Journal of Economics, Finance and Accounting Studies (JEFAS)
ISSN: 2709-0809
DOI: 10.32996/jefas
Journal Homepage: www.al-kindipublisher.com/index.php/jefas

JEFAS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Predicting Profitability Deterioration in U.S. Commercial Banks Using Explainable Machine Learning

Aftab Hossain Sazu
College of Business Administration, Central Michigan University, Mount Pleasant, Michigan, United States
Corresponding Author: Aftab Hossain Sazu, **E-mail:** aftab.worklife@gmail.com

| ABSTRACT

This study develops an explainable machine-learning framework for predicting next-quarter profitability deterioration in U.S. commercial banks. Public regulatory records yield 277,959 bank-quarter observations from 7,344 banks. Deterioration is defined as a decline of at least one percentage point in annualized quarterly return on assets. L2 logistic regression, Random Forest, and XGBoost are evaluated through an expanding window covering 15 out-of-sample quarters from 2022Q1 to 2025Q3. XGBoost achieves mean ROC AUC of 0.8196, average precision (AP) of 0.3281, Brier score of 0.0303, and top-decile event capture of 54.09 percent. A restricted five-knot cubic-spline logistic model using only current ROA attains mean ROC AUC of 0.7593 and AP of 0.2247; full-predictor XGBoost improves these measures by 0.0603 and 0.1035 and raises top-decile capture by 7.71 percentage points (all paired $p < 0.001$). When the original out-of-sample scores are evaluated against persistent deterioration over two quarters, XGBoost attains mean ROC AUC of 0.9202 and AP of 0.3989. SHAP analysis consistently ranks current profitability, nonperforming loans, and bank size as the three leading predictors, followed by the compensation expense ratio and the net interest income-to-assets ratio. The framework can support early-warning prioritization but should complement, not replace, financial analysis and supervisory judgment.

| KEYWORDS

profitability deterioration, commercial banks, explainable machine learning, XGBoost, early-warning systems, bank risk
JEL classification: C53, G21, G32

| ARTICLE INFORMATION

ACCEPTED: 01 July 2026 **PUBLISHED:** 22 August 2026 **DOI:** 10.32996/jefas.2026.8.8.15

1. Introduction

Bank profitability supports loss absorption, capital formation, lending capacity, and financial stability. A material decline in profitability can signal emerging weakness even when an institution remains solvent and well capitalized. Predicting that decline is difficult because earnings reflect nonlinear relationships among current profitability, asset quality, capital, employee compensation burden, liquidity, funding composition, securities holdings, and bank size.

Logistic regression remains a common benchmark for bank-risk prediction because it is transparent and produces probability estimates. Its linear index, however, may not represent threshold effects or interactions among bank characteristics. Random Forest and XGBoost can capture such patterns without requiring them to be specified in advance (Breiman, 2001; Chen & Guestrin, 2016). Explainability methods can then show which reported financial characteristics contribute most strongly to an individual model prediction.

Most bank early-warning research focuses on failure, insolvency, supervisory distress, or broad financial vulnerability. Evidence from central-bank and peer-reviewed studies shows that machine-learning methods can improve bank-risk prediction and support risk-based prioritization (Bräuning et al., 2019; Suss & Treitel, 2019; Petropoulos et al., 2020; Hu et al., 2025). Profitability deterioration is an earlier and more frequent outcome than outright failure, yet relatively little evidence addresses whether a substantial next-quarter earnings decline can be forecast from public U.S. regulatory data.

This study asks whether explainable tree-based models improve the prediction of next-quarter profitability deterioration relative to L2 logistic regression and to restricted benchmarks based only on current ROA. The outcome is a decline of at least one

percentage point in annualized quarterly return on assets. The analysis reconstructs quarterly earnings from cumulative Call Report values, applies training-only preprocessing, and evaluates all models through 15 sequential out-of-sample quarters rather than a random split. A separate persistence check asks whether the same risk scores identify declines that remain adverse one additional quarter.

The study contributes a reproducible bank-level early-warning design; exact source-field definitions, sample attrition, and missing-data information; a broad evaluation combining discrimination, AP, calibration, and risk concentration; formal paired comparisons with full-model and ROA-only benchmarks; robustness tests across outcome thresholds, persistence, and bank-size groups; and SHAP analyses of both global importance and prediction direction. The results show that XGBoost consistently adds predictive information beyond a flexible function of current ROA, although Random Forest remains a competitive full-predictor benchmark.

2. Literature Review

2.1 Bank Profitability, Machine Learning, and Explainability

Bank profitability reflects an institution's capacity to absorb losses, accumulate capital, and sustain financial intermediation. Return on assets is commonly used to measure how effectively a bank generates earnings from its asset base. Profitability depends on interconnected bank-specific and external conditions, including capitalization, credit quality, expense control, liquidity, interest-generating activity, and the macroeconomic environment.

Athanasoglou et al. (2008) identify capitalization, credit risk, operating expenses, and macroeconomic conditions as important profitability determinants. Dietrich and Wanzenried (2011) show that those relationships can change during financial disruption. For U.S. banks, Kanas et al. (2012) find nonlinear links between profitability and the business cycle, interest rates, credit risk, loan composition, capital, and financial structure; their semi-parametric model also outperforms a conventional linear model out of sample. These findings motivate flexible prediction methods rather than a purely linear specification.

Logistic regression is widely used for binary risk prediction because its coefficients and probabilities are comparatively easy to interpret. That convenience comes with a functional-form restriction: unless nonlinear transformations and interactions are added explicitly, the model assumes that predictors combine through a linear index. This can be restrictive when risk changes at different rates across the predictor distribution.

Random Forest averages many decorrelated decision trees and can represent nonlinearities and interactions without prespecification (Breiman, 2001). XGBoost builds trees sequentially so that each tree addresses residual prediction errors, while shrinkage, row and column sampling, and regularization constrain complexity (Chen & Guestrin, 2016). Both methods are therefore plausible alternatives when bank characteristics interact or exhibit threshold behavior.

Applications to bank supervision provide related evidence. Bräuning et al. (2019) develop a decision-tree early-warning system for small European banks and emphasize its role in prioritizing supervisory attention rather than triggering automatic action. Suss and Treitel (2019) compare classical and machine-learning methods for U.K. bank distress and find strong Random Forest performance on ranking and Brier-score measures. These studies examine supervisory distress in European banking systems and, in some cases, rely on confidential supervisory information, leaving a distinct role for a public-data study of U.S. profitability deterioration.

Recent peer-reviewed evidence broadens the comparison. Petropoulos et al. (2020) evaluate machine-learning methods for bank insolvency prediction; Kristóf and Virág (2022) compare decision trees and deep neural networks for EU-27 bank failures; and Citterio (2024) reviews the expanding bank-failure prediction literature. Hu et al. (2025) apply machine learning to U.S. bank failures and stress testing. These studies reinforce the value of flexible algorithms but focus chiefly on failure or insolvency rather than a more frequent profitability deterioration outcome. The Federal Reserve's review of Silicon Valley Bank further illustrates the need for timely identification and escalation of emerging vulnerabilities (Board of Governors of the Federal Reserve System, 2023).

Rare-event evaluation requires more than ROC AUC. Precision-recall measures are particularly informative when events are uncommon because precision responds directly to false positives (Saito & Rehmsmeier, 2015). Probability quality also matters if risk scores are used to set review thresholds; calibration can differ materially across learning algorithms even when ranking performance is similar (Niculescu-Mizil & Caruana, 2005). Finally, model transparency is important in financial-risk applications. SHAP provides additive feature-level contributions that can be aggregated for global importance or examined across feature values to describe model behavior (Lundberg & Lee, 2017).

Explainability and governance are especially important when machine-learning scores inform financial-risk decisions. Bussmann et al. (2021) demonstrate the role of explainable machine learning in credit-risk management, while Černevičienė and Kabašinskas (2024) document the rapid expansion of explainable AI in finance. The Bank for International Settlements (2024) identifies early-warning indicators as one useful central-bank application of AI. Revised Federal Reserve model-risk guidance emphasizes risk-proportionate validation, documentation, controls, monitoring, and human oversight (Board of Governors of the Federal Reserve System, 2026).

Recent evidence further strengthens the case for combining nonlinear predictive methods with explainability in financial-risk applications. Özkul et al. (2026) apply explainable machine learning to banking-sector profitability and show that profitability reflects interacting bank-specific, macroeconomic, and institutional factors, reinforcing the usefulness of flexible models for analyzing complex banking outcomes. Tuge and Msweli (2026), in a systematic review of explainable artificial intelligence in banking, emphasize that interpretability, governance, and human oversight remain important barriers to operational adoption of complex models. Rejeb et al. (2026) similarly document the growing role of explainable artificial intelligence in financial risk assessment, credit scoring, and related decision-support applications. Related evidence also indicates that Random Forest and XGBoost can outperform logistic regression in financial-distress prediction (Sizan et al., 2025), while machine-learning forecasting research highlights both the predictive advantages of nonlinear models and continuing concerns regarding explainability and data quality (Islam et al., 2025).

2.2 Research Gap and Contribution

Existing profitability studies primarily estimate average relationships with profitability levels, whereas bank early-warning studies generally predict failure or broad distress. Less is known about whether an individual U.S. commercial bank will experience a large profitability decline in the next quarter, whether nonlinear models improve that forecast consistently over time, whether apparent performance merely reflects mean reversion from current ROA, and whether the resulting scores identify deterioration that persists.

This study addresses those gaps by defining a forward-looking deterioration event, reconstructing quarterly earnings from publicly available cumulative Call Report values, enforcing chronological training and evaluation, comparing full models with linear and nonlinear current-ROA-only benchmarks, validating the original risk scores against a persistent two-quarter outcome, examining calibration and decision-curve net benefit, and evaluating both the magnitude and direction of SHAP contributions. The design is intended as a reproducible screening framework rather than a causal model of bank profitability.

3. Data and Methodology

3.1 Data, Outcome, and Predictors

The study uses publicly available quarterly regulatory data compiled from historical U.S. Call Reports and FDIC BankFind Suite uninsured-deposit records (Federal Deposit Insurance Corporation, n.d.; Federal Financial Institutions Examination Council, n.d.). The raw merged file contains 1,013,170 observations, 46 variables, and 15,683 banks from 1992Q1 through 2025Q3, with no duplicate bank-quarter records. Sequential eligibility restrictions reduce the file to 277,959 bank-quarter observations from 7,344 banks. Predictor quarters begin in 2011Q2, and associated outcome quarters extend through 2025Q3. Appendix Tables A1 and A2 report exact source fields, formulas, sample attrition, and missingness.

Annualized quarterly return on assets (ROA) equals four times quarterly net income divided by average total assets. Because Call Report income variables are cumulative within each calendar year, first-quarter income is used directly, while quarterly income for the second through fourth quarters is obtained by subtracting the preceding quarter's year-to-date value. Only consecutive bank-quarter observations are used.

The primary outcome equals one when annualized quarterly ROA declines by at least one percentage point in the following quarter and zero otherwise:

$$D_{i,t+1} = 1[ROA_{i,t+1} - ROA_{i,t} \leq -0.01].$$

Alternative thresholds of 0.5 and 1.5 percentage points are evaluated to determine whether the results depend on the chosen severity definition.

The predictors cover current profitability, earnings structure, asset quality, capital, employee compensation burden, liquidity, funding composition, securities holdings, and bank size. They are annualized quarterly ROA, the net interest income-to-assets ratio, nonperforming-loan ratio, loan-loss-reserve ratio, capital-to-assets ratio, compensation expense ratio, liquidity ratio, brokered-deposit ratio, securities-to-assets ratio, uninsured-deposit share, and the natural logarithm of total assets.

Quarterly net interest income, noninterest income, and salary-and-employee-benefit expense are reconstructed from cumulative Call Report values using the same differencing procedure applied to net income. The compensation expense ratio is quarterly salary-and-employee-benefit expense divided by the sum of quarterly net interest income and quarterly noninterest income. This field is a compensation-burden measure, not total noninterest expense or a comprehensive efficiency ratio. Within every rolling estimation period, continuous predictors are winsorized at the training-sample first and ninety-ninth percentiles. Logistic-regression predictors are then standardized using training-sample means and standard deviations. No evaluation-quarter or future-quarter information enters these transformations.

3.2 Models, Evaluation, and Performance Measures

Three models are compared. Logistic regression uses L2 regularization with C = 1.0. Random Forest uses 250 trees, maximum depth eight, minimum terminal-leaf size fifteen, and the square root of the available predictors at each split. XGBoost uses 300 trees, maximum depth three, learning rate 0.05, minimum child weight fifteen, row and column sampling rates of 0.80, L1 penalty 0.10, and L2 penalty 5.0. The random seed is 2026.

Performance is evaluated with an expanding window across 15 sequential out-of-sample outcome quarters from 2022Q1 through 2025Q3. For each evaluation quarter, training includes only observations whose outcomes occurred before that quarter. The fitted model then predicts deterioration during the evaluation quarter using predictors observed in the immediately preceding quarter. This design preserves the information sequence available to a real-time analyst and avoids the leakage associated with random observation-level splitting.

The finalized analysis was reproduced under Python 3.14.6 with NumPy 2.4.6, pandas 3.0.3, scikit-learn 1.9.0, and XGBoost 3.4.0 on Windows 11. The run generated a sample audit, quarter-level metrics, observation-level out-of-sample probabilities, paired comparison files, SHAP arrays, and a software manifest. The audit reproduced all prespecified sample counts exactly. Appendix A records the variable dictionary and attrition path needed to reconstruct the analysis sample.

Discrimination is measured by ROC AUC, average precision (AP), AP lift, and the Kolmogorov–Smirnov (KS) statistic. The implementation uses scikit-learn's average_precision_score; accordingly, the manuscript reports AP rather than an area-under-the-curve label. AP summarizes the ranked precision-recall tradeoff in a manner suited to rare outcomes (Saito & Rehmsmeier, 2015). Within each evaluation quarter, AP lift is AP divided by the quarter's event rate. The KS statistic is the maximum separation between event and nonevent score distributions.

Probability accuracy is measured by the Brier score (Brier, 1950) and Brier skill score. For each evaluation quarter, the reference forecast assigns every observation the event rate observed in the corresponding training sample; Brier skill equals one minus the model Brier score divided by the reference Brier score. Within each evaluation quarter, the highest-risk group contains 10 percent of observations ranked by predicted probability, with the group size rounded up to the next whole observation. Risk concentration is summarized by the event-capture rate and precision within this group. Table values are means across the 15 quarters, with standard deviations reported to show temporal variation.

Model differences are evaluated with paired two-sided Wilcoxon signed-rank tests across the 15 quarters (Wilcoxon, 1945). Tests compare XGBoost with both logistic regression and Random Forest. Statistical results are interpreted together with effect magnitudes and the number of quarters favoring each model; *p* values are not treated as a substitute for operational relevance.

3.3 Calibration, Explainability, and Robustness Checks

Calibration is assessed by pooling the strictly out-of-sample predictions after all rolling evaluations are complete. Reported diagnostics are the mean predicted probability, calibration intercept, calibration slope, and expected calibration error (ECE) across ten equal-frequency probability groups. Ideal intercept and slope values are zero and one, respectively. These pooled diagnostics are descriptive because banks may appear in multiple evaluation quarters.

Decision-curve analysis evaluates whether model-based risk scores improve the net benefit of selecting banks for review (Vickers & Elkin, 2006). At threshold probability p_t, net benefit equals $TP/N - FP/N \times p_t/(1 - p_t)$. Model curves are compared with review-none and review-all strategies over displayed thresholds from 2 to 20 percent. The threshold odds represent the relative weight assigned to an unnecessary review compared with the benefit of correctly identifying a deterioration event; they are not a direct monetary valuation.

To test whether results are concentrated in one segment of the size distribution, banks are divided into within-quarter quartiles based on log assets. ROC AUC and average precision are calculated separately by model, size quartile, and quarter and then averaged across the 15 quarters.

SHAP values are calculated for XGBoost within each evaluation quarter (Lundberg & Lee, 2017). Global importance is the sample-size-weighted mean absolute SHAP value across quarters. Quarter-specific ranks measure temporal stability. Dependence plots relate observed predictor values to their SHAP contributions to the model's log odds, using a random display sample and equal-frequency bin means. Importance and dependence describe the fitted prediction function; neither establishes causality.

To isolate whether current ROA alone explains performance through mean reversion, two restricted logistic benchmarks are estimated in every rolling window. The first includes current ROA linearly. The second transforms current ROA through a cubic-spline basis with five training-sample quantile knots. Both use L2 regularization with C = 1.0, the same training-sample winsorization, training-only scaling, and the identical expanding-window evaluation. Paired tests compare full-predictor XGBoost with each restricted benchmark across the 15 quarters.

A separate persistent-deterioration label equals one when the primary one-percentage-point decline occurs in quarter t + 1 and ROA in quarter t + 2 remains at least one percentage point below its quarter-t level. Requiring the second following quarter leaves 62,918 observations, 4,712 banks, and 1,056 events (1.68 percent) across 14 outcome quarters from 2022Q1 through 2025Q2. The original strictly out-of-sample probabilities from the primary-outcome models and ROA-only benchmarks are evaluated against this label without refitting. This is a persistence validation of the original flags, not optimization for a new target.

4. Results

4.1 Evaluation Sample and Model Performance

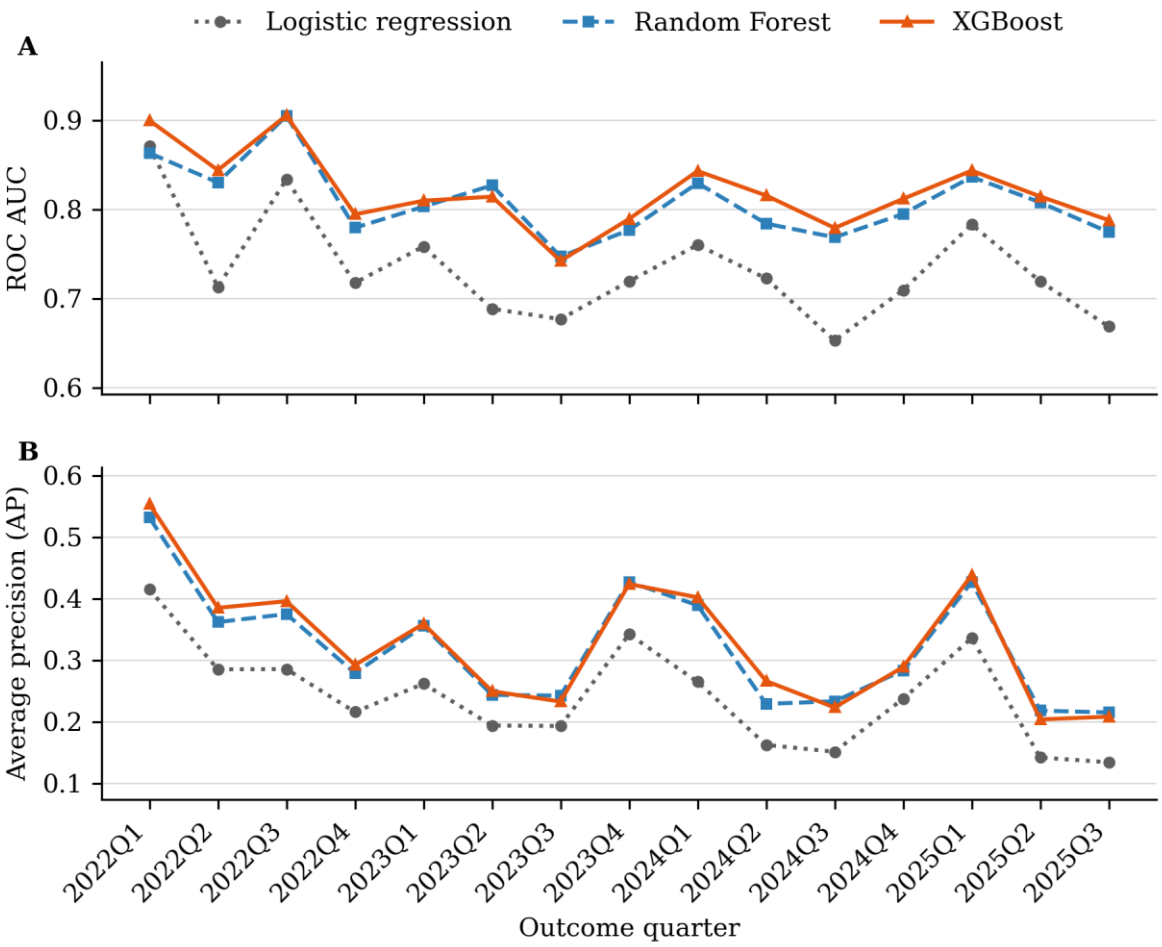
The modeling sample contains 277,959 bank-quarter observations from 7,344 banks. The rolling evaluation includes 67,660 observations from 4,742 banks across 15 quarters from 2022Q1 through 2025Q3. The evaluation period contains 2,551 profitability-deterioration events, an event rate of 3.77 percent. Of 313,218 records eligible immediately before the complete-case requirement, 35,257 lack a finite uninsured-deposit share and two have a nonfinite compensation expense ratio; no other predictor has a nonfinite value.

Both tree-based models outperform logistic regression on every mean performance measure. XGBoost has the highest mean ROC AUC (0.8196), AP (0.3281), AP lift (9.8986), KS statistic (0.5266), Brier skill score (0.1682), top-decile event capture (54.09 percent), and top-decile precision (19.61 percent), together with the lowest Brier score (0.0303). Random Forest is close to XGBoost but remains ahead of logistic regression by a substantial margin. The standard deviations in Table 1 and quarterly paths in Figure 1 show meaningful variation in absolute performance, especially for AP. Nevertheless, both tree-based models generally maintain an advantage over logistic regression, while the relative ordering of XGBoost and Random Forest varies across quarters.

Table 1. Mean Out-of-Sample Performance Across 15 Evaluation Quarters

Performance measure	Logistic regression	Random Forest	XGBoost
ROC AUC	0.7329 (0.0600)	0.8084 (0.0410)	0.8196 (0.0430)
Average precision (AP)	0.2413 (0.0824)	0.3207 (0.0966)	0.3281 (0.1032)
AP lift	7.1843 (3.1898)	9.6437 (4.0532)	9.8986 (4.3838)
KS statistic	0.4491 (0.0820)	0.5042 (0.0683)	0.5266 (0.0661)
Brier score	0.0325 (0.0146)	0.0306 (0.0136)	0.0303 (0.0136)
Brier skill score	0.1107 (0.0546)	0.1610 (0.0675)	0.1682 (0.0777)
Top-decile event capture	49.84% (9.64 pp)	52.69% (9.73 pp)	54.09% (9.57 pp)
Top-decile precision	18.18% (7.52 pp)	19.01% (7.20 pp)	19.61% (7.66 pp)

Note: Entries are quarterly means with standard deviations in parentheses. pp = percentage points. Higher values are better except for the Brier score, for which lower values are better.



Note: Points are quarter-specific out-of-sample results. Average precision varies partly with each quarter's event rate.

Figure 1. Out-of-Sample Performance Across 15 Evaluation Quarters

4.2 Statistical Comparison and Threshold Robustness

Every quarterly XGBoost-versus-logistic difference in ROC AUC, average precision, KS statistic, Brier score, Brier skill score, top-decile capture, and top-decile precision favored XGBoost. The four primary comparisons each produced an exact two-sided p value of 0.000061, reported as $p < 0.001$. Relative to Random Forest, XGBoost produced significantly higher ROC AUC and KS, a lower Brier score, and stronger top-decile concentration. Its average-precision advantage over Random Forest was positive in 10 of 15 quarters but was not statistically significant at conventional levels ($p = 0.095$); the Brier-skill difference was also not significant ($p = 0.055$). These results support XGBoost as the strongest overall model while showing that Random Forest remains a competitive benchmark.

Table 2. Paired XGBoost Performance Improvements Across Evaluation Quarters

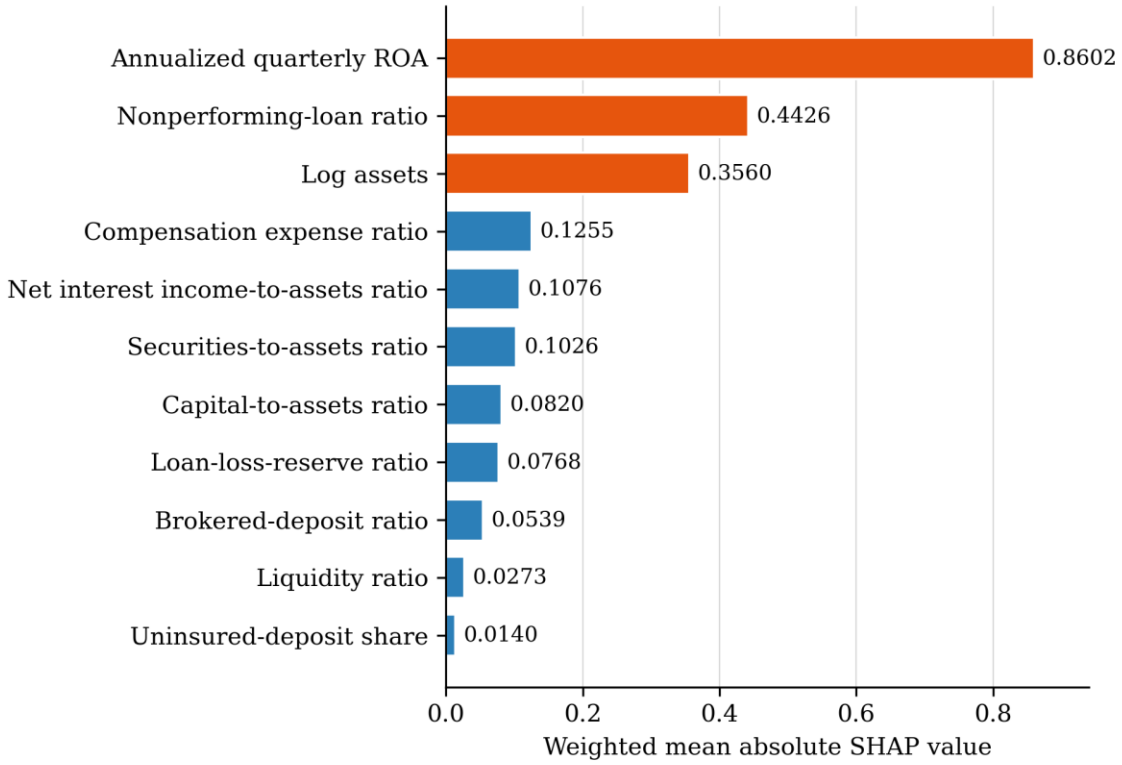
Performance measure	vs. logistic	p value	vs. Random Forest	p value
ROC AUC	0.0867	<0.001	0.0111	0.003
Average precision	0.0868	<0.001	0.0074	0.095
KS statistic	0.0776	<0.001	0.0225	<0.001
Brier-score reduction	0.0022	<0.001	0.0003	0.026
Brier skill score	0.0575	<0.001	0.0072	0.055
Top-decile capture	4.25 pp	<0.001	1.39 pp	0.011
Top-decile precision	1.43 pp	<0.001	0.60 pp	0.005

Note: Values are mean paired quarterly improvements. For all measures except the Brier score, improvement equals XGBoost minus the comparator; for the Brier score, it equals comparator minus XGBoost. p values are from paired two-sided Wilcoxon signed-rank tests. pp = percentage points.

At the 0.5-percentage-point threshold, the event rate increases to 9.09 percent. XGBoost achieves a ROC AUC of 0.7553 and average precision of 0.3242, compared with 0.7070 and 0.2775 for logistic regression. At the stricter 1.5-percentage-point threshold, the event rate falls to 2.25 percent; XGBoost achieves a ROC AUC of 0.8483 and average precision of 0.3435, compared with 0.7360 and 0.2313 for logistic regression. All XGBoost-versus-logistic differences are significant at $p < 0.001$. The ranking advantage therefore persists under both broader and more severe deterioration definitions.

4.3 Explainability, Calibration, and Decision Utility

The finalized SHAP aggregation ranks annualized quarterly ROA first with a weighted mean absolute value of 0.8602, the nonperforming-loan ratio second at 0.4426, and log assets third at 0.3560. These predictors rank first, second, and third in every evaluation quarter. The compensation expense ratio ranks fourth at 0.1255, followed by the net interest income-to-assets ratio at 0.1076, securities-to-assets ratio at 0.1026, capital-to-assets ratio at 0.0820, loan-loss-reserve ratio at 0.0768, brokered-deposit ratio at 0.0539, liquidity ratio at 0.0273, and uninsured-deposit share at 0.0140.

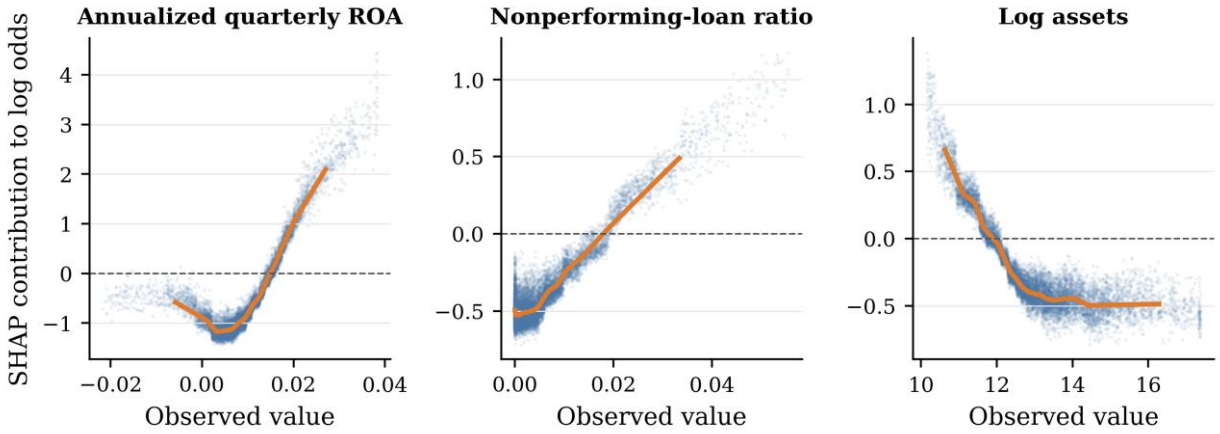


Note: Importance is the sample-size-weighted mean absolute SHAP value. It measures predictive contribution magnitude, not causality or direction.

Figure 2. XGBoost Predictor Importance Across 15 Evaluation Quarters

The dependence patterns add directional information about the fitted model. Current ROA has a nonlinear relationship: contributions are negative through much of the lower and middle range but increase sharply at high current profitability. This pattern is consistent with mean reversion or greater mechanical scope for a one-percentage-point decline from an unusually strong starting quarter; it does not imply that high profitability is intrinsically harmful. Higher nonperforming-loan ratios are

associated with increasingly positive deterioration contributions, while larger banks receive progressively more negative contributions until the size effect levels off. These patterns describe conditional model behavior rather than causal effects.



Note: Blue points are a reproducible random display sample from the pooled out-of-sample SHAP records after excluding the outer 1 percent of feature values for visualization. Orange lines are mean SHAP contributions within 20 equal-frequency bins.

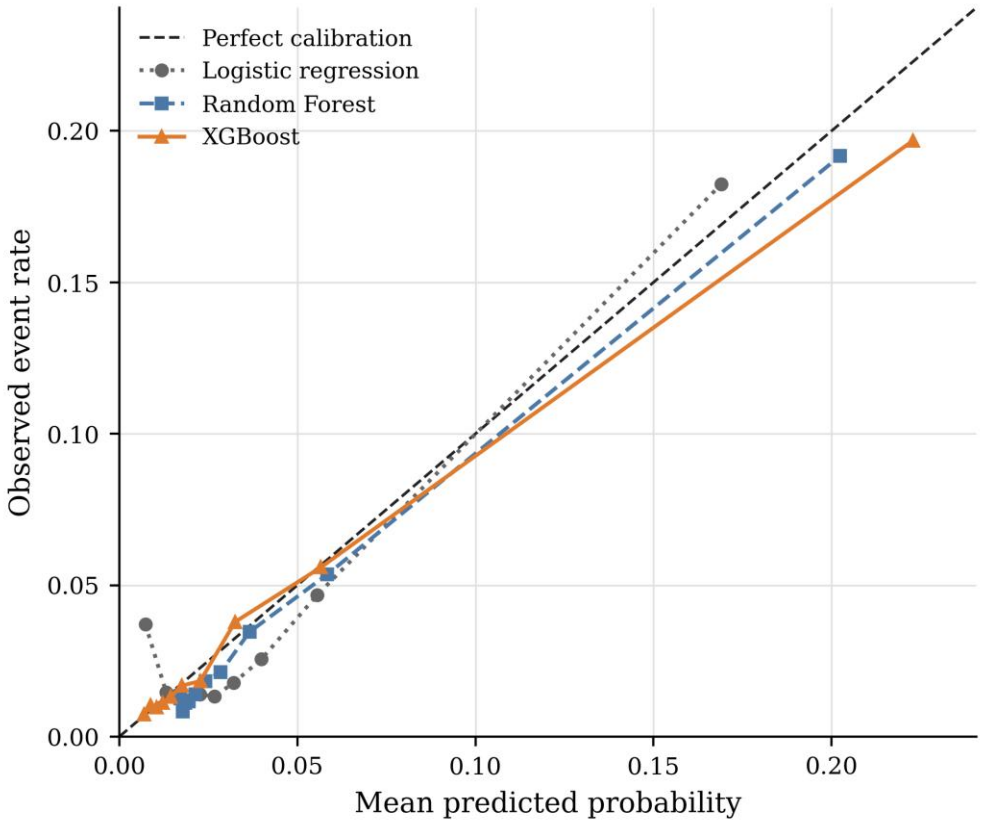
Figure 3. SHAP Dependence for the Three Leading Predictors

The pooled observed event rate is 0.0377. XGBoost's mean predicted probability is 0.0405, indicating modest overall overprediction, but it has the lowest ten-bin ECE (0.0042) and a calibration slope closest to one (0.9624). Random Forest has ECE 0.0067 and slope 1.1170, while logistic regression has ECE 0.0114 and slope 0.9319. XGBoost also has the lowest pooled Brier score. The reliability curves in Figure 4 show that no model is perfectly calibrated at every probability level, particularly in the highest-risk group, so operational deployment would still require ongoing recalibration checks.

Table 3. Pooled Out-of-Sample Calibration Diagnostics

Model	Mean predicted probability	Calibration intercept	Calibration slope	10-bin ECE
Logistic regression	0.0402	−0.2631	0.9319	0.0114
Random Forest	0.0444	0.1022	1.1170	0.0067
XGBoost	0.0405	−0.1840	0.9624	0.0042

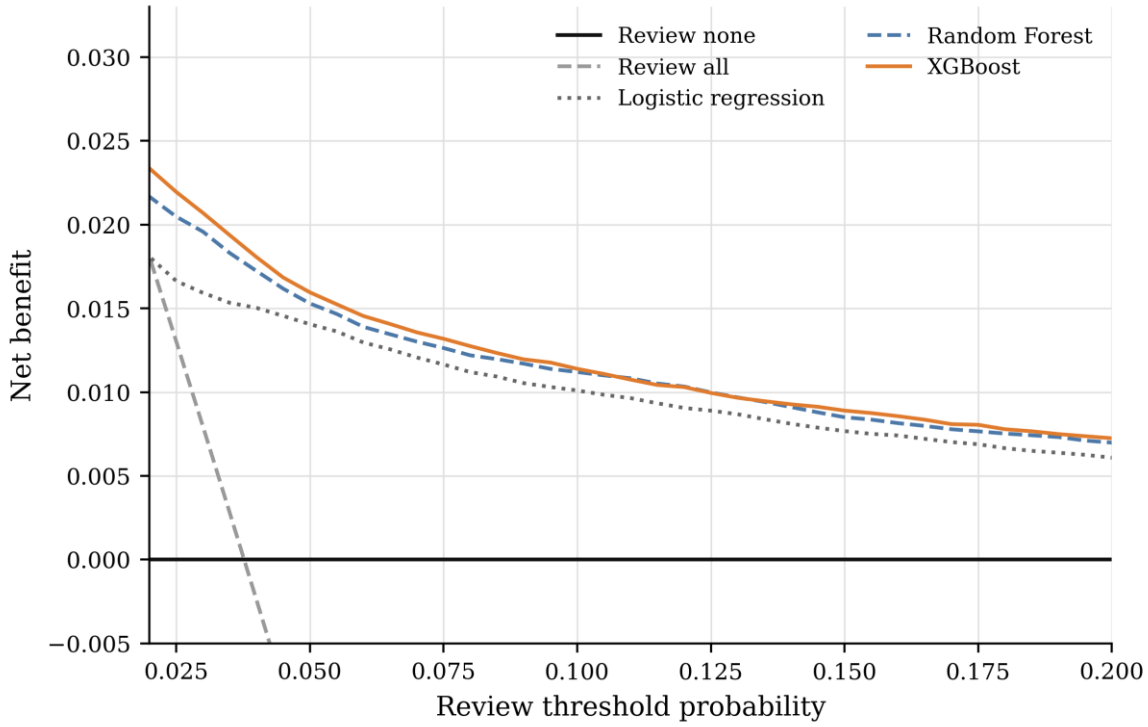
Note: The pooled observed event rate is 0.0377. Ideal calibration intercept and slope values are 0 and 1. ECE is expected calibration error across ten equal-frequency groups; lower values indicate closer agreement.



Note: Each point represents one equal-frequency group of pooled out-of-sample predictions. The dashed 45-degree line denotes perfect calibration.

Figure 4. Pooled Out-of-Sample Calibration

Decision-curve analysis shows that XGBoost has the highest pooled net benefit at most thresholds from 2 to 20 percent, with Random Forest marginally higher around 11–13 percent. At the 5-percent threshold, net benefit is 0.0160, 0.0153, and 0.0141 for XGBoost, Random Forest, and logistic regression, respectively, versus −0.0129 for reviewing all banks and 0 for reviewing none. These values reflect the tradeoff between detecting deterioration and unnecessary reviews, not direct financial gains.



Note: Net benefit equals TP/N − FP/N × p_t/(1 − p_t), where p_t is the review threshold probability. Curves pool strictly out-of-sample predictions.

Figure 5. Decision-Curve Net Benefit Across Review Thresholds

4.4 Bank-Size, Mean-Reversion, and Persistence Robustness

XGBoost's discrimination advantage over logistic regression appears in every within-quarter bank-size quartile. Mean XGBoost ROC AUC ranges from 0.7834 among the largest banks to 0.8590 in the third quartile, compared with 0.6962 and 0.7862 for logistic regression. Average precision is also higher for XGBoost in all four groups. The smallest-bank quartile has the highest event rate, 5.56 percent, but the model advantage is not confined to that segment.

Table 4. Out-of-Sample Performance by Within-Quarter Bank-Size Quartile

Size quartile	Event rate	Logistic ROC AUC	XGBoost ROC AUC	Logistic average precision	XGBoost average precision
Q1 (smallest)	5.56%	0.7011	0.8024	0.2713	0.3500
Q2	3.28%	0.7539	0.8098	0.2525	0.3451
Q3	2.94%	0.7862	0.8590	0.2727	0.3362
Q4 (largest)	3.29%	0.6962	0.7834	0.2271	0.3030

Note: Banks are assigned to log-asset quartiles separately within each evaluation quarter. Reported values are means across the 15 quarters.

Current ROA is a strong restricted predictor, but it does not account for the full-model result. The linear ROA-only logistic benchmark has mean ROC AUC of 0.6777, AP of 0.2151, and top-decile capture of 45.85 percent. Allowing a five-knot cubic spline raises those values to 0.7593, 0.2247, and 46.38 percent. Relative to the nonlinear benchmark, full XGBoost improves mean ROC AUC by 0.0603, AP by 0.1035, lowers the Brier score by 0.0030, and raises top-decile capture by 7.71 percentage points. Every quarterly difference for these four measures favors XGBoost (all paired $p < 0.001$). Thus, flexible mean reversion in current ROA explains part, but not all, of the predictive signal.

The original scores also identify deterioration that remains adverse one additional quarter. Across the 14 quarters with the second following ROA available, XGBoost attains mean ROC AUC of 0.9202, AP of 0.3989, and top-decile capture of 76.22 percent for the persistent label. The nonlinear ROA-only score attains 0.8709, 0.2495, and 70.05 percent. XGBoost's ROC AUC and AP are higher in all 14 quarters, while capture is higher in 13 of 14 (all paired $p < 0.001$). Because the probabilities were trained for the primary outcome, these results show persistence of the original risk ranking rather than performance of a separately optimized two-quarter model.

Table 5. Mean Performance of ROA-Only Benchmarks and Full XGBoost

Model or score	ROC AUC	AP	Top-decile capture
Panel A. Primary one-quarter deterioration (15 quarters; event rate 3.77%)			
ROA-only linear logistic	0.6777	0.2151	45.85%
ROA-only spline logistic	0.7593	0.2247	46.38%
Full XGBoost	0.8196	0.3281	54.09%
Panel B. Persistent two-quarter validation (14 quarters; event rate 1.68%)			
ROA-only linear score	0.8144	0.2464	69.67%
ROA-only spline score	0.8709	0.2495	70.05%
Full XGBoost score	0.9202	0.3989	76.22%

Note: Entries are unweighted means across outcome quarters. Panel B evaluates the original strictly out-of-sample probabilities against the persistent label; models were not refitted for that outcome.

5. Discussion

5.1 Interpretation and Practical Implications

The results show that substantial next-quarter profitability deterioration is predictable with meaningful out-of-sample discrimination using only public bank-level financial information. XGBoost's mean ROC AUC of 0.8196 and AP of 0.3281 are substantively informative relative to the 3.77-percent event rate. Across the 15 evaluation quarters, the highest-risk decile captures an average of 54.09 percent of deterioration events and has mean precision of 19.61 percent, indicating that the model is a screening instrument rather than a deterministic classifier.

XGBoost improves on logistic regression in every evaluation quarter for all primary paired measures and also outperforms the nonlinear ROA-only benchmark in every quarter on ROC AUC, AP, Brier score, and top-decile capture. This consistency supports the view that nonlinearities, interactions, and financial information beyond the starting ROA contribute to forecasting a discrete profitability decline. Random Forest performs nearly as well, and its AP difference from XGBoost is not statistically significant.

The calibration and decision-curve results extend the ranking evidence. XGBoost produces the smallest pooled calibration error and the highest net benefit over a broad range of review thresholds. These findings matter because a model can rank banks well yet still issue probabilities that are unsuitable for operational thresholding. The modest overprediction in the finalized sample remains a reason to monitor and recalibrate probabilities as banking conditions change.

The explainability analysis identifies a stable core of predictive information. Current profitability, nonperforming loans, and bank size occupy the first three importance positions in all 15 quarters. High current ROA is associated with a sharply higher predicted contribution to a subsequent large decline, a pattern consistent with mean reversion or greater mechanical scope for a fall. The restricted spline benchmark confirms that this mechanism is powerful, but the full XGBoost model's significant incremental performance shows that it does not explain the full result. None of these relationships should be interpreted causally. These patterns are also consistent with recent explainable-machine-learning evidence from Özkul et al. (2026), who find that bank profitability reflects interacting bank-specific, macroeconomic, and institutional factors.

Performance across size quartiles indicates that the main result is not driven solely by small or large institutions. The variation in event rates and absolute metrics nevertheless suggests that model monitoring should be segmented. A single overall metric can conceal differences in the frequency and predictability of deterioration across the bank-size distribution.

For regulators, supervisory agencies, and financial institutions, the model can support quarterly risk triage. Ranking banks by predicted deterioration risk allows analysts to concentrate initial review on a manageable subset; across the 15 evaluation quarters, the highest-risk decile captures an average of 54.09 percent of subsequent deterioration events. SHAP explanations can identify which reported characteristics contribute most strongly to a flag, including current profitability, asset quality, size, employee compensation burden, and earnings structure.

Operational thresholds should be selected from the relative costs of missed events and unnecessary reviews, not from predictive performance alone. The decision curves show that model-guided review adds value across plausible threshold probabilities, while the calibration analysis shows why probability monitoring is necessary. Consistent with current model-risk guidance, a production process should include data-quality checks, independent validation, documentation, quarterly performance and subgroup monitoring, periodic retraining, calibration review, change control, and recorded analyst review (Board of Governors of the Federal Reserve System, 2026). This emphasis on transparent, human-supervised use is also consistent with the banking-focused XAI review of Tuge and Msweli (2026), which highlights interpretability, governance, and human oversight as central implementation concerns.

The framework should remain an input to professional judgment. It does not incorporate qualitative supervisory information, strategic changes, local market developments, or events that occur between reporting dates. A high-risk score should initiate further analysis rather than automatic supervisory or managerial action.

5.2 Limitations

The primary outcome is a one-quarter change in annualized ROA reconstructed from regulatory reports. Events may therefore reflect temporary earnings volatility, accounting adjustments, unusual income or expense items, or mean reversion as well as sustained deterioration. The ROA-only benchmarks and persistent-outcome validation directly reduce this concern but do not eliminate sensitivity to the change-based definition. The persistent analysis evaluates scores trained for the primary outcome rather than models optimized for a two-quarter label.

The predictor set is limited to public financial ratios and bank size. It excludes macroeconomic conditions, geographic exposures, governance, supervisory assessments, market prices, and higher-frequency information. Results are specific to U.S. commercial banks and to an evaluation period from 2022Q1 through 2025Q3; structural change may weaken future performance.

Model hyperparameters are prespecified rather than selected through nested time-series tuning. This choice reduces optimization against the evaluation sample but may leave attainable performance unrealized. The paired Wilcoxon tests use 15 sequential quarters for the primary outcome and 14 for persistent validation and do not explicitly model serial dependence, so p values should be interpreted with effect sizes and sign consistency. Pooled calibration and decision curves are descriptive and do not adjust for repeated banks across quarters.

The low event rate and imperfect score separation contribute to false positives, as reflected in mean top-decile precision of 19.61 percent. SHAP values explain the fitted model rather than the data-generating process, and dependence plots may reflect correlated predictors and model interactions. Continued validation, recalibration, and analyst review are therefore necessary before operational use.

6. Conclusion

This study develops a reproducible, explainable machine-learning framework for predicting next-quarter profitability deterioration among U.S. commercial banks. In an expanding-window evaluation of 67,660 bank-quarter observations across 15 quarters, XGBoost achieves the strongest mean combination of discrimination, probability accuracy, and risk concentration. Its mean ROC AUC is 0.8196, AP is 0.3281, Brier score is 0.0303, and highest-risk-decile event capture is 54.09 percent. Relative to a nonlinear current-ROA-only benchmark, XGBoost raises mean ROC AUC by 0.0603, AP by 0.1035, and top-decile capture by 7.71 percentage points, with every quarterly difference favoring XGBoost.

Calibration, decision-curve, and SHAP dependence analyses provide information beyond mean ranking performance. The original risk scores also identify events that remain adverse two quarters after the predictor date: on that persistent label, XGBoost attains mean ROC AUC of 0.9202 and AP of 0.3989. Its explainability profile is centered on current profitability, nonperforming loans, bank size, the compensation expense ratio, and the net interest income-to-assets ratio. The framework can support prioritization for further review, but it should complement established financial analysis, supervisory indicators, and professional judgment rather than operate autonomously.

Statements and Declarations

Data Availability: The underlying Call Report data are publicly available from the Federal Financial Institutions Examination Council and the Federal Deposit Insurance Corporation. The analytical dataset was constructed by the author from these public regulatory records; no confidential or proprietary data were used. Replication code, the variable dictionary, sample-audit outputs, and derived nonconfidential result files are available from the author upon reasonable request.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

AI Use Disclosure: Generative artificial intelligence (ChatGPT) was used to assist with language editing. The author reviewed and verified the final manuscript and accepts full responsibility for its content.

ORCID iD: https://orcid.org/0009-0009-3964-9417

References

[1]. Athanasoglou, P. P., Brissimis, S. N., & Delis, M. D. (2008). Bank-specific, industry-specific and macroeconomic determinants of bank profitability. *Journal of International Financial Markets, Institutions and Money*, *18*(2), 121–136. https://doi.org/10.1016/j.intfin.2006.07.001

[2]. Bank for International Settlements. (2024). *Artificial intelligence and the economy: Implications for central banks* (BIS Annual Economic Report 2024). https://www.bis.org/publ/arpdf/ar2024e3.htm

[3]. Board of Governors of the Federal Reserve System. (2023). *Review of the Federal Reserve's supervision and regulation of Silicon Valley Bank*. https://www.federalreserve.gov/publications/review-of-the-federal-reserves-supervision-and-regulation-of-silicon-valley-bank.htm

[4]. Board of Governors of the Federal Reserve System. (2026). *Revised guidance on model risk management* (Supervision and Regulation Letter SR 26-2). https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm

[5]. Bräuning, M., Malikkidou, D., Scalone, S., & Scricco, G. (2019). *A new approach to early warning systems for small European banks* (ECB Working Paper Series No. 2348). European Central Bank. https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2348~351ba1be4c.en.pdf

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

[7]. 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%3C0001:VOFEIT%3E2.0.CO;2

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

[9]. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review*, *57*, 216. https://doi.org/10.1007/s10462-024-10854-8

[10]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[11]. Citterio, A. (2024). Bank failure prediction models: Review and outlook. *Socio-Economic Planning Sciences*, *92*, 101818. https://doi.org/10.1016/j.seps.2024.101818

[12]. Dietrich, A., & Wanzenried, G. (2011). Determinants of bank profitability before and during the crisis: Evidence from Switzerland. *Journal of International Financial Markets, Institutions and Money*, *21*(3), 307–327. https://doi.org/10.1016/j.intfin.2010.11.002

[13]. Federal Deposit Insurance Corporation. (n.d.). *BankFind Suite*. Retrieved August 13, 2026, from https://banks.data.fdic.gov/bankfind-suite

[14]. Federal Financial Institutions Examination Council. (n.d.). *Central Data Repository's public data distribution*. Retrieved August 13, 2026, from https://cdr.ffiec.gov/

[15]. Hu, W., Shao, C., & Zhang, W. (2025). Predicting U.S. bank failures and stress testing with machine learning algorithms. *Finance Research Letters*, *75*, 106802. https://doi.org/10.1016/j.frl.2025.106802

[16]. Islam, M. R., Wasif, M., Samia, S., Sourav, M. S. R., & Hossain, A. (2025). The role of machine learning in forecasting U.S. GDP growth after the COVID-19 pandemic. *Journal of Economics, Finance and Accounting Studies*, *7*(2), 163–175. https://doi.org/10.32996/jefas.2025.7.2.14

[17]. Kanas, A., Vasiliou, D., & Eriotis, N. (2012). Revisiting bank profitability: A semi-parametric approach. *Journal of International Financial Markets, Institutions and Money*, *22*(4), 990–1005. https://doi.org/10.1016/j.intfin.2011.10.003

[18]. Kristóf, T., & Virág, M. (2022). EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks. *Research in International Business and Finance*, *61*, 101644. https://doi.org/10.1016/j.ribaf.2022.101644

[19]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, 4765–4774. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[20]. 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

[21]. Özkul, G., Akçakanat, Ö., & Özdemir, O. (2026). Rethinking profitability dynamics in the EU banking sector: An explainable machine learning approach to bank sector-specific, macroeconomic, and institutional quality factors. *Journal of Risk and Financial Management*, *19*(8), 553. https://doi.org/10.3390/jrfm19080553

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

[23]. Rejeb, A., Rejeb, K., & Treiblmaier, H. (2026). Explainable artificial intelligence in finance: A bibliometric and topic modeling analysis using BERTopic. *Quality & Quantity*. https://doi.org/10.1007/s11135-026-02837-4

[24]. 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), e0118432. https://doi.org/10.1371/journal.pone.0118432

[25]. Sizan, M. M. H., Chouksey, A., Miah, M. N. I., Pant, L., Ridoy, M. H., Sayeed, A. A., & Khan, M. T. (2025). Bankruptcy prediction for US businesses: Leveraging machine learning for financial stability. *Journal of Business and Management Studies*, *7*(1), 01–14. https://doi.org/10.32996/jbms.2025.7.1.1

[26]. Suss, J., & Treitel, H. (2019). *Predicting bank distress in the UK with machine learning* (Bank of England Staff Working Paper No. 831). Bank of England. https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/predicting-bank-distress-in-the-uk-with-machine-learning.pdf

[27]. Tuge, N. B., & Msweli, N. T. (2026). Explainable artificial intelligence in the banking sector: A systematic literature review. *Applied Artificial Intelligence*, *40*(1), 2676353. https://doi.org/10.1080/08839514.2026.2676353

[28]. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

[29]. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Appendix A. Reproducibility Details

Table A1 maps every model predictor to the standardized Call Report or FDIC BankFind fields used in construction. Table A2 provides the complete attrition path from the raw merged file to the modeling sample.

Table A1. Predictor Definitions and Regulatory Source Fields

Predictor	Construction	Source code(s)
Annualized quarterly ROA	4 × quarterly net income / average assets	RIAD4340; RCFD2170
Net interest income-to-assets	4 × quarterly net interest income / average assets	RIAD4074; RCFD2170
Nonperforming-loan ratio	Nonperforming loans / gross loans	(RCFD1403 + RCFD1407) / (RCFD2122 + RCFD2123)
Loan-loss-reserve ratio	Loan-loss allowance / gross loans	RCFD3123 / (RCFD2122 + RCFD2123)
Capital-to-assets ratio	Total equity / total assets	RCFD3210 / RCFD2170
Compensation expense ratio	Quarterly salary and employee-benefit expense / (quarterly net interest income + quarterly noninterest income)	RIAD4135 / (RIAD4074 + RIAD4079)
Liquidity ratio	Cash and balances due / total assets	(RCFD0081 + RCFD0071) / RCFD2170
Brokered-deposit ratio	Brokered deposits / total deposits	RCON2365 / RCFD2200
Securities-to-assets ratio	Securities / total assets	(RCFD1754 + RCFD1773) / RCFD2170
Uninsured-deposit share	Estimated uninsured deposits / total deposits	FDIC BankFind DEPUNA / DEP
Log assets	Natural logarithm of total assets	ln(RCFD2170)

Note: RIAD fields are year-to-date income-statement values. Quarter 1 is used directly; quarters 2–4 are obtained by subtracting the preceding consecutive quarter. Average assets use current and preceding-quarter RCFD2170. Call Report records are linked by ID_RSSD; FDIC fields are matched by CERT and reporting quarter.

Table A2. Sample Attrition

Eligibility stage	Observations	Banks	Removed
Raw merged bank-quarter records	1,013,170	15,683	—
Source analysis-sample flag	1,006,040	15,576	7,130
Predictor quarter 2011Q2 or later	321,433	7,431	684,607
Consecutive prior quarter available	320,658	7,423	775
Consecutive following-quarter ROA available	313,218	7,345	7,440
All 11 predictors finite (final sample)	277,959	7,344	35,259

Note: Removed is relative to the preceding row. The stage sequence is cumulative.

Missing-data audit. Among the 313,218 records eligible before the complete-case filter, 35,257 (11.26 percent) have a nonfinite uninsured-deposit share and two (0.0006 percent) have a nonfinite compensation expense ratio. The other nine predictors have no nonfinite values. Overlap between the two exclusions leaves 35,259 records removed at the final stage. No imputation is used.