1 Introduction

Predicting corporate default is a central concern in modern finance, both for regulators and for financial institutions issuing credit. Accurate and transparent models are essential for capital requirements under frameworks such as Basel III as well as for internal risk management, pricing and lending decisions. Because of these regulatory demands and the need for explainable risk assessments, logistic regression has long been the standard in default prediction research (**??**). Its popularity is grounded on several well-documented advantages: the model is interpretable, transparent, and produces well-calibrated probability estimates.

However, logistic regression relies on strong assumptions about how predictors relate to default risk. It assumes that each variable has a linear and monotonic effect on the probability of default. Paradoxically, these assumptions are beneficial in many contexts—they simplify validation, allow for strong model-risk controls, and make it possible to clearly test whether the underlying data satisfy the model's requirements. This transparency is a key reason why regulators often prefer simpler models.

At the same time, predictive accuracy is central in default modelling. Small improvements in model discrimination can translate into meaningful financial gains. This has driven growing interest in more flexible machine-learning methods. Gradient-boosted decision trees (GBDT) have demonstrated strong performance in credit-risk settings, largely because they can capture nonlinear relationships, interaction effects, and other complex patterns that logistic regression cannot represent without extensive manual feature engineering (**?**). These capabilities make modern machine-learning models promising tools for advancing the predictive frontier in default prediction.

But these models introduce their own challenges. Their structure makes them difficult to interpret, complicating both regulatory compliance and theoretical understanding. This lack of interpretability is problematic in credit risk, where decisions must be explainable and defensible. A model that labels a firm as high-risk without a clear rationale is difficult to justify in a regulated environment. Recent developments in explainable AI (XAI) offer a potential way forward—methods such as SHAP provide consistent, model-agnostic explanations that reveal which factors drive predictions, making it possible to extract meaningful insights even from complex machine-learning models (**?**).

In this paper, we face an additional constraint: we do not have access to loan-level data, which is the standard in credit-scoring research. Instead, our dataset consists solely of publicly available financial statements and corporate registry information. This motivates the following research questions:

1. Can we construct a competitive default-prediction model using only publicly available accounting information?

2. What can a GBDT 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?

4. To what degree is a GBDT + XAI framework applicable beyond credit-risk modelling, and should these methods be explored more widely within financial research?

1.1 Overview of Data and Method

1.2 Main Results

1.3 Contribution

1.4 Literature Review

Credit-risk modelling has evolved gradually over several decades, influenced by regulatory standards, data availability, and methodological developments. Traditional statistical approaches still dominate practical credit-risk modelling, particularly in regulated banking environments, while machine-learning methods remain mostly complementary and are more firmly established in the academic literature than in regulatory practice.

Within this broader landscape, early models relied on financial ratios to construct simple, interpretable predictors of corporate failure. The Altman Z-score (**?**) combines financial ratios into a linear discriminant function that classifies firms as solvent or distressed. Later, Ohlson's O-score (**?**) introduced logistic regression to estimate bankruptcy probability using accounting and firm-specific variables. Logistic regression remains widely used due to its interpretability and regulatory acceptability. Though ratio-based and linear models impose fixed functional relationships and therefore struggle to capture non-linearities, interactions, or behavioural risk patterns. These limitations motivated the adoption of more flexible machine-learning approaches, for example tree-based methods and techniques designed to manage imbalanced credit datasets, such as SMOTE (**?**).

A major methodological shift occurred with **?** Gradient Boosting Machine, which underpins modern ensemble learners such as XGBoost and LightGBM. Empirical studies, including **?**, show that these boosting-based models consistently outperform traditional statistical approaches in predicting bank failures. Another is the bachelor thesis by **?**, who uses CatBoost to predict corporate credit ratings and concludes that boosting models offer strong predictive ability and meaningful interpretability for credit rating classification.

The growing empirical superiority of machine-learning techniques has intensified interest in explainable AI (XAI) as a way to overcome their inherent interpretability limitations. This broader issue is highlighted by **?**, who emphasize that many of the most accurate models function as black boxes, creating a tension between predictive performance and transparency. Their work demonstrates that while modern ensemble and gradient-boosted models achieve state-of-the-art accuracy across a wide range of domains, their opacity necessitates complementary XAI methods to ensure that model behaviour remains understandable and trustworthy.

The most relevant contribution for this thesis is the study by **?**, who apply XAI techniques to a Random Forest model predicting defaults among Italian non-financial firms. Their results show that predictor relevance varies across the eco-

nomic cycle, with liquidity and indebtedness becoming particularly influential during financial stress. By integrating permutation importance, Accumulated Local Effects (ALE), and Shapley values, they uncover the non-linear mechanisms driving firm-level distress and demonstrate that ML models not only outperform logistic regression but also provide deeper economic insight when paired with XAI. Complementary work by **?** and **?** further highlights how transparent ML frameworks and Shapley-based similarity networks can improve interpretability and decision-making in financial risk management.

1.5 Gap in the Literature

While **?** show that machine-learning models combined with XAI can uncover non-linear drivers of default, important gaps remain in the existing literature. Their paper focuses on Italian firms and there is limited evidence on how similar methods perform for Swedish companies. In addition, prior studies do not account for the more recent macro-financial environment, including the post-COVID period and the tightening of monetary policy. Although **?** demonstrate that predictor importance varies across the cycle, they provide only limited examination of the interaction structures that generate these patterns. Furthermore, earlier work relies on loan-level data, leaving open the question of how effectively modern ML and XAI methods can operate when only publicly available accounting information is used.

2 Data and Methodology

2.1 Data

2.1.1 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.

2.1.2 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.

2.1.3 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.

2.2 Methodological Framework

2.2.1 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).

2.2.2 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.

2.3 Model Specifications

2.3.1 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.

2.3.2 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.

2.4 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.

2.4.1 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.

2.4.2 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.

3 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. Third, we look at how feature SHAP importance varies over time to see if the model relies on different features during different regimes. Lastly, 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.

3.1 Model Performance Comparison

In Table 1, out-of-sample performance metrics for both models are presented. LightGBM achieves an AUC of 0.899, compared to 0.871 for the logistic regression, representing an overperformance of around 2.8 percentage points. The LogLoss improvement of around 7% likewise shows that the LightGBM model makes more accurate probability estimates. Critically, both models are well calibrated, with ECEs of below 0.5%, indicating taht predicted probabilities closely respond to observed default frequencies. This shows that the gains in distrimination and average probibalistic accuracy don't come at the expense of probibalistic reliability.

Table 1: Model Performance Comparison

Model	AUC	Log Loss	Brier Score	ECE
LightGBM	0.8990	0.0655	0.0158	0.0032
Logistic Regression	0.8712	0.0707	0.0165	0.0047
Δ (LightGBM − Logit)	+0.028	−0.005	−0.0006	−0.0015

Notes: Performance evaluated on 20% held-out test set (N = 60,830).

These gains in LogLoss and AUC are significant, but not groundbreaking, and both performance perform very well. Improvements at this level are consistent with gains made in benchmarking studies, such as Lessmann et al. (2015).

The logistic regression, which based on the previous section, is clearly a competitive model serving as a credible benchmark has the attractive property of being able to provide in-sample p-values and coefficients. Table **??** reports the coefficient estimates and p-values of this model using heteroscedasticity-consistent standard errors. Here we see values which one would expect to be significant appear, such as net margin, company age, and cash ratio. For example, we see that paying a dividend has a negative coefficient, suggesting lower probability of default (a value of 1 in the target variable corresponds to a default).

Table 2: Logistic Regression Coefficient Estimates

Feature	**Coefficient**	**Robust SE**	**Z-Score**	**P-Value**
Intercept	−5.905	0.117	−50.45	$< 10^{-300}$
Dividend Payer	−1.591	0.071	−22.41	3.06×10^{-111}
Industry (SNI)	38.108	1.777	21.45	4.99×10^{-102}
Net Profit Margin	−1.054	0.052	−20.39	2.22×10^{-92}
Interest Rate on Debt	0.472	0.031	15.16	7.00×10^{-52}
Company Age	−1.222	0.089	−13.76	4.34×10^{-43}
Days Payables Outstanding	0.395	0.029	13.50	1.60×10^{-41}
Log Cash & Bank	−1.006	0.083	−12.06	1.71×10^{-33}
Capital Turnover	0.565	0.053	10.70	9.81×10^{-27}
Depreciation Intensity	−0.688	0.066	−10.49	9.78×10^{-26}
Revenue Drawdown (5Y)	−0.384	0.040	−9.69	3.45×10^{-22}
Revenue CAGR (3Y)	0.209	0.026	8.03	9.45×10^{-16}
County	42.317	5.897	7.18	7.20×10^{-13}
Debt Ratio	0.182	0.029	6.22	4.93×10^{-10}
Days Sales Outstanding	−0.259	0.045	−5.78	7.67×10^{-9}
Revenue Growth (YoY)	0.111	0.020	5.66	1.48×10^{-8}
Cash Ratio	−1.842	0.326	−5.65	1.57×10^{-8}
Quick Ratio	−0.824	0.147	−5.59	2.27×10^{-8}
Retained Earnings / Equity	−0.115	0.024	−4.76	1.90×10^{-6}
Value Added per Employee	−0.511	0.109	−4.67	3.01×10^{-6}
Equity Ratio Δ (YoY)	−0.280	0.061	−4.63	3.61×10^{-6}
Term Spread	−0.306	0.070	−4.36	1.29×10^{-5}
Cash Ratio Δ (YoY)	0.288	0.084	3.43	6.14×10^{-4}
Cash Interest Coverage	0.505	0.184	2.74	6.08×10^{-3}
Log Total Equity	0.298	0.111	2.69	7.14×10^{-3}
Revenue per Employee	−0.097	0.053	−1.84	0.065
Profit CAGR (3Y)	−0.071	0.054	−1.32	0.188
Inventory Days Δ (YoY)	0.025	0.032	0.79	0.430
Equity Ratio	0.025	0.134	0.19	0.852
Return on Equity	0.001	0.028	0.05	0.958

Notably, several variables that financial theory would suggest should matter, perhaps most significantly the equity ratio and, fail to achieve statistical significance given the presence of the other variables in this model. Subsequent analysis reveals that the LightGBM will rely relatively more on these values, showing that different models end up discovering different signals in the dataset.

Several coefficients also appear to have the "wrong" sign, such as total equity and revenue growth, with higher values corresponding to a higher probability of default. This could for example be because of a mediating effect with another variable (or set of variables), or it could be a genuine signal.

3.2 Global Feature Importance

In Figure 1, we can see the difference in global mean SHAP values of differnet features between the two models. Because SHAP values decompose predictions into additive feature contributions on the log-odds scale, they permit direct comparison across model architectures. The features have been ordered by their importance ranking in the LightGBM model.

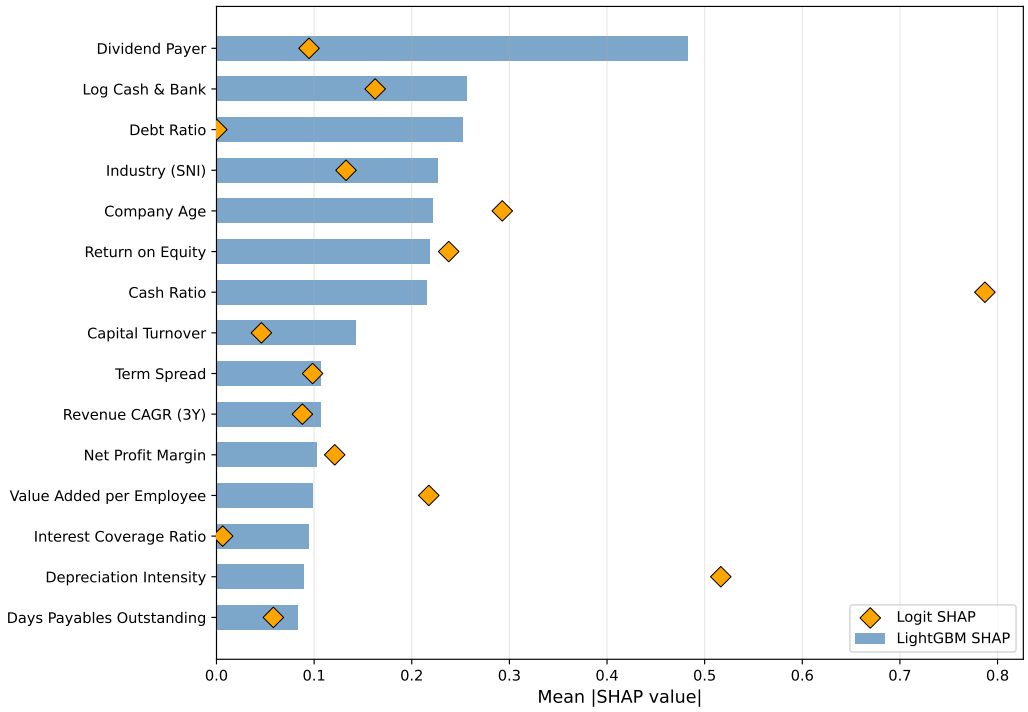

Figure 1: SHAP Feature Importance Comparison

The comparison reveals substantial divergence between the models. Whether or not a company pays a dividend emerges as the dominant predictor for LightGBM, with a mean absolute SHAP value approximately twice that of the next most important feature. The logit model instead prioritises cash ratio and the depreciation intensity of the company. The importance of cash ratio is obvious, but the prevalence of depreciation intensity might indicate that companies with larger investments are more prone to default, which is not entirely unreasonable.

It is generally difficult to find any clear pattern as to why one model seems to prefer one feature over another, likely partially reflecting the richness of the dataset. Patterns mapping to the same underlying signal might be discernable with multiple values.

If any structural differnece is to be discerned, it is that LightGBM generally distributes feature importance more evenly while the logit is more reliant on a few features. This is a rather logical consequence of the design of the two models, with the logistic regression imposing a linearity assumption on the data which often will not hold, while LightGBM is free to approximate any underlying distribution.

To examine not only importance magnitude but also effect direction and heterogeneity, Figure **??** presents SHAP summary plots for both models. Each point represents a firm-year observation, with the horizontal position representing the feature's contribution to the prediction (in log-odds), with positive values indicating it contributes to a default, while the colour of the point indicates the features value, low (blue) or high (red).

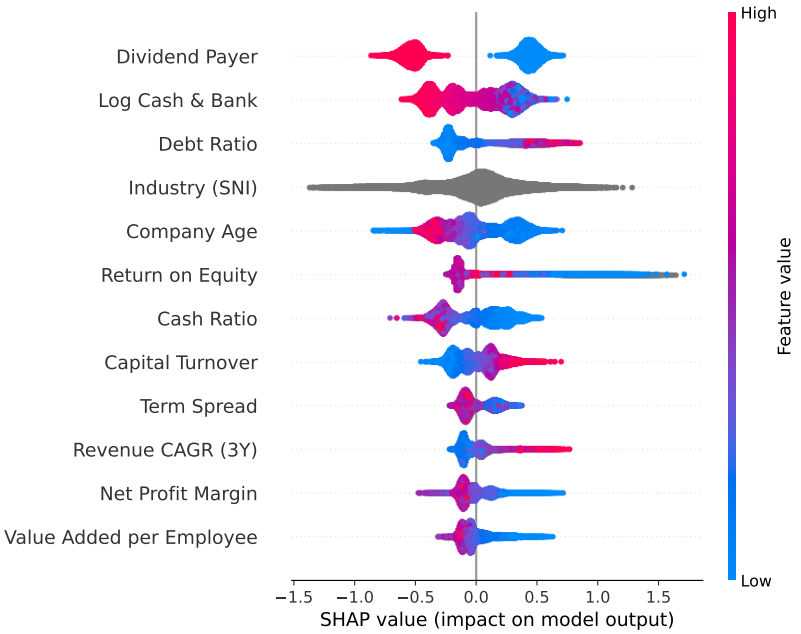

Figure 2: SHAP Summary Plot for LightGBM

Because the interacting and non-monotonic nature of the LightGBM model, we see that Figure 2 displayes considerably more heterogenous effects than Figure 3, which instead produces cleaner, more uniform gradients, a direct consequence of its linear structure. The greater dispersion in LightGBM reflects its capacity to learn context-dependent effects, that is the interaction of different features resulting in the same value having different implications depending on other firm characteristics.

The effect is clearly illustrated for company age, where we, in the LightGBM plot can clearly see low values contributing to both high and low predictions, and with the dividend payer dummy variable, which has a larger range of effects for the LightGBM model than the purely binary nature in the logsitic regression.

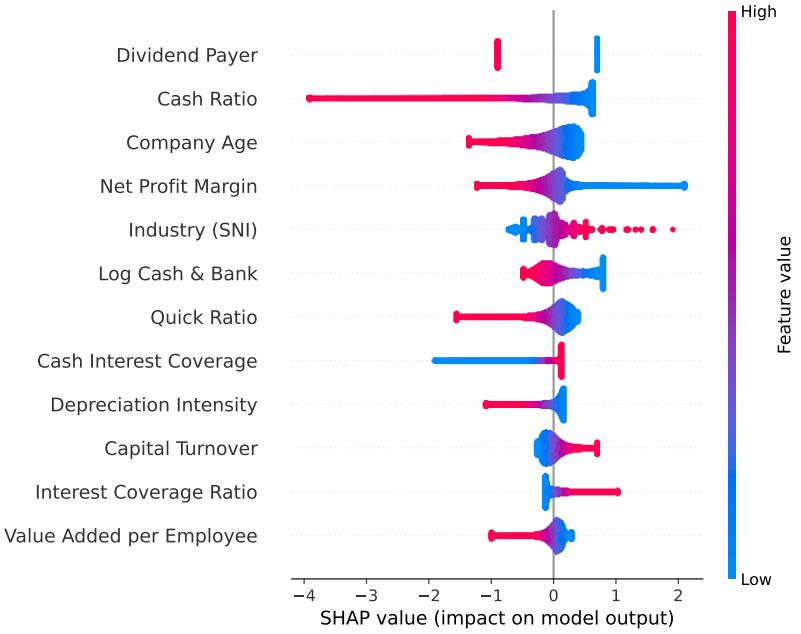

Figure 3: SHAP Summary Plot for Logistic Regression

3.3 Functional Form Case Studies

To examine how each feature influences default risk across its range, we employ Accumulated Local Effects (ALE) plots. We look at a few features which exhibit interesting ALE plots.

3.3.1 Net Profit Margin

Figure 4 presents ALE curves for the net profit margin, showing how this affects the probability of default, translated to be centered at zero.

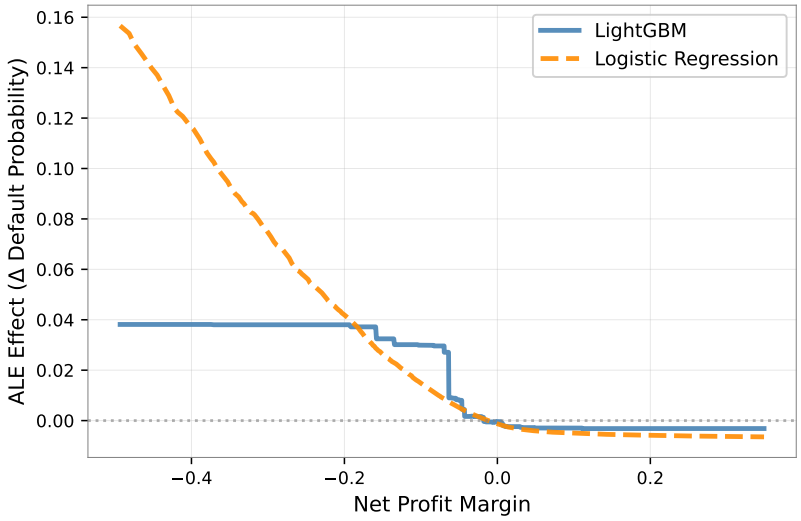

Figure 4: ALE Plot: Net Profit Margin

The logistic regression exhibits relatively predictable behaviour, with a negative

slope starting around a profit margin of 0, with worse profit margins translating to higher probabilities of default in a linear manner. LightGBM, on the other hand, seems to see the net profit margin as an almost binary value, with the maximum negative effect being achieved already around -10%, contributing a maximum of around 4bp to a next-year default.

This threshold behaviour is characteristic of tree-based models, which learn discrete splits rather than continuous relationships. However, this pattern may also reflect a genuine economic insight: once a firm is sufficiently unprofitable, the precise magnitude of losses becomes less informative for default prediction, as other indicators of distress—such as liquidity ratios, leverage, or payment behaviour—may become more salient. Since ALE plots are specifically designed to isolate the marginal effect of a single feature while accounting for correlations, this plateau likely represents diminishing marginal information content rather than a data artefact, though we cannot entirely rule out the latter interpretation.

3.3.2 Days Payable Outstanding

Figure 5 displays the ALE curves for days payable outstanding (DPO), measuring the average number of days a firm takes to pay its suppliers. The logistic regression

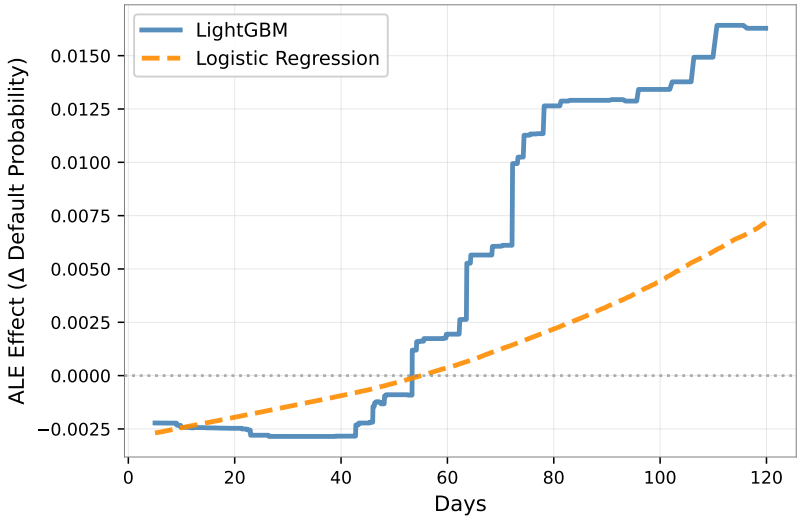

Figure 5: ALE Plot: Days Payable Outstanding

exhibits a gradual, approximately linear increase in default probability as DPO extends, consistent with the intuition that delayed supplier payments signal cash flow constraints. LightGBM reveals a more nuanced pattern: the effect remains relatively muted below 40 days, rises steeply between approximately 40 and 80 days, and then plateaus thereafter. The 40-day threshold may correspond to standard trade credit terms, beyond which delayed payment becomes a meaningful warning signal. The plateau after 80 days could reflect either a genuine diminishing marginal effect—once payment delays are severe, additional days provide little incremental information—or alternatively, data sparsity at extreme values reducing the model's ability to distinguish between varying degrees of severe delinquency.

3.3.3 Return on Equity

Figure 6 presents the ALE curves for return on equity (ROE), defined here as net profit divided by shareholders' equity. LightGBM exhibits a pronounced thresh-

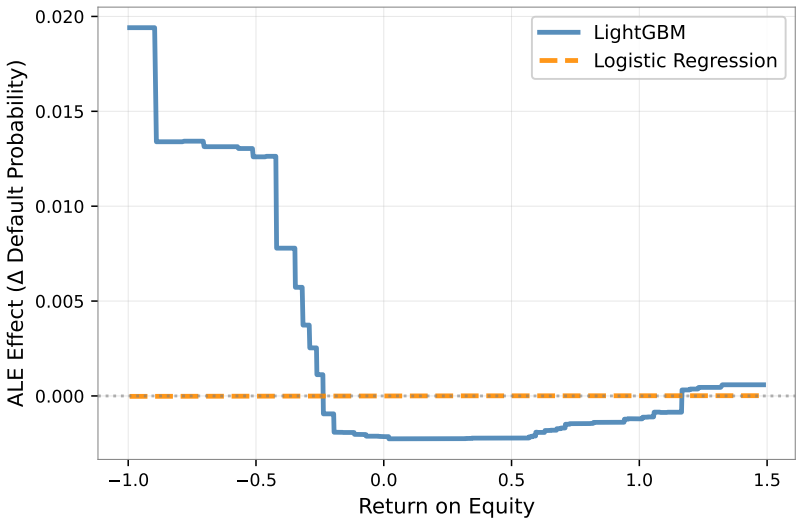

Figure 6: ALE Plot: Return on Equity

old effect centred around an ROE of approximately -0.25. Firms with ROE below this threshold face substantially elevated default risk, with the effect increasing sharply as ROE deteriorates further into negative territory. Above this threshold, the relationship flattens considerably, suggesting that once a firm achieves modest profitability relative to equity, further improvements contribute minimally to default risk reduction. A slight U-shaped uptick appears at very high ROE values (above approximately 1.0), which could indicate that exceptionally high returns are associated with volatile or leveraged business models carrying marginally elevated risk, though this pattern may also reflect data artefacts given the relative scarcity of observations at such extreme values. Notably, the logistic regression displays essentially no signal across the entire ROE range, likely because ROE is highly correlated with other profitability metrics in the model, with those alternative features capturing the relevant variation in the linear specification.

3.3.4 Company Age

Figure 7 shows the ALE curves for company age, measured in years since incorporation. LightGBM identifies firms between approximately 1 and 7 years old as carrying elevated default risk, consistent with the well-documented liability of newness in organisational ecology. After this period, the effect decreases with diminishing returns, stabilising for firms beyond roughly 30 years of age, reflecting survivorship bias, whereby firms that persist are systematically healthier. Strikingly, firms at year zero exhibit a pronounced negative effect on default probability, contrary to what one might initially expect. This likely reflects the fact that newly founded firms typically possess starting capital and have not yet had sufficient time to accumulate the operational difficulties or debt burdens that precipitate default.

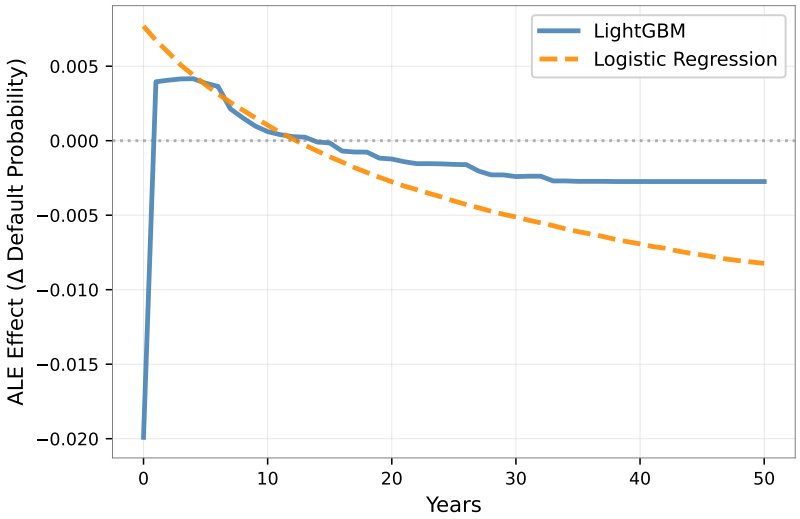

Figure 7: ALE Plot: Company Age

The logistic regression captures the general direction of this relationship but imposes a strictly linear decay, thereby missing the much lower initial value for firms with an age of 0, though one could use this insight to create a dummy variable to rectify this problem and thereby allow the logistic regression to capture this non-linearity.

3.4 Interaction Effects

A key theoretical advantage of tree-based models is their capacity to capture feature interactions without explicit specification. The preceding analysis documented LightGBM's ability to learn non-linear main effects; we now examine whether interactions contribute meaningfully to its predictions.

3.4.1 Pairwise Interaction Structure

Figure 8 presents a heatmap of mean absolute SHAP interaction values for the top 15 features.

The heatmap reveals a notable pattern: most cells remain light, indicating weak pairwise interactions. Even the more strongly interacting features display a mean interaction of around 0.08 log-odds unit, which is far less than one percent. This suggests that the model's predictions are predominantly driven by additive main effects rather than complex feature dependencies. The one feature which stands out is the industry code, with it interacting particularly strongly with the most significant variables (whether the firm pays dividends, and the log of its cash and bank).

This pattern is economically intuitive. Industry classification moderates the relationship between firm characteristics and default risk. Firms will have different characteristics, and be exposed to different risks in different industries. A young technology firm faces different survival dynamics than a young manufacturing firm, and capital turnover norms differ between retail and professional services. The

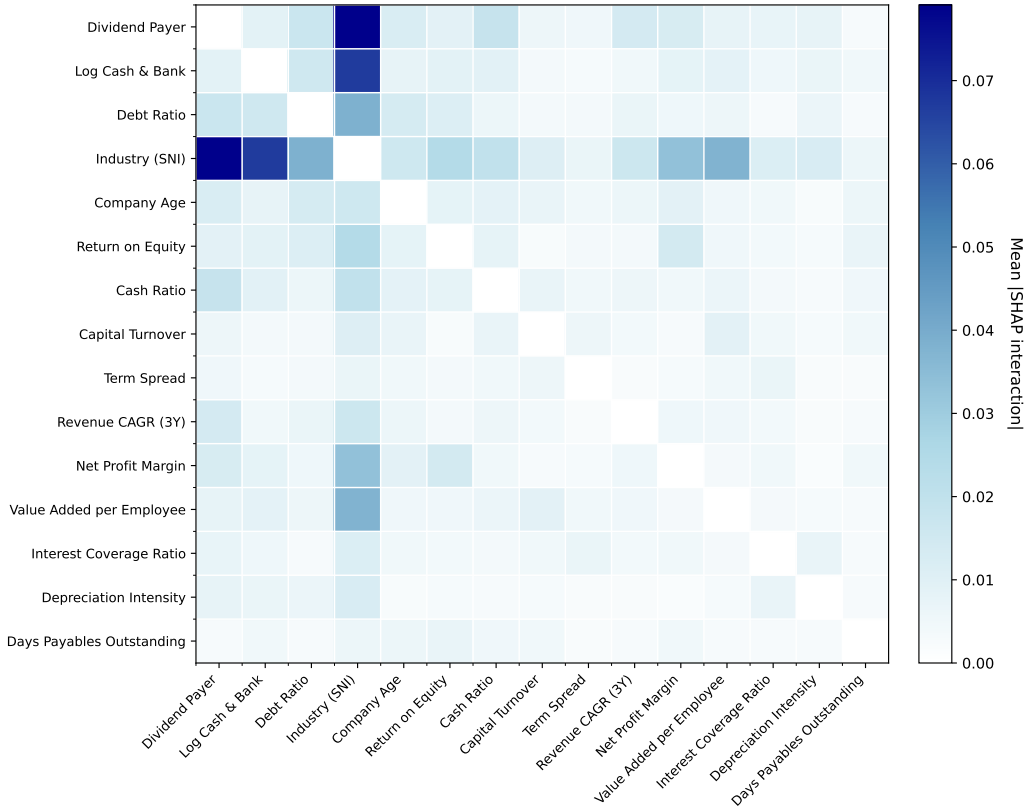

Figure 8: SHAP Interaction Heatmap (Top 15 Features)

LightGBM is able to capture these nuances, and is therefore able to go beyond simpler fixed effects, or target encoding used in logistic regression modelling.

3.4.2 The Collective Importance of Interactions

While individual pairwise interactions are generally weak, Figure 9 examines whether interactions matter in aggregate.

The left panel shows the distribution of interaction-to-main-effect ratios across all observations. The median ratio is 47.5%, meaning that for half of all firms, interaction effects are at least half as large as main effects. This is a substantial contribution that cannot be dismissed as negligible.

The center panel examines whether interactions become more important for high-risk predictions. The mean interaction ratio increases modestly from lower to higher risk deciles, but remains consistently above 50% throughout. Interactions contribute meaningfully across the entire risk spectrum, not merely for edge cases.

The right panel confirms that both main effects and interaction effects scale together: firms with larger main effect contributions also tend to have larger interaction contributions. Higher predicted default probabilities (darker points) cluster in regions with elevated effects of both types.

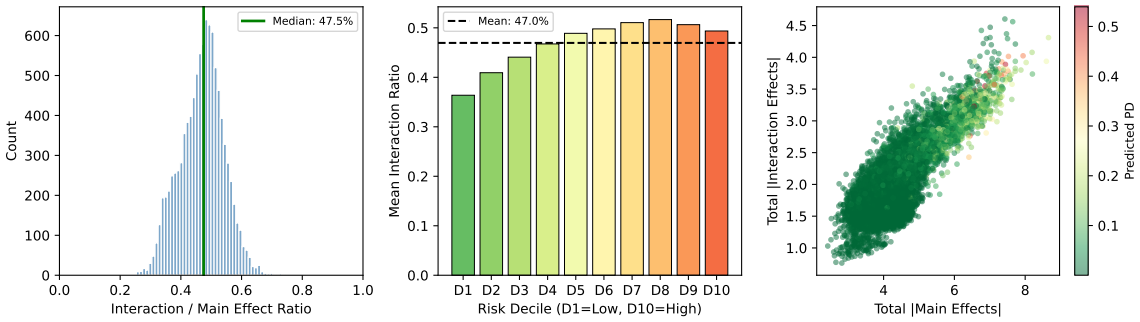

Figure 9: Distribution and Importance of SHAP Interactions
Left: Distribution of interaction-to-main-effect ratios across observations. Center: Mean interaction ratio by predicted risk decile. Right: Scatter of total main effects versus total interaction effects, colored by predicted probability.

3.4.3 Interpretation

These findings present a nuanced picture. Individual pairwise interactions are generally weak: no single feature pair dominates the model's behavior. Yet interactions collectively account for approximately half of the magnitude of main effects. The model captures many subtle, distributed interactions rather than a few dominant ones.

This structure has implications for model interpretability. For most practical purposes, the model can be understood as approximately additive: risk is roughly the sum of individual feature contributions. However, the interaction component is sufficiently large that treating predictions as purely additive would miss meaningful structure. The 2.8 percentage point AUC advantage over logistic regression reflects both the non-linear main effects documented in Section 3.3 and the distributed interaction effects quantified here.

From a regulatory perspective, the predominantly additive structure is reassuring. Simple feature attribution—explaining a prediction as the sum of individual contributions—remains defensible for most observations. The interaction component adds refinement without fundamentally restructuring how predictions should be understood.