



## Explainable Creditworthiness Scoring using a Calibrated Multimodal Framework

Ankur Agarwal, Shashi Prabha, Raghav Yadav

<sup>1,2,3</sup>Department of Computer Science and Information Technology,  
Sam Higginbottom University of Agriculture, Technology and Sciences, Prayagraj, Uttar Pradesh, India

\*[alld.ankur@gmail.com](mailto:alld.ankur@gmail.com)

### Abstract

*This research presents an explainable, calibrated multimodal framework for creditworthiness scoring. It integrates financial ratios with narrative risk disclosures by using the 823 firm SEC (Securities and Exchange Commission.) filings. In this approach we have used a class-weighted, L2-regularized logistic regression within a unified pipeline. It produces calibrated probabilities via isotonic regression. We have assessed the performance across discrimination (ROC-AUC, PR-AUC), reliability (Brier, calibration curves), operating thresholds (F1, asymmetric costs), and business diagnostics. We have also included the Decision Curve Analysis to cover the pattern of net benefit across practical cutoffs. Our implementation covers an ablation study to verify multimodal gains over financial-only and text-only baselines. The present research findings state that the LightGBM + SHAP benchmark depicts better interpretability along with improved prediction accuracy. Thus, the study provide an useful insights towards effective governance and transparent system requirements which makes this suitable for effective financial risk management.*

**Keywords:** credit risk scoring; explainable AI (XAI); financial disclosure; multimodal credit modeling

*How to Cite:* A. Agarwal, S. Prabha, and R. Yadav, "Explainable Creditworthiness Scoring using a Calibrated Multimodal Framework", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 3, pp. 677 - 688, Jun. 2026.  
*Permalink/DOI:* <https://doi.org/10.29207/resti.v10i3.7507>

*Received:* January 19, 2026

*Accepted:* April 10, 2026

*Available Online:* June 16, 2026

*This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)*

*Published by [Ikatan Ahli Informatika Indonesia](https://www.iaii.or.id/)*

### 1. Introduction

Credit risk models play a central role in capital allocation, loan pricing, and portfolio monitoring. These models are subjected to rigorous validation, transparency, calibration, and ongoing monitoring [1]. There is a well-known trade off in such models, for example, highly accurate models (boosted trees and deep networks) often lack transparency [2], [3]. However, traditional and generalized linear models remain interpretable but with limited performance [4]. In practice, many organizations require calibrated, decision-making models with strong interpretability so that they can withstand audit and governance scrutiny. Here, multimodal approaches which combine structured financial indicators and narrative disclosures are helpful and have become popular in recent years. In this field, prior research shows that accounting-based measures of liquidity, leverage, and profitability are strong predictors of financial distress [5-7], with market and structure-based models [8], [9]. At the same time, mandated financial disclosures, particularly risk factor

narratives, contain forward-looking information related to regulation, supply chains, cyber risk, and litigation [10-12]. Recent advances in explainable AI, ranging from coefficient-based interpretation to Shapley-value methods, allow predictions to be linked directly to underlying evidence [13-15]. However, several important challenges remain in existing credit risk modeling approaches, including leak-free integration of textual features, reliable probability estimation suitable for policy-driven decision thresholds, and generation of faithful explanations that satisfy regulatory and governance requirements [16], [17]. To address these challenges, we propose an explainable and calibrated multimodal framework for creditworthiness scoring that integrates structured financial variables with narrative risk disclosures while ensuring leak-free learning, decision-ready probability estimation, and governance-oriented explainability.

Early studies established that firm financial ratios contain strong predictive signals of financial distress. Beaver [18] demonstrated the discriminative power of

liquidity and profitability measures, while Altman's Z-score [19] combined multiple ratios into a parsimonious, auditable scoring model that remains influential in practice. Subsequent work replaced discriminant analysis with logistic regression to improve probabilistic interpretation and calibration, and extended static models to hazard-based frameworks that capture time-to-default dynamics. Market-based and structural models further complemented accounting signals, with evidence that market-implied default probabilities can outperform purely financial models in some settings [20].

Despite advances in machine learning, logistic regression and scorecards continue to dominate regulated credit environments due to their stability, interpretability, and governance acceptance [21]. Benchmarking studies show that tree ensembles and gradient-boosting methods often achieve superior discrimination, although performance rankings are sensitive to validation design, class imbalance, and cost-sensitive evaluation choices [22].

Parallel research demonstrates that textual disclosures provide incremental risk information. Analyses of 10-K filings show that narrative content can predict volatility and complement quantitative indicators [23], while finance-specific dictionaries improve the interpretation of tone and risk in corporate disclosures [24]. Risk Factor (Item 1A) narratives have been linked to post-filing volatility, trading activity, and dispersion, indicating that mandated disclosures contain forward-looking risk signals [25].

Explainable AI has become central to making risk models auditable. SHAP-based methods provide consistent local and global attributions for complex learners [26], while linear models remain attractive in regulated contexts because coefficients serve as exact, global explanations with additive local decompositions. In parallel, probability calibration methods such as isotonic regression improve reliability beyond rank-based metrics [27], [28], and Decision Curve Analysis translates predicted probabilities into net benefit under asymmetric costs and varying policy thresholds [29]. Model governance frameworks further emphasize validation, calibration, and ongoing monitoring, with drift detection commonly operationalized using metrics such as the Population Stability Index (PSI) [30], [31].

Recent studies have increased the range of application of Explainable Artificial Intelligence (AI) in Credit Risk Analysis beyond traditional single-modal Financial Scorecards. In addition, Kocoglu and Ersoz [32] compared SHAP and LIME for predicting Commercial Credit Limits, where they stated that although both models provide Local Interpretability, SHAP is significantly better than LIME because it offers a Theoretically Consistent and Stable Attribution Framework, which motivated us to use SHAP for explaining Non-Linear Models. Additionally, recent Multimodal Approaches have included Structured Financial Data and Textual Earnings Call Transcripts

and Management Discussion Sections. However, most studies have focused solely on Optimizing Predictive Accuracy of their proposed method without considering Information Leakage in Label Construction, Post-Hoc Probability Calibration, or Decision-Oriented Evaluation Under Asymmetric Misclassification Costs. Our proposed framework addresses each of these four aspects -- Multimodal Fusion, Leak-Free Learning, Calibrated Probability Estimation, and Governance Oriented Explainability -- in a Single Reproducible Pipeline Applied to SEC EDGAR Filings [32].

Despite growing interest in multimodal credit risk modeling, existing studies often focus on predictive accuracy, financial text analytics, or interpretability in isolation. Many approaches integrate financial indicators and textual disclosures without addressing information leakage, decision-ready probability calibration, or governance-oriented deployment. Moreover, limited work systematically evaluates the incremental contribution of narrative disclosures within a unified explainable framework [33].

To address these gaps, this study proposes an explainable and calibrated multimodal framework that integrates structured financial variables and narrative risk disclosures while ensuring interpretability and decision reliability. The framework introduces a leak-free learning design by constructing distill labels solely from financial variables, jointly incorporates multimodal feature fusion, probability calibration, and explainability, and supports deployment through stability monitoring and decision curve analysis. In addition, ablation experiments evaluate the marginal contribution of textual disclosures beyond traditional financial indicators. Accordingly, the main contributions of this study are summarized in the paragraph mentioned below.

A leak-free multimodal framework integrating financial ratios and narrative risk disclosures for creditworthiness prediction. A unified pipeline combining explainable modeling, probability calibration, and decision-oriented evaluation. An empirical analysis of modality contributions through ablation studies to quantify the value of multimodal integration. A governance-oriented evaluation framework incorporating calibration, explainability, and stability monitoring for real-world credit risk deployment.

The remainder of the paper is structured as follows. Section 2 describes the proposed methods, covering data extraction, leak-free labeling, and the modeling pipeline with calibration, thresholding, and evaluation procedures. Section 4 presents empirical results. Section 5 concludes with policy and governance implications, limitations, and directions for future research.

## 2. Methods

The framework aims to build a predictive and explainable creditworthiness scoring system. It starts

with data integration of financial ratios, binary disclosures, and narrative text, followed by weak supervision to define distress labels from financial data without text leakage. Preprocessing involves imputation, normalization, TF-IDF, and feature fusion. In the modelling stage, Logistic Regression serves as the explainable base-line, while LightGBM with SHAP provides a non-linear benchmark. Evaluation is performed using ROC-AUC, PR-AUC, precision, recall, F1, calibration, and decision curve analysis. Finally, explainability and robustness are ensured through SHAP values, temporal validation, PSI studies, and ablation experiments. The workflow of the proposed methodology is shown in Figure 1.

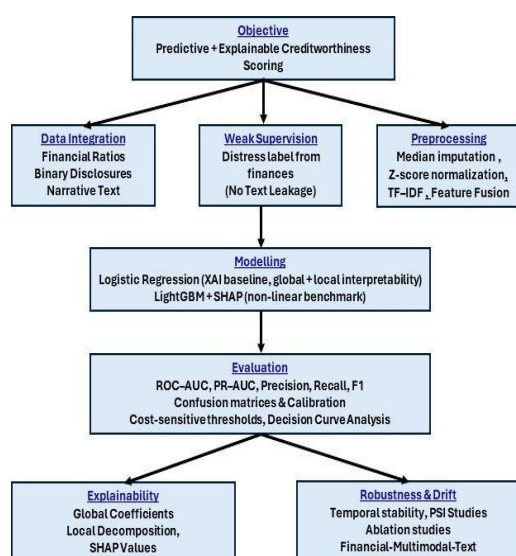


Figure. 1 Workflow of the proposed methodology.

## 2.1 Data Acquisition and Integration

The research utilizes U.S. Securities and Exchange Commission (SEC), EDGAR repositories to access corporate filings through the SEC EDGAR API [34]. We are using this API to automate the process of extracting financial disclosures. The extraction pipeline integrates structured parsing, exception handling, and multiple output formats. As shown in Figure 2, filings are retrieved from the SEC EDGAR database. Later it passed through the extraction engine for region and identifier detection. The extracted filings were exported in CSV format for easy processing. exported in formats such as XML, JSON, TXT, and CSV. Following are the extracted components: Narrative disclosures: Item 1A: Risk Factors, capturing qualitative risk statements. Structured financials: Inline XBRL (iXBRL) financial statements.

For each filing, metadata includes accession number, company name, Central Index Key (CIK), reporting period, and form type. Later, preprocessing of narrative sections was performed. This includes tokenization, lower casing, stop word removal. We also perform the keyword matching to flag risk categories such as liquidity, default, regulatory, cybersecurity, and geopolitical risks. Parallely, structured financial indicators such as assets, liabilities, equity, net income,

and cash flows are transformed into financial ratios like Debt-to-Equity, Current Ratio, Return on Equity and Return on Assets.

The heterogeneous features such as metadata, financial ratios, binary disclosure flags, and text embeddings are then integrated into a unified multimodal dataset, with each filing represented as a single row. To ensure that the 823 firm-period filings included in this study could be used appropriately within a multimodal model framework, a two-step (bifurcated) preprocessing methodology was developed. In the first step, all financial numeric data were normalized using z-scores so that each feature had an average value of 0 and a variance of 1. The application of z-scoring is important for both the computational efficiency and reliability of L2-regularized logistic regression models.

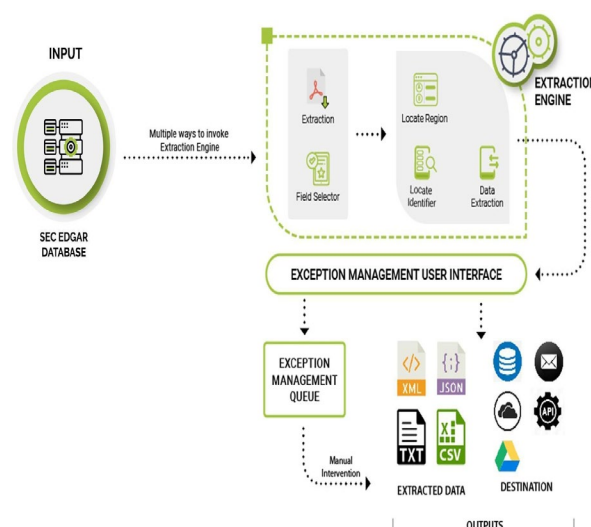


Figure 2. Automated SEC-EDGAR data extraction pipeline with exception management and output formats.

Numeric missingness within the same set of numeric variables were handled through the use of column medians for imputation purposes; because of the symmetric nature of the median, it is reasonable to expect that the overall distributions would remain preserved by this method and therefore less likely than other methods to introduce bias. A standard natural language processing (NLP) "cleaning" methodology was then applied to extract the risk factor information contained in Item 1A. Tokenization was performed on the Item 1A text along with converting all tokens to lowercase case. All stop words and tokens containing non-alphanumeric characters were removed. Finally, the clean text was vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) with a maximum vocabulary size of 1000 terms. After cleaning, we focus on ten U.S. large-cap issuers:

$T = \{AAPL, AMZN, AVGO, GOOGL, JPM, MSFT, TSLA, NVDA, META, NFLX\}$ ,

This includes 823 firm-period filings and the final dataset contains 24 columns, covering: (i) structured financials, (ii) binary disclosures, and (iii) compact narrative summaries derived from Item 1A. This

integration balances quantitative rigor with qualitative foresights to support multimodal modeling of financial distress [35], [36].

## 2.2 Label Construction

Equations 1–6 represent the key parameters considered for constructing the labels. To avoid label leakage, the distress label  $y \in \{0, 1\}$  is constructed exclusively from financial variables. Here, textual disclosures are not used in outcome definition.

$$y = I^h(\text{NI} \leq 0) \vee (\text{CR} < 1.0 \wedge \text{D/E} > 2.0) \vee (\text{CFO} \leq 0 \wedge \text{ROE} < 0)^i, \quad (1)$$

where:

$$\text{NI} = \text{Net Income}, \quad (2)$$

$$\text{CR} = \frac{\text{CA}}{\text{CL}}, \text{ Current Ratio}, \quad (3)$$

$$\text{D/E} = \frac{\text{T L}}{\text{SE}}, \text{ Debt-to-Equity Ratio}, \quad (4)$$

$$\text{CFO} = \text{Cash Flow from Operations}, \quad (5)$$

$$\text{ROE} = \frac{\text{NI}}{\text{SE}} \text{ Return on Equity}. \quad (6)$$

## 2.3 Financial Ratios as Predictive Features

Equation 1 captures three dimensions of distress such as profitability stress, liquidity- leverage interaction and weak profitability with negative operating cash flows in following way:

Profitability stress:  $\text{NI} \leq 0$

Illiquidity with leverage risk:  $\text{CR} < 1.0 \wedge \text{D/E} > 2.0$

Weak profitability with negative cash flows:  $\text{CFO} \leq 0 \wedge \text{ROE} < 0$

We compute the standard financial ratios presented in Equations 7 and 8, which are widely used in credit risk assessment and financial analysis [35, 36].

Efficiency (Asset Turnover):

$$\text{AT} = \frac{\text{SR}}{\text{TA}} \quad (7)$$

Solvency (Equity Ratio):

$$\text{ER} = \frac{\text{SE}}{\text{TA}}. \quad (8)$$

These ratios represent efficiency and solvency, and complementing qualitative disclosures.

## 2.4 Preprocessing and Feature Engineering

Equation 9 presents the z-score normalization method used to standardize the numerical features:

$$x' = \frac{x - \mu}{\sigma}, \quad (9)$$

$x$  is the original feature value,  $\mu$  is the mean of the feature, and  $\sigma$  is the standard deviation of the feature. This transformation scales the data to have a mean of zero and a standard deviation of one.

Equation 10 calculates the TF-IDF weight of each term, highlighting words that are important within a

document while reducing the influence of commonly occurring terms across the corpus [37].

$$w_{t,d} = \text{tf}(t, d) \cdot \log \frac{N}{1 + nt}. \quad (10)$$

Equation 11 combines the financial, binary, and textual features into a single multimodal feature vector, enabling the model to utilize information from multiple data sources simultaneously.

$$X = [x_{fin}, x_{bin}, x_{txt}], \quad (11)$$

$x_f$  in are ratios,  $x_{bin}$  are binary risk flags, and  $x_{txt}$  are TF-IDF embeddings.

Early Fusion, which can be defined as Feature-Level Concatenation, is utilized to merge the various forms of the data that will be fed into the system. During Early Fusion, the Normalized Financial Ratio Vector ( $X_{fin}$ ), Binary Risk Flag Vector ( $X_{bin}$ ), and Sparse TF-IDF Text Vector ( $X_{txt}$ ) are combined to form a High-Dimensional Multimodal Input Vector ( $X$ ). The vectors are merged into one long vector so they can then be sent to the Classifier for processing.

Early Fusion was chosen due to its ability to create Joint Representations for all Modalities during Training. When learning Joint Representations, the model learns how quantitative measures of financial stability can confirm or contradict qualitative risk warnings expressed in the Narrative Filings. By merging the Heterogeneous Signals at the Input Level, the Pipeline maintains Cross-Modal Interactions that a Single Modality Model would otherwise ignore.

## 2.5 Baseline Logistic Regression and Ablation Studies

We employ a class-weighted, L2-regularized logistic regression as the baseline predictive model. Logistic regression is a widely used benchmark in credit risk and bankruptcy prediction tasks. It is widely used techniques due to its transparency, interpretability, and regulatory acceptance [38]. Beginning with the application of class weights to handle the naturally occurring class imbalance between distressed and non-distressed firm-periods in the data; this will help prevent the model from weighting the majority class. The model estimates the probability of financial distress using the logistic function, as shown in Equation 12.

$$P(y = 1 | X) = \sigma(\beta^T X + b), \sigma(z) = \frac{1}{1 + e^{-z}} \quad (12)$$

where  $\beta$  are the model coefficients,  $b$  is the intercept, and  $\sigma(\cdot)$  is the logistic sigmoid function mapping real values into the (0, 1) probability range.

The optimization objective is presented in equation 13.

$$\min \beta - \sum_{i=1}^N [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] + \frac{\lambda}{2} \|\beta\|^2 \quad (13)$$

$\hat{p}_i$  is the predicted probability of distress for observation  $i$ , and  $\lambda$  is the L2- regularization parameter that penalizes large coefficients, thus preventing overfitting and improving generalization. To assess the

contribution of each data modality, we conduct ablation experiments using three logistic regression variants: Structured-only model which incorporates financial ratios and binary indicators (LR\_num); A text-only model based on TF-IDF representations of Item 1A risk disclosures (LR\_txt); A multimodal model which structured and textual features (LR\_all).

Model performance is evaluated using ROC-AUC for overall discrimination, PR-AUC to account for class imbalance, and the Brier Score to assess probability calibration [39]. All these metrics capture both discrimination (ranking of firms by risk) and calibration (reliability of predicted probabilities). The ablation results thus quantify the incremental benefit of multimodality, i.e., the extent to which textual risk disclosures provide predictive power beyond structured financial indicators.

## 2.6 Calibration, Thresholding, and Decision Analysis

Accurate probability estimates are essential for credit risk management, as decisions depend not only on ranking but also on the reliability of predicted risks [40]. While logistic regression typically yields well-calibrated probabilities, non-linear models may require post-hoc calibration [41], [42]; accordingly, isotonic regression is used to align predicted scores with observed event frequencies and evaluated using the Brier Score [39].

To support operational decision-making, classification thresholds are optimized under asymmetric misclassification costs, reflecting the higher penalty of missed defaults. Practical utility is further assessed using Decision Curve Analysis, which evaluates net benefit across thresholds by explicitly balancing true positive gains against false positive costs, providing a decision-relevant complement to discrimination metrics [43].

## 2.7 Explainability and Robustness

Explainability and robustness are extremely important for the deployment of credit risk models. In regulated environments, stakeholders require not only high predictive accuracy but also transparent reasoning behind each model decision and temporal stability across deployment periods [14][44]. In case of logistic regression baseline, interpretability is essential. Here, model coefficients provide global explanations with the help of quantification of the direction and magnitude of each feature's effect on distress risk.

Likewise local explanations are obtained through additive decompositions that attribute individual predictions to specific feature contributions. These types of attributions support auditability and regulatory review. For non-linear models, explainability is ensured using SHAP values, which provide consistent global and local attributions for complex learners such as LightGBM. SHAP analysis enables identification of dominant risk drivers across firms as well as detailed breakdowns of individual predictions. It preserves

interpretability despite model nonlinearity. In this study, we monitor distributional shifts between training and test data using the Population Stability Index (PSI) [44]. This quantifies changes in the distribution of predicted scores. It helps to detect mild to severe drifts. Here we combine coefficient-based explanations for linear models and SHAP-based attributions for non-linear models. As a result this framework supports both interpretability and temporal stability which makes it suitable for governance-grade credit risk deployment.

## 2.8 Nonlinear benchmarking using LightGBM

To complement the transparent but linear logistic regression baseline, we employ LightGBM [3] as a non-linear benchmark. LightGBM is a gradient boosting framework that builds additive tree-based models in a stage-wise manner, optimizing a differentiable loss function while incorporating regularization to control model complexity. This comparison allows us to assess the trade-off between predictive performance and interpretability when moving from linear to non-linear models under an identical feature set. The LightGBM training objective minimizes the empirical loss over all observations while incorporating a regularization term on the ensemble of trees, as shown in Equation 14.

$$L = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (14)$$

where  $l(\cdot)$  denotes the pointwise loss function (logistic loss for binary classification),  $\hat{y}_i$  is the model prediction for observation  $i$ , and  $\Omega(f_k)$  is a regularization term that penalizes the complexity of tree  $k$ .

This formulation enables LightGBM to capture non-linear feature interactions while mitigating overfitting through explicit regularization. LightGBM is well suited for structured tabular data such as financial ratios, as it effectively handles high-dimensional and heterogeneous features while capturing non-linear interactions among predictors. In this study, it is used as a non-linear benchmark to the transparent logistic regression baseline, enabling a direct comparison between interpretability and predictive performance under an identical feature set. While logistic regression provides governance-grade transparency through coefficient-based explanations, LightGBM offers greater modeling flexibility and the potential for improved accuracy. This comparison highlights the fundamental trade-off in credit risk modeling between interpretability required for regulatory acceptance and predictive performance needed for operational deployment.

# 3. Results and Discussions

## 3.1 Dataset Description

The dataset consists of 822 firm-period observations across 24 variables, combining identifiers, financial indicators, and narrative risk disclosures. Each record is associated with a company ticker, form type (10-K/10-Q), and filing metadata. Financial attributes include balance sheet items (total assets, total liabilities, equity,

current assets, and current liabilities), profitability measures (net income, return on equity, return on assets), and liquidity indicators (current ratio, debt to equity ratio, cash flow from operations). All financial variables are complete with no missing values, ensuring reliability for quantitative analysis. The narrative dimension is more sparse. The risk factors summary field, extracted from Item 1A of SEC filings, has approximately 52% missing entries. In addition, binary indicators were created from the narratives to flag specific themes: regulatory risk (5.7%), supply chain disruptions (5.2%), COVID-19 (2.2%), liquidity crisis (0.6%), and default risk (0.1%). These indicators are highly imbalanced, underscoring the rarity of certain risk mentions in disclosures. Table 1 presents descriptive statistics of numeric variables, while Table 2 reports the class distribution of binary narrative risk indicators. Together, these summaries highlight the heterogeneity of financial magnitudes across firms and the sparsity of rare narrative risk events.

Table 1. Descriptive statistics of numeric variables

| Variable          | Mean     | Std      | Min      | Max      |
|-------------------|----------|----------|----------|----------|
| cik               | 9.46e+05 | 5.29e+05 | 1.96e+04 | 1.73e+06 |
| Period of report  | 2.02e+07 | 5.68e+04 | 1.99e+07 | 2.03e+07 |
| Filed as of date  | 2.02e+07 | 5.67e+04 | 2.00e+07 | 2.03e+07 |
| Total assets      | 1.17e+11 | 1.72e+11 | 1.00e+00 | 6.82e+11 |
| Total liabilities | 4.62e+10 | 8.52e+10 | 1.00e+00 | 3.09e+11 |
| Net income        | 1.15e+10 | 2.22e+10 | 0.00e+00 | 9.98e+10 |

Table 2. Class balance of binary narrative risk indicator

| Variable                  | +ve | -ve | Total | Rate (%) |
|---------------------------|-----|-----|-------|----------|
| Mentions regulatory       | 47  | 775 | 822   | 5.72     |
| Mentions supply chain     | 43  | 779 | 822   | 5.23     |
| Mentions COVID-19         | 18  | 804 | 822   | 2.19     |
| Contains liquidity crisis | 5   | 817 | 822   | 0.61     |
| Contains default risk     | 1   | 821 | 822   | 0.12     |

Figure 3 further illustrates the distribution of firm-period observations across the top ten tickers, highlighting that the dataset is concentrated in a small number of large-cap issuers.

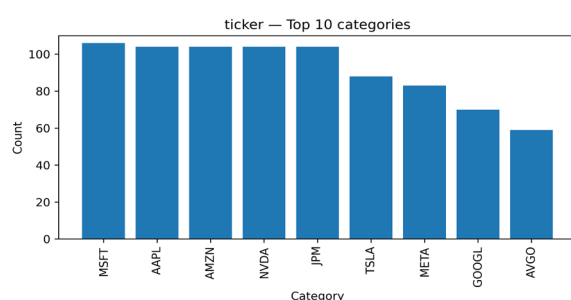


Figure 3. Top 10 tickers by filing frequency.

### 3.2 Ablation study results of Logistic Regression and its variants

Table 3 presents the core results of the logistic regression ablation study across different feature modalities, including numerical, boolean, and textual inputs, as well as their combinations under various regularization schemes. The results show that models based on numerical features consistently achieve the strongest performance, with the LR-L1 (num) configuration obtaining the highest ROC-AUC (0.983) and PR-AUC (0.963). It is also evident that adding boolean or textual features to numerical inputs leads to only marginal changes in performance. It reveals that the predictive signal is largely depends upon numerical attributes. In contrast, the models which depends on textual features exhibit a noticeable reduction in performance. Likewise, boolean-only models perform poorly across all the metrics and highlight their limited discriminative power in isolation.

Table 3. Logistic Regression Ablation Study – Core Performance Metrics

| Model                    | ROC-AUC | PR-AUC | F1@0.5 |
|--------------------------|---------|--------|--------|
| LR-L1 (num)              | 0.983   | 0.963  | 0.814  |
| LR-L2 (num+bool)         | 0.972   | 0.924  | 0.793  |
| LR-L2 (num+txt)          | 0.972   | 0.925  | 0.750  |
| LR-L2 (num+bool+txt)     | 0.972   | 0.924  | 0.793  |
| LR-ENet50 (num+bool+txt) | 0.972   | 0.924  | 0.814  |
| LR-L2 (txt)              | 0.864   | 0.720  | 0.682  |
| LR-L2 (bool)             | 0.527   | 0.369  | 0.470  |

Figure 4 presents the results of an ablation study which compares the predictive power of different feature modalities. It can be observed that numerical features (num) provide the strongest individual signal and achieve the highest PR-AUC and F1 scores among single modalities. Likewise, boolean disclosures (bool) and text features (txt) contribute in positive manner but they are less predictive on their own. It is important to note that the combination of modalities leads to consistent performance gains, with num+bool, num+txt, and the full multimodal setting (num+bool+txt) outperform single-modality baselines. This highlights the value of the integration of the heterogeneous data sources for creditworthiness prediction.

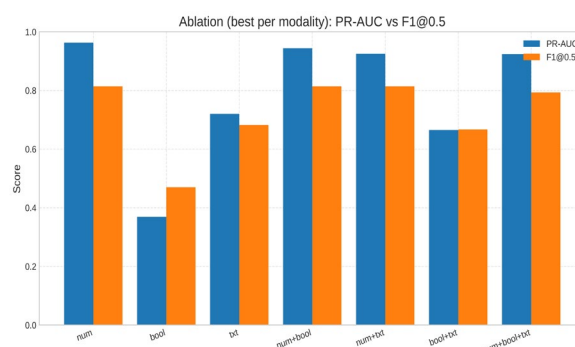


Figure 4. Ablation study results.

The Pearson correlation heatmap in Figure 5 illustrates strong positive correlations between total assets and total liabilities as well as a high degree of similarity between net income and return on equity based upon a direct relationship from definition. The data also show very little correlation between return on equity and the debt-to-equity ratio; this finding confirms that both leverage and profitability capture different forms of risk. The lack of inter-feature correlation among the major predictor variables also supports the notion that multicollinearity will have an extremely small impact on the stability of the logistic regression coefficients thereby supporting the use of coefficient-based explanations for model output.

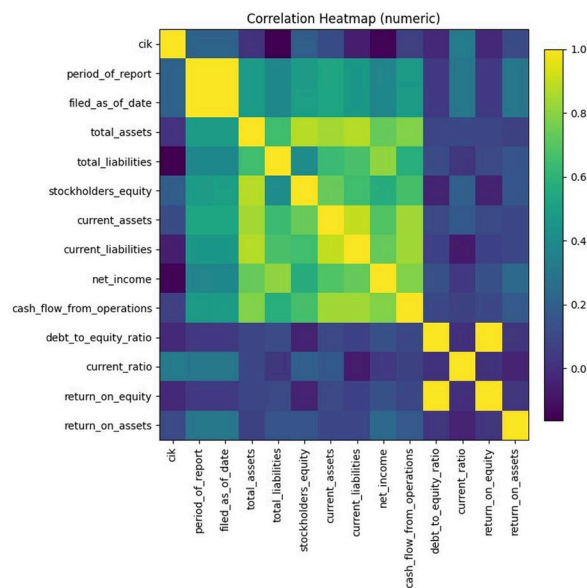


Figure 5. Correlation heatmap of numeric features.

It is clear from the ablation study results that the models relying solely on numeric features consistently achieve the highest discriminative power, with ROC-AUC  $\approx$  0.98 and PR-AUC above 0.92. Adding boolean disclosures or textual features yields marginal or mixed improvements, with certain combinations (e.g., num+bool) achieving slightly higher threshold-dependent performance ( $F1@best = 0.955$ ) due to better precision-recall trade-offs. In contrast, text-only and bool+text models perform substantially worse (ROC-AUC in the 0.73–0.86 range), underscoring the relatively weaker predictive contribution of unstructured narratives in isolation. The confusion matrices in Figure 6 illustrate the distribution of errors for representative models (LR-L1-num and LR-L1-num+bool). Numeric-driven models achieve lower false negative rates, while text-augmented variants exhibit higher misclassifications. These visualizations complement the ablation table by highlighting trade-offs across feature modalities. The evaluation is further illustrated through enhanced comparative visualizations. The Precision-Recall scatter plot in Figure 7 is plotted for a single fixed-threshold (threshold = 0.5) value. Each point on the scatter plot corresponds to one of the different configurations used

to train each of the different models. Points are colored based upon the input modality, while points are represented as triangles or circles depending upon whether L2 (ridge) or L1 (Lasso) penalty was applied to training. Contour lines corresponding to iso-F1 values were added to both plots. These contour lines correspond to pairs of precision and recall values that would result in the same F1 score. The best performing models in terms of F1 score are annotated with non-overlapping bubbles.

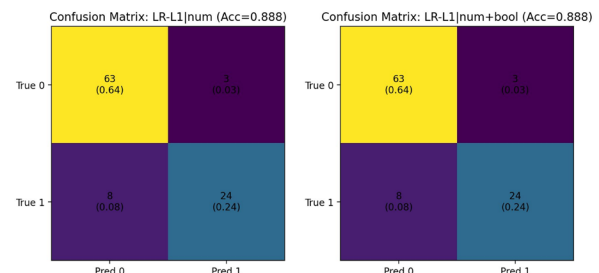


Figure 6. Confusion matrices for (left) LR-L1-num and (right) LR-L1-num+bool models.

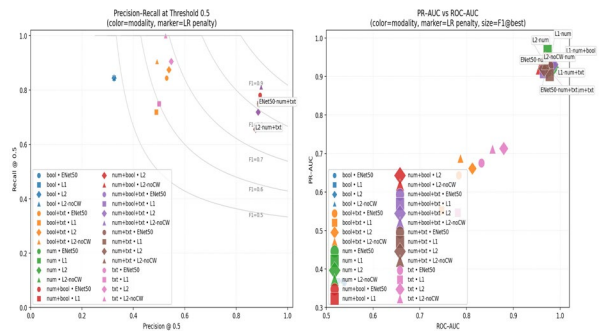


Figure 7. (Left) Precision-Recall scatter with iso-F1 contours. (Right) PR-AUC vs ROC-AUC bubble plot with  $F1@best$  as bubble size.

Two additional plots show the relationships between PR-AUC and ROC-AUC, once again using colors and shapes to distinguish between the input modalities and the penalty types. Bubble sizes have been set proportional to the  $F1@best$  metric, thus capturing how effective each model's ability to effectively classify samples is when evaluated under varying thresholds. As such, the four plots shown here collectively create an interpretable representation of the various behaviors exhibited by all of the different multimodal logistic regression-based models. In particular, they allow for easy comparisons regarding the tradeoffs inherent in achieving high levels of precision, high levels of recall, and high overall discriminative abilities among the different multimodal logistic regression variants.

### 3.3 Explainability: Global, Local, and Modal Attribution

Figure 8 presents the top-40 standardized logistic regression coefficients. Positive coefficients (red) increase predicted distress risk, while negative coefficients (blue) reduce it. The dominant risk-enhancing driver is leverage ratio, followed by current liabilities, net income, and cash flow from operations.

In contrast, strong protective signals are observed for return on assets and current assets, which substantially lower estimated risk.

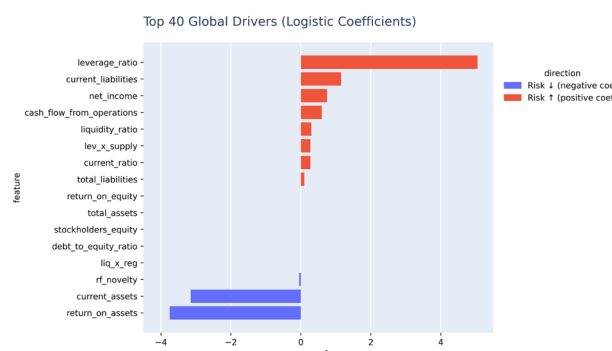


Figure 8 Top-40 logistic regression coefficients. Red bars increase predicted risk; blue bars are protective.

At the firm level, Figure 9 illustrates the local explanation for the highest-risk case (JPM). The log-odds are primarily driven upward by large positive contributions from leverage ratio, current assets, and return on assets. These effects are partially offset by negative contributions from current liabilities, net income, and cash flow from operations, as well as smaller stabilizing effects from liquidity and capital structure variables.

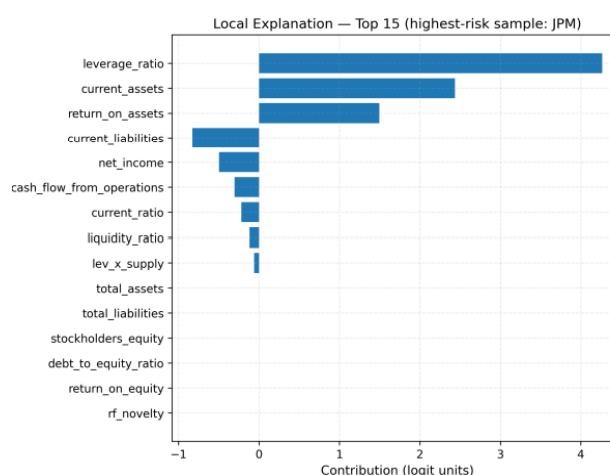


Figure 9. Local explanation for the highest-risk firm (JPM). Bars indicate feature contributions to logodds.

The cumulative effect of these contributions is shown in the local waterfall plot (Figure 10). Here, the net logit is strongly positive, confirming the high predicted distress probability. Here red bars shows increase in risk and blue bars decrease it. This figure shows how each financial ratio contributed to the model's decision for one firm. The leverage ratio and current assets were the strongest reasons the model classified it as distressed/high-risk with nearly 100% probability. To clarify directional asymmetry, Figures 11 and 12 separate protective versus risk-enhancing features. Figure 11 shows the features that most strongly reduced the predicted risk of distress. return on assets and current assets emerged as the dominant protective factors, followed by equity- and asset-based variables. These indicators suggest that firms with higher

profitability and strong asset positions are less likely to experience financial distress.

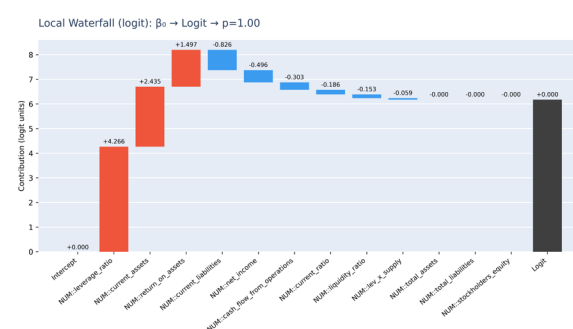


Figure 10. Local waterfall decomposition of a logistic-regression prediction in logit units.

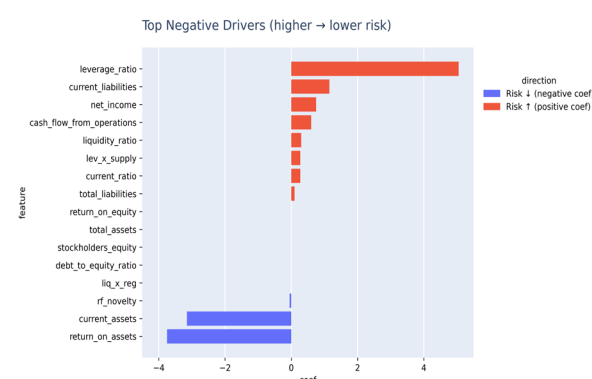


Figure 11. Top negative (protective) drivers. Higher values reduce predicted risk (blue coefficients).

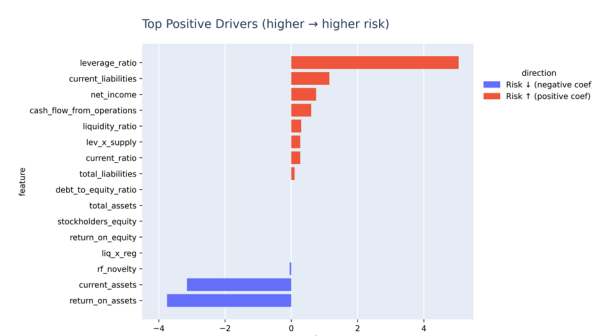


Figure 12. Top positive (risk-increasing) drivers. Higher values increase predicted risk (red coefficients).

A single firm's prediction is decomposed into its parts as shown in Figure 13 using the SHAP Force Plot. The red segments (such as the High Leverage Ratio and Strong Current Assets) are pushing the outcome toward Distress. In contrast, the Blue Segments are pulling the Outcome Downward. This results in a high probability of Financial Distress for this Case.

The global features of all firms are illustrated in Figure 14. The figure represents SHAP values of each individual firm-feature combination and colors represent their respective feature values (low – blue; high – red). Net income, current ratio and leverage ratio have the greatest effect globally on firms' performance, although they do so in opposite directions depending upon the magnitude of the firm's value.



Figure 13. SHAP force plot for a distressed firm.

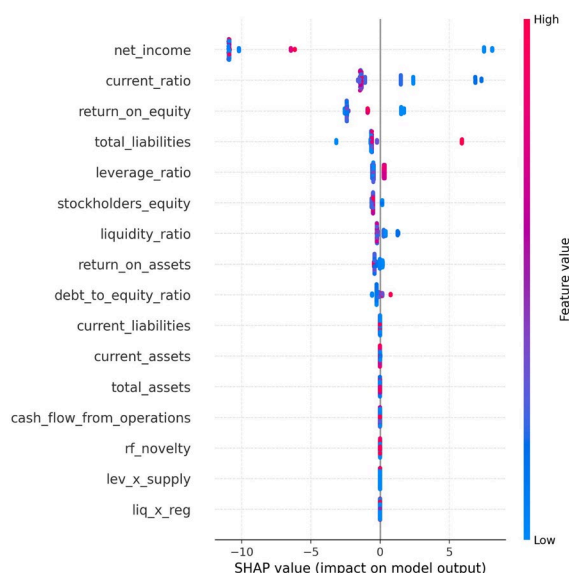


Figure 14. Global SHAP beeswarm summary across firms.

Figure 15 shows the global importance of features estimated using mean absolute SHAP values for the LightGBM model. The result highlights that the net income emerges as the most influential factor, followed by current ratio, return on equity, and total liabilities. thus it highlights the dominance of profitability, liquidity, and leverage indicators in creditworthiness prediction.

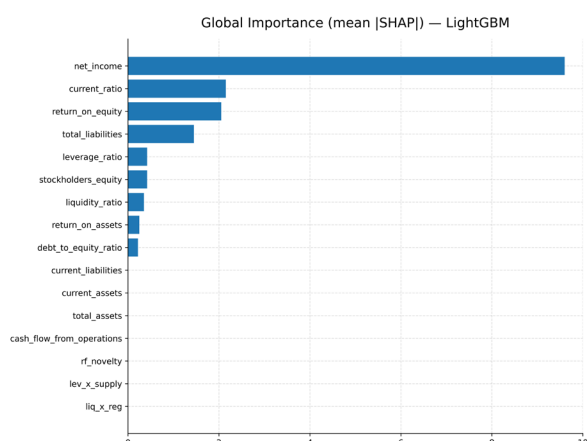


Figure 15. Global feature importance (LightGBM).

Similarly other variables such as equity, liquidity ratios, and debt-to-equity provide marginal contributions and confirmed that financial performance and solvency measures are the important parameters and affect the model's predictive power.

Figure 16 ranks the top-12 contributors (logit units) for the same firm. Here it is clearly visible that leverage ratio and current assets are the strongest positive drivers, while current liabilities, net income, and cash flow from operations provide offsetting negative contributions.

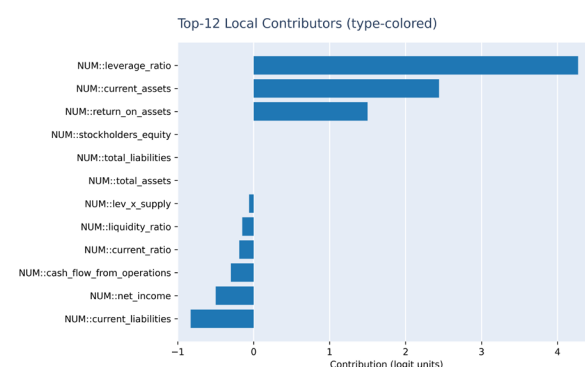


Figure 16. Top-12 local feature contributions for a firm.

As shown in Figure 17, the uncalibrated scores lie near to the identity line across probability bins, yielding a lower Brier score (0.063) compared to isotonic calibration (0.099). It means the uncalibrated logistic regression model already provide reliable probability estimates. This also indicates that post-hoc calibration causes overfitting and causes poor probability indicators. Therefore, we retain the uncalibrated outputs to have trustworthy downstream threshold optimization and aligned regulatory requirements for interpretable credit risk models.

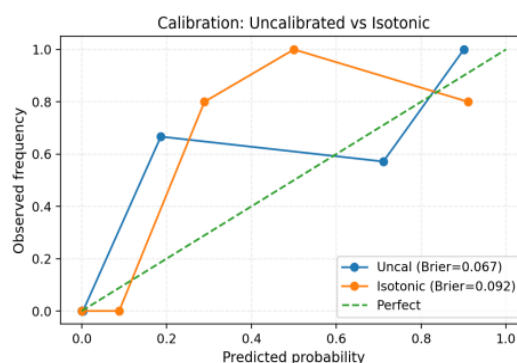


Figure 17. Reliability curves: uncalibrated vs. isotonic-calibrated logistic regression.

We compare the calibration of a logistic regression model that has been neither calibrated nor isotonically calibrated by comparing their respective reliability diagrams in Figure 17. The diagonal dashed line is the ideal point of calibration in which all predicted probabilities are equivalent to the observed frequency

of events. The logistic regression model's curve follows very close to this line (Brier score = 0.063) since the model already provides good alignment of the predicted values for distress levels with actual distress rates. However, the curve of the isotonic-calibrated model diverges from the diagonal at low probability levels (Brier score = 0.099) as the calibration process fit too much of the training data based on the relatively small number of cases used to train the model. Thus, we have sufficient justification for retaining the raw, uncalibrated output of the logistic regression model when performing subsequent analyses to determine thresholds for regulatory reporting.

The DCA shown in Figure 18 provides an objective comparison of our decision-making strategy with two naive alternatives: treating all as risky and treating none as risky. At threshold values typical of those found in practice (0.1 – 0.5) the proposed model results in greater net benefits than either of these naive approaches. As such, this provides empirical validation that the model's suitability goes beyond predictive accuracy and into areas where decision making regarding credit risk are typically governed by regulatory bodies.

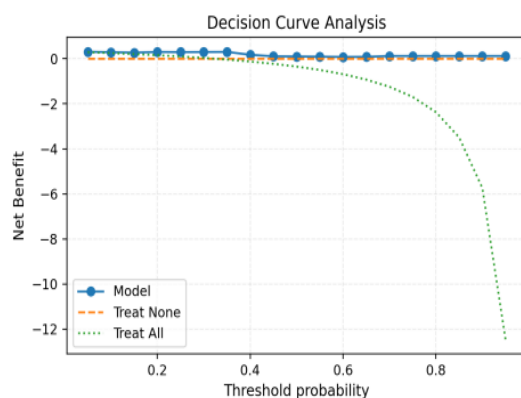


Figure 18. Decision Curve Analysis for the proposed model.

Table 4. Population Stability Index (PSI) Bucket

| Bucket         | Train Count | Train % | Test Count | Test % | Contribution(PSI) |
|----------------|-------------|---------|------------|--------|-------------------|
| (0.000, 0.100] | 229         | 0.229   | 253        | 0.253  | 0.004             |
| (0.100, 0.200] | 243         | 0.243   | 249        | 0.249  | 0.001             |
| (0.200, 0.300] | 187         | 0.187   | 167        | 0.167  | 0.002             |
| (0.300, 0.400] | 141         | 0.141   | 127        | 0.127  | 0.002             |
| (0.400, 0.500] | 97          | 0.097   | 92         | 0.092  | 0.000             |
| (0.500, 0.600] | 54          | 0.054   | 61         | 0.061  | 0.001             |
| (0.600, 0.700] | 26          | 0.026   | 30         | 0.030  | 0.001             |
| (0.700, 0.800] | 13          | 0.013   | 14         | 0.014  | 0.000             |
| (0.800, 0.900] | 6           | 0.006   | 5          | 0.005  | 0.000             |
| (0.900, 1.000] | 4           | 0.004   | 2          | 0.002  | 0.001             |
| Total PSI      | —           | —       | —          | —      | 0.012             |

Table 4 presents the Population Stability Index (PSI) analysis across probability score buckets. This indicates some distributional shift between training and testing samples, likely due to class imbalance in rare risk

regions. In the context of our study, these findings highlight the need for continuous monitoring of score stability, ensuring that creditworthiness predictions remain reliable over time without recalibration.

### 3.4 Results Interpretation and Implications

The experimental results provide important insights into multimodal credit risk modeling. The ablation analysis shows that structured financial variables provide the strongest predictive signal (ROC-AUC  $\approx$  0.98), while textual and binary disclosures offer only marginal improvements when combined with financial indicators. This finding refines prior studies by demonstrating that narrative disclosures, although forward-looking, have limited standalone predictive power due to sparsity but provide complementary contextual information that enhances interpretability and risk explanation.

The results further show that interpretable models such as logistic regression achieve performance comparable to complex non-linear methods, with limited additional gains from LightGBM. This supports the suitability of transparent and calibrated models for governance-sensitive financial applications, particularly where auditability and regulatory compliance are essential. In addition, the analysis highlights the importance of probability calibration and decision-oriented evaluation, as the logistic regression model produces reliable probability estimates and decision curve analysis confirms the practical utility of calibrated risk predictions for real-world credit decisions. These findings have both practical and theoretical implications. From a practical perspective, financial institutions may prioritize structured financial indicators as primary predictors while using narrative disclosures to improve contextual understanding of risk. From a theoretical perspective, the study provides new evidence on the relative contribution of heterogeneous data sources in credit risk prediction and emphasizes the importance of integrating predictive performance with explainability and decision reliability in multimodal learning frameworks.

### 3.5 Contribution to Existing Literature

The findings extend existing research in three ways. First, unlike prior studies that report significant gains from multimodal integration, the results show that structured financial indicators remain the dominant predictors of distress, while textual disclosures provide only complementary value. Second, the study demonstrates that explainability-driven models can achieve strong predictive performance, with leverage and liquidity variables identified as key risk drivers, reinforcing the role of explainable AI in regulated financial applications. Third, the proposed framework introduces a governance-oriented evaluation by integrating calibration, decision analysis, and stability monitoring, emphasizing decision reliability and deployment robustness beyond predictive accuracy. From a practical perspective, the results suggest that

financial institutions may prioritize structured financial indicators as primary predictors while using narrative disclosures to enhance interpretability and contextual understanding of risk. The findings also support the use of calibrated and interpretable models for auditability and regulatory compliance.

From a theoretical perspective, the study provides new evidence on the relative contribution of heterogeneous data sources in credit risk prediction and highlights the importance of integrating predictive performance with explainability and decision reliability in multimodal learning frameworks. The experimental results illustrate three key aspects of multimodal credit risk modeling. First, as shown by the ablation study, the most significant predictor of credit worthiness is provided by structured financial data, which was indicated by the high value of the Receiver Operating Characteristic Area Under the Curve (ROC-AUC), close to 0.98, obtained through the financial-only logistic regression. The reasons for the high values of this measure are due to the inclusion of well-established distress indicators such as net income, current ratio, leverage ratio, and cash flow from operations — that have been traditionally used in studies performed by Beaver [18], Altman [19], etc. Second, the improvements in performance achieved by incorporating both textual and binary disclosure features result primarily from two characteristics of the dataset analyzed: (i) the fact that around 48 % of summaries of Item 1A do not exist, thus limiting the use of textual information; (ii) the large imbalance in the binary risk flags, since their frequency of occurrence ranges from 4.9 % to 5.8 %. In spite of these limitations, the multimodal configurations (num+bool, num+txt, num+bool+txt) were able to perform at least as well as, if not slightly better than, the pure numerical configuration in terms of F1@best in those cases when the combination of both types of features helped to increase the probability of making an accurate prediction based solely on financial ratios. Thirdly, the similar performance levels reported for the logistic regression and LightGBM represent a key aspect from a governance standpoint. These results confirm that the potential benefits of using complex models, in this case the LightGBM model, are very small compared to the disadvantages associated with lower transparency and auditability of those same models. Therefore, they support the regulatory preference for developing simple, transparent and auditable models for evaluating credit risks. Finally, the Decision Curve Analysis shows that the framework has applicative value that extends beyond measures of predictive accuracy. Specifically, it generates higher net benefits than the 'treat-all' and 'treat-none' strategies for all decision thresholds from 0.1 to 0.5, confirming its readiness for being deployed in real-world environments characterized by different costs of Type I and Type II errors.

#### 4. Conclusions

This study presents an explainable multimodal

framework for credit risk prediction. It combines financial ratios and narrative disclosures from SEC filings. We have conducted experiments on 822 firm–period observations and identified that numeric financial indicators have more predictive power as compared to textual disclosures. However, textual disclosures provide limited but complementary forward-looking information. It is clear that the models using only financial features achieve strong performance (ROC–AUC  $\approx$  0.98), whereas text-only models perform substantially worse which comes up with strong interpretability. We have implemented global and local explanations techniques and identified that leverage and current liabilities are the key risk factors while returning on assets and current assets are protective signals. We also performed the comparisons with non-linear models and confirmed similar accuracy, with better explainability by the implementation of SHAP-based attributions. A Robustness analysis was also performed to highlight the idea that integrating numeric, Boolean, and textual features captures non-redundant information and improves decision quality. In this study we have also performed calibration, threshold optimization, and decision curve analysis for risk assessment. The present research can be extended by scaling larger datasets, richer text modeling, survival analysis, cost-aware objectives, and enhanced drift management for real-world deployment. Despite promising results, several limitations should be noted. The dataset is limited to a small number of large-cap firms, and the narrative disclosures exhibit high sparsity, which may restrict the predictive contribution of textual features. Future work may explore larger and more diverse datasets, advanced language models for richer text representation, and survival-based or cost-sensitive learning approaches to further enhance multimodal credit risk modeling.

#### References

- [1] Risk Management—Supervisory Guidance on Model Risk Management  
<https://www.federalreserve.gov/frbs/guidance/supervisory-guidance-on-model-risk-management.htm>
- [2] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29(5), 1189–1232 (2001)  
<https://doi.org/10.1214/aos/1013203451>
- [3] Ke, G., Meng, Q., Finley, T., et al.: Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems* (2017)  
<https://dl.acm.org/doi/10.5555/3294996.3295074>
- [4] Breiman, L.: Statistical modeling: The two cultures. *Statistical science* 16(3), 199–231 (2001)  
<https://doi.org/10.1214/ss/1009213726>
- [5] Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23(4), 589–609 (1968)  
<https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- [6] Ohlson, J.A.: Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18(1), 109–131 (1980) <https://doi.org/10.2307/2490395>
- [7] Shumway, T.: Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business* 74(1), 101–124 (2001)  
<https://doi.org/10.1086/209665>
- [8] Merton, R.C.: On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29(2), 449–470 (1974)  
<https://doi.org/10.2307/2978814>

- [9] Campbell, J.Y., Hilscher, J., Szilagyi, J.: In search of distress risk. *Journal of Finance* 63(6), 2899–2939 (2008) <https://doi.org/10.1111/j.1540-6261.2008.01416.x>
- [10] Loughran, T., McDonald, B.: When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance* 66(1), 35–65 (2011) <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- [11] Li, Feng, Textual Analysis of Corporate Disclosures: A Survey of the Literature (February 7, 2011). *Journal of Accounting Literature*, Forthcoming, Available at SSRN: <https://ssrn.com/abstract=1756926>
- [12] U.S. Securities and Exchange Commission: Regulation S-K, Item 105: Risk Factors. <https://www.sec.gov/rules/final/2019/33-10668.pdf> (2019)
- [13] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD*, pp. 1135–1144 (2016) <https://dl.acm.org/doi/10.1145/2939672.2939778>
- [14] Lundberg, S., Lee, S.-I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* (2017) <https://dl.acm.org/doi/10.5555/3295222.3295230>
- [15] Lundberg, S.M., Erion, G., et al.: From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence* 2, 56–67 (2020) <https://doi.org/10.1038/s42256-019-0138-9>
- [16] Roberts, D.R., Bahn, V., Ciuti, S., al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40(8), 913–929 (2017) <https://doi.org/10.1111/ecog.02881>
- [17] Vickers, A.J., Elkin, E.B.: Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making* 26(6), 565–574 (2006) <https://doi.org/10.1177/0272989X06295361>
- [18] Beaver, W.H.: Financial ratios as predictors of failure. *Journal of Accounting Research* 4, 71–111 (1966) *Empirical Research in Accounting: Selected Studies* <https://doi.org/10.2307/2490171>
- [19] Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23(4), 589–609 (1968) <https://doi.org/10.2307/2978933>
- [20] Hillegeist, S.A., Keating, E.K., Cram, D.P., Lundstedt, K.G.: Assessing the probability of bankruptcy. *Review of Accounting Studies* 9(1), 5–34 (2004) <https://doi.org/10.1023/B:RAST.0000013627.90884.b7>
- [21] Hand, D.J., Henley, W.E.: Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A* 160(3), 523–541 (1997) <https://doi.org/10.1111/j.1467-985X.1997.00078.x>
- [22] Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* 247(1), 124–136 (2015) <https://doi.org/10.1016/j.ejor.2015.05.030>
- [23] Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., Smith, N.A.: Predicting risk from financial reports. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 272–280 (2009) <https://doi.org/10.3115/1620754.1620794>
- [24] Loughran, T., McDonald, B.: When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance* 66(1), 35–65 (2011) <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- [25] Kravet, T., Muslu, V.: Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies* 18(4), 1088–1122 (2013) <https://doi.org/10.1007/s11142-013-9228-9>
- [26] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774 (2017) <https://dl.acm.org/doi/10.5555/3295222.3295230>
- [27] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 625–632 (2005) <https://dl.acm.org/doi/10.1145/1102351.1102430>
- [28] Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 609–616 (2001) <https://dl.acm.org/doi/10.5555/645530.655658>
- [29] Vickers, A.J., Elkin, E.B.: Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making* 26(6), 565–574 (2006) <https://doi.org/10.1177/0272989X06295361>
- [30] Siddiqi, N.: *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards* Wiley, (2017) <https://doi.org/10.1002/9781119282396.index>
- [31] Federal Reserve Board, Revised Guidance on Model Risk Management (2026) <https://www.federalreserve.gov/supervisionreg/srletters/SR2602.pdf>
- [32] Kocoglu, E., & Ersoz, F. (2025). Explainable Artificial Intelligence (XAI) for Commercial Credit Limit Prediction Model: SHAP-LIME Comparison. In: Durakbasa, N.M., et al. (eds) *Digitalization in Additive Manufacturing. ICPTDI 2024. Springer Tracts in Additive Manufacturing*. Springer, Cham. [https://doi.org/10.1007/978-3-031-84873-5\\_44](https://doi.org/10.1007/978-3-031-84873-5_44)
- [33] Gama, J., Zliobait'e, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. *ACM Computing Surveys* 46(4), 44–14437 (2014) <https://doi.org/10.1145/2523813>
- [34] U.S. Securities and Exchange Commission: SEC EDGAR Application Programming Interface (API). <https://www.sec.gov/edgar/sec-api-documentation>. Accessed: 2025-08-20 (2025)
- [35] Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance* 23(4), 589–609 (1968) <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- [36] Beaver, W.H.: Financial ratios as predictors of failure. *Journal of Accounting Research* 4, 71–111 (1966) <https://doi.org/10.2307/2490171>
- [37] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1), 11–21 (1972) <https://doi.org/10.1108/eb026526>
- [38] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: *Applied Logistic Regression*. Wiley, (2013) <https://doi.org/10.1002/9781118548387>
- [39] Brier, G.W.: Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78(1), 1–3 (1950) [http://dx.doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- [40] Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: *Proceedings of the ACM SIGKDD*, pp. 694–699 (2002) <https://doi.org/10.1145/775047.775151>
- [41] Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: *Proceedings of the 8th ACM SIGKDD*, pp. 694–699 (2002) <https://doi.org/10.1145/775047.775151>
- [42] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: *ICML*, pp. 625–632 (2005)
- [43] Vickers, A.J., Elkin, E.B.: Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 26(6), 565–574 (2006) <https://doi.org/10.1177/0272989X06295361>
- [44] Committee on Supervision and Implementation: Model risk management. Technical report, Federal Reserve Board (2017). Supervisory Guidance SR 11-7 <https://doi.org/10.1145/1102351.1102430>