World Journal of Management Science
Print ISSN: 2959-9628
Online ISSN: 2959-9636
DOI: https://doi.org/10.61784/wms4120

AN EXPLAINABLE HYBRID MACHINE LEARNING FRAMEWORK FOR CASH-FLOW-CONDITIONED MULTI-HORIZON LIQUIDITY RISK EARLY WARNING: AN SME-ORIENTED BENCHMARK STUDY

XinRan Yue[1*], RuFeng Zhu[2]
[1]*University of Denver, Denver 80208, CO 80208, United States.*
[2]*Brandeis University, Waltham 02453, MA 02453, United States.*
[*]*Corresponding Author: XinRan Yue*

Abstract: Small and medium-sized enterprises (SMEs) are particularly exposed to liquidity shocks, yet public, transaction-level SME cash-flow panels suitable for reproducible forecasting remain scarce. This paper therefore studies a narrower but operationally related task: cash-flow-conditioned, multi-horizon liquidity-risk early warning. A real public benchmark containing 43,405 financial-statement observations, 64 ratios, missing values, and bankruptcy labels one to five years ahead is used as an evidence base. The proposed framework combines three views: an L2-regularized logistic model, a LightGBM model using all ratios, and a dedicated LightGBM expert using 23 cash-flow and liquidity ratios. Non-negative horizon-specific weights are selected on a calibration split in log-odds space, followed by isotonic probability calibration, a recall-constrained warning threshold, and SHAP-based feature and semantic-group explanations. Across five held-out horizons, the calibrated hybrid achieved mean ROC-AUC 0.914, PR-AUC 0.617, and Brier score 0.0260. The uncalibrated ensemble had higher ranking performance (PR-AUC 0.663) but materially worse probability error (Brier 0.0424), exposing a genuine calibration-discrimination trade-off. The cash-flow expert received positive weight at the two- and five-year horizons, while the full-feature tree dominated the remaining horizons; the PR-AUC gain over LightGBM-all was not statistically significant (one-sided Wilcoxon p=0.50). The study supplies executable code, data hashes, result tables, figures, and verified references. Because the benchmark does not identify firms by a legal SME definition and is not a cash-flow time series, the findings support an SME-oriented methodology rather than a claim of direct SME field validation.
Keywords: SME; Liquidity risk; Financial distress; Cash-flow ratios; LightGBM; Probability calibration; SHAP; Early warning

1 INTRODUCTION

Liquidity failure is often the terminal mechanism through which otherwise viable firms enter distress: obligations become due before sufficiently liquid resources can be mobilized. The problem is especially consequential for SMEs because they typically have thinner cash buffers, greater customer and supplier concentration, less diversified funding, and more limited access to emergency capital. Classical failure-prediction research established that accounting ratios contain useful leading information, while later hazard and market-based studies emphasized that distress is dynamic and horizon dependent [1-6]. SME-specific evidence further shows that a model designed for large corporations need not transfer unchanged to smaller firms [7-8].

The practical target in this study is an early-warning system that estimates the probability of severe liquidity distress before bankruptcy. Such a system should do more than rank firms. It should provide probabilities that are usable for triage, maintain recall at an explicitly chosen operating point, and expose the financial drivers of each warning. These requirements create several methodological tensions. Rare-event class imbalance makes accuracy misleading [9-10]. Flexible tree ensembles frequently outperform linear models on tabular credit data, but their outputs can be poorly calibrated [11-14]. Post-hoc explanations are useful, yet high-stakes use also requires transparent model structure, bounded claims, and human review [15-18].

A direct implementation of the original topic—transaction-level cash-flow amount forecasting followed by liquidity-risk classification—would require firm-level cash balances, receipts, payments, credit lines, and dated distress outcomes. No openly downloadable dataset located for this project simultaneously provided those elements for a sufficiently large SME panel. Rather than fabricate cash-flow sequences or label synthetic output as empirical evidence, the paper pivots to a related and reproducible question: can cash-flow and liquidity ratios contribute an interpretable, multi-horizon distress-warning framework when evaluated on a real public benchmark? This pivot preserves the substantive direction while separating what the data support from what they do not.

The resulting contribution is fourfold. First, the study defines a leakage-resistant, horizon-specific evaluation protocol on five imbalanced datasets instead of pooling lead times. Second, it introduces an adaptive three-view ensemble that contains a dedicated cash-flow/liquidity expert but permits validation to assign any branch zero weight. Third, ranking, probability calibration, and the warning threshold are treated as distinct design problems. Fourth, SHAP values are

aggregated into auditable financial groups, while negative findings—especially the absence of a significant gain over the full LightGBM branch—are reported explicitly. All reported numbers are generated by the accompanying code from the downloaded data files.

2 RELATED WORK

Early corporate-failure models relied on univariate ratios, discriminant analysis, logit, and probit formulations [1-4]. Their enduring value is interpretability, but reviews identify restrictive distributional assumptions, unstable variable selection, sampling bias, and sensitivity to the definition and timing of failure [19-20]. Shumway's discrete-time hazard formulation and Campbell et al.'s distress-risk model helped shift attention toward time-to-event structure and horizon-specific prediction [5-6].

SME credit risk has distinct empirical characteristics. Altman and Sabato developed an SME-specific logit model and showed that specialized modeling can outperform a generic corporate specification [7]. Ciampi and Gordini demonstrated nonlinear neural models on Italian small enterprises [8]. These studies support segment-aware modeling, but proprietary data and heterogeneous legal definitions limit exact replication across countries.

Machine learning expands the functional forms available for distress prediction. Gradient boosting and related ensembles capture nonlinearities, missingness patterns, and feature interactions [21-22]. Comparative studies generally find strong performance from tree-based methods, although results vary by dataset, sampling, and metric [9, 11-12, 23-24]. The UCI Polish Companies Bankruptcy benchmark was introduced with an ensemble boosted-tree study and remains useful because it provides five lead-time datasets with identical ratio semantics [23, 25].

Explainability and probability quality are separate from discrimination. SHAP provides additive feature attributions with a game-theoretic foundation, and TreeSHAP makes them computationally practical for tree ensembles [15-16]. LIME offers a model-agnostic local alternative [17]. Calibration research shows that strong rankers can still produce distorted probabilities and that isotonic or sigmoid transformations can correct this behavior [13-14]. PR curves are preferable to accuracy and often more informative than ROC curves when the positive class is rare [10]. Brier score evaluates the squared error of probabilistic forecasts [26]. Finally, Rudin [18] and Kapoor and Narayanan [27] motivate restrained interpretation, transparent protocols, and explicit controls against leakage.

Cash-flow forecasting research also cautions against equating earnings with cash generation. Accrual components can help predict future cash flows, while the mapping between earnings and cash flows changes with the operating cycle and accrual process [28-29]. The present study does not forecast cash-flow amounts; instead, it uses ratios directly connected to working capital, cash earnings, current liabilities, and liquidity coverage as a specialized evidence view.

3 DATA AND PROBLEM FORMULATION

3.1 Data Source and Scope

The experiments use the Polish Companies Bankruptcy dataset from the UCI Machine Learning Repository [25], whose associated study is [23]. The source contains five ARFF files. Each file represents a different lead time: the file named 5year.arff contains observations one year before the outcome, whereas 1year.arff contains observations five years before the outcome. The bankrupt observations were drawn from 2000-2012 and the operating observations from 2007-2013. The five files contain 43,405 financial-statement observations in total, but the repository does not provide persistent firm identifiers; consequently, this count must not be interpreted as 43,405 unique firms.

Each observation contains 64 real-valued ratios and a binary bankruptcy label. The ratios span profitability, leverage, liquidity, working capital, turnover, and size. Missingness ranges from 1.23% to 1.87% of cells. Positive prevalence falls from 6.94% at one year to 3.86% at five years, making the long-horizon problem highly imbalanced, see Table 1.

Table 1 Dataset Composition by Warning Horizon

Lead	File	N	Pos.	Prev.	Missing
1 y	5year.arff	5,910	410	6.94%	1.23%
2 y	4year.arff	9,792	515	5.26%	1.40%
3 y	3year.arff	10,503	495	4.71%	1.47%
4 y	2year.arff	10,173	400	3.93%	1.87%
5 y	1year.arff	7,027	271	3.86%	1.30%

3.2 SME-Oriented Interpretation

The repository does not include employee count, revenue thresholds, or legal SME indicators. Therefore, the benchmark cannot validate performance specifically within a statutory SME population. It is used as a real corporate-distress testbed for an SME-oriented method. Any deployment claim requires external validation on a jurisdiction-specific SME panel.

3.3 Prediction Task

For each horizon h in {1,...,5}, let x_h in R^64 be a financial-ratio vector and y_h in {0,1} indicate bankruptcy at the end of the horizon. The objective is to estimate P(y_h=1 | x_h), then convert that probability into a warning decision using a threshold selected without test-label access. Horizons are modeled separately to prevent cross-horizon leakage and to allow branch weights to change with lead time.

3.4 Cash-Flow/Liquidity View

Twenty-three ratios are designated ex ante as the specialized view: working capital/assets, current ratio, liquidity-gap days, gross profit/current liabilities, cash earnings/sales, liability-coverage days, cash earnings/liabilities, gross profit/assets, cash earnings/liabilities including depreciation, liabilities less cash/sales, operating expenses/current liabilities, cash quick ratio, liabilities/cash operating capacity, acid-test ratio, EBITDA ratios, current assets/liabilities, current liabilities/assets, current-liability days, working capital, liquidity surplus/cash expense, current-liability sales days, and sales/current liabilities. This view is not claimed to be a cash-flow statement; it is a transparent proxy set tied to cash conversion and short-term funding pressure.

4 PROPOSED FRAMEWORK

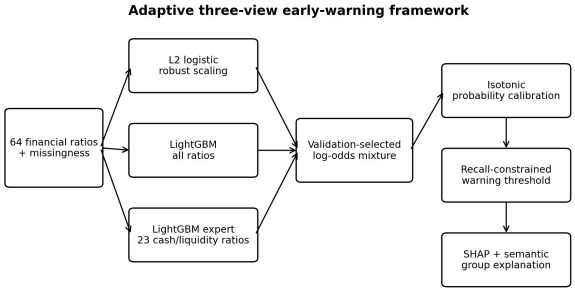

Figure 1 Adaptive Three-View Warning Framework

4.1 Data Partitioning and Preprocessing

Each horizon is split by stratified sampling into 60% training, 20% calibration, and 20% test sets, see Figure 1. The test set is untouched until final evaluation. For the linear branch, missing values are median-imputed with missingness indicators and transformed by robust scaling using the 10th and 90th percentiles. LightGBM branches receive the original numeric matrix and handle missing values internally.

4.2 Three Complementary Branches

The first branch is an L2-regularized logistic model with class-balanced loss. It supplies a low-variance linear perspective and an interpretable baseline. The second branch is a LightGBM classifier over all 64 ratios. The third is a LightGBM expert restricted to the 23 cash-flow/liquidity ratios. Both tree branches use 80 estimators, learning rate 0.08, 15 leaves, minimum child size 35, column subsampling, L1 regularization 0.15, L2 regularization 1.5, and balanced class weights. These fixed settings limit opportunistic tuning.

4.3 Adaptive Log-Odds Mixture

Let p_h,k(x) denote the probability from branch k at horizon h. The uncalibrated ensemble is

$$z_h(x) = \sum_k w_{h,k} \log\left[\frac{p_{h,k}(x)}{1-p_{h,k}(x)}\right], \; p_h^{\text{raw}}(x) = \frac{1}{1+\exp(-z_h(x))} \quad (1)$$

Weights are non-negative and sum to one. They are selected on the calibration set from a coarse simplex grid with step 0.25 to maximize PR-AUC; Brier score breaks ties. A coarse grid makes the fusion auditable and constrains variance. Crucially, a branch may receive zero weight. The architecture is therefore hybrid by available evidence views, not by forcing every view into every horizon.

4.4 Probability Calibration

An isotonic regression map g_h is fitted on the calibration scores, producing p_h(x)=g_h(p_h^raw(x)). Isotonic calibration can create tied scores and therefore can reduce ROC-AUC or PR-AUC even while improving probability error. The framework evaluates this trade-off rather than assuming calibration improves every metric [13-14].

4.5 Operating Threshold

The warning threshold tau_h is selected on the calibrated validation probabilities. Among thresholds with recall at least 0.80, the method chooses the one with highest precision, with ties favoring higher recall and then the higher threshold. This rule encodes an explicit missed-distress tolerance. Test precision, recall, F1, and balanced accuracy are then reported at tau_h.

4.6 Explanations

TreeSHAP values are computed for the full and specialized tree branches [15-16]. Linear coefficients and branch-specific SHAP magnitudes are normalized and combined using the selected mixture weights. Feature importance is also aggregated into profitability, liquidity/cash flow, leverage/solvency, efficiency/turnover, and size/capital-structure groups. Overlapping ratios are divided equally across their assigned groups to avoid double counting.

5 EXPERIMENTAL PROTOCOL

The study uses one pre-registered random seed (17) for the 60/20/20 stratified split at every horizon. This design was chosen to complete all five real-data experiments reproducibly within the execution budget; it is not a substitute for repeated nested validation. No synthetic observations, SMOTE samples, or manually edited labels are used. The raw ARFF files are distributed with SHA-256 hashes in the metadata.

Baselines are the L2 logistic branch, LightGBM-all, and LightGBM-cash. Two ablations are also computed: a two-branch mixture without the cash expert and the three-branch ensemble without isotonic calibration. PR-AUC is the primary ranking metric because positive prevalence is below 7% [10]. ROC-AUC is included for comparability. Brier score measures probability error [26]. Balanced accuracy, precision, recall, and F1 describe the selected warning operating point.

Uncertainty is summarized in two ways. First, mean and standard deviation across the five lead-time tasks describe horizon variability; these are not repeated-seed standard deviations. Second, each proposed-model test result receives a 95% percentile interval from 1,000 nonparametric bootstrap resamples, conditional on the held-out split. A one-sided Wilcoxon signed-rank test compares PR-AUC differences across the five horizons. With only five paired tasks, this test has low power and its p-values should be interpreted cautiously.

The software environment was Python 3.13.5, NumPy 2.3.5, pandas 2.2.3, LightGBM 4.6.0, SHAP 0.50.0, SciPy, and scikit-learn. The supplied script downloads missing data from a public mirror, recomputes all metrics and figures, and records file hashes and package versions.

6 RESULTS

6.1 Overall Discrimination and Calibration

The calibrated hybrid achieved mean ROC-AUC 0.9139, PR-AUC 0.6168, and Brier score 0.0260 over the five horizons. LightGBM-all produced ROC-AUC 0.9123, PR-AUC 0.6147, and Brier 0.0263. The mean PR-AUC difference was only 0.0021 and was not significant (one-sided Wilcoxon p=0.50). The calibrated hybrid significantly exceeded the L2 logistic and cash-only branches across the five tasks at the minimum attainable one-sided p-value of 0.03125, but the small number of horizons makes this evidence descriptive rather than definitive, see Table 2.

Table 2 Mean Performance Across Five Horizons

Model	ROC	PR	Brier	Prec.	Rec.	F1
L2 logistic	0.859	0.319	0.0372	0.169	0.836	0.263
LightGBM-all	0.912	0.615	0.0263	0.279	0.783	0.390
LightGBM-cash	0.780	0.207	0.0423	0.095	0.787	0.169
Proposed hybrid	0.914	0.617	0.0260	0.298	0.776	0.407
Hybrid raw	0.921	0.663	0.0424	0.304	0.768	0.414

The uncalibrated three-view ensemble achieved higher mean ROC-AUC (0.9213) and PR-AUC (0.6631), but its Brier score was 0.0424. Isotonic calibration lowered Brier error by 38.6% relative to the raw ensemble while reducing PR-AUC by 0.0463. This is not a contradiction: monotone isotonic regression can create ties and compress extreme scores. The result demonstrates why a deployment-oriented study should report both discrimination and probability quality, see Figure 2.

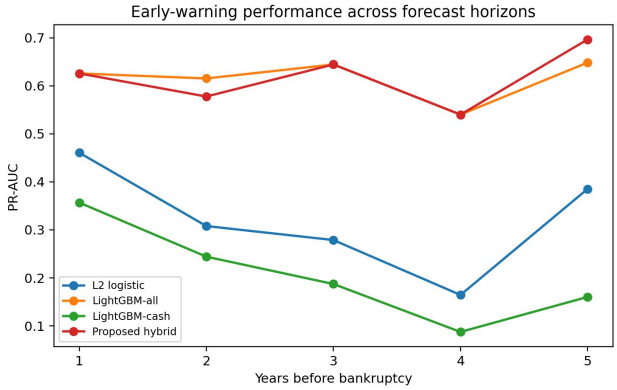

Figure 2 PR-AUC Across Warning Horizons

6.2 Horizon-Specific Results

Proposed-model PR-AUC ranged from 0.540 at four years to 0.696 at five years. The corresponding 95% split-conditional bootstrap intervals were [0.426, 0.649] and [0.568, 0.814]. The one-year model reached PR-AUC 0.626, ROC-AUC 0.913, Brier 0.0365, precision 0.481, recall 0.634, and F1 0.547. Recall on the test set can fall below the calibration target of 0.80 because the threshold rule is selected on a different sample; this is expected sampling variation, not a post-hoc threshold change, see Table 3 and Figure 3.

Table 3 Proposed Model by Horizon (95% Bootstrap CI)

Year	PR	PR 95% CI	ROC	Brier	P	R	F1
1	0.626	[0.524, 0.723]	0.913	0.0365	0.481	0.634	0.547
2	0.578	[0.478, 0.674]	0.928	0.0288	0.302	0.806	0.439
3	0.644	[0.555, 0.726]	0.919	0.0257	0.220	0.828	0.348
4	0.540	[0.426, 0.649]	0.876	0.0228	0.113	0.850	0.199
5	0.696	[0.568, 0.814]	0.933	0.0165	0.373	0.759	0.500

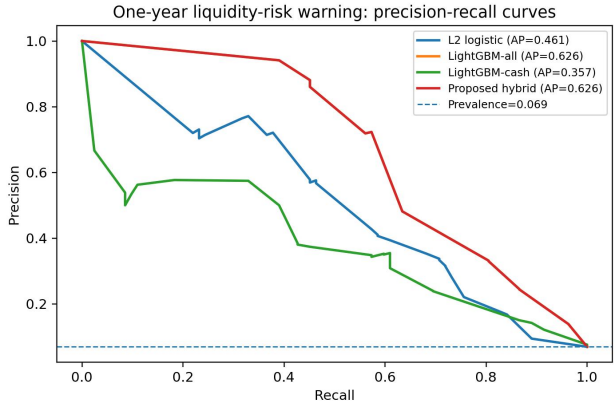

Figure 3 One-Year Held-Out Precision-Recall Curves

6.3 Adaptive Branch Selection

The L2 logistic branch received zero weight at all horizons. The full LightGBM branch received weight 1.0 at one, three, and four years. At two and five years, the validation procedure assigned 0.75 to LightGBM-all and 0.25 to the cash-flow/liquidity expert. The cash expert therefore contributed at specific longer or intermediate horizons, but it was not uniformly useful. This is a substantive result: a domain view can add complementary evidence without being strong enough as a standalone classifier. Conversely, forcing equal weights would have degraded performance, see Table 4 and Figure 4.

Table 4 Validation-Selected Branch Weights

Lead year	L2 logit	LGB-all	Cash expert
1	0.00	1.00	0.00
2	0.00	0.75	0.25
3	0.00	1.00	0.00
4	0.00	1.00	0.00
5	0.00	0.75	0.25

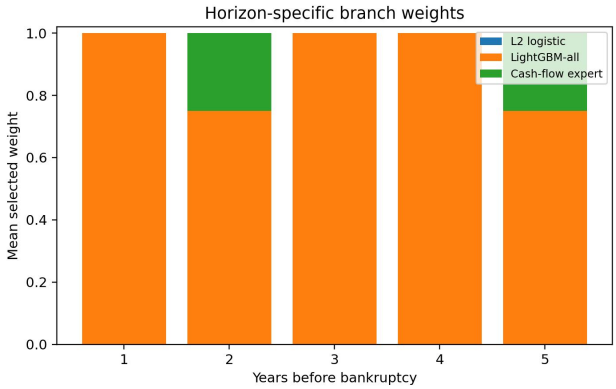

Figure 4 Validation-Selected Evidence-View Weights

6.4 Explanation Audit

For the one-year task, the selected weights were (0,1,0), so the reported hybrid attribution equals the full-tree SHAP attribution. The largest normalized drivers were sales growth (0.137), operating profit/financial expense (0.126), operating expenses/liabilities (0.075), acid-test ratio (0.065), and reserves/assets (0.053). At the semantic-group level, efficiency/turnover accounted for 31.9%, profitability 30.8%, liquidity/cash flow 18.6%, leverage/solvency 13.0%, and size/capital structure 5.7%. The presence of acid-test ratio, liquidity-gap days, cash earnings/sales, and cash quick ratio among leading features supports the relevance of liquidity information, but the dominant groups show that liquidity warnings cannot be reduced to cash ratios alone, see Table 5-6, and Figure 5.

Table 5 Top One-Year SHAP Drivers

Feature	Description	Importance
Attr21	Sales growth	0.137
Attr27	Operating profit / financial expense	0.126
Attr34	Operating expenses / liabilities	0.075
Attr46	Acid-test ratio	0.065
Attr25	Reserves / assets	0.053
Attr39	Profit on sales / sales	0.036
Attr6	Retained earnings / assets	0.033
Attr24	3-year gross profit / assets	0.031

Figure 5 Top One-Year Normalized Feature Importance

Table 6 One-Year Semantic-Group Importance

Financial group	Normalized share
Efficiency & turnover	31.9%
Profitability	30.8%
Liquidity & cash flow	18.6%
Leverage & solvency	13.0%
Size & capital structure	5.7%

6.5 Horizon-Level Calibration Trade-Off

Calibration reduced Brier score at every horizon: the absolute reduction ranged from 0.0096 at one year to 0.0291 at three years. At the same time, calibrated PR-AUC was lower at all five horizons, with reductions from 0.0290 to 0.0638. The direction is therefore consistent across tasks rather than being driven by one anomalous horizon. The result also clarifies the role of the calibration split: isotonic regression is retained because the intended output is a probability for triage, but the uncalibrated score should be preserved when ranking capacity is the sole objective. An operational system can store both outputs while ensuring that thresholds and expected-risk calculations use the calibrated probability, see Table 7 and Figure 6.

Table 7 Calibration-Discrimination Trade-Off by Horizon

Year	PR raw	PR cal.	Delta PR	Brier raw	Brier cal.	Delta Brier
1	0.689	0.626	-0.064	0.0461	0.0365	-0.0096
2	0.615	0.578	-0.037	0.0485	0.0288	-0.0197
3	0.673	0.644	-0.029	0.0548	0.0257	-0.0291
4	0.593	0.540	-0.053	0.0394	0.0228	-0.0166
5	0.745	0.696	-0.049	0.0234	0.0165	-0.0069

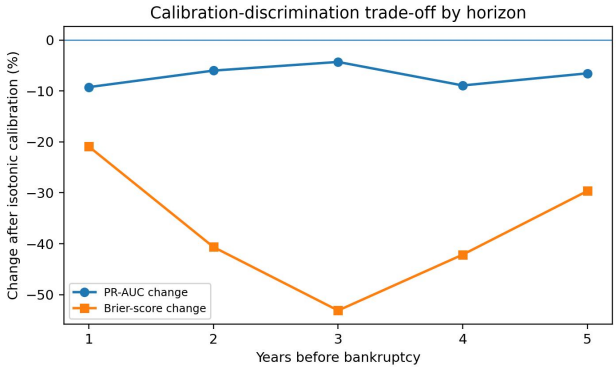

Figure 6 Relative Metric Changes After Isotonic Calibration

6.6 Reproducibility Checks

The five data-file hashes in the metadata match the distributed ARFF files. All tables in this paper are generated from CSV outputs rather than manually transcribed experimental claims. The result package also includes the calibration ablation, branch weights, per-horizon metrics, bootstrap intervals, SHAP feature importance, semantic-group importance, and figures. The exact calibration-versus-ranking values reported in Table 7 are read from the same per-run output used for the aggregate results.

7 DISCUSSION

The experiments support a restrained conclusion. A full-feature boosted tree is the principal predictive engine on this benchmark. The proposed framework adds value not by producing a large universal PR-AUC gain, but by organizing complementary evidence, allowing the data to reject unhelpful branches, calibrating probabilities, and linking each alert to financial drivers. This distinction matters for EI-style engineering research: an operational decision-support pipeline is more than a leaderboard classifier.

For an SME lender or treasury platform, the calibrated probability can serve as a triage signal rather than an automatic credit decision. A high-risk alert could trigger review of cash balances, aged receivables, supplier terms, tax liabilities, covenant headroom, and available credit lines. The semantic explanation can guide that review. Thresholds should be chosen from the institution's cost matrix and capacity, not copied mechanically from this study.

The nonzero cash-expert weight at two and five years suggests that a specialized view may be useful when the signal carried by working capital and short-term liabilities differs from the broader profitability and efficiency signal. However, the cash-only branch's mean PR-AUC was only 0.207. The evidence therefore rejects a simplistic claim that cash ratios alone are sufficient. It also rejects the stronger claim that the hybrid significantly outperforms LightGBM-all on ranking quality.

Several limitations bound validity. First, the benchmark has no legal SME identifier, so external SME validity is untested. Second, the data are ratios from annual statements, not dated transaction-level cash-flow sequences; the study performs risk forecasting, not cash-balance forecasting. Third, absent persistent firm identifiers, the five horizon files cannot be joined and potential firm overlap cannot be audited. Fourth, the observations predate current accounting, lending, and macroeconomic regimes. Fifth, a single split leaves model-selection uncertainty; bootstrap intervals are conditional on that split and do not capture dataset shift. Sixth, isotonic calibration is data hungry and may overfit when

positives are extremely scarce. Seventh, SHAP explains the fitted model, not causality. These issues require prospective validation before use in credit or liquidity governance.

8 VALIDITY AND DEPLOYMENT REQUIREMENTS

8.1 External Validation Design

A field study should use a legally identified SME cohort, preserve calendar order, and define the prediction timestamp before any outcome-related information becomes available. Candidate inputs include opening and closing cash balances, invoice and payment dates, aged receivables and payables, payroll and tax schedules, revolving-credit utilization, covenant headroom, and contemporaneous accounting ratios. Outcomes should distinguish temporary cash shortfalls, covenant breaches, arrears, restructuring, and bankruptcy rather than collapsing every event into one label. A temporal train-validation-test design is preferable to random splitting because the intended deployment faces accounting-policy, interest-rate, and macroeconomic drift.

8.2 Joint Cash-Flow and Warning Evaluation

The original research direction can be restored when transactional data are available by coupling a cash-balance forecasting model with the present warning layer. Forecast quality should then be evaluated with scale-aware errors and interval coverage, while the warning layer should retain PR-AUC, Brier score, calibration plots, and operating-point metrics. The two components should not be optimized by a single ranking metric: a numerically accurate cash forecast can still miss tail liquidity events, and a strong classifier need not produce usable cash amounts. A joint protocol should therefore report both forecast error and the decision consequences of thresholded warnings.

8.3 Governance Controls

Before operational use, thresholds must be tied to review capacity and asymmetric costs, probability calibration must be checked by calendar period and firm segment, and explanations must be treated as review aids rather than causal claims. Alerts should route to a human analyst with access to source transactions and current credit-line information. Monitoring should include input missingness, score drift, subgroup performance, calibration drift, alert volume, and realized intervention outcomes. Model versions, data snapshots, threshold changes, and analyst overrides should be logged so that warnings can be audited. These controls are part of the proposed engineering framework because transparency at model-training time is insufficient without transparency at decision time.

9 REPRODUCIBILITY ARTIFACT

9.1 Artifact Contents

The accompanying package is designed to separate raw evidence, executable analysis, derived outputs, and manuscript claims. The data directory contains the five unchanged ARFF files used in the experiments. The experiment script contains dataset loading, horizon mapping, preprocessing, branch fitting, validation-weight selection, isotonic calibration, recall-constrained thresholding, bootstrap intervals, significance tests, SHAP aggregation, and figure generation. The results directory contains machine-readable per-run and aggregate metrics, branch weights, confidence intervals, feature and group importance, figures, and an experiment metadata file. A separate verification table records the bibliographic identifier and checked source for every reference.

9.2 Verification Workflow

A reproducer can first compare the SHA-256 values in the metadata with the distributed data files, then run the experiment script with the fixed seed and inspect the regenerated CSV files. Aggregate claims in the abstract, results, and conclusion are derived from those CSV outputs. The one-year explanation is traceable to the stored feature-importance table, and the calibration discussion is traceable to the raw-versus-calibrated rows in the per-run metric file. Package versions and the Python runtime are recorded to make environment differences visible rather than silently treating them as equivalent.

9.3 Reuse Boundary

Reproduction of this benchmark does not establish external validity for a new jurisdiction, calendar period, or statutory SME segment. The artifact is intended to make the reported study auditable and to provide an implementation baseline for a later time-stamped SME study. Any reuse should preserve the separation between training, calibration, and test information, should document changes to the cash/liquidity feature set and threshold rule, and should report negative or nonsignificant comparisons alongside favorable results.

10 CONCLUSION

This paper presented an explainable, cash-flow-conditioned, multi-horizon liquidity-risk early-warning framework using a real public corporate-distress benchmark. The framework combines linear, full-feature tree, and cash/liquidity expert views through horizon-specific validation weights, isotonic calibration, recall-constrained thresholding, and SHAP group explanations. The calibrated model achieved mean ROC-AUC 0.914, PR-AUC 0.617, and Brier score 0.0260. Its ranking advantage over LightGBM-all was not significant, while calibration materially improved probability error at the cost of ranking metrics. The central contribution is therefore a transparent and reproducible engineering workflow with explicit trade-offs, not an overstated claim of universal superiority. A true next step is external validation on time-stamped SME bank-account and accounting data, where cash-balance forecasts and distress warnings can be evaluated jointly.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Beaver W H. Financial ratios as predictors of failure. Journal of Accounting Research, 1966, 4: 71-111. DOI: 10.2307/2490171.
[2] Altman E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 1968, 23(4): 589-609. DOI: 10.1111/j.1540-6261.1968.tb00843.x.
[3] Ohlson J A. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 1980, 18(1): 109-131. DOI: 10.2307/2490395.
[4] Zmijewski M E. Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 1984, 22: 59-82. DOI: 10.2307/2490859.
[5] Shumway T. Forecasting bankruptcy more accurately: A simple hazard model. The Journal of Business, 2001, 74(1): 101-124. DOI: 10.1086/209665.
[6] Campbell J Y, Hilscher J, Szilagyi J. In search of distress risk. The Journal of Finance, 2008, 63(6): 2899-2939. DOI: 10.1111/j.1540-6261.2008.01416.x.
[7] Altman E I, Sabato G. Modelling credit risk for SMEs: Evidence from the U.S. market. Abacus, 2007, 43(3): 332-357. DOI: 10.1111/j.1467-6281.2007.00234.x.
[8] Ciampi F, Gordini N. Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of Italian small enterprises. Journal of Small Business Management, 2013, 51(1): 23-45. DOI: 10.1111/j.1540-627X.2012.00376.x.
[9] Brown I, Mues C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 2012, 39(3): 3446-3453. DOI: 10.1016/j.eswa.2011.09.033.
[10] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 2015, 10(3): e0118432. DOI: 10.1371/journal.pone.0118432.
[11] Barboza F, Kimura H, Altman E. Machine learning models and bankruptcy prediction. Expert Systems with Applications, 2017, 83: 405-417. DOI: 10.1016/j.eswa.2017.04.006.
[12] Lessmann S, Baesens B, Seow H-V, et al. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 2015, 247(1): 124-136. DOI: 10.1016/j.ejor.2015.05.030.
[13] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning, 2005: 625-632. DOI: 10.1145/1102351.1102430.
[14] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002: 694-699. DOI: 10.1145/775047.775151.
[15] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, 2017. DOI: 10.5555/3295222.3295230.
[16] Lundberg S M. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2: 56-67. DOI: 10.1038/s42256-019-0138-9.
[17] Ribeiro M T, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 1135-1144. DOI: 10.1145/2939672.2939778.
[18] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1: 206-215. DOI: 10.1038/s42256-019-0048-x.
[19] Balcaen S, Ooghe H. 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. The British Accounting Review, 2006, 38(1): 63-93. DOI: 10.1016/j.bar.2005.09.001.
[20] Sun J, Li H, Huang Q-H, et al. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. Knowledge-Based Systems, 2014, 57: 41-56. DOI: 10.1016/j.knosys.2013.12.006.

[21] Friedman J H. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 2001, 29(5): 1189-1232. DOI: 10.1214/aos/1013203451.
[22] Ke G. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30, 2017. DOI: 10.5555/3294996.3295074.
[23] Zieba M, Tomczak S K, Tomczak J M. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 2016, 58: 93-101. DOI: 10.1016/j.eswa.2016.04.001.
[24] Geng R, Bose I, Chen X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. European Journal of Operational Research, 2015, 241(1): 236-247. DOI: 10.1016/j.ejor.2014.08.016.
[25] Tomczak S. Polish Companies Bankruptcy. UCI Machine Learning Repository, Dataset, 2016. DOI: 10.24432/C5F600.
[26] Brier G W. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 1950, 78(1): 1-3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
[27] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 2023, 4(9): 100804. DOI: 10.1016/j.patter.2023.100804.
[28] Barth M E, Cram D P, Nelson K K. Accruals and the prediction of future cash flows. The Accounting Review, 2001, 76(1): 27-58. DOI: 10.2308/accr.2001.76.1.27.
[29] Dechow P M, Kothari S P, Watts R L. The relation between earnings and cash flows. Journal of Accounting and Economics, 1998, 25(2): 133-168. DOI: 10.1016/S0165-4101(98)00020-2.