Journal of Mathematical Finance and Risk Management
ISSN: 3080-9983 (Print), ISSN: 3080-9991 (Online) | Vol. 1, No. 3, 2026

MAC-Stack: Calibrated Multi-Model Ensemble Learning with Exact SHAP Attribution for Explainable Financial Distress Prediction

Mira Lam *, Caroline Sullivan

Joseph M. Katz Graduate School of Business, University of Pittsburgh, USA

* **Corresponding author**: Mira Lam

Abstract: Multi-model ensembles of gradient-boosted trees are among the most accurate tools for financial distress prediction, but they are usually deployed as three loosely coupled steps: ad hoc fusion weights, post-hoc probability calibration and SHAP explanations of individual members. This paper proposes MAC-Stack (margin-additive calibrated stacking), which combines heterogeneous learners (regularized logistic regression, LightGBM, XGBoost, CatBoost and a liquidity-view expert) through one non-negative logistic stacker fitted on cross-fitted log-odds margins. Because fusion and calibration form a single affine map in log-odds space, calibration preserves the ensemble ranking, the ensemble Shapley attributions follow in closed form from the members' TreeSHAP values and sum exactly to the calibrated log-odds, and no attribution can contradict a unanimous sign of the members. On six tasks from two public benchmarks (five Polish forecasting horizons and a Taiwanese panel), evaluated on deduplicated data with nested stratified cross-validation, MAC-Stack achieved the best average rank in PR-AUC, ROC-AUC, Brier score and log-loss (mean PR-AUC 0.709, ROC-AUC 0.948). It exceeded a re-implementation of a recent log-odds/isotonic early-warning framework with its published settings by 0.111 PR-AUC on average (0.086–0.120 per task), whereas adding isotonic recalibration to MAC-Stack reduced PR-AUC by 0.025 without lowering the Brier score. The ensemble attributions were exact to within 5.0×10^{-6} in log-odds, showed no sign-consistency violations (versus up to 2.6% for unconstrained stacking) and recovered the leading distress drivers previously reported for the same horizon.

Keywords: Financial distress prediction, bankruptcy prediction, stacked generalization, gradient boosting, probability calibration, SHAP, explainable artificial intelligence.

1. Introduction

Financial distress prediction supports credit approval, loan pricing, provisioning and supervisory monitoring. Since the discriminant model of Altman [1], ratio-based prediction has moved toward machine learning, and large comparative studies consistently find that tree ensembles, gradient boosting in particular, outperform classical statistical models on financial-ratio data [2–5]. Combining heterogeneous classifiers by voting or stacking can add further accuracy [6], [7], and SHapley Additive exPlanations (SHAP) [8–10] have become the standard way to justify individual credit decisions [11].

In deployment, a distress model has to do three things at once: rank firms well under severe class imbalance, output probabilities that can be compared with a risk appetite or an expected-loss threshold, and explain every alert in terms that an analyst can audit. Multi-model ensembles usually meet these needs with three separate components (fusion weights, a post-hoc calibration map and member-level SHAP values) whose interaction is rarely analyzed. Yue and Zhu [12] recently made this design space explicit in a multi-horizon early-warning framework that fuses a logistic model, a full-feature LightGBM and a dedicated cash-flow/liquidity expert in log-odds space, recalibrates the fused score with isotonic regression, applies a recall-constrained warning threshold and aggregates SHAP values into semantic financial groups. Their study also documented a calibration–discrimination trade-off: isotonic recalibration improved the Brier score but lowered PR-AUC. Two structural properties explain this trade-off. First, an isotonic map is piecewise constant, so it ties firms whose scores differ and discards within-step ranking information. Second, after a non-linear calibration map, attributions computed for the fused score no longer add up to the calibrated probability on which alerts are issued.

This paper shows that both problems disappear when fusion and calibration are confined to one affine map in log-odds (margin) space. We propose MAC-Stack (margin-additive calibrated stacking): heterogeneous learners whose native outputs are log-odds margins are combined by a non-negative logistic stacker fitted on cross-fitted out-of-fold margins, in the tradition of stacked generalization and the super learner [13–15]. The stacker estimates the fusion weights together with a Platt-type temperature and intercept [16] by minimizing a strictly proper scoring rule [17], uses every training observation and, through the linearity of the Shapley value, yields exact ensemble attributions in closed form. The contributions are as follows.

1) Method: MAC-Stack unifies fusion and calibration in one convex, non-negativity-constrained problem over cross-fitted margins; it builds on the horizon-specific design, liquidity view and operating rule of the three-view framework of Yue and Zhu.

2) Theory: we prove rank preservation of the calibrated output, exact additive attributions at feature, group and member level, and sign consistency with unanimous members, properties that isotonic or unconstrained fusion lack.

3) Evidence: a leakage-controlled benchmark on six tasks from two public datasets compares MAC-Stack with six single learners, a re-implementation of that framework with its published settings, seven alternative fusion schemes and five leave-one-learner-out ablations, using corrected

resampled tests.

4) Explanation audit: we quantify attribution exactness, sign consistency, the collapse of risk levels under isotonic calibration, cross-fold stability and cross-model agreement, and we report horizon-wise semantic-group profiles.

Section II reviews related work, Section III presents the method and its properties, Section IV describes the experimental design, Section V reports the results, Section VI discusses implications and limitations, and Section VII concludes.

2. Related Work

2.1. Machine Learning for Distress Prediction

Accounting-ratio models date back to the multiple discriminant analysis of Altman [1]. Benchmarks on credit and bankruptcy data have since shown that tree ensembles dominate single classifiers [2, 3]. On the Polish bankruptcy panel used here, Zięba et al. [4] obtained strong results with extreme gradient boosting and synthetic features, and Ben Jabeur et al. [5] reported that CatBoost improves corporate-failure prediction one to three years ahead. On the Taiwanese panel, Liang et al. [18] found solvency and profitability ratios to be the most informative financial indicators.

2.2. Ensembles and Stacking

Classifier ensembles for bankruptcy prediction have been compared systematically [6], and stacking with a meta-learner has been applied to financial distress with financial ratios and corporate-governance indicators [7]. Stacked generalization [13] trains a combiner on cross-validated predictions; Breiman [14] found non-negativity constraints on the combination weights essential for stable stacking, and the super learner [15] provides oracle guarantees for cross-validated weighting. Yue and Zhu fused three views in log-odds space with weights chosen on a coarse simplex grid by PR-AUC on a calibration split. Their framework is the natural reference for this work: we retain its log-odds fusion, liquidity view and warning rule, but replace grid search and isotonic recalibration by a constrained likelihood-based stacker on cross-fitted margins.

2.3. Calibration and Explanation

Platt scaling [16] fits a sigmoid to classifier scores, whereas isotonic regression fits a monotone step function; boosted trees benefit from either [19]. Proper scoring rules such as the Brier score [20] and the logarithmic score reward calibrated probabilities [17]. SHAP [9] attributes a prediction to features by Shapley values [8], and TreeSHAP computes them exactly for tree ensembles [10]; Bussmann et al. [11] used Shapley values to explain the credit risk of small and medium enterprises. To our knowledge, no previous distress study has derived exact attributions for a calibrated heterogeneous ensemble or measured how the calibration step affects the faithfulness of the explanations. Sun et al. combine Transformer-based feature-interaction learning with SHAP in insurance risk assessment, offering a valuable example of evaluating predictive quality alongside feature-level interpretability [29].

3. Methodology

Fig. 1 summarizes MAC-Stack. Its three ingredients, namely margin-additive members, a cross-fitted non-negative logistic stacker and closed-form attributions, are described in turn.

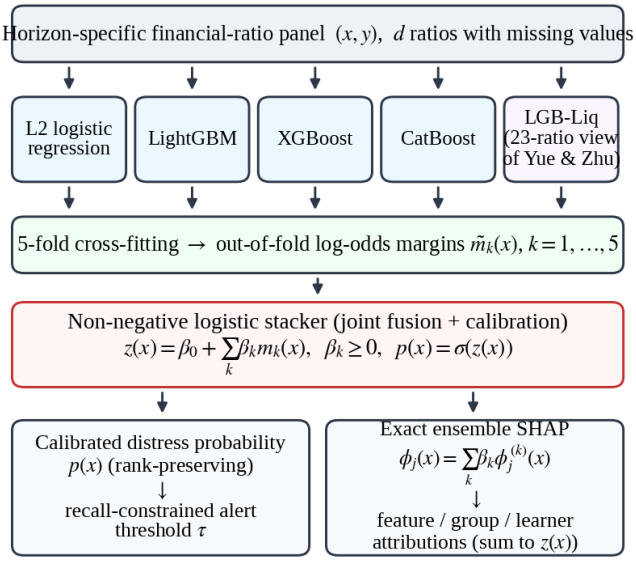


Figure 1. Overview of MAC-Stack. Heterogeneous members produce log-odds margins; a non-negative logistic stacker fitted on cross-fitted margins performs fusion and calibration jointly, which makes the ensemble SHAP attribution exact and additive.

3.1. Problem Setting

Each task provides observations (x_i, y_i), $i = 1, \ldots, n$, where $x_i \in \mathbb{R}^d$ collects the financial ratios of a firm-year (some possibly missing) and $y_i \in \{0, 1\}$ indicates bankruptcy at the end of a fixed horizon. As in the three-view framework, every horizon is modeled separately. The goal is a probability of distress $p(x)$, an alert rule based on $p(x)$ and an attribution vector that decomposes the model output into ratio contributions.

3.2. Margin-Additive Members

MAC-Stack uses $K = 5$ members whose native output is a log-odds margin $m_k(x) = \text{logit}\, p_k(x)$ and whose attributions are exact in margin space: (i) L2-regularized logistic regression on median-imputed ratios with missingness indicators and robust 10th–90th percentile scaling, i.e., the preprocessing of [12], followed by winsorization at ±5; (ii) LightGBM [21]; (iii) XGBoost [22]; (iv) CatBoost [23]; and (v) a LightGBM liquidity expert restricted to a cash-flow and liquidity view. For the Polish data this view is the 23-ratio set defined by Yue and Zhu; for the Taiwanese data it consists of the 27 indicators that refer to cash flow, current or quick assets, working capital or current liabilities. Tree members handle missing values natively. For each member, attributions $\varphi^{(k)}(x) \in \mathbb{R}^d$ and a base value $\varphi_0^{(k)}(x)$ satisfy local accuracy:

$$\phi_0^{(k)}(x) + \sum_{j=1}^{d} \phi_j^{(k)}(x) = m_k(x) \qquad (1)$$

For the tree members we use path-dependent TreeSHAP [10] as implemented natively in each library. For the logistic member every transformed column c depends on a single raw ratio, so its interventional linear SHAP value $w_c(t_c(x) - E[t_c])$ is assigned to the source ratio, which keeps (1) exact. Random forests are not used as members because their averaged probabilities are not additive in log-odds.

3.3. Cross-Fitted Non-Negative Logistic Stacking

The training rows are split into $V = 5$ stratified folds. For

row i in fold v(i), the out-of-fold margin is

$$\tilde{m}_k(x_i) = m_k^{(-v(i))}(x_i) \tag{2}$$

Where $m_k(-v)$ denotes member k trained without fold v. Collecting $\tilde{m}_i = (\tilde{m}_1(x_i), \ldots, \tilde{m}_K(x_i))^\top$, the stacker solves

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0,\, \beta \geq 0} \frac{1}{n} \sum_{i=1}^{n} \ell\,(y_i, \beta_0 + \beta^\top \tilde{m}_i) + \lambda \|\beta\|^2 \tag{3}$$

With the logistic loss ℓ and a small ridge term ($\lambda = 10^{-4}$) for numerical stability. Problem (3) is convex and is solved by L-BFGS-B. The members are then refitted on all training rows, and the deployed model is

$$p(x) = \sigma(z(x)), \quad z(x) = \beta_0 + \alpha \sum_{k=1}^{K} w_k\, m_k(x) \tag{4}$$

Where $\alpha = \Sigma_k \beta_k$ and $w_k = \beta_k/\alpha$. The simplex weights w state how much each member is trusted, whereas α and β_0 act as a temperature and an intercept: $\alpha < 1$ shrinks over-confident margins and β_0 corrects the base rate. Compared with the grid-plus-isotonic scheme of the three-view framework, (3) estimates continuous weights and the calibration map jointly under a strictly proper scoring rule [17] and uses all training rows instead of reserving a calibration split.

3.4. Properties

Proposition 1 (rank preservation): If $\alpha > 0$, p(x) in (4) is a strictly increasing function of the fused score $s(x) = \Sigma_k w_k m_k(x)$; hence the ROC-AUC and PR-AUC of p equal those of s.

Proof: σ is strictly increasing and $z = \beta_0 + \alpha s$ with $\alpha > 0$. An isotonic map, in contrast, is piecewise constant: test firms whose scores fall into the same step receive identical probabilities, and their relative order is lost.

Proposition 2 (exact additive attribution): Define

$$\phi_j(x) = \sum_{k=1}^{K} \beta_k\, \phi_j^{(k)}(x) \tag{5}$$

$$\phi_0(x) = \beta_0 + \sum_{k=1}^{K} \beta_k\, \phi_0^{(k)}(x) \tag{6}$$

Then

$$\phi_0(x) + \sum_{j=1}^{d} \phi_j(x) = z(x) = \operatorname{logit} p(x) \tag{7}$$

Moreover, if each $\varphi^{(k)}$ is the Shapley value of a coalition game v_k, then φ is the Shapley value of the game $\beta_0 + \Sigma_k \beta_k v_k$.

Proof: Multiply (1) by β_k, sum over k and add β_0; the second statement follows from the linearity axiom of the Shapley value [8]. Consequently, for any partition of the ratios into semantic groups G, $\Phi_G(x) = \Sigma_{j \in G} \varphi_j(x)$ is an exact group decomposition, and $B_k(x) = \beta_k(m_k(x) - \varphi_0^{(k)}(x))$ decomposes the same log-odds across members.

Proposition 3 (sign consistency): If $\beta \geq 0$ and all members with a non-zero attribution for ratio j at x assign it the same sign, then $\varphi_j(x)$ has that sign. Proof: $\varphi_j(x)$ is a non-negative combination of terms with that sign.

Unconstrained stacking may assign negative weights to highly correlated members and thereby reverse their contributions, while any non-linear post-calibration g breaks (7) because logit $g(\sigma(s))$ is not additive in the ratios.

3.5. Alert Threshold

We adopt the recall-constrained operating rule of Yue and Zhu: on the out-of-fold probabilities of the training data, the threshold τ is the value with the highest precision among thresholds that achieve a recall of at least 0.80. Test labels are never used.

4. Experimental Setup

4.1. Datasets

Table I summarizes the six tasks. The Polish Companies Bankruptcy data [4, 24] contain 64 ratios of manufacturing firms observed one to five years before the bankruptcy status; task Ph uses the file whose label is observed h years ahead (P1 = 5year.arff, …, P5 = 1year.arff), which reproduces the horizon-specific setting of the reference framework. The Taiwanese Bankruptcy Prediction data [18], [25] contain 95 ratios of listed firms from 1999–2009; the constant "Net Income Flag" was removed. Each Polish file contains 60–90 exact duplicate rows (identical ratios and labels, never with conflicting labels); they were removed before splitting so that no observation can appear in both a training and a test fold [26]. Every file was checked against an independent public copy (value by value for the Polish files and by SHA-256 hash for the Taiwanese file).

Table I. Characteristics of the Six Prediction Tasks

Task	Source	N	Dup.	d	Liq.	Pos.	Prev. (%)	Miss. (%)
P1	5year.arff	5,850	60	64	23	408	6.97	1.24
P2	4year.arff	9,710	82	64	23	513	5.28	1.40
P3	3year.arff	10,416	87	64	23	493	4.73	1.48
P4	2year.arff	10,083	90	64	23	398	3.95	1.87
P5	1year.arff	6,945	82	64	23	271	3.90	1.30
TW	data.csv	6,819	0	94	27	220	3.23	0.00

N: firm-year observations after removing exact duplicates (Dup.); Liq.: size of the liquidity view; Prev.: share of bankrupt firms; Miss.: share of missing ratio values.

4.2. Protocol

Each task is evaluated by stratified 5-fold cross-validation with a fixed seed, giving 30 test folds in total. Inside each outer training fold, the members are cross-fitted with an inner stratified 5-fold split to produce the margins in (2), so the stacker, the threshold and all weights are learned without the outer test fold. Hyperparameters were fixed a priori and not tuned: LightGBM and XGBoost use 300 trees with learning rate 0.05 (31 leaves and depth 5, respectively), CatBoost uses 400 iterations of depth 6 with learning rate 0.08, and the logistic model uses C = 0.1. No resampling or synthetic minority samples are used.

4.3. Baselines and Ablations

Single learners: the five members and a random forest (RF, 200 trees, median imputation). Reference: YZ-3V, our re-implementation of the three-view framework of [12] with its published settings, i.e., a class-balanced logistic model without winsorization; class-balanced LightGBM with 80 trees, learning rate 0.08, 15 leaves, minimum child size 35, L1 = 0.15 and L2 = 1.5; a liquidity-view LightGBM with the same settings; a 75/25 split of each outer training fold that reproduces their 60/20/20 design; simplex-grid weights (step 0.25) selected by PR-AUC with Brier tie-breaking; and isotonic calibration. YZ-3V (raw) denotes the same fused score before isotonic calibration. Fusion ablations on our five members: equal-weight soft voting; equal-weight logit averaging with Platt scaling; non-negative stacking of probabilities instead of margins; unconstrained stacking; MAC-Stack followed by isotonic calibration (MAC+Iso); MAC-Stack fitted on the 25% hold-out split instead of cross-fitted margins (MAC hold-out); and the grid-plus-isotonic rule of YZ-3V applied to our members (Grid+Iso hold-out). Leave-one-learner-out variants refit (3) without one member.

4.4. Metrics and Statistical Tests

PR-AUC (average precision) is the primary metric because distressed firms are rare [27]; ROC-AUC, the Brier score [20], log-loss (LL) and the expected calibration error (ECE, ten equal-mass bins) complement it. Recall@10% (R@10) is the share of distressed firms captured by a review list containing the riskiest 10% of firms. At the threshold τ we report F1 and balanced accuracy (BAcc). MAC-Stack is compared with every alternative by a one-sided Wilcoxon signed-rank test over the 30 paired test folds and, within each task, by the corrected resampled t-test of Nadeau and Bengio [28], which accounts for overlapping training sets. All experiments ran on a single CPU thread with Python 3.12, scikit-learn 1.8.0, LightGBM 4.7.0, XGBoost 3.4.1 and CatBoost 1.2.10.

5. Results

5.1. Predictive Performance

Table II and Fig. 2 report PR-AUC by task. MAC-Stack attains the highest PR-AUC on five of the six tasks and the best average rank (1.17), followed by CatBoost (2.33) and LightGBM (3.33). The gains over individual boosting members are small but systematic: relative to CatBoost, the strongest single learner, the mean PR-AUC gain is 0.006 and is positive on five of the six tasks (Wilcoxon $p = 0.010$), although it reaches significance within no single task under the corrected t-test; relative to LightGBM and XGBoost the gain is positive on all six tasks and significant within four and three tasks, respectively.

Table II. PR-AUC by Task (Mean of Five Test Folds)

Model	P1	P2	P3	P4	P5	TW	Avg.	Rank	p
LR	0.589	0.439	0.295	0.267	0.475	0.410	0.413	7.50	<.001
RF	0.540	0.405	0.385	0.398	0.562	0.454	0.457	7.17	<.001
LightGBM	0.809	0.725	0.667	0.680	0.779	0.488	0.691	3.33	<.001
XGBoost	0.807	0.726	0.666	0.687	0.789	0.480	0.692	3.50	<.001
CatBoost	0.817	0.747	0.684	0.693	0.797	0.479	0.703	2.33	.010
LGB-Liq	0.400	0.255	0.223	0.207	0.178	0.307	0.261	9.00	<.001
YZ-3V	0.731	0.633	0.571	0.582	0.699	0.373	0.598	6.33	<.001
Soft voting	0.780	0.717	0.659	0.666	0.759	0.486	0.678	4.67	<.001
MAC-Stack (ours)	0.817	0.748	0.688	0.702	0.805	0.493	0.709	1.17	—

Rank: average rank over tasks; p: one-sided Wilcoxon signed-rank test over the 30 folds (alternative: MAC-Stack better). Best value per column in bold.

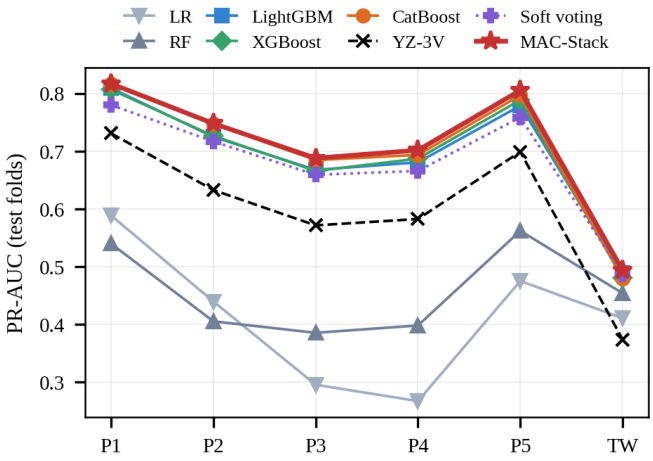


Figure 2. Test-fold PR-AUC by task. P1–P5: Polish data, label observed one to five years ahead; TW: Taiwanese data.

Table III shows that the advantage is larger for probability quality: MAC-Stack has the lowest Brier score and the lowest log-loss on all six tasks, and its Brier score is below that of CatBoost on all six tasks ($p < 0.001$). With the recall-constrained threshold it reaches an F1 of 0.478 and a balanced accuracy of 0.869, and a review list of the riskiest 10% of firms captures 79.9% of the distressed firms. The Taiwanese task is the hardest (PR-AUC 0.493); there the logistic member and the random forest come closer to the boosted models, as Fig. 2 shows.

Table III. Average Performance Over the Six Tasks

Model	ROC	PR	Brier ×100	LL	ECE	R@10	F1	BAcc
LR	0.865	0.413	3.41	0.133	0.014	0.618	0.281	0.790
RF	0.890	0.457	3.37	0.129	0.015	0.641	0.299	0.810
LightGBM	0.944	0.691	2.37	0.106	0.017	0.789	0.457	0.866
XGBoost	0.945	0.692	2.30	0.094	0.014	0.791	0.450	0.863
CatBoost	0.945	0.703	2.24	0.091	0.013	0.794	0.452	0.867
LGB-Liq	0.811	0.261	4.04	0.174	0.022	0.472	0.181	0.717
YZ-3V	0.924	0.598	2.45	0.109	0.010	0.744	0.375	0.842
Soft voting	0.932	0.678	2.47	0.098	0.017	0.773	0.419	0.855
MAC-Stack (ours)	0.948	0.709	2.14	0.083	0.007	0.799	0.478	0.869

LL: log-loss; R@10: recall within the riskiest 10% of firms; F1 and BAcc at the recall-constrained threshold. Best value per column in bold.

5.2. Fusion Design and the Calibration–Discrimination Trade-Off

Table IV and Fig. 3 compare fusion schemes. Our re-implementation of YZ-3V reproduces the trade-off reported in [12]: isotonic calibration lowers its Brier score from 3.87 to 2.45 ($\times 10^{-2}$) but reduces PR-AUC from 0.647 to 0.598. MAC-Stack exceeds YZ-3V by 0.111 PR-AUC on average, on all six tasks, and significantly within each task. Because YZ-3V keeps its published, lighter member settings and reserves a calibration split, its absolute level lies below that of our individual boosted members; the gap accumulates in three steps: stronger members under the original grid-plus-isotonic rule (Grid+Iso hold-out, PR-AUC 0.636), replacing that rule by our stacker on the same hold-out data (MAC hold-out, 0.684), and cross-fitting on all training rows (MAC-Stack, 0.709). Isotonic recalibration of MAC-Stack (MAC+Iso) lowers PR-AUC by 0.025 on all six tasks and does not improve the Brier score (difference $0.016 \times 10-2$ in favor of MAC-Stack). The calibration–discrimination trade-off is therefore a property of the non-linear calibration map rather than of calibrated ensembles as such.

Margin space also matters: stacking probabilities instead of margins costs 0.015 PR-AUC, and equal weights cost 0.031 (soft voting) and 0.016 (logit averaging with Platt scaling). Non-negativity costs no accuracy: unconstrained stacking is statistically indistinguishable from MAC-Stack (Wilcoxon $p = 0.415$), yet it placed at least one negative weight in 22 of 25 Polish folds.

Table IV. Fusion Schemes, Reference Framework and Leave-One-Learner-Out Ablations (Means Over Tasks)

Method	PR	Brier ×100	LL	ΔPR	Wins	p_{PR}	p_{Brier}
MAC-Stack (ours)	0.709	2.14	0.083	—	—	—	—
YZ-3V	0.598	2.45	0.109	+0.111	6/6	<.001	<.001
YZ-3V (raw)	0.647	3.87	0.144	+0.062	6/6	<.001	<.001
Grid+Iso (hold-out)	0.636	2.30	0.104	+0.073	6/6	<.001	<.001
MAC (hold-out)	0.684	2.25	0.088	+0.025	6/6	<.001	<.001
MAC+Iso	0.684	2.15	0.086	+0.025	6/6	<.001	.037
Soft voting	0.678	2.47	0.098	+0.031	6/6	<.001	<.001
Logit avg.+Platt	0.693	2.22	0.087	+0.016	6/6	<.001	<.001
Prob. stacking	0.694	2.33	0.100	+0.015	6/6	<.001	<.001
MAC (unconstr.)	0.708	2.14	0.084	+0.001	2/6	.415	.619
w/o LR	0.709	2.14	0.083	+0.000	3/6	.693	.642
w/o LightGBM	0.708	2.14	0.083	+0.001	4/6	.190	.214
w/o XGBoost	0.708	2.14	0.084	+0.001	5/6	.111	.118
w/o CatBoost	0.696	2.21	0.086	+0.012	5/6	<.001	<.001
w/o LGB-Liq	0.709	2.14	0.083	−0.000	2/6	.762	.820

ΔPR: PR-AUC of MAC-Stack minus that of the method; Wins: tasks on which MAC-Stack has the higher PR-AUC; p: one-sided Wilcoxon tests over 30 folds. YZ-3V: re-implementation of the three-view framework (Section IV-C).

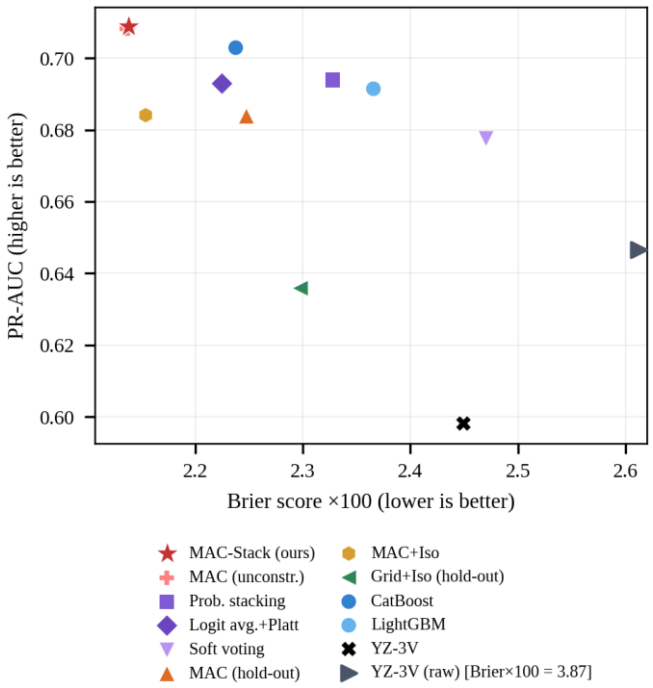


Figure 3. Discrimination–calibration plane (means over six tasks); upper left is better. Isotonic variants lose PR-AUC; the raw YZ-3V score lies outside the plotted range (arrow).

5.3. Learned Weights and Member Contributions

Fig. 4 shows the stacking weights. CatBoost receives the largest weight on every Polish horizon (w = 0.51–0.70), XGBoost the second largest, and the logistic member is excluded in every P4 fold; on the Taiwanese task XGBoost dominates (w = 0.45). The liquidity expert receives at most 0.06 and is excluded in every P1 and P4 fold, which is consistent with the horizon-dependent use of the cash-flow expert in the three-view framework (unused at the one-year horizon, active at the two- and five-year horizons). The leave-one-learner-out rows of Table IV confirm that only CatBoost is individually indispensable (removing it costs 0.012 PR-AUC), whereas removing any other member changes PR-AUC by less than 0.001: the members are largely redundant, and the stacker exploits this by concentrating weight instead of averaging. The temperature α ranges from 0.74 (TW) to 1.05, so the fused margin is shrunk mainly on the smaller Taiwanese panel.

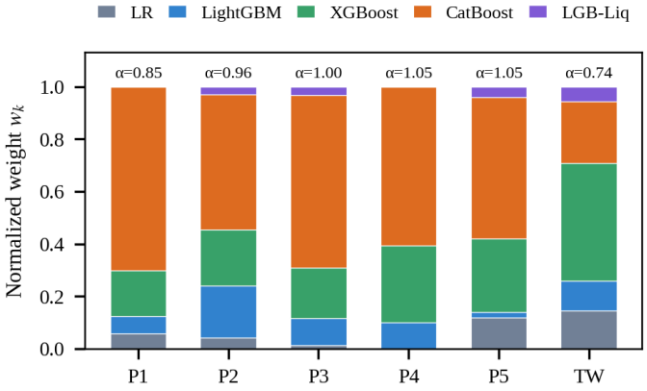


Figure 4. Normalized stacking weights wk (mean over folds) and temperature α per task.

5.4. Fidelity of the Ensemble Explanations

Table V checks the theory empirically. Over all 30 folds the largest local-accuracy error of the ensemble attributions is 5.0×10^{-6}, i.e., floating-point precision, and no sign-consistency violation occurs, whereas unconstrained stacking contradicts a unanimous member sign in up to 2.64% of consensus pairs (P1). Isotonic calibration compresses the 1,170–2,083 test firms of a fold into only 1.5–2.3% distinct probability levels (MAC+Iso) or 1.2–1.9% (YZ-3V); within each level the member attributions still vary while the probability acted upon does not. For YZ-3V, the probability implied by its additive explanation differs from the calibrated output by 0.5–4.5 percentage points at the median and by 7.7–31.9 points at the 90th percentile, so an analyst reading the explanation would see a materially different risk from the one that triggered the alert.

Table V. Fidelity of the Ensemble Explanations per Task

Task	Max err.	Sign viol. (%)		Distinct levels (%)			p90 gap (pp)	
		MAC	Unc.	MAC	+Iso	YZ-3V	+Iso	YZ-3V
P1	2.9e-6	0.00	2.64	99.9	2.3	1.9	2.0	14.5
P2	3.8e-6	0.00	0.03	100.0	1.8	1.7	1.2	26.8
P3	3.4e-6	0.00	0.13	100.0	1.7	1.3	1.2	31.9
P4	3.5e-6	0.00	2.53	100.0	1.5	1.2	0.9	24.9
P5	4.1e-6	0.00	0.15	100.0	2.1	1.7	0.8	10.1
TW	5.0e-6	0.00	0.00	100.0	1.8	1.5	1.6	7.7

Max err.: largest $|\varphi_0 + \Sigma\varphi_j - \text{logit } p|$ over test firms; Sign viol.: share of consensus pairs whose sign is reversed (MAC-Stack vs. unconstrained stacking); Distinct levels: distinct predicted probabilities as a share of test firms; p90 gap: 90th percentile of the difference between the calibrated probability and the probability implied by the additive explanation (MAC+Iso and YZ-3V).

5.5. Content of the Explanations

Fig. 5 shows cross-fitted ensemble attributions for P1, where every firm is explained by a model trained without it. The four leading drivers, namely operating expenses to total liabilities (X34), operating profit to financial expenses (X27), sales growth (X21) and the acid-test ratio (X46), coincide with the leading drivers reported for the same one-year horizon in the three-view study, although they were obtained with a different model, explanation pipeline and validation protocol. The attributions of most leading ratios are close to monotone: the predicted risk rises with lower sales growth (X21), a lower acid-test ratio (X46), lower retained earnings to assets (X6) and higher operating expenses relative to liabilities (X34) (absolute Spearman correlation between ratio value and attribution of at least 0.68), whereas the effect of interest coverage (X27) is non-monotone (correlation 0.24). X34 and X27 occupy the first two positions at every Polish horizon, X46 is always among the top four, and the liquidity-

gap measure X5 enters the top six from P2 onward.

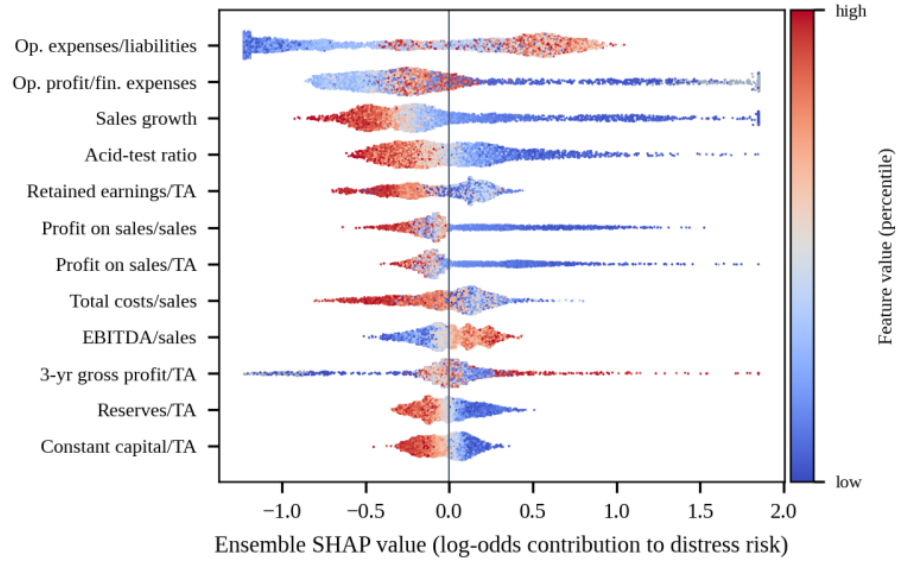


Figure 5. Cross-fitted ensemble SHAP values for P1 (12 most important ratios); each firm is explained by the model trained without it. Color encodes the within-sample percentile of the ratio.

Fig. 6 aggregates the attributions into the five semantic groups of the three-view framework. For P1, the efficiency/turnover (32.7%), liquidity/cash-flow (19.2%) and size (4.9%) shares lie within one percentage point of the previously reported shares; the profitability and leverage shares differ mainly because we classify the interest-coverage ratio X27 as a solvency measure. Liquidity and cash-flow ratios carry 19.0–25.4% of the attribution mass at every Polish horizon and peak two to three years ahead, whereas profitability dominates on the Taiwanese panel (33.5%), in line with [18].

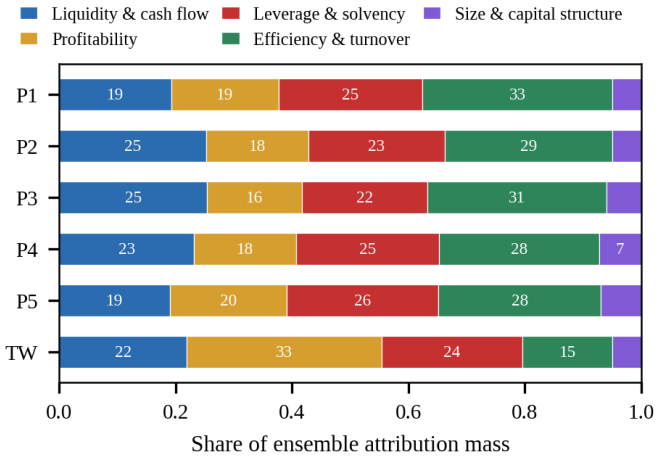


Figure 6. Share of ensemble attribution mass by semantic group (three-view taxonomy) for each task; numbers are percentages.

5.6. Stability and Cross-Model Agreement

Table VI asks whether the global explanations are reproducible across folds. The rank stability of MAC-Stack (Spearman $\rho = 0.82$ between fold pairs) clearly exceeds that of CatBoost (0.69), its dominant member, and of the logistic model (0.77), and its importance-weighted similarity (cosine 0.962) matches LightGBM and XGBoost (0.963 and 0.965). LightGBM and XGBoost remain more stable in rank terms ($\rho = 0.86$ and 0.84; top-10 overlap 0.73 and 0.78 versus 0.63) because they concentrate attribution on fewer ratios. Semantic-group shares change little between folds (standard deviation 1.3–1.9 percentage points), which supports group-level reporting. The agreement column reveals a Rashomon effect: the logistic member and the tree members rank ratios very differently ($\rho = 0.36$), and even CatBoost agrees with LightGBM and XGBoost only at $\rho = 0.78$. MAC-Stack agrees with each tree member at $\rho = 0.90$–0.93; its attribution is a consensus that no single member provides.

Table VI. Explanation Stability Across Folds and Agreement With Other Learners (Means Over Tasks)

Model	ρ	J@10	cos	Grp. s.d.	Agree.
LR	0.766	0.599	0.937	1.57	0.357
LightGBM	0.857	0.734	0.963	1.29	0.876
XGBoost	0.843	0.775	0.965	1.29	0.876
CatBoost	0.686	0.543	0.945	1.86	0.782
MAC-Stack (ours)	0.818	0.627	0.962	1.75	0.917

ρ, J@10, cos: mean pairwise Spearman correlation, top-10 Jaccard index and cosine similarity of global importance vectors between folds; Grp. s.d.: fold-to-fold standard deviation of semantic-group shares (percentage points); Agree.: mean Spearman correlation with the other tree learners (with all tree learners for LR).

5.7. Computational Cost

On one CPU thread, fitting MAC-Stack (25 cross-fitted member fits, five refits and the stacker) took 45–54 s per fold, about 11 times a single CatBoost fit; the stacker itself took about 0.01 s. Exact ensemble attributions for the 1,170–2,083 test firms of a fold took 5.4–9.0 s, i.e., at most 4.6 ms per firm. Scoring needs the five members plus six fusion parameters, so the stacker adds no measurable latency.

6. Discussion

Implications: Practitioners who deploy calibrated ensembles often assume that they must trade ranking power for reliable probabilities. Our results indicate that this trade-off is avoidable when calibration is affine in log-odds and estimated jointly with the fusion weights: in this benchmark MAC-Stack has the best average rank both in PR-AUC and ROC-AUC and in Brier score and log-loss. Exact attributions

also have practical value. An alert, its probability and its explanation refer to the same number; the explanation can be decomposed by ratio, by semantic group or by member model; and sign consistency guarantees that the ensemble never contradicts all of its members. These properties can be tested automatically by a model-risk validator, as Table V does. Symmetry-aware credit-risk modeling demonstrates the value of combining structured financial representations with invariant feature relationships in risk-sensitive predictive evaluation [30].

Limitations: First, each task was evaluated with one repetition of 5-fold cross-validation (30 test folds); per-task corrected tests have only four degrees of freedom, so the small gains over CatBoost are significant only when pooled across tasks. Second, both panels are cross-sections from 1999–2013 without firm identifiers, so temporal validation and validation on legally defined SME populations remain to be done. Third, path-dependent TreeSHAP attributions are exact with respect to their own coalition games but inherit the usual caveats of Shapley values under correlated ratios; they describe the model, not causal mechanisms. Fourth, hyperparameters were fixed rather than tuned; tuning would strengthen all tree-based methods but is unlikely to change the comparison of fusion rules, which share the same members.

7. Conclusion

This paper presented MAC-Stack, a calibrated multi-model ensemble for explainable financial distress prediction in which fusion and calibration form a single non-negative affine map in log-odds space. The design preserves ranking under calibration, yields exact feature-, group- and member-level SHAP attributions of the calibrated output, and guarantees sign consistency with unanimous members. Across six public tasks MAC-Stack achieved the best average rank in discrimination and probability quality, improved PR-AUC over a re-implementation of a recent log-odds/isotonic early-warning framework by 0.111 on average, and showed that the calibration–discrimination trade-off disappears when isotonic recalibration is replaced by joint affine calibration. Future work will address temporal and SME-specific validation, cost-sensitive alert policies and interaction-level attributions. Both datasets are publicly available from the UCI repository; the experimental code and fold-level results are available from the authors upon reasonable request.

References

[1] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[2] 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

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

[4] Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 58, 93–101. https://doi.org/10.1016/j.eswa.2016.04.001

[5] Teng, D. (2025). TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows. World Journal of Engineering Research, 3(6), 55–66.

[6] Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing, 24, 977–984. https://doi.org/10.1016/j.asoc.2014.08.047

[7] Liang, D., Tsai, C.-F., Lu, H.-Y., & 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

[8] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. Eurasia Journal of Science and Technology, 5(2), 34–48.

[9] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[10] 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(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[11] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

[12] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72. https://doi.org/10.61784/wms4120

[13] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[14] Breiman, L. (1996). Stacked regressions. Machine Learning, 24(1), 49–64. https://doi.org/10.1007/BF00117832

[15] van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1), Article 25. https://doi.org/10.2202/1544-6115.1309

[16] Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers (pp. 61–74). MIT Press.

[17] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378. https://doi.org/10.1198/016214506000001437

[18] Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. European Journal of Operational Research, 252(2), 561–572. https://doi.org/10.1016/j.ejor.2016.01.012

[19] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625–632. https://doi.org/10.1145/1102351.1102430

[20] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[21] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

[22] Teng, D. (2025). RSRBench: Reproducibility and performance benchmarking for regional seismic risk workflows. Academic Journal of Architecture and Civil Engineering, 3(1), 40–52.

[23] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.

[24] Tomczak, S. (2016). Polish companies bankruptcy. UCI Machine Learning Repository. https://doi.org/10.24432/C5F600

[25] UCI Machine Learning Repository. (2020). Taiwanese bankruptcy prediction. https://doi.org/10.24432/C5004D

[26] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804

[27] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. Journal of Trends in Finance and Economics, 1(1), 85–100.

[28] Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281. https://doi.org/10.1023/A:1024068626366

[29] Sun, T., Liu, X., Zhang, L., & Wang, H. (2024). Enhancing auto insurance risk evaluation with Transformer and SHAP. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3446179

[30] Han, X., Chen, J., Zhang, X., & Yang, Y. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. Symmetry, 17, Article 341. https://doi.org/10.3390/sym17030341