JAFM: Journal of Accounting and Finance Management
E-ISSN: 2721-3013
P-ISSN: 2721-3005
https://dinastires.org/JAFM dinasti.info@gmail.com +62 811 7404 455
DOI: https://doi.org/10.38035/jafm.v7i2

Hybridizing Structural Credit Risk and Machine Learning for Corporate Distress Prediction: Evidence from Indonesian Non-Financial Public Firms

Nakula Senchaki[1], Rofikoh Rokhim[2]
[1]Universitas Indonesia, Jakarta, Indonesia, nakula.senchaki@office.ui.ac.id
[2]Universitas Indonesia, Jakarta, Indonesia, rofikoh.rokhim@ui.ac.id

Corresponding Author: nakula.senchaki@office.ui.ac.id[1]

Abstract: This study develops an explainable early-warning framework for predicting corporate distress among non-financial firms listed on the Indonesia Stock Exchange. Using a firm-month panel of 107,448 observations from 2014 to 2024, the study constructs a 12-month forward distress label based on PKPU and bankruptcy events. The analysis compares Logistic Regression, Random Forest, XGBoost, and a hybrid XGBoost model incorporating Merton-based structural indicators, evaluated using ROC-AUC, PR-AUC, precision, recall, and F1-score under a time-based split. The results show that tree-ensemble models outperform Logistic Regression, with XGBoost achieving the strongest standalone rare-event performance, including PR-AUC of 0.151 and F1-score of 0.217. Adding Merton structural indicators does not improve aggregate ROC-AUC or PR-AUC, but improves recall and F1-score, indicating incremental detection value at the operational threshold. SHAP analysis shows that distress predictions are mainly driven by solvency, leverage, retained earnings, debt-servicing capacity, profitability, asset structure, and market signals. The model captures 66.7% of distress events with an average lead time of 8.5 months. The study contributes an interpretable hybrid framework for corporate distress early warning in an emerging-market setting.

Keywords: Corporate Distress, XGBoost, Distance-to-Default, SHAP, Emerging Markets

INTRODUCTION

The Corporate distress remains material in Indonesia despite stable headline economic growth. Between 2020 and 2024, the Directorate General of General Courts recorded a substantial number of incoming PKPU and bankruptcy cases each year, with PKPU consistently dominating formal restructuring proceedings (Direktorat Jenderal Badan Peradilan Umum (Badilum), 2024). This pattern matters because PKPU and bankruptcy are not merely legal outcomes; they are visible endpoints of financial deterioration that usually begins much earlier. The bond market also reflects persistent credit-risk concerns. Wider credit spreads on lower-rated corporate bonds and continuing corporate debt defaults indicate that investors still demand meaningful compensation for issuer risk (Credit Guarantee and Investment Facility, 2024; PT Pemeringkat Efek Indonesia, 2024). For creditors, institutional investors, and corporate financial managers, the implication is practical: they need an early-warning tool that

can identify deteriorating issuers months before distress becomes formalized in court proceedings.

The academic toolkit for corporate distress prediction has evolved through several waves, but these approaches are still often applied separately. Earlier accounting-based distress studies, including Beaver (1966), Altman (1968), Ohlson (1980), and Zmijewski (1984), show that financial ratios contain meaningful information about corporate failure risk. Hazard and market-based models later strengthened the role of time-varying and market information in default prediction (Shumway, 2001; Bauer & Agarwal, 2014). Accounting-ratio models remain widely used because their inputs are accessible and their logic is transparent. Their limitation is timing. Accounting data are reported periodically and may fail to capture risk that develops between reporting dates.

The structural credit-risk framework introduced by Merton (1974), and later operationalized through Distance-to-Default (DtD) measures such as Bharath and Shumway (2008), addresses this limitation by incorporating market-based information on firm value, volatility, and debt obligations. Yet standalone structural models do not always dominate empirically. In an emerging-market setting, Peykani et al. (2023) examine Iranian listed companies from 2016 to 2021 and show that machine-learning models outperform standalone structural models: Random Forest records ROC-AUC of 0.97 and Gradient Boosted Decision Tree records 0.91, while Merton and Geske record only 0.53 and 0.58, respectively. These findings suggest that structural signals contain useful economic information but may require a more flexible modeling architecture to become effective in noisy and heterogeneous capital markets.

Machine learning has expanded the predictive frontier in bankruptcy and distress modeling. Gradient-boosted decision trees, particularly XGBoost, can capture nonlinear relationships and high-order interactions that are difficult for Logistic Regression to model. Alanis et al. (2022) provide a useful global benchmark. Using a comprehensive U.S. sample of 131,261 firm-years and 2,585 bankruptcies, they show that tree-ensemble methods outperform competing models in one-year-ahead bankruptcy forecasts. Their full-sample results show that XGBoost achieves ROC-AUC of 0.837 when only Distance-to-Default and industry indicators are used, increases to 0.907 after accounting variables are added, and reaches 0.916 after market variables are included. This evidence supports the use of XGBoost as a serious benchmark for bankruptcy prediction. Prior studies show that machine-learning models, particularly tree-ensemble methods, can improve bankruptcy prediction relative to traditional statistical and structural approaches (Alanis et al., 2022; Barboza et al., 2017; Peykani et al., 2023). XGBoost is used in this study because it is a gradient-boosted tree algorithm capable of capturing nonlinear relationships and interaction effects among accounting, market, and structural risk indicators.

However, two issues remain unresolved. First, machine-learning models are often treated as substitutes for structural credit-risk models, rather than as frameworks that can absorb theory-guided structural signals. Second, high-performing machine-learning models are difficult to use in institutional settings if their predictions cannot be explained. SHAP addresses the second issue by attributing model predictions to individual features, allowing users to understand which financial or market variables drive predicted distress risk (Lundberg & Lee, 2017; Lundberg et al., 2020). The integration of Merton's structural logic, XGBoost's nonlinear prediction, and SHAP-based explainability is therefore a natural next step for distress prediction, but it remains underexplored in emerging-market settings.

This gap is particularly relevant for Indonesia. Existing domestic distress-prediction studies remain heavily influenced by accounting-ratio models, while studies that apply machine learning often do not combine it with structural market-based risk indicators or explainable diagnostics. At the same time, Indonesia has features that make model design more demanding: distress events are rare, equity-market liquidity is uneven, firm fundamentals are highly heterogeneous, and decision-makers require signals that can be justified to credit committees,

investment committees, regulators, and corporate management. Under these conditions, neither standalone accounting models, standalone structural models, nor black-box machine learning is fully satisfactory. A useful model must be predictive, forward-looking, and interpretable.

This article develops and evaluates an explainable early-warning framework for corporate distress prediction among non-financial firms listed on the Indonesia Stock Exchange. Using a firm-month panel of 107,448 observations from 2014 to 2024 and a 12-month forward distress label based on PKPU or bankruptcy events, the study compares Logistic Regression, Random Forest, XGBoost, and a Merton-augmented XGBoost model. Distress is genuinely rare in the sample: only 907 firm-month observations, or 0.84 percent, carry a positive label. For this reason, the evaluation emphasizes PR-AUC, recall, and F1-score alongside ROC-AUC. Three findings stand out. First, tree-ensemble models outperform Logistic Regression, with XGBoost delivering the strongest rare-event performance, reflected in PR-AUC of 0.151 and F1-score of 0.217, compared with Logistic Regression's PR-AUC of approximately 0.018. Second, embedding the Merton DtD signal into XGBoost does not materially improve overall ranking performance, but it improves distress detection as reflected in recall and F1-score. Third, SHAP identifies solvency, retained earnings, debt-servicing capacity, and leverage as dominant risk drivers, while the high-risk bucket captures 66.7 percent of distress events with an average lead time of 8.5 months.

Corporate Distress and Structural Credit-Risk Theory

Corporate distress refers to a condition in which a firm's financial capacity deteriorates to the point where it struggles to meet its obligations, potentially leading to restructuring, default, or bankruptcy. Altman's (1968) Z-Score and Ohlson's (1980) O-Score show that balance-sheet and income-statement variables such as leverage, profitability, working capital, and retained earnings can distinguish financially weak firms from healthier firms. The Merton (1974) framework provides a more forward-looking view by treating equity as a call option on firm assets, with default occurring when asset value falls below the default boundary. Bharath and Shumway (2008) later operationalize this logic through Distance-to-Default, a market-based measure that combines firm value, volatility, and debt obligations.

However, standalone structural models do not always perform strongly outside deep and liquid capital markets. Peykani et al. (2023) show that machine-learning models outperform structural models in the Iranian capital market. Alanis et al. (2022) also show that DtD is useful, but its predictive value becomes stronger when combined with accounting and market variables inside more flexible predictive models. These findings suggest that the Merton framework remains theoretically important, but its signal may be more effective when embedded in a broader modeling architecture rather than used as a standalone classifier.

Machine Learning in Bankruptcy Prediction

Machine learning has expanded the predictive frontier in bankruptcy and distress modeling. Unlike Logistic Regression, tree-ensemble models can capture nonlinear relationships, threshold effects, and interactions among predictors. This is important because corporate distress rarely emerges from one variable alone. It often reflects combinations of leverage pressure, declining profitability, weak retained earnings, deteriorating market signals, and reduced debt-servicing capacity.

The Indonesian setting requires careful evaluation because distress is a rare event. ROC-AUC is useful for measuring overall discrimination, but it may overstate model usefulness when the positive class is sparse. Saito and Rehmsmeier (2015) argue that precision-recall evaluation is more informative than ROC analysis in imbalanced binary classification. Precision, recall, and F1-score are reported to capture the operational quality of distress classification at a given threshold, particularly the trade-off between detecting actual distress cases and limiting false alarms (Powers, 2011). Based on this literature, tree-ensemble models

are expected to outperform Logistic Regression in predicting distress among Indonesian non-financial firms.

H1: Tree-ensemble models outperform Logistic Regression in out-of-sample corporate distress prediction among Indonesian non-financial public firms, with XGBoost expected to deliver the strongest performance on rare-event metrics such as PR-AUC, precision, and F1-score.

Hybrid Structural-Machine Learning Framework

Prior studies often treat structural models and machine-learning models as competing alternatives. In this substitution view, structural models such as Merton and Geske are compared with machine-learning models such as Random Forest and Gradient Boosted Decision Tree, and the best-performing model is selected. Peykani et al. (2023), for example, compare structural and machine-learning models in the Iranian capital market and show that machine-learning models outperform standalone structural models in out-of-sample ROC-AUC evaluation. This evidence is valuable, but it leaves a deeper question unresolved: should structural credit-risk theory be discarded when machine learning performs better, or should its signal be embedded inside machine learning?

The integration view is more relevant for this study. Merton's (1974) structural credit-risk framework links default risk to the position of firm value relative to its debt obligations, while Distance-to-Default operationalizes this logic using market value, volatility, and debt information (Bharath & Shumway, 2008). In this study, Merton Distance-to-Default is not treated as a competing standalone model but as a theory-guided feature within a nonlinear classifier. This design preserves the economic logic of structural credit-risk theory while allowing XGBoost to capture nonlinear interactions among DtD, leverage, equity volatility, market capitalization, profitability, and other financial variables.

The empirical caveat is that DtD may overlap with variables already present in the model. Alanis et al. (2022) show that bankruptcy-prediction performance improves when Distance-to-Default is combined with accounting and market variables in tree-ensemble models, but the incremental value of additional information may depend on how much new signal it contributes beyond existing firm-level predictors. Since DtD is partly driven by leverage, market value of equity, and volatility, a feature-rich XGBoost model may already capture much of the same information through individual accounting and market variables. Therefore, adding DtD may not materially improve aggregate ranking metrics such as ROC-AUC or PR-AUC. Its value may instead appear at the operational decision threshold, where the model classifies firms into distress or non-distress signals.

H2: Embedding Merton Distance-to-Default into XGBoost provides incremental distress-detection value among Indonesian non-financial public firms, reflected in higher recall and F1-score, even when aggregate ROC-AUC and PR-AUC do not materially improve.

Explainability and Risk Drivers

High-performing machine-learning models are difficult to use in institutional settings if their predictions cannot be explained. Credit committees, investment committees, regulators, and corporate management need to understand why a firm is classified as high risk. SHAP provides a solution by decomposing model predictions into feature-level contributions (Lundberg & Lee, 2017). For tree-based models such as XGBoost, SHAP is particularly useful because it can translate complex nonlinear predictions into interpretable risk drivers.

In bankruptcy prediction, economically meaningful drivers should reflect established finance theory. Solvency indicators, such as equity ratio and leverage, capture the firm's capital buffer and proximity to financial fragility. Retained earnings summarize accumulated profitability and internal funding capacity, a key component in Altman's (1968) framework. Interest coverage reflects debt-servicing capacity and directly measures whether operating income can sustain financial obligations. Asset-structure variables, such as PPE ratio, may also

matter because they reflect the composition of assets that can support operations or collateral value. Market-based variables, including volatility and DtD, complement accounting information by capturing risk that may emerge between reporting dates.

H3: SHAP explanations of the best-performing model identify economically interpretable indicators of solvency, retained earnings, debt-servicing capacity, leverage, asset structure, and market risk as dominant contributors to predicted corporate distress among Indonesian non-financial public firms.

METHOD

This study uses a firm-month panel of non-financial public firms listed on the Indonesia Stock Exchange (IDX) from January 2014 to December 2024. Financial firms, including banks, insurers, multifinance companies, and securities firms, are excluded because their balance-sheet structures, regulatory capital requirements, and reporting conventions differ materially from those of non-financial firms. The final dataset consists of 107,448 firm-month observations from 814 non-financial firms. The dependent variable is highly imbalanced: 907 observations are classified as distress and 106,541 observations are classified as non-distress, meaning the positive distress class represents only 0.84 percent of the sample.

Financial-statement data are obtained from S&P Capital IQ, while market data are collected from daily stock-price and market-information records from the Indonesia Stock Exchange. The risk-free rate used in the structural model is based on Bank Indonesia data. Distress events are identified from publicly verifiable legal records, particularly PKPU and bankruptcy cases from the commercial court system through the Sistem Informasi Penelusuran Perkara and related public sources. These event records are then mapped into firm-month format to construct the 12-month forward distress label.

Table 1. Sample Characteristics and Distress Distribution

Parameter	Definition	Value
Observation period	Sample window	2014-2024
Unit of analysis	Panel observation level	Firm-month
Sample universe	Non-financial IDX-listed firms	814 firms
Final observations	Final firm-month panel observations	107.448
Non-distress observations	No PKPU/bankruptcy event within the next 12 months	106.541
Distress observations	PKPU and/or bankruptcy event within the next 12 months	907
Distress rate	Share of final observations	0.84%
Prediction horizon	Forward-looking distress window	12 months
Panel structure	Balance type	Unbalanced

Source: Author's calculation based on S&P Capital IQ, IDX market data, Bank Indonesia data, and public PKPU/bankruptcy records

Distress Definition and Prediction Horizon

The dependent variable is a binary indicator that equals one if firm i experiences PKPU or bankruptcy within 12 months after month t, and zero otherwise. Formally:

$$y_{i,t} = \begin{cases} 1, & \text{if firm experiences PKPU or bankruptcy within } (t, t+12) \\ 0, & \text{otherwise} \end{cases}$$

This definition has three advantages. First, PKPU and bankruptcy are hard legal events that can be verified through public records. This reduces the subjectivity that may arise when distress is defined using soft accounting proxies such as negative equity, consecutive losses, or declining profitability. Second, the 12-month horizon makes the model forward-looking. The model is not designed to identify firms that are already legally distressed; it is designed to flag firms that may enter distress within the next year. Third, the firm-month structure allows the model to produce a more granular early-warning signal than a firm-year design.

To reduce look-ahead bias, predictors at month t use only information that would reasonably have been available by the end of that month. Market variables are calculated using

price and trading information up to the end of month t. accounting variables are mapped into the monthly panel using the most recently available financial statements, so that future information is not used to predict past distress.

Variable Construction

The predictor variables are organized into three main groups: accounting-based variables, market-based variables, and structural credit-risk variables. The accounting variables capture firm fundamentals, the market variables capture forward-looking investor signals, and the structural variables capture default-risk information derived from the Merton framework.

Table 2. Variable Definitions and Measurement

Variable block	Variable	Measurement	Interpretation
Dependent variable	Distress indicator	Equals 1 if a firm experiences PKPU and/or bankruptcy >t 12 months; 0 otherwise	Forward-looking legal distress label
Accounting fundamentals	Profitability	Net income relative to total assets	Captures earnings-generating capacity
Accounting fundamentals	Liquidity	Current assets relative to current liabilities	Captures short-term financial flexibility
Accounting fundamentals	Asset turnover	Sales or revenue relative to total assets	Captures operating efficiency
Accounting fundamentals	Retained earnings ratio	Retained earnings relative to total assets	Captures accumulated profitability and internal funding capacity
Capital structure and solvency	Leverage	Total debt or liabilities relative to total assets	Captures debt burden
Capital structure and solvency	Equity ratio	Total equity relative to total assets	Captures capital buffer
Capital structure and solvency	Debt-to-equity	Total debt relative to total equity	Captures capital-structure pressure
Capital structure and solvency	Interest coverage	EBIT relative to interest expense	Captures debt-servicing capacity
Asset structure	PPE ratio	Property, plant, and equipment relative to total assets	Captures fixed-asset intensity and asset composition
Market-based indicators	ret_1m	Stock return over the previous 1 month	Captures short-term market performance
Market-based indicators	ret_3m	Stock return over the previous 3 months	Captures recent market performance
Market-based indicators	ret_6m	Stock return over the previous 6 months	Captures medium-term market performance
Market-based indicators	ret_12m	Stock return over the previous 12 months	Captures annual market performance
Market-based indicators	vol_3m	Equity-return volatility over the previous 3 months	Captures short-term market risk
Market-based indicators	vol_6m	Equity-return volatility over the previous 6 months	Captures medium-term market risk
Market-based indicators	vol_12m	Equity-return volatility over the previous 12 months	Captures annual market risk
Market-based indicators	log_mcap	Natural logarithm of market capitalization	Captures firm size and market valuation
Market-based indicators	turnover_1m	Trading turnover over the previous 1 month	Captures stock liquidity and trading activity
Market-based indicators	ihsg_ret_1m	IHSG return over the previous 1 month	Captures broad market movement
Market-based indicators	excess_ret_1m	Stock return minus IHSG return over the previous 1 month	Captures firm-specific market performance
Structural credit-risk indicators	Distance-to-Default	Merton-based distance between firm value and default boundary	Captures proximity to default boundary

Structural credit-risk indicators	Merton probability of default	Probability of default derived from Distance-to-Default	Captures market-implied default probability

Source: Author's construction based on Alanis et al. (2022), Peykani et al. (2023), and the study's data sources

Model Specification

This study compares five model specifications: Logit V1, Logit V2, Random Forest, XGBoost, and Hybrid XGBoost with Merton-based structural indicators, Logistic Regression is used as the statistical baseline because it remains a standard probability model for binary outcomes (Hosmer & Lemeshow, 2000).

Logit V1 is an accounting-based statistical baseline, while Logit V2 extends the baseline by adding market variables. Logistic Regression is used because it is transparent, widely applied in bankruptcy and distress prediction, and directly comparable with traditional accounting-based models. The logit model estimates the probability of distress as:

$$P(y_{i,t} = 1 \mid x_{i,t}) = \frac{1}{1 + cxp[-(\beta o + \beta' x_{i,t})]}$$

Where Xi,t is the vector of predictors for firm i at month t.

Random Forest is included as a tree-ensemble comparator. It builds multiple decision trees using bootstrapped samples and averages their predictions. This model is useful because it captures nonlinear relationships and interactions among variables while reducing overfitting through bagging.

XGBoost is the primary machine-learning model. It builds decision trees sequentially, where each new tree corrects errors from previous trees. This makes XGBoost suitable for distress prediction because it can capture nonlinear patterns, threshold effects, and interactions among accounting and market variables. The XGBoost objective function can be written as:

$$\mathcal{L} = \sum_{i,t} \iota(y_{i,t}, \mathcal{P}_{i,t}) + \sum_{k=1}^{K} \Omega(fk)$$

Where $l(y_{i,t}, \hat{p}_{i,t})$ is the classification loss, $\hat{p}_{i,t}$ is the predicted distress probability, and $\Omega(f_k)$ is the regularization term that penalizes tree complexity. The Hybrid XGBoost model uses the same accounting and market predictors as the standalone XGBoost model but adds DtD and Merton probability of default as structural features. This design tests whether a theory-guided structural signal provides incremental value when embedded inside a nonlinear classifier. The hybrid model does not treat Merton as a competing model; it treats Merton as an additional risk signal inside machine learning.

Class imbalance is handled through model-level weighting. Logistic Regression and Random Forest use balanced class weighting, while XGBoost uses scale_pos_weight to penalize misclassification of the minority distress class more heavily. The test set retains the natural class distribution, so performance reflects realistic deployment conditions rather than artificially balanced data.

Table 3. Model Specifications and Evaluation Design

Model / evaluation component	Specification / output	Purpose
Logit V1	Accounting variables only	Serves as the accounting-based linear baseline for distress prediction
Logit V2	Accounting and market-based variables	Tests whether market information improves the linear benchmark
Random Forest	Accounting and market-based variables using a bagged tree-ensemble model	Provides a nonlinear tree-ensemble comparator
XGBoost	Accounting and market-based variables using gradient-boosted trees	Serves as the primary machine-learning model for rare-event distress prediction

Hybrid XGBoost	Accounting, market-based, and Merton structural indicators	Tests whether Distance-to-Default and Merton probability of default add incremental detection value
Class imbalance treatment	Balanced class weight for Logit and Random Forest; scale_pos_weight for XGBoost specifications	Reduces bias toward the majority non-distress class
Time-based validation	Training period: 2014–2020; test period: 2021–2024	Reduces data leakage and approximates real-world forecasting
Performance metrics	ROC-AUC, PR-AUC, precision, recall, F1-score, and confusion matrix	Evaluates overall discrimination, rare-event prediction, and classification quality
Explainability analysis	SHAP global importance and summary plot	Identifies economically interpretable drivers of predicted distress
Practical early-warning analysis	Risk-bucket analysis and lead-time analysis	Translates predicted probabilities into actionable risk monitoring outputs

Source: Author's construction based on the research design

Data split and model evaluation

The dataset is split chronologically to reduce data leakage and approximate real-world forecasting. Observations from 2014 to 2020 are used as the training set, while observations from 2021 to 2024 are used as the test set. This time-based split ensures that the model is trained on earlier information and evaluated on later observations. Unlike random splitting, chronological splitting prevents information from future periods from influencing model estimation.

Model performance is evaluated using ROC-AUC, PR-AUC, precision, recall, F1-score, and confusion matrix. ROC-AUC measures the model's overall ability to distinguish distress from non-distress observations across thresholds. PR-AUC is emphasized because the positive class is rare; in imbalanced classification, precision-recall evaluation provides a more informative view of how well the model identifies the minority class. Precision measures the share of predicted distress observations that are truly distressed, while recall measures the share of actual distress observations successfully detected by the model. F1-score summarizes the balance between precision and recall.

The main classification metrics are reported using a probability threshold of 0.50 for comparability across models. In addition, Youden's J statistic is used as a supplementary threshold analysis to identify the point on the ROC curve that balances the true positive rate and false positive rate. This supplementary analysis helps interpret the trade-off between detecting more distress events and generating additional false alarms.

Beyond standard classification metrics, the study evaluates practical early-warning usefulness through risk-bucket analysis and lead-time analysis. Risk buckets classify firm-month observations into low-, medium-, and high-risk groups based on predicted distress probability. The high-risk bucket is then evaluated by its ability to capture future distress events relative to the base distress rate. Lead-time analysis measures how many months before the legal distress event the model first flags the firm as high risk. These analyses translate model output into a more actionable form for creditors, investors, and corporate decision-makers. Figure 1 summarizes the analytical workflow of the study, from data integration and firm-month panel construction to feature engineering, distress labeling, model evaluation, SHAP interpretation, risk-bucket formation, and lead-time analysis.

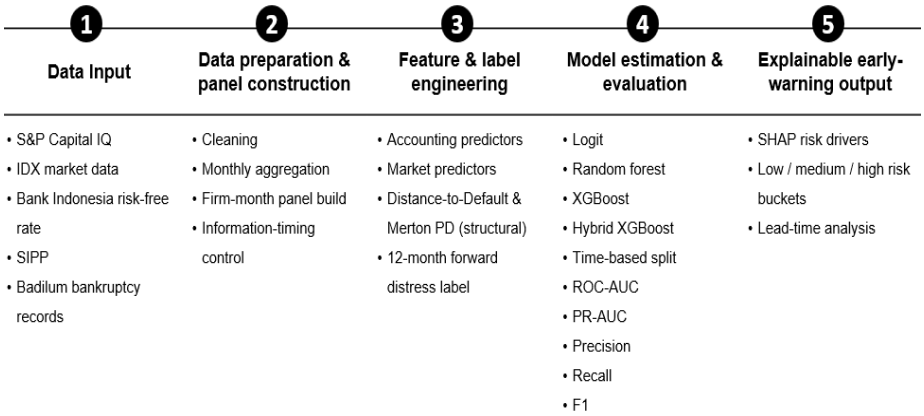

Figure 1. Research Framework and Analytical Workflow
Source: Authors' methodological design

SHAP interpretation

After identifying the best-performing model, this study applies SHAP to interpret the drivers of predicted distress. SHAP decomposes each prediction into feature-level contributions, allowing the model's output to be interpreted both globally and locally. For each firm-month prediction, the SHAP representation can be written as:

$$p_{i,t} = E[p] + \sum_{j=1}^{J} \emptyset_{j,i,t}$$

Where $\hat{p}_{i,t}$ is the predicted distress probability, $E[\widehat{p}]$ is the model's expected prediction, and $\emptyset_{j,i,t}$ is the SHAP contribution of feature j for firm i at month t. SHAP is used as a model-interpretation tool, not as a causal inference method. All analyses are conducted using Python in Google Colab, with pandas and numpy for data processing, scikit-learn for Logistic Regression and Random Forest, xgboost for XGBoost, and SHAP for model interpretation.

RESULTS AND DISCUSSION

As reported in the Method section, the final sample consists of 107,448 firm-month observations from 814 non-financial firms listed on the Indonesia Stock Exchange during 2014-2024. Among these observations, 907 firm-months are classified as distress, while 106,541 firm-months are classified as non-distress. The distress class therefore represents only 0.84 percent of the sample. This confirms that the empirical setting is a rare-event prediction problem rather than a balanced classification task.

This rare-event structure shapes the interpretation of the model results. ROC-AUC remains useful because it measures broad discriminatory power, but it cannot be the sole basis for evaluating an early-warning system. In a sample where fewer than one percent of observations are classified as distress, PR-AUC, precision, recall, and F1-score provide a more decision-relevant view of model performance. These metrics evaluate whether the model can identify future PKPU or bankruptcy cases in the high-risk tail of the predicted-probability distribution. The results should therefore be read through two lenses: statistical discrimination and operational usefulness.

Table 4 presents the out-of-sample performance of all models on the test set. The results show a clear performance gap between Logistic Regression and tree-ensemble models. Logit V1 records ROC-AUC of 0.657 and PR-AUC of 0.019, while Logit V2 records ROC-AUC of 0.649 and PR-AUC of 0.018. The addition of market variables to the linear Logit specification does not improve out-of-sample performance.

Tree-ensemble models deliver stronger results. Random Forest records the highest ROC-AUC at 0.817, while XGBoost records a nearly identical ROC-AUC of 0.816. However, ROC-AUC does not tell the full story. Random Forest achieves higher recall, but its precision and

F1-score remain low. XGBoost delivers the strongest rare-event performance among the standalone models, with PR-AUC of 0.151, precision of 0.196, recall of 0.244, and F1-score of 0.217. This means that XGBoost provides a better balance between detecting true distress cases and limiting false alarms.

Table 4. Model Performance Comparison on Test Set

Model	ROC-AUC	PR-AUC	Precision	Recall	F1-Score
Logit V1	0.657	0.019	0.020	0.431	0.038
Logit V2	0.649	0.018	0.015	0.379	0.029
Random Forest	0.817	0.092	0.058	0.496	0.103
XGBoost	0.816	0.151	0.196	0.244	0.217
Hybrid XGBoost + Structural Indicators	0.812	0.134	0.198	0.265	0.227

Source: Authors' estimation using Python

These results support H1. Tree-ensemble models outperform Logistic Regression in out-of-sample corporate distress prediction among Indonesian non-financial public firms. More specifically, XGBoost is the most decision-relevant standalone model because it performs best on PR-AUC, precision, and F1-score. This finding is consistent with the view that corporate distress is shaped by nonlinear relationships among solvency, leverage, profitability, market volatility, and other firm-level signals.

Incremental Value of Merton Structural Indicators

The Hybrid XGBoost model tests whether Merton-based structural indicators add value when embedded into a nonlinear classifier. The results are nuanced. Hybrid XGBoost records ROC-AUC of 0.812 and PR-AUC of 0.134, both lower than standalone XGBoost. This means that adding Distance-to-Default and Merton probability of default does not improve aggregate ranking performance.

The operational classification metrics tell a different story. Hybrid XGBoost improves precision from 0.196 to 0.198, recall from 0.244 to 0.265, and F1-score from 0.217 to 0.227. This indicates that the structural indicators help the model detect more true distress observations at the classification threshold, even though they do not improve overall ranking metrics. In other words, the structural signal is not useful as a broad AUC booster, but it contributes to distress detection where the early-warning decision is actually made.

This pattern is economically intuitive. Distance-to-Default is partly constructed from market value, volatility, and debt obligations. These components overlap with variables already available to XGBoost, such as leverage, market capitalization, equity volatility, and return-based indicators. Once those variables are included separately, the Merton signal may not add much new information across the full score distribution. Its contribution appears more clearly near the high-risk boundary, where small changes in classification can determine whether a distressed firm is flagged or missed.

These findings support H2 in its operational form. Embedding Merton-based structural indicators into XGBoost provides incremental distress-detection value, reflected in higher precision, recall, and F1-score, even when ROC-AUC and PR-AUC do not improve. The appropriate interpretation is not that the hybrid model dominates XGBoost across all metrics. Rather, the evidence shows that Merton-based indicators provide theory-guided detection value within a machine-learning early-warning framework.

SHAP-Based Interpretation of Distress Drivers

The SHAP analysis explains why the XGBoost model classifies firms as high risk. Figure 2 reports the SHAP contribution by risk-driver category. The largest contribution comes from leverage and solvency variables, which account for 35.5 percent of total SHAP importance. This group includes equity ratio, interest coverage, leverage, and debt-to-equity. Profitability

and efficiency contribute 21.5 percent, followed by asset structure at 15.3 percent, volatility and liquidity at 13.7 percent, market signals at 11.9 percent, and macro variables at 1.9 percent.

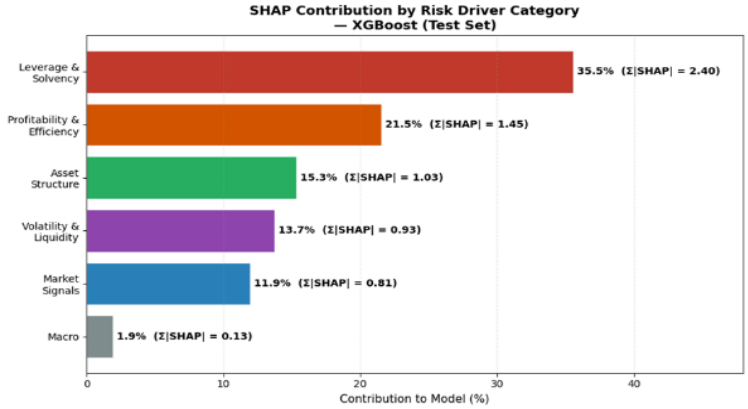

Figure 2. SHAP Contribution by Risk Driver Category

The dominance of leverage and solvency variables is consistent with corporate finance theory. Distress risk rises when a firm's equity buffer weakens, leverage increases, and operating earnings become less sufficient to cover interest obligations. Profitability and efficiency variables add an operating dimension: firms with weaker earnings generation and lower asset productivity are more likely to move toward distress. Asset structure also matters because fixed-asset intensity may affect operating flexibility, collateral value, and the firm's ability to adjust under financial pressure.

Market variables play a complementary role. Volatility, stock return, market capitalization, turnover, and excess return provide faster signals of changing investor expectations. However, they do not dominate the model. This is an important finding for the Indonesian setting. In markets where liquidity and information quality vary across firms, accounting fundamentals remain central, while market variables act as forward-looking complements.

These findings support H3. SHAP identifies economically interpretable indicators of solvency, leverage, profitability, efficiency, asset structure, and market risk as dominant contributors to predicted corporate distress. This reduces the black-box concern around XGBoost. The model is not only more accurate than Logistic Regression; it also learns risk patterns that are consistent with established financial logic.

Risk-Bucket and Early-Warning Performance

Model performance metrics are useful for comparing algorithms, but they do not fully answer the practitioner's question: which firms should be reviewed first? To address this issue, the predicted distress probabilities are translated into Low, Medium, and High-Risk buckets. This converts the model output into a watchlist format that is easier to use for credit monitoring, portfolio screening, and early-warning review.

Table 5 reports the early-warning performance of the risk-bucket framework. The high-risk bucket captures 66.7 percent of actual distress events, equivalent to 26 of 39 distress events in the evaluation window. The high-risk bucket also produces a risk lift of 4.8 times the sample average, meaning that firms in this group have a materially higher concentration of future distress events than the overall population. The average lead time is 8.5 months, while the median lead time is 11.0 months.

Table 5. Risk-Bucket and Early-Warning Performance

Metric	Result	Interpretation
High-risk bucket coverage	66.70%	Captures 26 of 39 actual distress events in the evaluation window

High-risk lift bucket	4.8x	Distress risk in the high-risk bucket is 4.8 times above the sample average
Average lead time	8.5 months	Model flags distress events, on average, 8.5 months before PKPU or bankruptcy
Median lead time	11.0 months	Many signals appear well before the legal distress event
Practical use	High-risk watchlist	Supports monitoring, credit review, portfolio screening, and early-warning prioritization

Source: Authors' early-warning analysis

These results show that the model has practical early-warning value. The model does not merely classify distress after the fact; it identifies a substantial share of future PKPU or bankruptcy events several months before they occur. This lead time is meaningful because it gives decision-makers time to review exposures, reassess risk limits, monitor liquidity pressure, or initiate further credit and investment analysis.

The risk-bucket result also strengthens the role of SHAP. A high-risk classification can be interpreted together with the firm's main SHAP drivers. This creates a practical workflow: identify the high-risk firm, understand the financial drivers behind the signal, and determine whether further review is needed. In this sense, the model functions not only as a prediction engine, but also as an explainable risk-monitoring tool.

Discussion

The empirical evidence supports the article's central argument: corporate distress prediction in Indonesia benefits from combining nonlinear machine learning, structural credit-risk signals, and explainable diagnostics. Logistic Regression remains useful as a transparent benchmark, but it performs weakly under rare-event metrics. Random Forest provides strong recall and broad discrimination, but its low precision and F1-score make it less attractive as an operational watchlist model. XGBoost provides the strongest standalone balance across rare-event metrics and is therefore the most useful model for early-warning purposes.

The hybrid model adds a more nuanced contribution. The addition of Merton-based structural indicators does not improve ROC-AUC or PR-AUC. This suggests that the structural signal partly overlaps with accounting and market features already captured by XGBoost. However, the hybrid model improves precision, recall, and F1-score. This means the structural signal contributes at the operational classification level, where the model decides which firms should be flagged. For rare-event risk monitoring, that distinction matters.

The SHAP results confirm that the model's predictions are economically credible. The main drivers of predicted distress are not arbitrary variables. They are solvency, leverage, profitability, efficiency, asset structure, and market signals. These are precisely the dimensions that finance theory and credit-risk practice would expect to matter. This strengthens the case for using explainable machine learning in corporate distress prediction, especially in institutional settings where model outputs must be interpretable and defensible.

The findings also clarify the role of structural credit-risk theory in modern machine-learning models. The evidence does not suggest that Merton-based indicators should replace machine learning, nor that they automatically improve all performance metrics. Their value is more specific: they provide a theory-guided signal that can improve detection at the decision threshold. This supports the article's broader framing that Merton should be embedded inside machine learning, rather than treated only as a competing standalone model.

Overall, the contribution of this study is not simply that XGBoost produces a higher AUC than Logit. The stronger contribution is the development of an explainable early-warning framework for Indonesian corporate distress. XGBoost provides the predictive engine. Merton-based indicators provide structural risk information. SHAP explains the drivers. Risk buckets translate predicted probabilities into an actionable monitoring tool. Together, these elements make the framework relevant not only for academic bankruptcy prediction research, but also

for practitioners who need to identify, understand, and prioritize corporate distress risk before legal distress events occur.

CONCLUSION

This article develops an explainable early-warning framework for predicting corporate distress among non-financial firms listed on the Indonesia Stock Exchange. Using a firm-month panel of 107,448 observations from 2014 to 2024 and a 12-month forward distress label based on PKPU and bankruptcy events, the study compares Logistic Regression, Random Forest, XGBoost, and a Merton-augmented XGBoost model. The results show that tree-ensemble models outperform Logistic Regression in rare-event distress prediction. XGBoost provides the strongest standalone early-warning performance, with PR-AUC of 0.151 and F1-score of 0.217, compared with Logit's PR-AUC of around 0.018. The hybrid model does not improve aggregate ROC-AUC or PR-AUC, but it improves recall and F1-score, indicating that Merton-based structural indicators add detection value at the operational threshold. SHAP analysis further shows that the model's predictions are driven by economically meaningful indicators, particularly solvency, retained earnings, debt-servicing capacity, leverage, profitability, efficiency, asset structure, and market signals.

The article contributes to the literature by repositioning the relationship between structural credit-risk theory and machine learning. Rather than treating Merton-based structural models and machine-learning models as competing alternatives, this study embeds Merton Distance-to-Default and Merton probability of default as theory-guided features inside a nonlinear classifier. In the Indonesian sample, the structural signal does not raise aggregate ranking performance, but it improves distress detection at the classification threshold. The implication is that the relevant question is not whether structural models should replace machine learning, or whether machine learning should discard structural theory. The more useful question is how structural credit-risk theory can improve specific decision-relevant outputs when embedded in a flexible and explainable predictive architecture.

The practical implication is also clear. The model is not only evaluated as a statistical classifier but translated into a usable early-warning tool. The high-risk bucket captures 66.7 percent of actual distress events, or 26 of 39 events in the evaluation window, with an average lead time of 8.5 months and a median lead time of 11 months. The high-risk bucket also has a 4.8 times risk lift relative to the sample average. For creditors, investors, and corporate risk managers, this means the model can help prioritize firms for deeper review before PKPU or bankruptcy becomes legally observable. SHAP further strengthens the practical value of the framework by explaining why a firm is classified as high risk. A risk flag can therefore be linked to specific financial drivers, such as weak equity buffer, low retained earnings, poor interest coverage, high leverage, or deteriorating market signals. The framework should not be read as a substitute for expert judgment, but as a structured tool to support credit monitoring, portfolio screening, and early risk diagnosis.

Several limitations should be acknowledged. First, the distress label is based on PKPU and bankruptcy events. This provides a hard and publicly verifiable definition, but it excludes firms that experience severe financial distress without entering formal court proceedings, such as out-of-court restructuring, covenant breaches, or informal creditor negotiations. Second, the structural indicator is based on a Merton-style Distance-to-Default construction; alternative structural specifications, such as iterative KMV estimation, Geske-type models, or different volatility assumptions, may produce different incremental effects. Third, the sample is limited to non-financial firms, so the findings cannot be generalized to banks, insurers, multifinance companies, or other regulated financial institutions with different balance-sheet structures and capital requirements. Fourth, the Indonesia Stock Exchange contains firms with highly uneven liquidity and market capitalization. Market-based predictors may be less reliable for thinly traded firms, and future research should test whether model performance differs across liquidity or size groups. Fifth, the study uses a time-based split to reduce data leakage, but it does not

implement full walk-forward validation across multiple rolling windows. Future studies can strengthen external validity by testing the framework across different market regimes. Sixth, SHAP is used to interpret model predictions, not to establish causal relationships. Future research can extend this work by testing SHAP stability across retrained models, adding calibration analysis, incorporating textual information from annual reports or audit opinions, and applying the framework to other ASEAN capital markets.

REFERENCES

Alanis, E., Chava, S., & Shah, A. (2022). Benchmarking machine learning models to predict corporate bankruptcy (SSRN Scholarly Paper No. 4249412). Social Science Research Network. https://doi.org/10.2139/ssrn.4249412

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609. https://doi.org/10.2307/2978933

Bank Indonesia. (2024). *BI rate*. Bank Indonesia.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405-417. https://doi.org/10.1016/j.eswa.2017.04.006

Bauer, J., & Agarwal, V. (2014). Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. *Journal of Banking & Finance, 40*, 432-442. https://doi.org/10.1016/j.jbankfin.2013.12.013

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, 4*, 71-111. https://doi.org/10.2307/2490171

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton Distance to Default model. *The Review of Financial Studies, 21*(3), 1339-1369. https://doi.org/10.1093/rfs/hhn044

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance, 63*(6), 2899-2939. https://doi.org/10.1111/j.1540-6261.2008.01416.x

Credit Guarantee and Investment Facility. (2024). *ASEAN+3 corporate bond market research 2024*. Asian Development Bank.

Direktorat Jenderal Badan Peradilan Umum. (2024). *Laporan pelaksanaan kegiatan Direktorat Jenderal Badan Peradilan Umum*. Mahkamah Agung Republik Indonesia.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/0471722146

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765-4774). Curran Associates.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 56-67. https://doi.org/10.1038/s42256-019-0138-9

Mahkamah Agung Republik Indonesia. (n.d.). *Sistem Informasi Penelusuran Perkara Pengadilan Niaga*. Retrieved from Sistem Informasi Penelusuran Perkara websites of Indonesian commercial courts.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance, 29*(2), 449-470. https://doi.org/10.2307/2978814

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109-131. https://doi.org/10.2307/2490395

Peykani, P., Salehi, M., & Karimi, A. (2023). The application of structural and machine learning models to predict the default risk of listed companies in Iranian capital market. *Financial Innovation, 9*(1), Article 70. https://doi.org/10.1186/s40854-023-00494-3

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

PT Bursa Efek Indonesia. (2024). *IDX market data*. Indonesia Stock Exchange.

PT Pemeringkat Efek Indonesia. (2024). *The default study: Period of 2007-2024*. PEFINDO.

S&P Global Market Intelligence. (2024). *S&P Capital IQ database*.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business, 74*(1), 101-124. https://doi.org/10.1086/209665

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32-35. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research, 22*, 59-82. https://doi.org/10.2307/2490859