ORIGINAL RESEARCH ARTICLE

J. F. Mkt. Gov. (2025) Vol. 4, Issue. 2
DOI: https://doi.org/10.54728/JFMG.202512.00121

Md Jahir Uddin Palas, PhD
University of Dhaka
Email: jahir@du.ac.bd

Benazir Imam Majumder*
University of Dhaka
Email: benazir.majumder@du.ac.bd

Keywords:
Financial Distress Prediction, Non-Life Insurance, Explainable Machine Learning, Early-Warning Systems

JEL Classification:
C53; G22; G33

Manuscript number
JFMG-202512-00121

Received: 08 December, 2025
First Revision: 19 February, 2026
Second Revision: 23 February, 2026
Accepted: 27 April, 2026
Published Online: 10 May, 2026
Published in Print: 20 May, 2026

ISSN (Online): 2958-9290
ISSN (Print): 2958-9282

*Corresponding Author

Predicting non-life insurers' financial distress: Evidence from Bangladesh

This study develops an interpretable early-warning framework to predict financial distress among non-life insurers in Bangladesh. A 2014–2024 firm-year panel is utilized to compare penalized logistic regression, random forest, and gradient-boosted trees, apply class-balance remedies, and map model decisions with SHAP. Gradient-boosted trees deliver the best out-of-time recall; SHAP consistently identifies management expense ratio, lagged underwriting performance, and reinsurance intensity as the strongest predictors. Robustness checks across resampling schemes and feature reductions confirm the stability of these operational signals. The results imply that supervisors in thin-premium markets should prioritize expense and underwriting monitoring and calibrate alarm thresholds to favor sensitivity. This research provides a compact, policy-ready pipeline that balances predictive performance with transparency.

1. Introduction

This research examines the prediction of financial distress in non-life insurance firms and develops an explainable early warning framework that is suitable for small emerging markets. Financial distress in insurers disrupts claims settlement, weakens risk pooling, and creates contingent fiscal obligations, which motivates supervisory interest in timely, actionable signals (Ayinaddis & Tegegne, 2023; Kebede et al., 2024). Recent methodological advances show that modern supervised learners, when combined with domain-aware explainability tools, improve the detection of rare distress events while preserving the interpretability required by regulators (Abrahamsen et al., 2024; Lokanan & Ramzan, 2024). This study, therefore, focuses on a compact set of accounting-based predictors and a transparent modeling pipeline that balances predictive power with operational transparency.

This topic is pursued for three practical and theoretical reasons. First, insurance markets in many emerging economies are characterized by thin premium bases, concentrated exposures, and incomplete reinsurance coverage, which magnify the solvency impact of underwriting losses and expense overruns (Kebede et al., 2024; Grize et al., 2020). Second, the distress-

prediction literature has broadened beyond classical discriminant and logistic frameworks to include ensemble trees, boosting, and deep-learning hybrids that typically deliver higher out-of-sample performance on structured financial panels (Dhamo et al., 2025; El Madou et al., 2023). Third, policymakers require interpretable diagnostics that map model outputs onto familiar accounting ratios, which favors explainable AI solutions such as SHAP and related decomposition methods (Bussmann et al., 2024; Hajek & Munk, 2023).

Three gaps motivate this study's empirical program. First, existing ML-based distress studies emphasize banking, manufacturing, or large publicly listed firms while dedicated examinations of non-life insurers in smaller markets are rare (Gavurová et al., 2022; Shetty et al., 2022). Second, many empirical implementations report nominally high accuracy but provide limited assessment of class-imbalance remedies, time-based holdouts, and sensitivity to feature selection, which undermines confidence in operational deployment (Wu et al., 2022; Kuizinienė, 2022). Third, interpretability remains underdeveloped for regulator-ready models because global feature rankings do not always reveal how accounting dynamics drive decisions at the firm-year level (Bussmann et al., 2024). This study explicitly addresses these gaps by combining a theory-driven predictor set, multiple imbalance strategies, and firm-level SHAP explanations evaluated on out-of-time holdouts.

Bangladesh's non-life insurance sector is selected for both empirical and policy reasons. The market's concentrated premium base, sensitivity to underwriting and expense shocks, evolving regulation, and high welfare costs from delayed claims make it ideal for early-warning research (Abrahamsen et al., 2024; Bussmann et al., 2024; Kebede et al., 2024; Lokanan & Ramzan, 2024). These features represent thin-premium emerging markets, offering a useful setting to test whether compact, interpretable models add supervisory value. Prior studies indicate that expense control, reserving, and reinsurance drive distress more than capital ratios, a pattern expected in Bangladesh, enhancing external validity (Ayinaddis & Tegegne, 2023; Shetty et al., 2022). From a methodological standpoint, Bangladesh provides a realistic data environment to assess parsimonious models and robust imbalance treatment relevant for regulators in similar markets (Ashraf & Vincent, 2021; Grize et al., 2020; Liang et al., 2020). Finally, focusing on one market allows careful linkage between model explanations, local accounting ratios, and supervisory frameworks, strengthening the practical transferability of this study's findings (Abrahamsen et al., 2024; Bussmann et al., 2024).

The economic motivation behind this study is centered on the real-world consequences of insurer distress in Bangladesh. The non-life insurance market frequently experiences operational failures, which are evident in delayed claims payments, opaque practices, and the erosion of policyholder trust. This crisis reached a critical point in July 2025, when the regulator (IDRA) classified 17 non-life insurers as "at risk" due to weak governance and fragile financial conditions (Halder, 2026). In this thin-premium market, a sharp distinction exists between underwriting distress and formal insolvency. Additionally, by the end of 2024, nearly 47% of the non-life insurance claims remain unsettled, which puts the entire industry in distress. Underwriting distress is the primary driver of market instability, whereas insolvency represents only the final stage of collapse. The existing supervisory framework relies heavily on static capital adequacy ratios and therefore moves slowly in detecting the operational shifts that cause financial distress. Consequently, there is a need for

a dynamic early-warning system to detect distress before the occurrence of substantial welfare loss. By focusing on these early signals, the model provides a proactive tool for regulatory intervention, which is currently absent in lag-based accounting thresholds.

Labeling and predictor choices on accounting and actuarial grounds are justified prior to modeling. This study measures distress with an economically meaningful underwriting proxy that captures loss and expense pressure relative to premiums, consistent with actuarial practice and with recent empirical work that uses combined-ratio thresholds to identify underwriting failure (Kebede et al., 2024; Ayinaddis & Tegegne, 2023). Candidate predictors comprise profitability measures, capital adequacy indicators, management expense ratios, reinsurance reliance, reserve metrics, growth, and scale. Choosing interpretable accounting variables reduces the risk that predictive performance derives from spurious correlations and facilitates immediate policy translation of model alarms (Liang et al., 2020; Tang et al., 2020).

This study adopts a modeling strategy that balances accuracy with transparency and operational readiness. Penalized logistic regression, k-nearest neighbors, support vector machines, random forests, and gradient-boosted trees, tune hyperparameters are evaluated with cross-validation, and assess performance on a strictly out-of-time holdout to emulate supervisory deployment. Class imbalance is addressed using synthetic oversampling procedures such as SMOTE and ADASYN, and compares these to no-resampling baselines. This research selects metrics that prioritize recall and the F1 score because the supervisory loss from failing to flag a truly distressed insurer typically exceeds the operational cost of false positives (Lokanan & Ramzan, 2024; Dhamo et al., 2025).

The empirical results reveal three robust patterns. First, ensemble tree methods, notably gradient-boosted machines, achieve the highest recall and competitive ROC-AUC across tuned specifications, in line with recent evidence on structured financial data (Dhamo et al., 2025; Lessmann et al., 2015). Second, SHAP-based decomposition consistently attributes the largest predictive influence to management expense ratios, lagged underwriting performance, and reinsurance intensity, while textbook capital ratios and firm size show weaker marginal contribution in this study's sample. Third, oversampling raises recall substantially at the expense of precision, while no-resampling favors precision but reduces sensitivity. These patterns suggest that underwriting inefficiency and expense mismanagement are the principal early-warning signals in thin-premium markets (Kebede et al., 2024; Grize et al., 2020).

Extensive robustness exercises are carried out to assess operational stability. This study varies resampling schemes, reduces feature sets based on SHAP rankings, and implements rolling time validation. The precision–recall tradeoff across imbalance strategies is stable, and feature reduction that removes low-importance capital variables retains most predictive performance, which supports a compact model design for supervisory use. Time-based validation confirms that gains are not artifacts of in-sample overfitting, reinforcing the practical viability of deploying an interpretable early-warning model within regulatory workflows (Wu et al., 2022; Petropoulos et al., 2020).

This study contributes to the literature and to policy practice in three ways. Methodologically, it supplies a replicable pipeline that integrates class-balance remedies, tuned ensemble learners, and SHAP explainability under time-validated evaluation. To ensure the study findings are robust enough for policy use, this paper

moves away from standard randomized testing and instead implements a rigorous, time-validated pipeline that accounts for the volatile data reporting typical of emerging economies. This study also helps bridge the gap between high-level machine learning and supervisory feasibility.

Empirically, this research delivers one of the first focused machine learning assessments of non-life insurer distress for a small emerging market. We shift the focus from traditional solvency-based research to the unique "operational decay" of thin-premium markets. Prior research in banking or developed insurance sectors often emphasizes liquidity or capital assets for detecting distress. The results of this study reveal that in a market like Bangladesh, the true leading indicators of distress are operational and underwriting metrics. These operational signals dominate early distress detection in Bangladesh, proving that failures are often visible in daily management long before they appear on a capital adequacy balance sheet.

Practically, this study translates model diagnostics into regulator-friendly triggers and monitoring heuristics, recommending explicit supervisory attention to management expense ratios, lagged underwriting outcomes, and calibrated thresholds that privilege recall. This provides a compact framework that translates complex SHAP diagnostics into clear, transparent triggers, allowing regulators to justify early interventions before a firm reaches the point of total collapse. These contributions advance academic understanding and provide a tangible surveillance tool for supervisors in comparable jurisdictions.

The findings of this study can be explicitly helpful in redesigning the supervisory strategy of the regulator by facilitating a shift from reactive policing to proactive, risk-based surveillance. At present, the Insurance Development and Regulatory Authority (IDRA) is forced to play its role only when an insurer is labeled unviable, and claims-settlement paralysis has already set in. This study provides the empirical evidence needed to change the established system and widen the window of intervention much earlier, before a collapse occurs. By prioritizing the monitoring of management expense trends and underwriting concentrations, a supervisor can move toward soft interventions, such as conducting early audits and placing restrictions on management payouts. Ultimately, this shifts the regulatory focus from just documenting failure to actively preventing the welfare loss associated with unpaid claims present in the Bangladesh non-life insurance industry.

The structure of the paper is organized as follows. Section 2 provides a review of the relevant literature. Section 3 outlines the data sources and research methodology. Section 4 presents empirical results, followed by a discussion of the findings in Section 5. Finally, Section 6 offers concluding remarks.

2. Literature review

Financial distress is defined as a firm's inability to meet its normal obligations without transformative remedial actions, where failure may arise from sustained operational losses, capital depletion, adverse asset–liability mismatches, or covenant breaches that precede formal insolvency. Recent work increasingly treats distress as a multi-dimensional state that combines short-run underwriting and liquidity pressures with longer-run solvency erosion (Dhamo et al., 2025; Wang et al., 2025). Empirical treatments extend classical probabilistic and hazard models by emphasizing early-warning indicators observable at the firm-year level, e.g., combined-ratio deterioration, negative operating cash flows, and escalating expense ratios, that reliably precede formal failure events (Kebede et al., 2024; Ayinaddis & Tegegne, 2023).

Foundational conceptualizations (Altman, 1968; Ohlson, 1980; Shumway, 2001) remain useful as they link observable accounting ratios to hazard-style failure risk. Consistent with these approaches, this study adopts a definition that privileges near-term operational indicators while recognizing solvency pathways that evolve through accounting measures and market signals.

To keep the theoretical discussion prediction-oriented and to generate clear, observable accounting implications, the review focuses on three complementary frameworks that directly imply testable measurement choices. First, the agency/operational governance framework emphasizes how managerial incentives, monitoring gaps, and organizational capability produce cost overruns, reserve under provisioning, and aggressive growth strategies; the proximate accounting signals are rising expense ratios, abnormal acquisition costs, and reserve shortfalls (Bussmann et al., 2024; Liang et al., 2020). These observable flow measures should respond earlier to managerial failure than static capital ratios when governance is weak. Second, the signaling/information-asymmetry framework highlights how disclosure quality and discretionary accounting affect the informativeness of contemporaneous capital measures. If firms can manage reported capital or reserves through pricing, reserving, or accounting choices, then flow-based and harder-to-manipulate indicators (realized underwriting outcomes, operating cash flows, expense trajectories) will be relatively more reliable short-horizon signals of distress (Abrahamsen et al., 2024; Ashraf & Félix, 2019; Samitas et al., 2020). Third, the market/regime and contingent-claims perspective underscores nonlinearity and conditionality: insolvency risk depends on underwriting volatility, reinsurance structure, leverage, and market cycles. Capital ratios, therefore, interact with underwriting risk and cycle state to determine marginal predictive content; this implies that predictive performance must be evaluated across regimes and with interaction terms (Petropoulos et al., 2020; Bragoli et al., 2021; Wang et al., 2025). These three frameworks jointly generate concrete, observable implications: (i) expense-based and underwriting flow measures should provide incremental predictive power for near-term distress, (ii) contemporaneous capital measures may be noisy absent adequate lags and institutional controls, and (iii) the predictive value of any measure is regime-dependent and may vary with governance and disclosure quality (Valaskova et al., 2018; Grize et al., 2020; Balasubramanian et al., 2019).

Methodological lessons from the predictive-modeling literature further constrain empirical choices. Ensemble and hybrid learners capture nonlinearities and interactions present in accounting panels and often improve raw classification performance, but gains can shrink under strict out-of-time validation and when pooled cross-country heterogeneity is present (Dhamo et al., 2025; Wu et al., 2022; Petropoulos et al., 2020). Explainability methods such as SHAP and LIME help translate algorithmic outputs into supervisory heuristics but are sensitive to correlated predictors and retraining instability. Thus, attribution results require robustness checks before policy use (Bussmann et al., 2024; El Madou et al., 2023; Kuizinienė, 2022). Finally, the rare-event nature of insolvency makes class-imbalance treatment and policy-aligned thresholding essential: oversampling may raise recall but also false alarms, whereas cost-sensitive calibration better aligns model outputs with supervisory loss functions (Bussmann et al., 2024; Wu et al., 2022).

Empirical findings that are established versus those that remain unresolved. Several strands of evidence are now

reasonably well established. First, both capital-based and flow-based indicators possess predictive content for insurer distress; classical ratio models and hazard-style specifications remain informative in many settings (Altman, 1968; Ohlson, 1980; Shumway, 2001). Second, ensemble machine-learning methods commonly improve in-sample classification performance and can add value when validated with appropriate holdouts (Dhamo et al., 2025; Wu et al., 2022). Third, interpretability techniques permit economically meaningful mappings from features to model outputs, improving the plausibility of algorithmic early-warning systems (Dhamo et al., 2025; Bussmann et al., 2024).

At the same time, important matters remain unresolved. Prominent empirical studies in thin-premium and small markets report that expense ratios, reserve adequacy, and reinsurance reliance sometimes outperform textbook capital ratios in predicting distress (Kebede et al., 2024; Ayinaddis & Tegegne, 2023; Grize et al., 2020). However, these findings are sensitive to label construction, lag strategy, and institutional heterogeneity: combined-ratio based distress labels can embed predictor information if not lagged adequately, and cross-country pooling can obscure governance and accounting differences that materially affect predictor performance (Zizi et al., 2021; Hanafy & Ming, 2021; Balasubramanian et al., 2019). Likewise, while alternative data (textual disclosures, sentiment proxies) add incremental value in large, transparent markets, their marginal benefit is limited where accounting data already capture short-term stress signals (Zhao et al., 2023; Hajek & Munk, 2023; Liang et al., 2020). Finally, the operational feasibility of model adoption, mapping feature attributions to supervisor actions, and quantifying the administrative cost of false alarms, remains underdeveloped (Petropoulos et al., 2020; Samitas et al., 2020; Kuizinienė, 2022).

Taken together, the literature establishes three points with reasonable consensus: first, both expense-based operational indicators and capital adequacy measures contain information relevant to insurer distress; second, machine-learning models can enhance predictive performance when subjected to rigorous out-of-time validation; and third, explainability mechanisms are necessary for translating predictive outputs into supervisory use. What remains unresolved, and what motivates this study, is whether expense-based, flow-oriented indicators systematically dominate contemporaneous capital ratios in near-term early-warning applications for thin-premium insurance markets once methodological concerns such as label construction, lag structure, class imbalance, and institutional heterogeneity are properly addressed. The absence of clarity on this relative importance limits both theoretical inference and practical regulatory guidance.

To bridge the gap between theory and prediction, this paper formalizes the predictive framework into explicit, testable hypotheses that directly map the reviewed literature to the empirical design, with particular relevance to the Bangladesh non-life insurance market. The first proposition, labeled operational primacy, posits that operational efficiency metrics, most notably the management/expense ratio, provide greater incremental predictive power for near-term insurer distress than contemporaneous capital adequacy measures in thin-premium environments, conditional on appropriate lagging and institutional controls. The second proposition, labeled model preference, posits that ensemble-based machine-learning models outperform traditional logistic regression,

KNN, and SVM benchmarks in identifying distressed insurers when evaluated using strict out-of-time validation and policy-aligned threshold calibration. In addition, the analysis tests whether lagged underwriting performance indicators exhibit stronger short-horizon predictive power than static capital ratios, and whether post-hoc explainability methods consistently identify operational and underwriting variables as dominant predictors in a stable and policy-relevant manner.

Empirically, these hypotheses are operationalized by comparing parsimonious ratio-based models with ensemble learners under alternative label constructions, lag structures, and class-imbalance strategies, and by evaluating performance using out-of-time holdouts and supervisory-feasible alarm rates. Feature-attribution results are subjected to permutation and retraining stability checks to assess their reliability for regulatory interpretation. By explicitly testing the relative predictive importance of expense-based versus capital-based indicators under rigorous validation and operational calibration, this study directly addresses the central unresolved issue in the literature: whether operational inefficiency is inherently more informative for early warning in thin-premium markets, or whether prior findings primarily reflect methodological and institutional contingencies.

3. Methodology

A predictive framework is developed to classify financial distress among non-life insurers in Bangladesh using firm-level accounting ratios and supervised machine learning. This section outlines the labeling strategy, feature selection, model set, evaluation protocol, and robustness checks. This study designs the pipeline to reflect regulatory needs: it emphasizes out-of-time validation, interpretable predictors, and sensitivity analyses that inform operational deployment.

All variables used in this study are constructed directly from audited financial statements of non-life insurance companies and from the Bangladesh Insurance Association (BIA) yearbooks. The measures follow standard accounting and actuarial definitions commonly used in insurance and financial analysis. No survey data, subjective assessments, or constructed indices are used. Apart from standard normalization for model implementation, the original accounting definitions are preserved.

The dataset comprises all non-life insurance companies operating in Bangladesh from 2014 to 2024. The final dataset consists of 506 company-year observations across 46 companies, including 45 private and one public insurer. However, the data for 3 companies were unavailable for the year 2024, as their annual reports had not been published as of 31st August 2025. The number of missing values consists of less than 1% of the total observations.

3.1 Detection of financial distress

Financial distress is defined with a transparent, economically meaningful underwriting rule: a firm-year is distressed if the combined ratio exceeds 1, shown in Table 1. This criterion directly captures when incurred claims plus management expenses exceed earned premium and therefore reflects underwriting failure. This study operationalizes the dependent variable as a binary indicator that equals one for distress and zero otherwise; it is applied independently to each firm-year observation.

Table-1 Distress labeling

Label	Status	Condition
1	Financial Distress	$\text{Combined Ratio}_t > 1$
0	Non-Distress	$\text{Combined Ratio}_t \leq 1$

3.2 Feature selection

Ten predictors are selected on accounting and actuarial grounds. These variables include Return on Assets, Equity Ratio, Reinsurance Ratio, Management Expense Ratio, Investment Yield, Unexpired Risk Reserve Ratio, Company Size, Leverage, Premium Growth, and Lagged Combined Ratio. Each variable maps to a clear economic channel in assessing solvency, profitability, operational efficiency, and risk exposure in the insurance industry. Table 2 summarizes the variables this study keeps in consideration for predicting financial distress.

Table-2 List of variables

Symbol	Variable	Definition	Economic Interpretation
CR	Combined Ratio	(Net Claims + Management Expense + Agency Commissions) / Net Premium	Underwriting performance
ROA	ROA	Net Profit / Total Assets	Overall profitability
ER	Equity Ratio	Total Equity / Total Assets	Capital strength and solvency
RR	Reinsurance Ratio	Reinsurance Ceded / Gross Premium	Risk transfer efficiency
MER	Management Expense Ratio	Management Expense / Net Premium	Operational cost efficiency
IY	Investment Yield	Investment Income / Total Investment	Asset portfolio performance
URR	URR Ratio	Unexpired Risk Reserve / Net Premium	Technical reserve adequacy
SIZE	Company Size	$\text{Log (Total Assets)}$	Economies of scale and market power
LEV	Leverage	Total Liability / Total Assets	Financial risk and debt burden
PG	Premium Growth	$(\text{Premium}_t - \text{Premium}_{t-1})/\text{Premium}_{t-1}$	Market share and business growth
LCR	Lagged Combined Ratio	$\text{Combined Ratio}_{t-1}$	Underwriting persistence and cycles

Source: Authors' contribution

3.3 Data splitting, preprocessing, and balancing

The panel is partitioned into two sets: the training set (2014 to 2022) and the test set (2023 to 2024). The training set is used for learning from observations, and the test set for simulating out-of-sample prediction.

Due to the inherent class imbalance related to the small number of distress insurers, the Synthetic Minority Oversampling Technique (SMOTE) is applied on the training dataset to balance the classes and improve the models' ability to learn from limited distress observations without creating bias of data leakage. Additionally, all features are standardized using the

Standard Scaler from the scikit-learn library in Python to ensure fair and consistent model training. By using this normalization process, this research transforms all the features' means to zero and standard deviations to 1. This step is crucial for models sensitive to feature scale.

Moreover, the combined ratio of the current year is excluded from the feature set to prevent the target leakage. Instead, a one-period lagged combined ratio is used as a predictor of financial distress in the main models. Therefore, this adjustment reduces the total number of observations to 460 spanning from 2015 to 2024.

3.4 Model development

Five supervised machine learning models are implemented to predict the financial distress of the insurers, as summarized in Table 3. The models include a baseline logistic regression, two classical algorithms, and two ensemble methods, which allow us to compare linear, non-linear, and ensemble-based approaches. This study includes KNN and SVM primarily as benchmark non-linear classifiers. However, the paper's substantive conclusions rely on the ensemble models, which deliver both superior out-of-sample performance and economically interpretable diagnostics.

Table-3 Model development

Baseline Model	Logistic Regression	Linear model for binary classification
Classical Machine Learning Models	K-Nearest Neighbors (KNN)	Instance based classifier; Classify based on nearest neighbors
	Support Vector Machine (SVM)	Find the optimal margin to separate classes
Ensemble Models	Random Forest	Incorporates many decision trees by averaging their votes to improve accuracy and reduce overfitting
	XGBoost (Extreme Gradient Boosting)	Builds trees sequentially, each correcting previous errors for strong predictive power

Initially, this study trains all models using their default hyperparameters. To prevent data leakage and ensure robust preprocessing, a pipeline is constructed incorporating mean imputation for missing values, feature scaling, and oversampling the minority class using SMOTE. Further, to imitate a real-world forecasting scenario, the models are trained on data from 2014 to 2022 and tested on data from 2023 to 2024.

3.5 Hyperparameter tuning and cross-validation

This research optimizes hyperparameters using a Grid Search and a 5-fold cross-validation approach on the training data to boost model performance. This approach divides the training data into five folds, trains the models in four folds, and validates the model's outcome on the remaining fold. This process repeats across all combinations in the specified hyper-parameter grid to detect the best hyper-parameters based on the F1 score. Later, each model is retained with the optimal hyperparameters on the full training set with a similar processing pipeline, including mean imputation, feature scaling, and SMOTE.

3.6 Evaluation metrics

Model performance is evaluated by using several classification metrics: Accuracy, Precision, Recall, F1-Score, Macro F1, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Table 4 presents the evaluation metrics used for measuring model performance.

Table-4 Evaluation metrics

Metric	Definition	Formula
Accuracy	Captures overall proportion of correct predictions.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	Measures proportion of correctly predicted distressed insurers out of all predicted as distressed.	$\frac{TP}{TP + FP}$
Recall	Calculates proportion of actual distressed firms correctly identified.	$\frac{TP}{TP + FN}$
F1-Score	Presents harmonic mean of Precision and Recall, balancing false positives and false negatives.	$2 \times \frac{Precision \times Recall}{Precision + Recall}$
Macro F1	Provides average of F1-Scores calculated independently for each class, treating all classes equally.	$\frac{F1_{Class\ 1} + F1_{Class\ 2}}{2}$

To evaluate these metrics, this study uses the confusion matrix, which summarizes prediction outcomes by comparing predicted labels against actual outcomes. It consists of four key components as shown in Table 5. This matrix forms the foundation for computing evaluation metrics.

Table-5 Components of Confusion Matrix

True Positives (TP)	Correctly predicted distressed insurers
True Negatives (TN)	Correctly predicted non-distressed insurers
False Positives (FP)	Non-distressed insurers incorrectly predicted as distressed
False Negatives (FN)	Distressed insurers incorrectly predicted as non-distressed

This research puts preference on Recall, F1-Score, and AUC-ROC because distressed firms are relatively rare in the dataset. These metrics are essential for evaluating the models' ability to discover rare but economically important distress events, where omitting a distressed firm can have severe consequences. Further, to ensure the reliability of our metrics, this study also employs a bootstrapping procedure with 1,000 iterations. The procedure allows us to calculate 95% confidence intervals for evaluation metrics and provides a more conservative and stable assessment of the model's performance beyond a single test-set split.

Additionally, to interpret the feature contributions of the best-performing model, the Shapley Additive Explanations (SHAP) value is used. SHAP is a game-theoretic approach that assigns an importance score to the features based on their contribution to the model's prediction. SHAP is applied exclusively to the XGBoost model, which demonstrated the best performance among all models. Therefore, SHAP values enable us to quantify the percentage contribution of each indicator to the final prediction and thereby help categorize the most and least influential features in predicting the financial distress of non-life insurers in Bangladesh.

3.7 Robustness check

To assess the reliability of the best predictive model, this study applies some robustness checks. First, alternative resampling methods are employed on the best-performing model. Here, Adaptive Synthetic Sampling (ADASYN) and no re-sampling are tested, in addition to SMOTE, to assess the impact of different techniques of handling class imbalance. Second, feature reduction is performed based on SHAP Values. The model is re-run using a reduced set of features to check the

outcome of the models after excluding low-importance and highly correlated variables. Third, consistent time-based validation is maintained. Performance is compared across these variations using standard metrics, with results discussed in the results section and visuals provided in the appendix.

Additionally, to ensure the best model's predictive logic is generalizable and robust against overfitting, a SHAP-based stability analysis was also performed. This involves calculating the mean absolute SHAP values for each feature in both the training and testing phases. By comparing the feature attribution rankings and standard deviations between these subsets, this study investigates whether the model relies on the same economic signals regardless of the specific data split.

4. Results

This section presents the findings from the overall empirical analysis. It includes summary statistics of the distressed and healthy insurers, model performance of the default and tuned models, feature importance analysis, and a robustness check to evaluate the reliability and validity of the best-performing model. Additional visual results, such as Confusion Matrix, ROC Curve, and Feature Importance Plots, are provided in the Appendix (Figures A1 to A6).

However, it is important to clarify the objective and interpretive scope of the analysis. The primary aim of this study is prediction rather than causal inference: the models are designed to identify reliable early-warning signals of insurer distress that can support supervisory monitoring, not to estimate structural relationships or policy-invariant causal effects. Accordingly, coefficient estimates, marginal effects, and feature-importance measures should be interpreted as indicators of predictive usefulness. The results reflect how consistently a variable helps distinguish distressed from non-distressed insurers in out-of-time data, rather than as evidence of underlying causal mechanisms. The use of ensemble learners and post-hoc explainability tools such as SHAP further reinforces this distinction: feature attributions summarize contributions to predictive accuracy within the trained model, conditional on the data and validation design, but do not imply that changes in these variables would mechanically reduce distress risk.

4.1 Summary statistics and preliminary analyses

The dataset comprises a total of 506 company-year observations from 46 non-life insurers of Bangladesh over the period of 2014 to 2024. Among 506 observations, three of the company's data for the year 2024 were imputed with mean values due to unavailable data. By using the criteria of a combined ratio greater than 1, Table 6 shows that 404 observations (79.84%) correspond to non-distressed insurers, while 102 (20.16%) observations are grouped as distressed insurers. Moreover, the yearly distribution of distress shows notable year-to-year fluctuations in the proportion of the distressed insurers in Bangladesh from 2014 to 2024. Table 6 exhibits the percentage of distress company-year observations along with their actual count with regard to the combined ratio criteria.

Table-6 Yearly distribution of financial distress

Year	Non-distressed	Distressed	Distressed (%)
2014	31	15	32.61
2015	34	12	26.09
2016	38	8	17.39
2017	41	5	10.87
2018	34	12	26.09
2019	36	10	21.74
2020	37	9	19.57
2021	43	3	6.52
2022	36	10	21.74
2023	36	10	21.74
2024	38	8	17.39
Total	**404**	**102**	**100%**

Source: Authors' calculation

Yearly distributions of distress show that the year 2014 experienced the highest proportion of distress (32.61%), followed by the years 2015 and 2018. In contrast, the year 2021 experienced the lowest percentage of distress (6.52%). This variability over time emphasizes the impact of several drivers, such as macroeconomic shifts, regulatory changes, and company-level management strategies, on financial outcomes. Moreover, it also justifies the inclusion of a time-variant factor (LCR) in predicting financial distress. Table 7 portrays the comparative analysis of group means along with the result of the t-test conducted to compare the mean values of the features between distressed and non-distressed insurers. The distressed insurers exhibit a lower ROA, Equity ratio, and premium growth, along with a small company size. Moreover, they have a high management expense ratio, greater dependency on reinsurance transfer, high leverage, and a higher combined ratio in the lag form. Additionally, the result of the t-test implies that ROA, reinsurance ratio, and management expense ratio basically hold statistically significant differences in their mean values between the distressed and the non-distressed insurers. The results suggest that poor profitability and high operational costs with intense dependency on reinsurance facilities can be associated with financial distress.

Table-7 Comparative analysis of group means

Feature	Mean (Distressed)	Mean (non-distressed)	Mean difference	P-value
ROA	0.0585	0.0708	-0.0123	0.0046*
ER	0.3294	0.3335	-0.0042	0.8528
RR	0.4645	0.4010	0.0635	0.0000*
MER	0.6203	0.4192	0.2011	0.0000*
IY	0.499	0.4461	0.0528	0.3346
URR	0.4181	0.4055	0.0126	0.2105
Size	3.1922	3.1961	-0.0038	0.9119
LEV	0.6706	0.6665	0.0042	0.8528
PG	0.0626	0.1299	-0.0674	0.1136
LCR	0.8198	0.7843	0.0355	0.4581

*Source: Authors' calculation (*Indicates statistical significance at the 1% level)*

The Variance Inflation Factor (VIF) analysis is conducted to assess the existence of multicollinearity among the features chosen for predicting financial distress. Table 8 shows that two important features, Leverage and Equity ratio, exhibit high VIF values. However, this study retains these variables because the goal is to get predictive accuracy for modeling financial distress rather than interpreting coefficients. Additionally, logistic regression is used as a benchmark model, and the tree-based models (Random Forest and XGBoost) are robust to multicollinearity by design. As these models form the core of this study's analysis, it is expected that multicollinearity has no adverse effect on the predictive performance of the models.

Table-8 Multicollinearity

Feature	VIF
ROA	1.109765
ER	15.72185
RR	1.172303
MER	1.236439
IY	1.090808
URR	1.14585
Size	1.644624
LEV	89.15425
PG	1.015793
LCR	1.020458

Source: Authors' calculation

In summary, these descriptive statistics and preliminary analyses provide a solid foundation for subsequent predictive modeling with benchmark, classical, and ensemble models.

4.2 Model performance using default settings

Following the preliminary data analysis and preprocessing of data, five supervised machine learning models are employed. The models include Logistic regression, KNN, SVM, Random Forest, and XGBoost. All models are implemented with their default hyperparameters. Missing values are implied through their mean values. SMOTE is applied to address the class imbalance issue, as the distressed companies are a minority here (20.12%). Therefore, this study trains the models on the 2014 to 2022 dataset and evaluates them on the 2023 to 2024 test set. Table 9 summarizes the results of the models with their default hyperparameters. Detailed classification results are presented in the form of a Confusion Matrix for all models in Figure A1 of the Appendix

Table-9 Default model performance summary

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.7065	0.3714	0.7222	0.4906	0.7538
K-Nearest Neighbors	0.7609	0.4231	0.6111	0.5000	0.7395
Support Vector Machine	0.7935	0.4815	0.7222	0.5778	0.8071
Random Forest	0.8152	0.5238	0.6111	0.5641	0.8431
XGBoost	0.7717	0.4545	0.8333	0.5882	0.8468

Source: Authors' calculation

The result of Table 9 shows that Logistic regression, as a benchmark model, achieves a test accuracy of 70.65% and a recall of 72.22% for detecting financially distressed insurers. The ROC score of 0.75 and limited precision (0.37) show a moderate discriminatory power in detecting financial distress, with a possibility of detecting some healthy insurers as distressed.

Among the KNN and SVM models, the Support Vector Machine holds a balanced performance with 79% accuracy and a recall of 72%. Moreover, the ROC-AUC of 0.81 signifies strong sensitivity to distressed cases. In comparison, K-Nearest Neighbors (KNN) demonstrated slightly lower performance than SVM, with an accuracy of 76% and a recall of 61%. The ROC-AUC also represents a moderate predictive competence of the KNN model.

Between two ensemble models, Random Forest exhibits robust performance, reaching 82% accuracy, with a precision and recall of 0.52 and 0.61, respectively. The ROC-AUC of 0.84 suggests effective distinction between distressed and non-distressed insurers. However, XGBoost outperforms other models with the highest recall of 0.83, along with an accuracy of 77%. As visualized in the ROC curves in Figure A3 of the Appendix, the ensemble models, Random Forest and XGBoost, achieve higher true positive rates across a range of thresholds compared to other default models.

Overall, all models hold reasonable predictive capacity with ROC-AUC values above 0.70. Notably, all the models have lower precision than recall, and this is expected with a dataset having a small number of distressed firms. Therefore, according to the F1 score and ROC-AUC value, the XGBoost model outperforms other models with its default settings.

4.3 Model tuning and evaluation

5-fold cross-validation and grid search hyper-parameter tuning are performed as a part of the robustness check. Grid search allows us to systematically test different combinations of hyperparameters for each model to find the best-performing combinations. Whereas, the 5-fold cross-validation splits the training data into five parts, and later the model trains on four parts and validates the model performance on the fifth. With five repeating processes, this study uses the average F1 score to select the hyperparameters to ensure the most reliable performance. Table 10 summarizes the best hyperparameters of the models along with their average F1 score.

Table-10 Model performance and best hyper-parameters

Model	Best hyper-parameters	CV F1 score	Default model F1 score	Tuned model F1 score
Logistic Regression	Regularization strength (C) = 0.01, Penalty type (L1/L2) = L2, Solver choice = *lbfgs*	0.5095	0.4906	0.4815
K-Nearest Neighbors	Number of neighbors = 9, Distance metric = Manhattan, Weight function = Distance	0.5601	0.5000	0.5833
Support Vector Machine	Regularization parameter (C) = 1, Gamma = Scale, Kernel type = RBF	0.5215	0.5778	0.5000
Random Forest	Number of trees = 200, Maximum tree depth = None, Minimum samples per split = 2	0.5724	0.5641	0.5714
XGBoost	Learning rate = 0.01, Maximum tree depth = 3, Number of Estimator = 50, Subsample ratio =1	0.6300	0.5882	0.5614

Source: Authors' calculation

Table 11 exhibits mixed yet meaningful results of the model performance after tuning with the best hyperparameters. Logistic regression maintains almost similar accuracy around 70% and ROC-AUC of 0.75. This indicates that, despite tuning, this benchmark model has limited improvement in predictive capacity. Tuned KNN model achieves 78% accuracy with 78% recall and ROC-AUC value above 0.80. The overall gain demonstrates that a tuned KNN model gains better predictive power in detecting distress. However, the tuned SVM model underperforms the default SVM models with an accuracy rate of 72% and a similar recall value of 72%

Table-11 Tuned model performance summary

	(Mean ± 95% Confidence Interval)				
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	0.694 (0.598–0.783)	0.363 (0.216–0.515)	0.719 (0.500–0.923)	0.477 (0.308–0.632)	0.751 (0.595–0.874)
K-Nearest Neighbors	0.781 (0.696–0.859)	0.469 (0.286–0.650)	0.777 (0.571–0.952)	0.580 (0.409–0.735)	0.801 (0.666–0.924)
Support Vector Machine	0.717 (0.620–0.804)	0.385 (0.226–0.548)	0.719 (0.500–0.923)	0.496 (0.326–0.655)	0.754 (0.600–0.872)
Random Forest	0.803 (0.717–0.880)	0.502 (0.290–0.714)	0.664 (0.444–0.867)	0.566 (0.375–0.732)	0.827 (0.692–0.934)
XGBoost	0.728 (0.630–0.815)	0.413 (0.250–0.568)	0.885 (0.714–1.000)	0.559 (0.385–0.704)	0.862 (0.754–0.943)

Source: Authors' calculation

Hyperparameter tuning improves the performance of both tree-based models. On one hand, the Random Forest model achieves better recall and F1 score with the cost of a slight drop in ROC-AUC. On the other hand, the post-tuned XGBoost model sustains strong results with the highest recall and ROC-ACU, which confirms its effectiveness. The confusion matrices after tuning present additional insights into how hyperparameter optimization affected the classification outcomes, particularly improving the identification of distressed firms (see Figure A2 in the Appendix).

Overall, tuning with hyperparameters has a varying effect across models, an increment in recall and precision with a slight drop in accuracy. After tuning, the ensemble models maintain strong performance across all evaluation measures, while XGBoost maintains its supremacy in detecting financial distress as reflected in its ROC-AUC. The ROC curves for the tuned models (see Figure A4 in the Appendix) demonstrate the improved discriminatory power of XGBoost and KNN models, especially in detecting distressed insurers. Therefore, these findings emphasize optimizing hyperparameters to increase the reliability and accuracy in predicting the financial distress in the Bangladesh non-life insurance market.

Furthermore, to address the risk of working with a small distress sample, this paper reports the mean performance metrics alongside 95% confidence intervals derived from 1000 bootstrap iterations in Table 11. This step adds reliability to the model's result by acting as a statistical stress test. For instance, the XGBoost model shows a high mean recall of 0.885; the interval (0.714 -1.000) acknowledges that results can be sensitive to the specific mix of insurers in the sample. Ultimately, these interval ranges of the key performance matrices offer a more conservative assessment of the model's power and ensure that the generalization of the result is based on consistent trends rather than an isolated data split.

4.4 Feature importance in predictive analysis

The XGBoost model shows better consistency in both the default and tuned models across all evaluation measures. Subsequently, feature importance is examined using an XGBoost model to understand the key financial drivers behind financial distress prediction in non-life insurers of Bangladesh. For identifying feature importance, SHAP is used, which has wide use in explaining the output of machine learning models. Table 12 shows the Mean SHAP values, whereas Figure A5 in the Appendix presents the feature importance plot generated from the XGBoost model, highlighting management expense ratio, reinsurance ratio, and lagged combined ratio as leading predictors.

Table-12 Shapley additive explanations

Feature	Mean SHAP value	Importance %
MER	0.89	43.53
RR	0.32	15.63
LCR	0.25	12.25
ROA	0.20	9.82
URR	0.12	6.03
Size	0.10	4.84
PG	0.08	3.78
IY	0.07	3.45
ER	0.01	0.65
LEV	0.00	0.02

Source: Authors' calculation

The sorted values of the mean SHAP values demonstrate that the management expense ratio is the most powerful indicator of financial distress in the Bangladesh non-life insurance market, which dominates nearly 44% of the model's decision. This suggests that higher operating costs are positively related to financial distress. Further, the reinsurance ratio and combined ratio with a one-period lag together hold almost 28% of the decision of the model. This indicates that, in the non-life insurance market, risk management practice and past underwriting performance play a critical role. These three indicators dominate approximately 70% of the decisions made by the XGBoost model. However, profitability measures and reserve adequacy ratio also improve the model performance in identifying distressed firms. Other variables such as company size, premium growth, and investment yield have a moderate influence, while two other indicators, namely equity ratio and leverage, show minimal impact on model performance. To further interpret the model decisions, SHAP value-based feature importance is presented in Figure A6 of the Appendix, showing how each feature influences the model's output in predicting distress.

This feature analysis highlights that expense management, prior underwriting outcomes, and risk transfer strategies serve as critical early warning signals for predicting financial distress among non-life insurers in Bangladesh.

4.5 Robustness check

Several robustness checks are conducted to ensure the reliability and stability of the XGBoost model. The reliability test includes evaluating the model performance under varying class imbalance handling strategies with a reduced feature set by eliminating financial indicators having lower mean SHAP values. Table 13 shows the result of the robustness check.

Table-13 Robustness check

Re-sampling	Feature Set	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SMOTE	Full Features	0.793	0.485	0.889	0.627	0.865
ADASYN	Full Features	0.783	0.471	0.889	0.615	0.865
None	Full Features	0.837	0.571	0.667	0.615	0.856
SMOTE	Reduced Features	0.761	0.444	0.889	0.593	0.884
ADASYN	Reduced Features	0.750	0.429	0.833	0.566	0.860
None	Reduced Features	0.837	0.579	0.611	0.595	0.843

Source: Authors' calculation

This study compares the model performance under three different re-sampling strategies: SMOTE, ADASYN, and no re-sampling. Both SMOTE and ADASYN improved the performance of recall (around 88.89%) with a low precision score. In contrast, no resampling provides the highest accuracy and precision score with a low recall (61%) score. These trade-offs demonstrate the challenges of identifying minor events in a limited dataset and thereby support the use of SMOTE to identify the distressed companies. Further, the XGBoost model is tested with a reduced set of eight features by removing the equity ratio and leverage due to low SHAP values. The results show a minimal drop in performance, confirming that these two features have low predictive power.

To verify the robustness of feature attributions, a stability analysis of SHAP values is conducted. This paper checked the consistency of the model's logic by comparing the mean absolute SHAP values across both the training and testing subsets. As outlined in Table 14, the feature attributes show noteworthy stability. The management expense ratio remains the significant driver, followed by the reinsurance ratio in both sets. In fact, the mean SHAP value increased for the lagged combined ratio and made it a robust predictor of insurer distress. Remarkably, the 95% confidence interval for the test set largely overlaps with the training averages and suggests that the model has captured the generalizable economic signal rather than statistical noise. Furthermore, the consistency in standard deviation across both the test and train sets indicates that the distribution of feature impacts is uniform.

Table-14 Stability of SHAP based feature importance

	Train			Test		
Feature	Mean SHAP	Std. Dev SHAP	95% CI	Mean SHAP	Std. Dev SHAP	95% CI
MER	0.92	0.45	0.88-0.97	0.89	0.42	0.80-0.97
RR	0.34	0.28	0.32-0.37	0.32	0.23	0.27-0.37
LCR	0.20	0.17	0.18-0.22	0.25	0.22	0.21-0.29
ROA	0.20	0.14	0.19-0.21	0.20	0.13	0.17-0.23
URR	0.11	0.12	0.10-0.13	0.12	0.11	0.10-0.14
Size	0.13	0.19	0.11-0.15	0.10	0.10	0.08-0.12
PG	0.08	0.06	0.07-0.08	0.08	0.06	0.07-0.09
IY	0.06	0.07	0.05-0.06	0.07	0.08	0.05-0.09
ER	0.01	0.02	0.01-0.01	0.01	0.02	0.01-0.02
LEV	0.00	0.00	0.00-0.00	0.00	0.00	0.00-0.00

Source: Authors' calculation

Overall, these robustness checks confirm the stability and predictive validity of the XGBoost model under various preprocessing scenarios, reinforcing its suitability for early warning systems in out-of-sample applications.

5. Discussion

The superior performance of the XGBoost model provides empirical support for the model preference proposition. Furthermore, the dominance of MER in the SHAP attribution (43.53%) validates the second proposition of operational primacy.

The results can be interpreted as evidence that operational inefficiencies and underwriting performance carry primary predictive power for non-life insurer distress in small emerging markets. Across specifications, the tuned gradient-boosted ensemble yields superior out-of-time recall while SHAP attributions consistently point to management expense ratios, lagged underwriting outcomes, and reinsurance intensity as the dominant contributors to distress probability. This pattern aligns with recent machine learning applications that emphasize ensembles for structured financial panels (Dhamo et al., 2025; Lessmann et al., 2015) and with insurance-specific studies that highlight expense control and reserve practice as proximate drivers of insurer stress (Kebede et al., 2024; Ayinaddis & Tegegne, 2023). The practical inference is straightforward: early-warning rules that focus on operational ratios catch stress earlier than rules that rely primarily on capital ratios in thin-premium environments, because underwriting deterioration and expense drift produce observable stress well before formal solvency erosion.

This study finds important departures from some conventional expectations about capital measures and firm size. In this study's sample, capital ratios and scale contribute less marginal predictive information than expense and underwriting metrics. This finding contrasts with corporate-finance narratives that treat capital adequacy as the core early-warning signal, but it resonates with theoretical accounts of thin-premium markets and with empirical insurance work showing that underwriting-cycle and expense dynamics dominate failure pathways where premium bases are narrow (Grize et al., 2020; Valaskova et al., 2018). An implication is that capital ratios are necessary but not sufficient indicators in these settings. They can remain outwardly adequate while operational problems silently accumulate. This divergence also signals potential measurement and signaling issues: firms may manage reserves or premium recognition in ways that mute the contemporaneous signal of true economic risk, a concern emphasized by signaling and disclosure theories (Abrahamsen et al., 2024; Ashraf & Félix, 2019).

The tradeoff between sensitivity and false alarms is central for policy translation. The experiments with SMOTE and ADASYN show consistent gains in recall, but material increases in false positives, which mirrors prior reports on class-imbalance remedies (Bussmann et al., 2024; Kuizinienė, 2022). For supervisors, this entails two practical choices: first, to choose operating thresholds that reflect the regulator's loss function and administrative capacity; second, to use model outputs as triage signals rather than single-step triggers for enforcement. In other words, models should feed into a staged supervisory workflow in which high-recall model flags prompt targeted inspections or request-for-information procedures rather than immediate market action. This approach reduces the social cost of misclassification while retaining the early detection benefits documented in this research and other studies (Petropoulos et al., 2020; Samitas et al., 2020).

Beyond the model's statistical accuracy, its real-life application lies in assisting regulators

in day-to-day decision-making. This study demonstrates this practical application by analyzing the relationship between the XGBoost model's probability scores and the number of insurers flagged for intervention (see Figure A7 in the Appendix). In practice, the probability threshold acts as an 'intervention trigger' that can be adjusted based on staff capacity and risk tolerance. For example, a regulator adopting a conservative attitude can choose a lower threshold to create a broader safety net and capture a large portion of the sample to ensure high recall. In contrast, a regulator with limited resources can raise the threshold to focus simply on companies with the highest chance of distress. This trade-off ensures the model is a flexible tool that allows supervisory authorities to explicitly define their loss function and prioritize the early detection of financial distress.

This research emphasizes the methodological inference that explainable ensemble models are operationally useful but require rigorous stability checks. SHAP attributions are valuable because they map statistical signals onto accounting concepts that supervisors understand, but attribution stability varies with model specification, correlated predictors, and retraining frequency (El Madou et al., 2023; Bussmann et al., 2024). This study, therefore, recommends routine attribution diagnostics, periodic retraining with rolling holdouts, and feature-reduction exercises that isolate a compact, robust predictor set. These precautions align with the statistical learning literature, which warns that algorithmic gains are conditional on validation protocols and on alignment of model objectives with policy desiderata (Wang et al., 2025; Wu et al., 2022).

Finally, this research considers external validity and comparative evidence. Studies that exploit alternative data, such as textual disclosures or vocal cues, report gains in large, transparent markets where such signals are rich and reliable (Zhao et al., 2023; Hajek & Munk, 2023). This study's results, together with other insurance-focused work, suggest that in small emerging markets where alternative data are scarce and accounting captures short-run stress, parsimonious ratio-based models may provide the most pragmatic surveillance gains (Kebede et al., 2024; Grize et al., 2020). This is not a claim that alternative data are universally unhelpful. Rather, it is a contextual observation: model design must account for data ecology and institutional features if it is to deliver policy-relevant early warnings (Liang et al., 2020; Petropoulos et al., 2020).

6. Conclusion

The objective of this study is to build an explainable, operationally viable early-warning framework for non-life insurer distress in a small emerging market context. Motivated by the social costs of delayed claim settlement and by a policy need for intelligible supervisory tools, this study evaluates a range of classifiers, experiments with class-imbalance remedies, and employs SHAP-based explanations to map predictions back to accounting indicators. These objectives are both scholarly and practical: to test whether modern supervised learners materially improve early-warning detection and to translate model outputs into supervisor-friendly diagnostics. The findings support both aims. Tuned gradient-boosted models deliver superior out-of-time recall and SHAP explanations consistently single out management expense ratios, lagged underwriting performance, and reinsurance reliance as primary early-warning signals.

This study's contributions are threefold. First, it provides empirical evidence that operational ratios matter more than textbook capital metrics for early detection in thin-premium insurance markets, thereby refining theoretical expectations from capital-structure and underwriting-cycle

perspectives. Second, it offers a replicable pipeline that balances predictive performance with explainability, showing how ensemble models combined with feature-attribution tools can be calibrated to policy loss functions. Third, it bridges method and practice by demonstrating how compact models based on accounting variables can be robust to feature reduction and to time-based validation, thereby lowering barriers to supervised adoption in resource-constrained regulatory environments.

The findings of this study offer a solid roadmap for modernizing insurance supervision in the Bangladesh non-life insurance market. The results will help the regulators to shift the focus from lagging capital-based metrics to leading operational indicators. In a market dominated by low claim settlement and public distrust, identifying a falling insurer is more important than accuracy. Therefore, the regulator should prioritize the timely identification of deteriorating insurers through the systematic monitoring of underwriting and expense ratios. Beyond the regulatory sphere, this framework serves as a critical diagnostic for insurance company management. The study also highlights that insurers' internal expense management and reinsurance adequacy are significant for maintaining long-term resilience. Additionally, this study provides a set of performance metrics to the large-scale policyholders and investors to evaluate the true stability of their risk-carriers. Ultimately, by adopting these transparent diagnostics, the industry can move away from merely documenting institutional failure and toward a proactive model that prevents the significant welfare loss associated with claim-settlement paralysis.

This research has several limitations. The analysis relies on standard accounting disclosures and on a distress labeling rule anchored in underwriting performance. Label construction and lag choices can materially affect measured predictor importance, which requires care to avoid information leakage and circularity. Institutional heterogeneity across markets also limits unconditional generalization: regulatory frameworks, accounting standards, and reinsurance market depth shape which predictors exert early-warning power. Lastly, model-driven supervisory adoption raises governance questions about retraining cadence, threshold governance, and administrative bandwidth to investigate model flags.

Three directions can be proposed for future research and for practical implementation. First, future studies can extend the pipeline to integrate stress scenarios and macro-linked covariates so models can anticipate systemwide episodes rather than idiosyncratic failures alone. Second, further research can evaluate hybrid systems that selectively incorporate alternative data sources where available, while preserving a parsimonious accounting backbone when data are limited. Third, researchers can undertake pilot implementations with regulators to calibrate alarm thresholds against administrative capacity and to observe how model-informed triage affects supervisory outcomes empirically. This study concludes that compact, explainable early-warning models are promising tools for insurer surveillance in small markets, provided that model design, validation, and operational integration proceed with rigorous attention to institutional context and policymaker objectives.

References

Abrahamsen, N.-G. B., Nylén-Forthun, E., Møller, M., de Lange, P. E., & Risstad, M. (2024). Financial distress prediction in the Nordics: Early warnings from machine learning models. Journal of Risk and Financial Management, 17(10), 432. https://doi.org/10.3390/jrfm17100432

Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609.

Ashraf, S. H., & Vincent, T. N. (2021). The role of board independence and ownership structure in improving the efficacy of corporate financial distress prediction model: Evidence from India. Journal of Risk and Financial Management, 14(7), 333. https://doi.org/10.3390/jrfm14070333

Ashraf, S., & Félix, A. (2019). Do traditional financial distress prediction models predict the early warning signs of financial distress? Journal of Risk and Financial Management, 12(2), 55. https://doi.org/10.3390/jrfm12020055

Ayinaddis, S. G., & Tegegne, H. G. (2023). Uncovering financial distress conditions and their determinant factors on insurance companies in Ethiopia. PLOS ONE, 18(10), e0292973. https://doi.org/10.1371/journal.pone.0292973

Balasubramanian, S. A., GS, R., P, S., & Natarajan, T. (2019). Modeling corporate financial distress using financial and non-financial variables: The case of Indian listed companies. International Journal of Law and Management, 61(3–4), 457–477. https://doi.org/10.1108/IJLMA-04-2018-0078

Bragoli, D., Ferretti, C., Ganugi, P., Marseguerra, G., & colleagues. (2021). Machine-learning models for bankruptcy prediction: Do industrial variables matter? Journal of Financial Studies, 17(2). https://doi.org/10.1080/17421772.2021.1977377

Bussmann, N., et al. (2024). Explainable machine learning to predict the cost of capital. Frontiers in Artificial Intelligence. https://doi.org/10.3389/frai.2024.1466321

Dhamo, Z., Gjeçi, A., Zibri, A., & Prendi, X. (2025). Business distress prediction in Albania: An analysis of classification methods. Journal of Risk and Financial Management, 18(3), 118. https://doi.org/10.3390/jrfm18030118

El Madou, K., Marso, S., El Kharrim, M., & El Merouani, M. (2023). Evolutions in machine learning technology for financial distress prediction: A comprehensive review and comparative analysis. Expert Systems, 40(5). https://doi.org/10.1111/exsy.13485

Gavurová, B., Jencová, S., Bačík, R., & Miskufová, M. (2022). Artificial intelligence in predicting the bankruptcy of non-financial corporations. Oeconomia Copernicana, 13(4), 1215–1251. https://doi.org/10.24136/oc.2022.035

Grize, Y.-L., Fischer, W., & Lützelschwab, C. (2020). Machine learning applications in nonlife insurance. Applied Stochastic Models in Business and Industry, 36(5), 523–537. https://doi.org/10.1002/asmb.2543

Hajek, P., & Munk, M. (2023). Speech emotion recognition and text sentiment analysis for financial distress prediction. Neural Computing and Applications, 35, 21463–21477. https://doi.org/10.1007/s00521-023-08470-8

Halder, S. (2026, January 4). Uncertainty hit life insurance in 2024, non-life grew. The Business Standard. https://forums.swift.org/t/if-vs-available-vs-if-available/40266

Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. Risks, 9(2), 42. https://doi.org/10.3390/risks9020042

Kebede, T. N., Tesfaye, G. D., & Erana, O. T. (2024). Determinants of financial distress: Evidence from insurance companies in Ethiopia. Journal of Innovation and Entrepreneurship, 13, Article 17. https://doi.org/10.1186/s13731-024-00369-5

Kuizinienė, D. (2022). Systematic review of financial distress identification using artificial intelligence methods. International Journal of Financial Studies, 10(4). https://doi.org/10.1080/08839514.2022.2138124

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Liang, D., Tsai, C.-F., Lu, H.-Y. R., & Chang, L.-S. (2020). Combining corporate governance indicators with stacking ensembles for financial distress prediction. Journal of Business Research, 120, 137–146. https://doi.org/10.1016/j.jbusres.2020.07.052

Lokanan, M. E., & Ramzan, S. (2024). Predicting financial distress in TSX-listed firms using machine learning algorithms. Frontiers in

Artificial Intelligence, 7, Article 1466321. https://doi.org/10.3389/frai.2024.1466321

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109–131.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. International Journal of Forecasting, 36(3), 1092–1113. https://doi.org/10.1016/j.ijforecast.2019.11.005

Samitas, A., Kampouris, E., & Kenourgios, D. (2020). Machine learning as an early warning system to predict financial crisis. International Review of Financial Analysis, 71, 101507. https://doi.org/10.1016/j.irfa.2020.101507

Shetty, S., Musa, M., & Brédart, X. (2022). Bankruptcy prediction using machine learning techniques. Journal of Risk and Financial Management, 15(1), 35. https://doi.org/10.3390/jrfm15010035

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. The Journal of Business, 74(1), 101–124.

Tang, X., Li, S., Tan, M., & Shi, W. (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. Journal of Forecasting, 39(5), 769–787. https://doi.org/10.1002/for.2661

Valaskova, K., Kliestik, T., Svabova, L., & Adamko, P. (2018). Financial risk measurement and prediction modelling for sustainable development of business entities using regression analysis. Sustainability, 10(7), 2144. https://doi.org/10.3390/su10072144

Wang, C., Gong, P., Li, J., & Wang, Z. (2025). Corporate financial distress prediction with multiperiod annual report data: A fusion deep neural network model. PLOS ONE, 20(9). https://doi.org/10.1371/journal.pone.0333064

Wu, D., Yang, H., & colleagues. (2022). Financial distress prediction using integrated Z-score and multilayer perceptron neural networks. Decision Support Systems, 159, 113814. https://doi.org/10.1016/j.dss.2022.113814

Zhao, Q., Xu, W., & colleagues. (2023). Predicting financial distress of Chinese listed companies using machine learning: To what extent does textual disclosure matter? International Review of Financial Analysis, 89, 102770. https://doi.org/10.1016/j.irfa.2023.102770

Zizi, Y., Jamali-Alaoui, A., El Goumi, B., Oudgou, M., & El Moudden, A. (2021). An optimal model of financial distress prediction: A comparative study between neural networks and logistic regression. Risks, 9(11), 200. https://doi.org/10.3390/risks91102001

Predicting Non-Life Insurers' Financial Distress: Evidence from Bangladesh
Appendix

Figure A1: Confusion Matrix (Default Models)

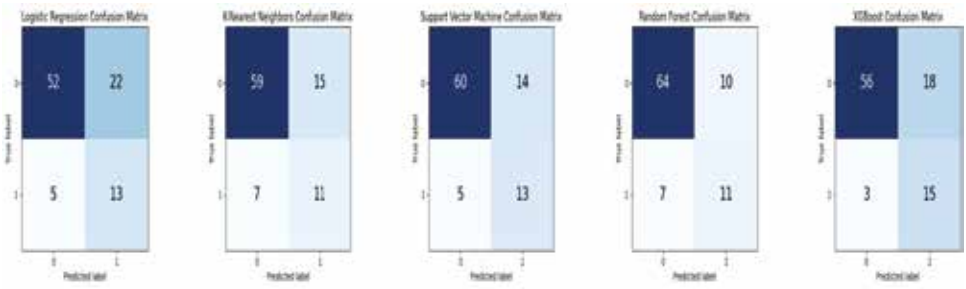

Figure A2: Confusion Matrix (Tuned Models)

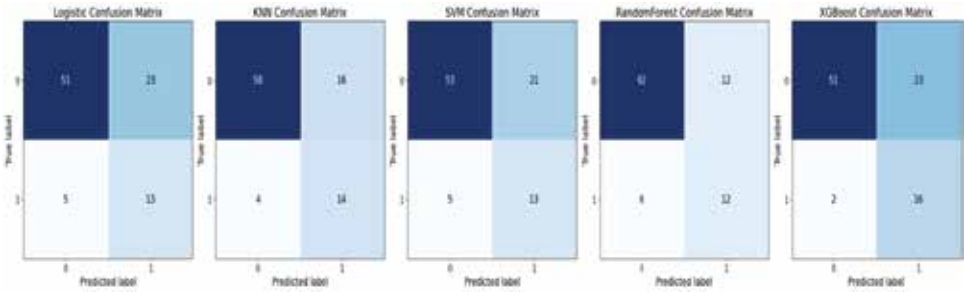

Figure A3: ROC Curve (Default Models)

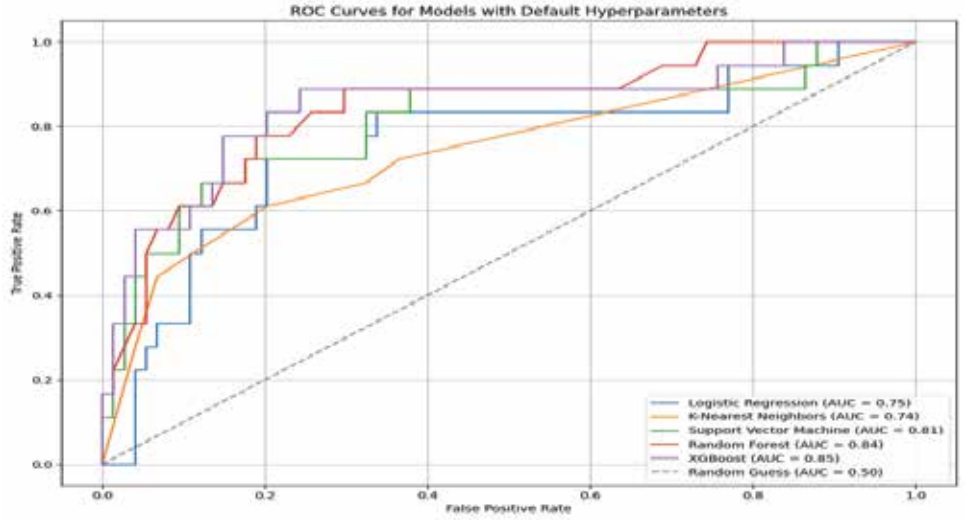

Figure A4: ROC Curve (Tuned Models)

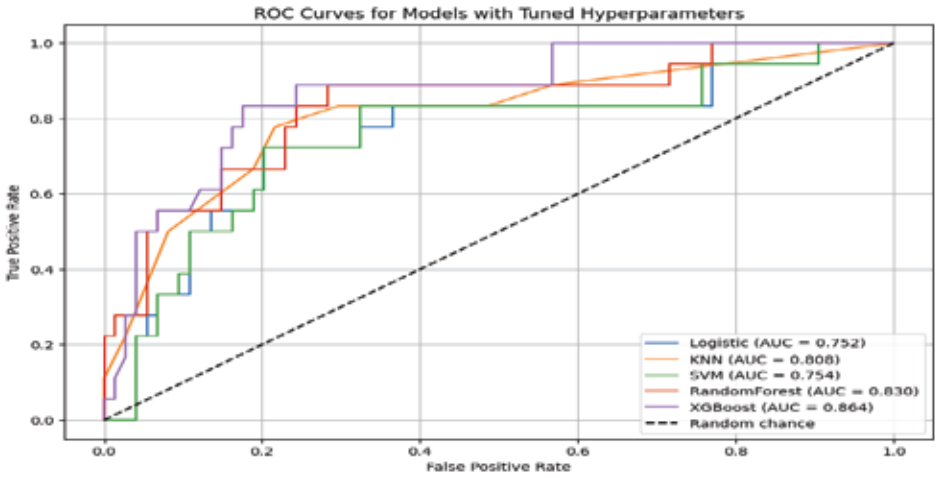

Figure A5: Feature importance plots for XGBoost

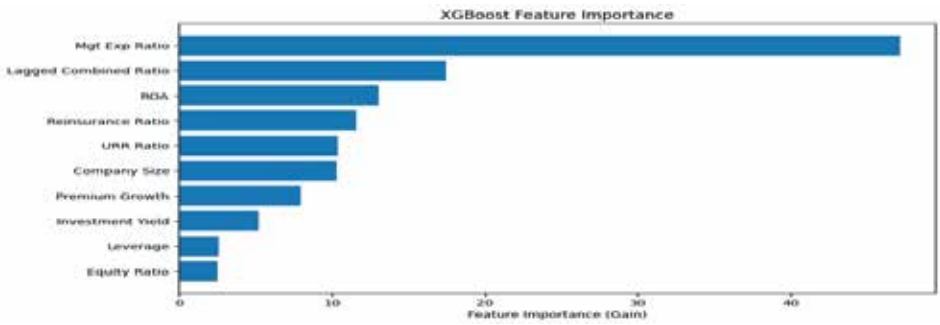

Figure A6: Feature importance for XGBoost (SHAP Values)