emphasizes, predictive models can inform and refine theory by revealing patterns that simpler specifications miss. The challenge is methodological: standard machine learning models are "black boxes" whose internal mechanics resist interpretation. This opacity conflicts with regulatory requirements for model transparency (EBA, 2021) and with the scientific objective of understanding *why* certain firms default. Explainable AI (XAI) techniques address this tension by providing post-hoc methods that decompose predictions into interpretable components. By applying SHAP values and Accumulated Local Effects to a high-performing GBDT model, this study aims to recover economic insights—the functional relationships between financial ratios and default risk—without sacrificing predictive accuracy. This approach aligns with an emerging literature demonstrating that XAI can maintain or even improve interpretability relative to traditional linear models while preserving the predictive gains of complex algorithms (Bussmann et al., 2021).

1.2.2 Evaluation Criteria

The distinction between prediction and inference also shapes how models should be evaluated. Standard econometric practice emphasizes in-sample fit and coefficient significance—metrics designed to assess whether estimated relationships are statistically robust. Predictive modelling, by contrast, requires metrics that assess performance on unseen data and that reflect the operational objectives of credit risk management. We therefore evaluate models along three complementary dimensions: discrimination, probabilistic accuracy, and calibration.

Discrimination: ROC-AUC The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) 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 0.70 and 0.90. ROC-AUC evaluates only rank-ordering, not the accuracy of probability levels. This is a meaningful limitation: a model that perfectly separates defaulters from survivors but assigns arbitrary probability values would achieve ROC-AUC of 1.0 while being useless for pricing or capital allocation. We therefore supplement this metric with measures of probabilistic accuracy.

Probabilistic Accuracy: LogLoss and Brier Score Since loan pricing and regulatory capital requirements under Basel III are direct functions of the estimated probability of default, the accuracy of these estimates—not merely their rank ordering—determines economic outcomes. Both LogLoss (Equation 1) and the Brier Score penalize confident but incorrect predictions more heavily than tentative ones. The Brier Score measures the mean squared error of probability forecasts:

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

where $y_i \in \{0, 1\}$ is the observed outcome and p_i is the predicted probability. LogLoss operates on a logarithmic scale, imposing steeper penalties for overconfident errors—precisely where credit losses concentrate. These metrics 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 model is well-calibrated if predicted probabilities correspond to empirical default frequencies: among firms assigned a 5% default probability, approximately 5% should actually default. This property is essential for provisioning, stress testing, and regulatory compliance. We assess calibration using the Expected Calibration Error (ECE) 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. Systematic miscalibration has direct financial consequences: underestimating default probabilities leads to underpriced risk and insufficient capital reserves, while overestimation results in uncompetitive pricing and forgone profitable lending.

1.3 Model Specifications

1.3.1 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 prior to estimation. Predictors exhibiting extreme outliers or heavy tails (e.g., interest coverage, profit margins) are first Winsorized at the 1st and 99th percentiles, clipping values beyond these thresholds to reduce the influence of extreme observations. All numeric features are then scaled using a Robust Scaler based on the median and interquartile range (IQR), which is less sensitive to remaining outliers than standard normalization. Categorical controls are processed via Target Encoding with smoothing, where levels are replaced by the smoothed posterior probability of the target. In industry practice, credit risk models often use Weight of Evidence (WOE) binning to enable logistic regression to capture non-linear patterns (Siddiqi, 2006). However, for this comparison, we deliberately use minimal preprocessing to isolate the architectural differences between linear and tree-based models.

Following preprocessing, the model estimates the log-odds of default as a linear combination of input features. To control for potential heteroscedasticity, we report HC3 (Heteroscedasticity-Consistent) robust standard errors. Implementation is performed in Python using scikit-learn for feature transformation and statsmodels for GLM estimation.

1.3.2 Gradient Boosted Decision Trees

To capture non-linearities and interaction effects without manual specification, we employ Gradient Boosted Decision Trees (GBDT). Unlike logistic regression, tree-based models are invariant to monotonic transformations of input features, so predictors are left in their raw form to preserve interpretability; only categorical variables undergo encoding.

Decision Trees The fundamental building block of this approach is the classification tree. Decision trees approximate complex relationships by recursively partitioning the feature space into J distinct, non-overlapping regions (leaves) R_j. For an observation vector x, the tree T predicts a constant value c_j corresponding to the region into which x falls:

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

where $\Theta = \{R_j, c_j\}_1^J$ represents the model parameters and $I(\cdot)$ is the indicator function. The regions are constructed via a greedy algorithm: at each step, the model identifies the single variable x_k and split point s that partition the data into two half-planes to maximize the reduction in impurity. This process allows the model to capture non-linearities and interaction effects natively.

Gradient Boosting Single decision trees are high-variance estimators prone to overfitting. We therefore employ Gradient Boosting (Friedman, 2001), which combines many weak learners (shallow trees) into a single predictive model $F(x)$. The ensemble is built 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, a regularization parameter scaling the contribution of each tree.

Our implementation minimizes the negative log-likelihood (LogLoss), defined as:

$$L(y, p) = -\left[y \log(p) + (1 - y) \log(1 - p)\right] \tag{1}$$

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.

Implementation We elect to use LightGBM (Ke et al., 2017). This choice is motivated by computational efficiency; LightGBM utilizes histogram-based algorithms and a leaf-wise tree growth strategy, allowing for faster training speeds and lower memory consumption on large tabular datasets without compromising the model's ability to minimize LogLoss.

1.4 Explainable AI Methods

To recover interpretable insights from the non-linear GBDT model, we employ two complementary post-hoc explanation methods. SHAP provides local explanations—decomposing individual predictions into feature contributions—while ALE reveals global patterns—the average functional relationship between each predictor and default risk across the entire dataset.

1.4.1 SHapley Additive exPlanations (SHAP)

The fundamental challenge in interpreting any predictive model is attribution: when a model predicts that a firm has a 15% probability of default, which features drove this prediction, and by how much? In linear models, coefficients provide a natural answer. In non-linear models with interactions, however, a feature's contribution depends on the values of other features, making attribution ambiguous.

SHAP resolves this ambiguity by borrowing a solution from cooperative game theory (Lundberg & Lee, 2017). The intuition is as follows: imagine each feature as a "player" in a coalition game where the "payout" is the model's prediction. We want to fairly distribute this payout among the players based on their individual contributions. The Shapley value—originally developed to allocate payoffs in coalitional games—provides the unique allocation satisfying certain fairness axioms (efficiency, symmetry, linearity, and null player).

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:

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

where F is the set of all features, S is a subset excluding feature j, and $f_x(S)$ denotes the model's prediction using only features in S. The weighting term $\frac{|S|!(|F|-|S|-1)!}{|F|}!$ accounts for all orderings in which the coalition could have formed. The resulting value ϕ_j can be interpreted as the average marginal contribution of feature j to the prediction, controlling for all possible interactions with other features.

For a firm predicted to have elevated default risk, SHAP values reveal which features pushed the prediction upward (positive ϕ_j) and which pushed it downward (negative ϕ_j). In econometric terms, these values represent the local marginal effect of each variable on the log-odds of default for that specific firm. Aggregating SHAP values across observations yields global importance measures: the mean absolute SHAP value for a feature indicates its average impact on predictions across the dataset.

Computing exact Shapley values requires evaluating $2^{|F|}$ subsets, which becomes prohibitive for models with many features. We therefore utilize the TreeSHAP algorithm (Lundberg et al., 2020), which exploits the hierarchical structure of decision trees to calculate exact Shapley values in polynomial time rather than exponential.

SHAP Interaction Values Beyond main effects, SHAP can decompose predictions into 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 that cannot be separated into the sum of their individual contributions. Formally, this is computed by distributing the Shapley value of feature j across its interactions with all other features. A non-zero interaction value $\phi_{j,k}$ indicates that the effect of feature j depends on the value of feature k, revealing conditional relationships that linear models cannot capture. For instance, a positive interaction between leverage and firm age might indicate that high leverage is particularly risky for young firms but less so for established ones. These interaction values enable the detection of economically meaningful contingencies that would otherwise remain hidden within the ensemble's structure.

1.4.2 Accumulated Local Effects (ALE)

While SHAP explains individual predictions, practitioners often require a global view: how does default risk change, on average, as a given financial ratio 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. PDPs compute the average prediction across all observations while artificially varying one feature, which can create unrealistic combinations (e.g., a firm with very low assets but very high revenue) that the model was never trained on. This extrapolation produces misleading estimates of feature effects.

Accumulated Local Effects (Apley & Zhu, 2020) avoid this extrapolation problem by computing effects conditionally. Rather than averaging across the entire dataset, ALE considers only observations that naturally occur near each feature value. The method works in three steps: (1) partition the feature's range into small intervals; (2) for observations within each interval, compute the local effect as the difference in predictions at the interval's upper and lower boundaries; (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 is taken conditionally on $X_j = z_j$, meaning we average only over observations near that feature value. The constant C centers the plot so that the average effect is zero. This conditional averaging ensures that ALE curves reflect the true effect of a feature as it varies within the actual data distribution, rather than in extrapolated regions.

The resulting ALE plot displays the feature's effect on default probability (relative to the baseline) across its range. For example, an ALE plot for leverage might reveal that default risk increases sharply as leverage exceeds a certain threshold, or that the relationship is non-monotonic—insights that a single linear coefficient cannot convey. Because ALE conditions on the observed data distribution, these curves provide an unbiased view of how a financial ratio influences default risk, independent of its correlation with other predictors.