Journal of Trends in Finance and Economics
Print ISSN: 3007-6951
Online ISSN: 3007-696X
DOI: https://doi.org/10.61784/jtfe4078

EARLY WARNING OF LIQUIDITY STRESS IN U.S. BANKS WITH SEASONALLY ALIGNED FRAILTY HAZARDS AND DELAYED-FEEDBACK CONFORMAL ALARMS

JiYuan Wang[1*], Yao Ge[2], JingYi Liu[3], BingYing Jiang[4]
[1]*Duke University, Durham, NC, USA.*
[2]*University of Miami, Coral Gables, FL, USA.*
[3]*Cornell University, Ithaca, NY, USA.*
[4]*University of Wisconsin-Madison, Madison, WI, USA.*
[*]*Corresponding Author: JiYuan Wang*

Abstract: Bank liquidity crises unfold within days, but the balance-sheet vulnerabilities behind them build up over several quarters. This paper develops SAFE (Seasonally Aligned Frailty Early-warning), a machine-learning early-warning system for liquidity stress in U.S. banks. Using the complete Call Reports for 2011–2023 (298,996 bank-quarters), we define a liquidity stress event as a quarterly deposit outflow of at least 7.5% of assets that the bank absorbs through wholesale borrowing or the liquidation of assets, or as a bank failure. SAFE forecasts the probability of such an event over horizons of one to four quarters with a censoring-aware discrete-time hazard model in which each horizon-specific hazard is a gradient-boosted ensemble. Because each hazard refers to a single calendar quarter, its inputs are aligned with the quarter being forecast through empirical-Bayes estimates of the bank's propensity to recurrent and seasonal funding runs. A delayed-feedback adaptive conformal procedure converts the forecasts into supervisory alarms and guarantees that the long-run share of missed events tracks a prescribed target under arbitrary distribution shift. In an out-of-time test on the 2021Q4–2023Q3 monetary-tightening episode, SAFE raises the area under the precision–recall curve relative to LightGBM from 0.205 to 0.232 at the one-quarter horizon and from 0.382 to 0.400 at the four-quarter horizon, outperforms direct classifiers given identical features at all multi-quarter horizons and produces a monotone risk term structure for every bank. SAFE is the most accurate model in each year of a 2016–2023 rolling backtest, and its alarms maintain a recall of 0.80 through the 2022–2023 deposit outflows, whereas a static threshold falls to 0.69.
Keywords: Early warning systems; Liquidity risk; Bank runs; Discrete-time hazard models; Gradient boosting; Conformal prediction; Call Reports

1 INTRODUCTION

In March 2023 Silicon Valley Bank lost a quarter of its deposits in a single day and failed, followed within weeks by Signature Bank and, in May, by First Republic Bank. The three institutions held more than $500 billion in assets, and their failures were preceded by a year in which aggregate U.S. bank deposits declined as the Federal Reserve raised policy rates [1, 2]. The supervisory review of Silicon Valley Bank concluded that the vulnerabilities that brought the bank down (a concentration of uninsured deposits, large unrealized losses on long-duration securities and rapid deposit growth) were visible in regulatory data, but that their severity was not appreciated in time [3]. The episode renewed a practical question for supervisors and risk managers: can regularly filed bank data be turned into a reliable, quarter-by-quarter early-warning signal of *liquidity* stress, and can such a signal be operated under the abrupt regime changes that characterize funding crises?

The early-warning literature offers strong tools but leaves this question open. Bank-level models predict failures or supervisory distress flags from CAMELS-type ratios [4, 5], and machine-learning studies show that gradient boosting and related learners improve such predictions [6, 7]. Three bottlenecks limit their use for liquidity early warning. First, failures are rare, lagging and, in the post-2011 period, mostly credit driven; a liquidity system needs a target that records funding runs when they happen. Second, supervisors need the whole *term structure* of risk, from the next quarter to a year ahead. The common practice of training one classifier per horizon produces term structures that can decrease with the horizon, discards banks that merge before the end of the horizon, and ignores the strong recurrence and seasonality of deposit flows at the bank level. Third, early-warning systems are judged by the alarms they issue. Alarm thresholds calibrated on past data drift when the distribution of risk scores changes, so the share of distressed banks that is caught can collapse precisely when a new type of crisis arrives [8].

This paper addresses the three bottlenecks with an end-to-end system that we call SAFE (*Seasonally Aligned Frailty Early-warning*). We construct a bank-quarter panel of all U.S. Call Reports from 2011Q2 to 2023Q4 (298,996 bank-quarters, 7,644 institutions) and define a liquidity stress event (LSE) as a quarterly deposit outflow of at least 7.5% of assets that the bank absorbed by borrowing or by liquidating assets, or a failure. SAFE estimates the term structure of LSE risk with a censoring-aware landmark hazard model in which each horizon-specific hazard is a gradient-boosted ensemble. The hazard decomposition assigns every horizon to one calendar quarter, which makes it

© By the Author(s) 2024, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

possible to feed each hazard with a *target-aligned seasonal frailty* (TASF) representation: an empirical-Bayes estimate of the bank's propensity to runs in general and in the calendar quarter being forecast, together with its deposit behaviour and run history in the same season of the two previous years. On top of the scores, delayed-feedback adaptive conformal alarms (DF-ACA) set the alarm threshold each quarter so that the long-run share of missed events tracks a supervisory target, with a finite-sample guarantee that holds under arbitrary distribution shift and accounts for the delay with which labels are revealed.

The empirical evaluation is strictly out of time. Models trained on 2011–2021 are tested on the 2021Q4–2023Q3 rate-shock regime, and a rolling-origin experiment re-estimates all models for each year from 2016 to 2023. SAFE is compared with nine baselines, including XGBoost and LightGBM on a rich 90-variable representation and the same learners equipped with exactly the information used by SAFE. On the rate-shock test set SAFE raises the one-quarter AUPRC from 0.205 for LightGBM to 0.232 (+13%) and the four-quarter AUPRC from 0.382 to 0.400. Against direct classifiers that receive the same TASF features, it attains higher AUPRC at every horizon from two to four quarters while producing term structures that are monotone for every bank by construction.

The paper makes four contributions.

Problem and data. We formulate bank liquidity early warning as a multi-horizon forecasting problem for run-type liquidity stress events and build the target, a liquidity-mechanism feature set (LCR-style stressed coverage, mark-to-market capital and run-induced capital depletion) and the evaluation protocol from the complete public Call Report files for 2011–2023, a period that contains the 2023 regional bank runs.

Target-aligned landmark hazard boosting. We show that the discrete-time hazard decomposition is not only a device for coherence and censoring: by tying each horizon to a calendar quarter, it enables a target-aligned empirical-Bayes frailty representation of recurrent, seasonal funding runs that direct classifiers cannot align. The resulting model is coherent by construction and outperforms all baselines, including those that receive the same features.

Alarms with guarantees under drift. We derive DF-ACA, an extension of adaptive conformal inference to batches of banks and to labels revealed with a delay, and prove that its long-run miss rate deviates from the target by at most $(1 + 2\gamma d)/(\gamma T)$ for any sequence of data.

Evidence. We provide a comprehensive validation: ablations of every component, rolling-origin backtests over eight years, sensitivity to the event definition and hyperparameters, bank-cluster bootstrap confidence intervals, size subgroups, case studies of the 2023 failures and Shapley-value explanations.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data, the event definition and the features. Section 4 presents SAFE and DF-ACA. Section 5 describes the experimental design, Section 6 reports the results and Section 7 discusses implications before Section 8 concludes.

2 RELATED WORK

2.1 Bank-Level Early Warning

Bank distress prediction has a long tradition in supervisory off-site monitoring. For U.S. commercial banks, [4] show that CAMELS-type ratios and real-estate exposures explain the post-2008 wave of failures, while [9] document the protective role of capital during crises. European evidence is provided by [5], who build a bank-level early-warning model with bank- and country-level indicators, and by [10], who add network linkages. Machine learning has entered this literature through gradient boosting [6, 11], support vector machines and related learners [12], neural networks and random forests [7, 13]. Methodological work on early-warning evaluation stresses that models must be judged by out-of-sample usefulness under policymakers' preferences [14-16]. At the level of systemic crises, [17] find that logit models are hard to beat in recursive out-of-sample tests, whereas [18] report gains from nonlinear learners and Shapley-value attributions. Two features are common to almost all of these studies: the target is failure (or a supervisory distress flag) rather than a liquidity event, and each forecast horizon is handled by a separate classifier.

2.2 Liquidity Risk and Deposit Runs

The theory of bank runs and empirical studies of depositor behaviour identify uninsured, relationship-poor and rate-sensitive deposits as the run-prone part of bank funding [19-22]. [23] show that the 2007–2009 crisis was a crisis of banks as liquidity providers, and [24] propose a liquidity mismatch index that outperforms the Basel III ratios in measuring liquidity risk. [25] compute Call-Report proxies of the Liquidity Coverage Ratio (LCR) and the Net Stable Funding Ratio and relate them to U.S. bank failures. The 2022–2023 tightening cycle renewed interest in the interaction between interest-rate risk and runs: [26] explain why deposits normally hedge rate risk, [1] show that this hedge disappears when deposits are uninsured and runnable, and [2] mark bank assets to market and show that unrealized losses combined with uninsured leverage exposed many banks to self-fulfilling solvency runs. [27] trace the contagion of the Silicon Valley Bank (SVB) run to uninsured deposits and held-to-maturity losses, and the supervisory post-mortem documents how these vulnerabilities were visible in regulatory data before the failure [3]. This literature supplies the economic mechanisms that our mechanism features encode; it does not provide a forecasting system that can be operated quarter by quarter.

2.3 Learning Under Distribution Shift and Calibrated Alarms

Tree ensembles remain the strongest general-purpose learners on tabular data [28], and gradient boosting implementations such as XGBoost and LightGBM are the de facto standard in credit and distress modelling [29, 30]. Recurrent-event and frailty models from survival analysis capture persistent unobserved heterogeneity across units, but they have not been combined with gradient boosting for bank early warning. Concept drift is a central concern for deployed predictive systems [8]. Conformal prediction offers distribution-free calibration [31]; extensions handle covariate shift [32], non-exchangeable data [33] and arbitrary online drift through adaptive conformal inference (ACI) [34]. The Neyman–Pearson paradigm controls one type of classification error at a user-chosen level [35]. None of these tools has been adapted to the specific structure of supervisory early warning, where labels arrive with a delay equal to the forecast horizon and where the error that matters most is a missed distress event. Shapley-value explanations are used to interpret our model [36].

3 DATA AND PROBLEM FORMULATION

3.1 Call Report Panel

We use the complete quarterly Consolidated Reports of Condition and Income (Call Reports, forms FFIEC 031, 041 and 051) published by the FFIEC Central Data Repository. The bulk files for 2011Q2–2020Q3 and 2020Q4–2023Q4 were obtained from two public mirrors of the official "Call Bulk All Schedules" releases (repositories poc1673/Call_Report_Analysis and mattrwyrick/bank-comparison-toolkit on GitHub). We parse ten schedules (RC, RC-B, RC-C Part I, RC-E, RC-L, RC-M, RC-N, RC-O, RC-R Part I and RI) for every filer and quarter, which yields a panel of 298,996 bank-quarters covering 7,644 insured institutions over 51 quarters. For each item we use the consolidated (RCFD) value when it is reported and the domestic-office (RCON) value otherwise; income items, which are reported year-to-date, are converted into quarterly flows. Failures are taken from the FDIC Failed Bank List and matched to filers through the FDIC certificate number; 167 of the 171 failures between 2011Q3 and 2023Q4 are matched, and 165 of them occur in the quarter that immediately follows the bank's last Call Report.
Aggregates computed from the panel reproduce the known evolution of the U.S. banking system: total assets grow from \$12.7 to \$23.7 trillion, deposits peak at \$19.9 trillion in 2022Q1 and decline thereafter, accumulated other comprehensive income reaches $-\$0.35$ trillion in 2022Q3, and wholesale borrowing jumps from \$1.23 to \$1.76 trillion in 2023Q1, the quarter of the SVB, Signature and First Republic runs. Uninsured deposits are estimated for every bank from Schedule RC-O as the balance of accounts above the \$250,000 limit net of the insured portion; for the institutions that also report the estimate directly (item RCON5597) the cross-sectional correlation of the two uninsured-deposit shares is 0.978.

3.2 Liquidity Stress Events

Failures alone are too rare and too late to discipline a liquidity early-warning system, and the post-2011 failures were largely credit driven [4]. We therefore define a *liquidity stress event* (LSE) as a funding run that the bank had to absorb on its balance sheet. Let $D_{i,s}$, $A_{i,s}$, $W_{i,s}$ and $L_{i,s}$ denote total deposits, total assets, wholesale borrowing (federal funds purchased, repurchase agreements and other borrowed money) and liquid assets (cash, balances due, federal funds sold, reverse repos and investment securities) of bank i at the end of quarter s. Bank i experiences an LSE in quarter s, $e_{i,s} = 1$, if it fails in s or if

$$-\Delta D_{i,s} \geq \tau\, A_{i,s-1} \quad \text{and} \quad \max\left(\Delta W_{i,s}, 0\right) + \max\left(-\Delta L_{i,s}, 0\right) \geq \kappa\left(-\Delta D_{i,s}\right), \tag{1}$$

with $\tau = 7.5\%$ and $\kappa = 0.5$ in the baseline. The first condition requires a deposit outflow of at least 7.5% of assets within one quarter; the second requires that at least half of the outflow was met by new wholesale borrowing or by running down liquid assets, which separates forced liquidity adjustments from orderly balance-sheet shrinkage. The definition is evaluated on consecutive quarterly reports only. Under this definition the unconditional one-quarter event rate is 1.54% and the four-quarter cumulative rate is 5.81%. In normal years the quarterly rate fluctuates between 1% and 2%; it rises to 3.1–3.7% in 2022Q4–2023Q2, when rate-sensitive and uninsured deposits left the banking system (Figure 1). Sensitivity to (τ, κ) is reported in Section 6.9.

3.3 Forecasting Target

For an origin quarter t and horizon $h \in \{1, \ldots, H\}$, $H = 4$, the supervisory question is the probability that bank i experiences at least one LSE within the next h quarters,

$$F_h\left(\mathbf{x}_{i,t}\right) = \Pr\left(\exists\, s \in (t, t+h]:\ e_{i,s} = 1 \middle| \mathbf{x}_{i,t}\right), \tag{2}$$

where $\mathbf{x}_{i,t}$ is computed from Call Reports up to and including quarter t. Two features of the panel make the direct estimation of (2) non-trivial. First, banks leave the panel through mergers and voluntary liquidations; a bank that disappears before $t+h$ without an event has an unknown label, and treating it as a non-event biases long-horizon models. Second, the term structure $h \mapsto F_h$ must be non-decreasing, a property that separately trained classifiers do not guarantee.

3.4 Features

The feature vector contains four groups (90 variables in total). *Balance-sheet ratios* (29) cover funding structure (deposits, non-interest-bearing, uninsured, brokered and large time deposits, wholesale funding, Federal Home Loan Bank advances and their short-term share), asset liquidity (cash, securities, liquid assets, pledged securities, loan-to-deposit ratio, unused commitments), market-value exposure (unrealized losses on available-for-sale and held-to-maturity securities relative to equity, AOCI, the share of securities with maturity above five years, the held-to-maturity share) and CAMELS-type indicators (capital, profitability, net interest margin, provisions, loss allowance, commercial real-estate concentration, cost of deposits, efficiency). *Mechanism features* (6) translate the run mechanism of [19], the LCR run-off logic of [37] and the mark-to-market argument of [2] into bank-level quantities (Section 4.1). *Trajectory features* (15) summarise the last four to eight quarters: deposit growth over one and four quarters, deposit-growth volatility, the drawdown of deposits from their recent peak, asset and loan growth, four-quarter changes of the uninsured share, wholesale funding, liquidity, deposit cost, unrealized losses, capital and loan-to-deposit ratio, and the number of LSEs in the last four quarters. *Peer-relative features* (39) are within-quarter percentile ranks of the balance-sheet and mechanism variables and of four dynamic variables. A calendar variable (quarter of year) is included. The target-aligned frailty features of Section 4.3 are added on top of this representation, see Table 1.

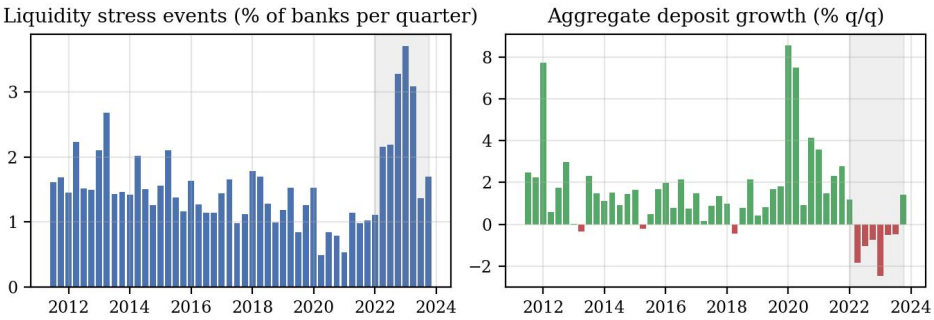


Figure 1 Quarterly Incidence of Liquidity Stress Events (Left) and Aggregate Deposit Growth of U.S. Banks (Right), 2011Q3–2023Q4

Note: The shaded area marks 2022–2023.

Table 1 Descriptive Statistics of Selected Features (All Bank-Quarters with an Observed Next-Quarter Outcome) and Their Means Conditional on a Liquidity Stress Event in the Next Quarter

Variable	Mean	Median	SD	No LSE in $t+1$	LSE in $t+1$
log total assets	12.412	12.246	1.467	12.412	12.059
Deposits/assets	0.828	0.855	0.122	0.828	0.864
Uninsured deposits/deposits	0.203	0.176	0.137	0.200	0.303
Brokered deposits/deposits	0.030	0.000	0.080	0.029	0.043
Wholesale funding/assets	0.039	0.018	0.056	0.039	0.023
Liquid assets/assets	0.339	0.307	0.178	0.337	0.462
Loans/deposits	0.750	0.764	0.256	0.752	0.582
Unrealized losses/equity	0.033	-0.002	0.220	0.028	0.068
Equity/assets	0.124	0.106	0.100	0.124	0.104
ROA	0.011	0.009	0.028	0.011	0.004
Cost of deposits	0.006	0.005	0.005	0.006	0.005
Deposit growth (4q)	0.060	0.040	0.156	0.061	0.069
Stressed liquidity coverage	1.083	1.016	0.523	1.083	1.200
Mark-to-market capital	0.124	0.106	0.100	0.124	0.103
Run-induced capital depletion	0.006	0.000	0.060	0.005	0.009
EB frailty (all quarters)	0.014	0.005	0.027	0.013	0.052
Seasonal frailty of target quarter	0.014	0.006	0.036	0.013	0.080

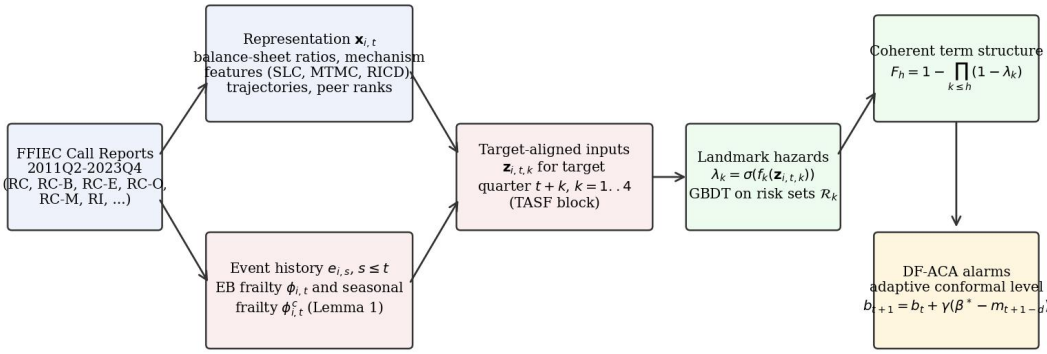


Figure 2 Architecture of SAFE and DF-ACA

4 METHODOLOGY

The proposed early-warning system, SAFE (*Seasonally Aligned Frailty Early-warning*), has a prediction layer and a decision layer. The prediction layer estimates the whole term structure $\{F_h\}_{h=1}^{H}$ with a censoring-aware landmark hazard booster whose horizon-specific models receive features aligned to the calendar quarter they forecast. The decision layer, *delayed-feedback adaptive conformal alarms* (DF-ACA), turns scores into supervisory alarms whose long-run miss rate is controlled under arbitrary distribution shift. Figure 2 summarizes the architecture.

4.1 Mechanism Features

Let C denote cash, balances due, federal funds sold and reverse repos, S^{FV} the fair value of available-for-sale and held-to-maturity securities, P pledged securities, U uninsured deposits, B brokered deposits, R federal funds purchased plus repos, $F^{\leq 1y}$ FHLB advances maturing within a year, M unused loan commitments, E book equity and $\mathrm{UL}^{\mathrm{AFS}}$, $\mathrm{UL}^{\mathrm{HTM}}$ the unrealized losses of the two securities portfolios. Following the LCR structure we define a stressed stock of liquid assets and a stressed 30-day outflow [37],

$$\begin{aligned} Q &= C + 0.85\max\left(S^{\mathrm{FV}} - P, 0\right), \\ O &= 0.40\,U + 0.05\left(D - U - B\right) + 0.25\,B + R + 0.25\,F^{\leq 1y} + 0.05\,M, \end{aligned} \tag{3}$$

and use the stressed liquidity coverage $\mathrm{SLC} = \log(1 + Q/O)$ and the stressed liquidity gap $\mathrm{SLG} = (Q - O)/A$. Following [2] we mark held-to-maturity securities to market and compute the mark-to-market capital ratio $\mathrm{MTMC} = (E - \mathrm{UL}^{\mathrm{HTM}})/A$ and the uninsured leverage of mark-to-market equity $U/\max(E - \mathrm{UL}^{\mathrm{HTM}}, 0.01A)$. Finally, we quantify the channel through which a run forces the sale of securities and the realization of their losses. If a fraction ρ of uninsured depositors withdraws, the liquidity need beyond cash is $N = \max(\rho U + R - C, 0)$, the share of the securities portfolio that must be sold is $\pi = \min(N/S^{\mathrm{FV}}, 1)$ and the run-induced capital depletion is

$$\mathrm{RICD} = \pi\,\frac{\max\left(\mathrm{UL}^{\mathrm{AFS}} + \mathrm{UL}^{\mathrm{HTM}}, 0\right)}{E}, \tag{4}$$

with $\rho = 0.3$. The run gap $\left(N - \max\left(S^{\mathrm{FV}} - P, 0\right)\right)/A$ completes the group.

4.2 Censoring-Aware Landmark Hazard Boosting

Define the discrete-time hazard of the first event after origin t,

$$\lambda_k\left(\mathbf{x}_{i,t}\right) = \Pr\left(e_{i,t+k} = 1 \mid e_{i,t+1} = \ldots = e_{i,t+k-1} = 0,\ \mathbf{x}_{i,t}\right), \qquad k = 1, \ldots, H. \tag{5}$$

The term structure follows from the product-limit identity

$$F_h(\mathbf{x}) = 1 - \prod_{k=1}^{h}\left(1 - \lambda_k(\mathbf{x})\right), \tag{6}$$

so that $F_1 \leq F_2 \leq \ldots \leq F_H$ holds by construction for every bank. For each k, the risk set $\mathcal{R}_k$ contains the origins (i, t) with no event in $(t, t + k - 1]$ and an observed outcome at $t + k$ that lies inside the training window. A separate gradient-boosted tree ensemble f_k [30] is fitted on $\mathcal{R}_k$ by maximizing

$$\mathcal{L}_k = \sum_{(i,t)\in\mathcal{R}_k} w_{i,t}\left[e_{i,t+k}\log\lambda_k + \left(1 - e_{i,t+k}\right)\log\left(1 - \lambda_k\right)\right], \qquad \lambda_k = \sigma\left(f_k\left(\mathbf{z}_{i,t,k}\right)\right), \tag{7}$$

where $\mathbf{z}_{i,t,k}$ is the horizon-specific input defined below and $w_{i,t}$ are optional inverse-probability weights for down-sampled non-events. Three properties follow. First, banks that merge or are liquidated contribute exactly the periods in which they were observed, so no label is imputed; a direct classifier for "event within h quarters" must instead drop every origin whose h-quarter window is incomplete. Second, the forecasts are coherent across horizons for every bank. Third, and central to SAFE, each λ_k refers to a single, known calendar quarter $c(t + k)$, which allows the inputs of f_k to be aligned with the quarter being forecast. A direct cumulative classifier mixes up to h target quarters with different seasonal profiles and cannot exploit this alignment. For $h = 1$ the landmark model and the direct classifier solve the same problem; the decomposition matters for $h \geq 2$.

4.3 Target-Aligned Seasonal Frailty

Liquidity stress events are recurrent at the bank level and strongly seasonal (Figure 1): public-fund, agricultural and tax-related deposit flows repeat in the same calendar quarter, and some funding models are persistently more run-prone than others. We represent this heterogeneity by an empirical-Bayes frailty. Let bank i have a latent event propensity $\pi_i \sim \mathrm{Beta}\left(ap_0, a(1 - p_0)\right)$, where p_0 is the unconditional quarterly event rate and a the prior strength, and let quarterly events be conditionally $\mathrm{Bernoulli}(\pi_i)$.

Lemma 1. *Given $S_{i,t}$ events in $N_{i,t}$ observed quarters up to t, the posterior mean of π_i is*

$$\phi_{i,t} = \mathbb{E}\left[\pi_i \mid S_{i,t}, N_{i,t}\right] = \frac{S_{i,t} + a\,p_0}{N_{i,t} + a}, \tag{8}$$

a convex combination of the bank's empirical event rate and p_0 with weight $N_{i,t}/\left(N_{i,t} + a\right)$ on the former.

Proof. By Beta–Bernoulli conjugacy the posterior is Beta$\big(ap_0 + S_{i,t},\ a(1-p_0) + N_{i,t} - S_{i,t}\big)$, whose mean is (8); rearranging gives $\phi_{i,t} = \frac{N_{i,t}}{N_{i,t}+a}\frac{S_{i,t}}{N_{i,t}} + \frac{a}{N_{i,t}+a}p_0$. □

Equation (8) shrinks short histories, such as those of newly chartered banks, towards the population rate and lets long histories speak for themselves. The *seasonal frailty* of calendar quarter c applies (8) to the quarters $s \le t$ with $c(s) = c$ (prior strength $a/2$). For the k-th hazard the target quarter is $t+k$, and the input of f_k is

$$\mathbf{z}_{i,t,k} = \Big(\mathbf{x}_{i,t},\ \phi_{i,t},\ N_{i,t},\ \phi_{i,t}^{c(t+k)},\ g_{i,t+k-4},\ g_{i,t+k-8},\ e_{i,t+k-4},\ e_{i,t+k-8}\Big), \tag{9}$$

where $g_{i,s}$ is the log deposit growth in quarter s and $e_{i,s}$ the event indicator. Because $k \le H = 4$, every element of (9) is observable at t: the same-season quarter one year before the target is $t+k-4 \le t$. We call the last five elements the target-aligned seasonal frailty (TASF) block. Its interpretation is direct: f_k is told how often the bank experienced a funding run in quarters of the same type as $t+k$, how its deposits behaved in that season one and two years earlier, and how often it experienced runs overall. All TASF quantities are computed from past events only and are recomputed at each origin, so no information from the evaluation window enters the features. In the experiments $p_0 = 0.014$, the training-period event rate, and $a = 8$; Section 6.9 varies a.

4.4 Delayed-Feedback Adaptive Conformal Alarms

A supervisor converts scores into alarms by a threshold. Because examination capacity is limited but missed distress is costly [14], we target the *miss rate* among distressed banks: in each quarter the share of banks that experience an LSE without having been flagged should be close to a target β^*, so that recall stays close to $1-\beta^*$. A static threshold calibrated on historical positives fails when the score distribution of distressed banks shifts, which is what happens when a new type of stress arrives. Moreover, the label of an h-quarter alarm is revealed only h quarters later.

Let $s_{i,t} = \hat{F}_h(\mathbf{x}_{i,t})$ and let $\mathcal{P}_t$ be the scores of the positives whose labels are revealed by quarter t within a rolling window of W origins. For a level b, the conformal lower quantile is $\hat{q}_b(\mathcal{P}_t) = s_{(\lfloor b(n+1)\rfloor)}$, the $\lfloor b(n+1)\rfloor$-th smallest element of $\mathcal{P}_t$ with $n = |\mathcal{P}_t|$, set to $-\infty$ if $b \le 0$ or $\lfloor b(n+1)\rfloor < 1$ and to $+\infty$ if $b \ge 1$. Under exchangeability, flagging $s \ge \hat{q}_b$ misses a new positive with probability at most b [31]. DF-ACA replaces the fixed level by an adaptive level b_t:

$$\begin{aligned} \theta_t &= \hat{q}_{b_t}(\mathcal{P}_t), \qquad \text{alarm}_{i,t} = \mathbb{1}\big[s_{i,t} \ge \theta_t\big], \\ m_t &= \frac{1}{\big|\{i: y_{i,t}=1\}\big|}\sum_{i:\, y_{i,t}=1} \mathbb{1}\big[s_{i,t} < \theta_t\big], \\ b_{t+1} &= b_t + \gamma\,(\beta^* - m_{t+1-d}), \end{aligned} \tag{10}$$

where $d = h$ is the label delay and the update uses the most recent miss rate observable at $t+1$. Equation (10) extends adaptive conformal inference to batches of banks and to delayed feedback [34].

Proposition 1. *Let $b_1 \in [0,1]$, $\gamma > 0$ and let the score and label sequences be arbitrary, with at least one positive in each quarter. Then for every $T \ge 1$ the miss rates produced by (10) satisfy* $\left|\frac{1}{T}\sum_{t=1}^{T} m_t - \beta^*\right| \le \frac{1+2\gamma d}{\gamma T}$.

Proof. If $b_t \le 0$ the threshold is $-\infty$ and $m_t = 0$; if $b_t \ge 1$ the threshold is $+\infty$ and $m_t = 1$. Each update changes b by at most $\gamma\max(\beta^*, 1-\beta^*) \le \gamma$. Suppose b crosses above 1 at quarter t_0, so that $b_{t_0} \le 1+\gamma$. For $t \ge t_0 + d - 1$ the update uses $m_{t+1-d} = 1$ as long as b remains above 1, so b decreases; before that, b can increase in at most $d-1$ further steps. Hence $b_t \le 1 + \gamma d$ for all t, and symmetrically $b_t \ge -\gamma d$. The update that uses m_u produces b_{u+d}, so summing (10) over $u = 1, \ldots, T$ gives $\gamma\sum_{u=1}^{T}(\beta^* - m_u) = b_{T+d} - b_d$, whose absolute value is at most $1 + 2\gamma d$. Dividing by γT yields the claim. □

Proposition 1 holds without any assumption on the data-generating process; in particular it holds through the transition from calm years to a run episode. The step size γ trades the speed of adaptation against the volatility of the threshold; we report $\gamma \in \{0.01, 0.05, 0.10\}$.

4.5 Algorithm and Complexity

Algorithm 1 below summarizes the system. With n origins, H horizons and T_b boosting rounds, training costs $O(H\,T_b\,n)$ for histogram-based boosting; the TASF block requires one pass over each bank's history with cumulative sums by calendar quarter, $O(n)$. Scoring a bank requires H ensemble evaluations and the product (6). The alarm layer stores the revealed positive scores of the last W quarters and performs one sort per quarter. On a single CPU core, fitting all four hazards on 2011–2021 takes about 100 seconds.

Algorithm 1 (SAFE with DF-ACA alarms).

Compute $\phi_{i,t}$, $N_{i,t}$ and the seasonal frailties $\phi_{i,t}^{c}$ from past events by Eq. (8).

For $k = 1, \ldots, H$: build the risk set $\mathcal{R}_k$ and the aligned inputs $\mathbf{z}_{i,t,k}$ of Eq. (9); fit f_k by gradient boosting on Eq. (7).

For each new quarter t: compute $\hat{\lambda}_k = \sigma\left(f_k\left(\mathbf{z}_{i,t,k}\right)\right)$ and $\hat{F}_h = 1 - \prod_{k \le h}(1 - \hat{\lambda}_k)$; reveal the labels of origin $t - d$, add their positive scores to $\mathcal{P}$ (window W), update b_t by Eq. (10); issue alarms for banks with $\hat{F}_h \ge \hat{q}_{b_t}(\mathcal{P})$.

5 EXPERIMENTAL DESIGN

5.1 Out-of-Time Protocol

All evaluations are strictly out of time. In the main experiment the models are trained on all records whose outcome quarter is 2021Q4 or earlier and are evaluated on the origins 2021Q4–2023Q3, the rate-shock regime that contains the 2023 runs. A horizon-h forecast is evaluated only if its whole outcome window lies within the sample (outcome quarter $\le$ 2023Q4) and its label is observed; for $h = 4$ this leaves the origins 2021Q4–2022Q4. The test set contains 38,194 bank-quarters. In the rolling-origin experiment, for each test year $Y \in \{2016, ..., 2023\}$ all models are re-estimated on outcomes up to $(Y - 1)$ Q4 and evaluated on the origins $(Y - 1)$ Q4– Y Q3, which covers calm years, the 2020 pandemic deposit surge and the 2022–2023 outflows.

5.2 Baselines

We compare against nine baselines that cover the supervisory, econometric and machine-learning traditions. On the standard 90-variable representation of Section 3 we estimate (i) CAMELS-Logit, a logistic regression on ten CAMELS-type ratios in the spirit of [4]; (ii) Logit-All, an ℓ_2-regularized logistic regression on all features; (iii) Hazard-Logit, a discrete-time hazard logit with horizon dummies, which shares the censoring-aware likelihood but is linear and pooled across horizons; (iv) a random forest; (v) a multilayer perceptron with two hidden layers (128 and 64 units); (vi) XGBoost [29]; and (vii) LightGBM [30]. To separate the contribution of the representation from that of the hazard structure, we also give the two strongest learners exactly the information used by SAFE: (viii) XGBoost + TASF and (ix) LightGBM + TASF are direct classifiers for "event within h quarters" that receive the frailty features and the aligned TASF blocks of all target quarters $t + 1, ..., t + h$. Direct classifiers are estimated separately for each horizon with censored origins removed.

5.3 Hyperparameters

Hyperparameters are fixed a priori and shared where applicable. All gradient-boosting models, including the four hazards of SAFE, use 500 trees, learning rate 0.05, 31 leaves, at least 100 records per leaf, row subsampling 0.8, column subsampling 0.7 and ℓ_2 regularization 5; XGBoost uses depth 6 with the same learning rate, number of trees and subsampling. The random forest has 150 trees, at least 50 records per leaf and bootstrap fraction 0.25; the MLP uses weight decay 10^{-3}, batch size 512 and early stopping; logistic models are standardized after median imputation. In the main experiment all models are trained on the full data. The ablation, rolling-origin and sensitivity experiments, which require many re-estimations, down-sample non-events with rate $q = 0.3$ and inverse-probability weights for every machine-learning model alike.

5.4 Metrics

Because events are rare, the primary metric is the area under the precision–recall curve (AUPRC). We also report the area under the ROC curve (AUROC), the Brier score, recall at a 10% false-alarm rate (R@FPR10), precision among the 5% highest-ranked banks (P@5%) and the relative usefulness of [14] with preference parameter $\mu = 0.8$, which weights missed events four times as heavily as false alarms. Uncertainty is quantified by a bank-level cluster bootstrap with 300 replications and by re-estimating the stochastic learners with different random seeds. Alarm rules are evaluated by the realized per-quarter recall, its mean absolute deviation from the target, the share of quarters with recall more than 10 percentage points below target, the worst quarter, the share of banks flagged and the precision of alarms.

5.5 Implementation

All models are implemented in Python with LightGBM 4.7, XGBoost 3.4 and scikit-learn 1.8 and run on a single CPU core. Data construction, feature engineering, all experiments, tables and figures are produced by the released code from the raw Call Report files.

6 RESULTS

6.1 Out-of-Time Performance in the 2022–2023 Rate-Shock Regime

Table 2 reports the main out-of-time comparison. The one-quarter test set contains 37,923 forecasts with 882 liquidity stress events, and the four-quarter test set contains 23,494 forecasts with 2,221 events. CAMELS-Logit, which summarizes capital, asset quality, earnings and basic liquidity, attains an AUPRC of only 0.056 at $h = 1$ and 0.176 at

$h = 4$: funding runs are poorly described by the ratios that drive solvency-oriented early warning. Linear models on the full representation reach 0.148 (Logit-All) and 0.124 (Hazard-Logit) at $h = 1$. Among the standard learners, gradient boosting is strongest (LightGBM 0.205 and 0.382, XGBoost 0.208 and 0.375), ahead of the random forest and the MLP, in line with the benchmark evidence of [28].

SAFE attains the highest AUPRC at both horizons, 0.232 at $h = 1$ and 0.400 at $h = 4$. Relative to LightGBM this is a gain of 13.2% and 4.7%, respectively. The gains extend to the operating metrics that matter for supervision: at a 10% false-alarm rate SAFE detects 63.6% of next-quarter events (LightGBM 61.5%), the 5% of banks with the highest four-quarter risk contain 51.0% eventual LSEs (49.9%), and the relative usefulness at $h = 1$ rises from 0.093 to 0.112. Brier scores of the boosting models are close to each other at both horizons. The lower block of Table 2 isolates the source of the gains. Giving XGBoost and LightGBM the same TASF information raises their one-quarter AUPRC to 0.230 and 0.232: at $h = 1$ the representation carries the improvement, and SAFE and LightGBM + TASF coincide because the landmark model and the direct classifier solve the same problem. At $h = 4$, however, the direct classifiers with TASF reach 0.382 and 0.388, whereas SAFE reaches 0.400, so the hazard decomposition adds value beyond the features. Figure 3 shows that the SAFE precision–recall curve lies above those of the baselines over most of the recall range at both horizons.

Table 2 Out-of-Time Performance on the Rate-Shock Test Set (Origins 2021Q4–2023Q3; Models Trained on Outcomes up to 2021Q4)

Model	AUROC	AUPRC	Brier	R@FPR10	P@5%	$U_{0.8}$
Panel A: one-quarter horizon ($h = 1$)						
CAMELS-Logit	0.681	0.056	0.023	0.295	0.082	0.001
Logit-All	0.834	0.148	0.022	0.551	0.171	0.037
Hazard-Logit	0.835	0.124	0.024	0.533	0.157	0.008
RF	0.856	0.153	0.021	0.578	0.168	0.029
MLP	0.815	0.134	0.022	0.490	0.155	0.045
XGBoost	0.869	0.208	0.021	0.615	0.190	0.090
LightGBM	0.870	0.205	0.021	0.615	0.194	0.093
XGBoost + TASF	0.872	0.230	0.021	0.622	0.198	0.114
LightGBM + TASF	0.877	0.232	0.021	0.636	0.208	0.112
SAFE (ours)	0.877	0.232	0.021	0.636	0.208	0.112
Panel B: four-quarter horizon ($h = 4$)						
CAMELS-Logit	0.664	0.176	0.085	0.268	0.252	0.040
Logit-All	0.801	0.329	0.076	0.481	0.456	0.246
Hazard-Logit	0.795	0.321	0.079	0.474	0.462	0.238
RF	0.820	0.359	0.073	0.495	0.440	0.258
MLP	0.787	0.290	0.083	0.453	0.425	0.215
XGBoost	0.825	0.375	0.074	0.511	0.483	0.274
LightGBM	0.829	0.382	0.074	0.523	0.499	0.287
XGBoost + TASF	0.823	0.382	0.077	0.510	0.499	0.274
LightGBM + TASF	0.828	0.388	0.077	0.511	0.506	0.275
SAFE (ours)	0.831	0.400	0.077	0.524	0.510	0.286

Note: R@FPR10: recall at a 10% false-alarm rate; P@5%: precision in the top 5% of banks; $U_{0.8}$: relative usefulness with $\mu = 0.8$. Upper block: standard 90-variable representation; lower block: models with the TASF information.

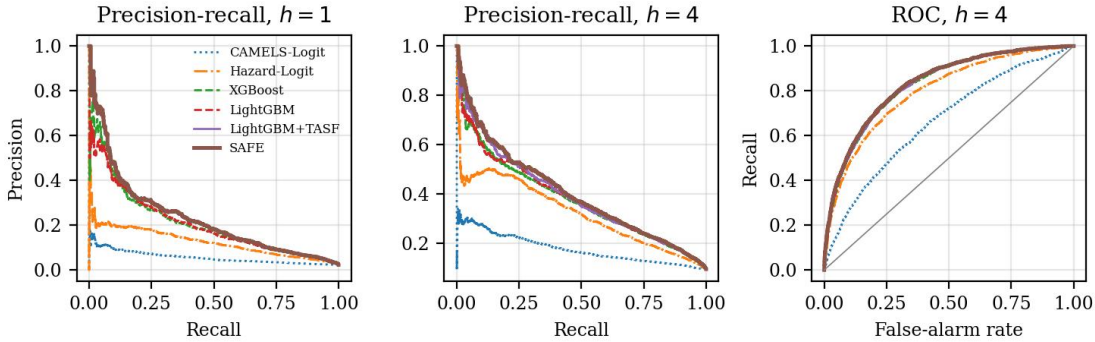


Figure 3 Precision–Recall Curves at $h = 1$ and $h = 4$ and ROC Curves at $h = 4$ on the Rate-Shock Test Set

6.2 Coherent Term Structures

Table 3 reports AUROC and AUPRC for all four horizons together with the share of test banks whose forecast term structure decreases at least once between consecutive horizons. Separately trained direct classifiers violate monotonicity frequently: 26.2% of the term structures produced by LightGBM and 33.4% of those produced by LightGBM + TASF contain at least one decreasing step, which means that for these banks the reported probability of a run within a year is

lower than the probability of a run within a shorter period. SAFE and Hazard-Logit are coherent for every bank by construction. Among the coherent models SAFE dominates at every horizon, and against the strongest direct competitor, LightGBM + TASF, its AUPRC advantage grows with the horizon, from 0.005 at $h = 2$ to 0.009 at $h = 3$ and 0.012 at $h = 4$. This pattern is the one predicted by Section 4.2: the more distinct calendar quarters a forecast window spans, the more a direct classifier has to infer which aligned block belongs to which part of its window, whereas each SAFE hazard receives exactly the block of its own target quarter.

Table 3 Term Structure of Out-of-Time Performance and Coherence

Model	AUROC				AUPRC				Non-mon.
	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 1$	$h = 2$	$h = 3$	$h = 4$	
CAMELS-Logit	0.681	0.674	0.670	0.664	0.056	0.100	0.142	0.176	0.0%
Hazard-Logit	0.835	0.814	0.804	0.795	0.124	0.208	0.275	0.321	0.0%
LightGBM	0.870	0.844	0.835	0.829	0.205	0.280	0.336	0.382	26.2%
LightGBM + TASF	0.877	0.850	0.837	0.828	0.232	0.304	0.351	0.388	33.4%
SAFE (ours)	0.877	0.852	0.840	0.831	0.232	0.309	0.360	0.400	0.0%

Note: Non-monotone: share of test bank-quarters for which the forecast $\widehat{F}_h$ decreases at least once in h.

6.3 Ablation Study

Table 4 removes one component at a time. All variants are re-estimated under the same down-sampling ($q = 0.3$), so they are compared with the full model estimated under the same setting. Three findings stand out. First, both parts of the TASF block matter and they are complementary: removing the seasonal frailties lowers the four-quarter AUPRC from 0.397 to 0.386, removing the same-season deposit history lowers the one-quarter AUPRC from 0.227 to 0.210, and removing the whole block lowers the two metrics to 0.205 and 0.380. Second, target alignment matters exactly where the theory says it should. Feeding every hazard with the block of the next quarter instead of its own target quarter leaves $h = 1$ unchanged by construction but lowers the AUPRC at $h = 2$ from 0.303 to 0.291. Replacing the hazard decomposition by direct classifiers with identical features gives a comparable one-quarter accuracy, since both solve the same problem at $h = 1$, and a lower four-quarter AUPRC (0.387). Third, the generic representation remains important: removing peer ranks costs 0.042 of one-quarter AUPRC, the largest single loss, and removing the trajectory features costs 0.030, whereas the six mechanism features add a smaller but consistent increment at all horizons. Peer ranks are robust to the common shifts in deposit growth and funding costs that separate the 2020–2021 deposit surge from the 2022–2023 outflows, which explains their importance out of time.

Table 4 Ablation Study (AUPRC by Horizon and AUROC at $h = 1,4$; All Variants Trained with Down-Sampling $q = 0.3$)

Variant	AUPRC $_1$	AUPRC $_2$	AUPRC $_3$	AUPRC $_4$	Δ_1	Δ_4
SAFE (full model)	0.227	0.303	0.353	0.397	+0.000	+0.000
−target alignment	0.227	0.291	0.347	0.394	+0.000	-0.003
−seasonal frailty	0.224	0.297	0.347	0.386	-0.004	-0.011
−same-season deposit history	0.210	0.292	0.347	0.389	-0.018	-0.007
−TASF features	0.205	0.283	0.337	0.380	-0.022	-0.017
−trajectory features	0.197	0.280	0.336	0.380	-0.030	-0.017
−peer ranks	0.185	0.279	0.338	0.386	-0.042	-0.010
−mechanism features	0.225	0.298	0.349	0.393	-0.003	-0.004
−hazard (direct, same features)	0.232	0.301	0.354	0.387	+0.004	-0.009

Note: Δ_h: change in AUPRC relative to the full model.

6.4 Rolling-Origin Backtests, 2016–2023

Table 5 and Figure 4 report the rolling-origin experiment, in which every model is re-estimated each year on all earlier data. SAFE attains the highest AUPRC in each of the eight test years at both horizons. Averaged over the years, its one-quarter AUPRC is 0.274 against 0.228 for LightGBM and 0.055 for CAMELS-Logit, a relative gain of 20%; at four quarters the averages are 0.381, 0.355 and 0.134. The advantage is present in calm years (2016–2018), in 2020, when the pandemic deposit surge reversed the usual relation between funding conditions and outflows (0.309 against 0.217), and in the 2022–2023 tightening cycle (0.203 against 0.177 in 2022 and 0.268 against 0.234 in 2023). The event counts in Table 5 show the doubling of LSE incidence in 2022–2023 relative to 2020–2021, which every model faces out of sample.

Table 5 Rolling-Origin Backtests

AUPRC, $h = 1$	AUPRC, $h = 4$

Test year	AUPRC, $h = 1$ Events	CAMELS	LightGBM	SAFE	AUPRC, $h = 4$ Events	CAMELS	LightGBM	SAFE
	1				4			
2016	316	0.062	0.265	0.289	1128	0.131	0.372	0.402
2017	303	0.052	0.245	0.284	1153	0.123	0.358	0.379
2018	321	0.050	0.296	0.352	1010	0.131	0.374	0.416
2019	257	0.057	0.242	0.307	901	0.113	0.344	0.371
2020	187	0.053	0.217	0.309	617	0.100	0.336	0.343
2021	183	0.044	0.148	0.182	882	0.092	0.290	0.314
2022	419	0.041	0.177	0.203	1821	0.171	0.371	0.394
2023	463	0.082	0.234	0.268	400	0.210	0.390	0.432
Mean		0.055	0.228	0.274		0.134	0.355	0.381

Note: For each test year Y the models are trained on outcomes up to $(Y-1)$Q4 and evaluated on the origins $(Y-1)$Q4–YQ3 (down-sampling $q = 0.3$ for both learners). Events: number of LSEs in the evaluated windows.

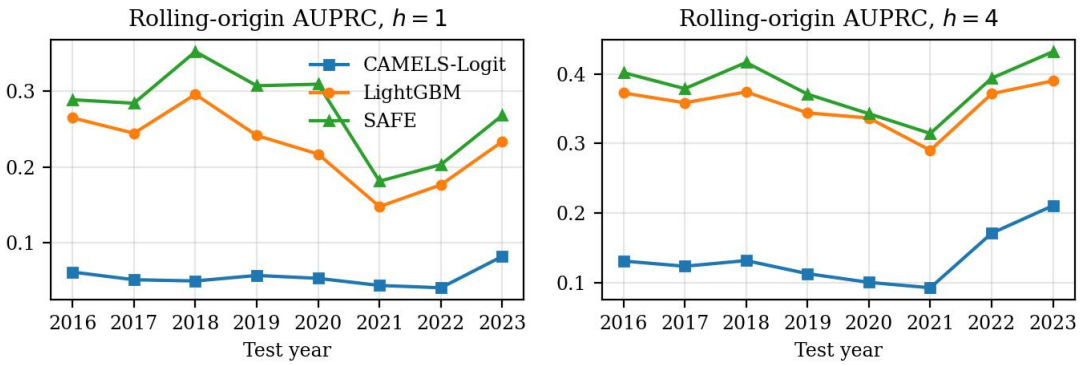


Figure 4 Rolling-Origin AUPRC by Test Year at $h = 1$ (Left) and $h = 4$ (Right)

6.5 Alarm Calibration Under Drift

We operate the one-quarter SAFE scores of the rolling-origin experiment as a live alarm system from 2016Q4 to 2023Q3 (28 quarters), using the first four out-of-sample quarters (2015Q4–2016Q3) as the initial calibration pool. The target recall is 0.8 ($\beta^* = 0.2$), the pool window is $W = 8$ quarters, and labels are revealed with a one-quarter delay. Table 6 and Figure 5 compare DF-ACA with a static split-conformal threshold, a rolling-window recalibration without the adaptive level and a fixed examination budget that flags the 10% riskiest banks every quarter.

The static threshold, calibrated once, achieves a mean recall of 0.733 over the 28 quarters, falls more than ten points below target in 43% of the quarters and reaches a worst quarter of 0.423. Its performance deteriorates in the rate-shock period, with a mean recall of 0.689 and half of the quarters below 0.7. The fixed budget performs worse still in that period (0.612), because it cannot raise the number of alarms when stress becomes widespread. DF-ACA with $\gamma = 0.05$ attains a mean recall of 0.802 over the full period and 0.821 in the rate-shock period, reduces the share of quarters below 0.7 to 17.9% and 12.5%, and raises the worst-quarter recall to 0.614. It does so by adapting the alarm rate: the share of flagged banks rises from about 12% on average before the rate-shock period to 27% during 2021Q4–2023Q3 (Figure 5, right). Rolling recalibration without the adaptive level performs similarly in this sample; DF-ACA adds the distribution-free guarantee of Proposition 1 at no additional cost. The results are insensitive to the step size, and Table 7 shows that DF-ACA also tracks other targets (0.697 for a target of 0.7, against 0.587 for the static rule; 0.902 for a target of 0.9). The quality of the scores matters for the cost of a given recall: with LightGBM scores DF-ACA reaches the same recall (0.805) but must flag 17.3% of the banks on average, against 16.1% with SAFE scores, and the precision of its alarms falls from 0.082 to 0.073.

Table 6 Alarm Calibration with One-Quarter SAFE Scores (Target Recall 0.8)

Threshold rule	Mean recall	|Recall−0.8|	Share < 0.7	Worst quarter	Alarm rate	Precision
			Full period, 2016Q4–2023Q3 (28 quarters)			
Static conformal	0.733	0.106	0.429	0.423	0.105	0.106
Rolling conformal	0.796	0.088	0.179	0.579	0.155	0.084
DF-ACA, $\gamma = 0.01$	0.798	0.089	0.179	0.579	0.157	0.083
DF-ACA, $\gamma = 0.05$	0.802	0.089	0.179	0.614	0.161	0.082
DF-ACA, $\gamma = 0.10$	0.804	0.092	0.214	0.632	0.164	0.081
Budget (top 10%)	0.727	0.117	0.429	0.462	0.100	0.103

Threshold rule	Mean recall	$\lvert\text{Recall}-0.8\rvert$	Share < 0.7	Worst quarter	Alarm rate	Precision
Rate-shock period (8 quarters)						
Static conformal	0.689	0.140	0.500	0.423	0.144	0.117
Rolling conformal	0.814	0.105	0.125	0.641	0.258	0.082
DF-ACA, $\gamma = 0.01$	0.816	0.104	0.125	0.641	0.261	0.081
DF-ACA, $\gamma = 0.05$	0.821	0.110	0.125	0.641	0.273	0.078
DF-ACA, $\gamma = 0.10$	0.830	0.112	0.125	0.641	0.287	0.075
Budget (top 10%)	0.612	0.188	0.750	0.462	0.100	0.137

Note: Share < 0.7: share of quarters with realized recall below 0.7. Alarm rate: average share of banks flagged.

Table 7 Alarm Calibration for Other Recall Targets and for LightGBM Scores (Full Period, 28 Quarters)

Target recall	Rule	Mean recall	$\lvert\text{Recall}-\text{target}\rvert$	Worst quarter	Alarm rate
0.7	Static split-conformal	0.587	0.133	0.295	0.056
	DF-ACA ($\gamma = 0.05$)	0.697	0.109	0.509	0.101
0.9	Static split-conformal	0.891	0.047	0.712	0.261
	DF-ACA ($\gamma = 0.05$)	0.902	0.052	0.750	0.296
0.8	Static (LightGBM scores)	0.769	0.065	0.526	0.138
	DF-ACA (LightGBM scores)	0.805	0.082	0.577	0.173

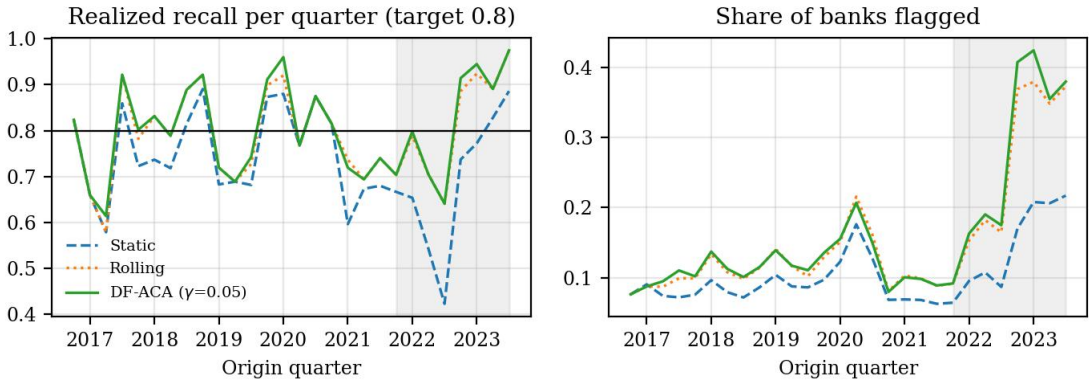


Figure 5 Realized Recall per Quarter (Left) and Share of Banks Flagged (Right) for Static, Rolling-Window and DF-ACA Thresholds

Note: The shaded area marks the rate-shock period.

6.6 The 2023 Episode: Case Studies

Table 8 follows the banks at the centre of the 2023 turmoil through the test period, reporting the percentile of their four-quarter risk among all banks in the same quarter. Silvergate Bank, whose deposits collapsed in 2022Q4, is ranked at the 97.5th percentile already at 2022Q3, one quarter before the outflow. PacWest and Western Alliance, which lost deposits during 2023Q1, rise to the 94th and 75th percentiles at that report. SVB and Signature Bank are ranked at the 79th and 75th percentiles at 2022Q3 and at the 79th and 82nd percentiles at 2022Q4, their last reports before failure, against the 11th and 5th percentiles under CAMELS-Logit. First Republic rises from the 24th percentile at 2022Q4 to the 73rd at 2023Q1, the quarter of its first run. Small banks with a history of recurrent runs, such as Citizens Bank of Sac City (failed November 2023), carry high frailties and are ranked above the 99th percentile throughout. Table 9 reports the one-quarter percentiles of all failed banks in the test window at their last report: SAFE places each of them above the 65th percentile, whereas CAMELS-Logit ranks First Republic at the 2nd and Signature at the 13th percentile. The three large-bank runs of March–May 2023 unfolded within days after events that postdate the preceding quarterly reports; quarterly information placed them well above the median risk but not at the very top of the distribution, which is consistent with the self-fulfilling character of these runs documented by [2].

Table 8 Case Studies of the 2023 Episode

Bank	Origin	Assets	Percentile of $\hat{F}_4$			Unins.	UL/	LSE
		($bn)	SAFE	LightGBM	CAMELS	share	equity	$t+1..t+4$
SVB	2022Q3	210.2	78.7	89.7	9.4	0.85	1.24	01--
SVB	2022Q4	209.0	78.7	85.7	11.2	0.84	1.14	1---
Signature	2022Q3	114.5	74.6	77.2	5.1	0.88	0.43	11--
Signature	2022Q4	110.4	82.2	73.5	4.6	0.87	0.40	1---

Bank	Origin	Assets	Percentile of $\hat{F}_4$			Unins.	UL/	LSE
First Republic	2022Q3	205.1	13.1	7.7	2.2	0.68	0.35	011-
First Republic	2022Q4	212.6	23.8	25.7	2.2	0.68	0.30	11--
First Republic	2023Q1	232.9	73.1	69.2	1.7	0.49	0.25	1---
Silvergate	2022Q3	15.3	97.5	97.2	71.1	0.89	0.92	1111
Silvergate	2022Q4	11.4	99.8	99.7	99.7	0.58	0.20	1111
Silvergate	2023Q1	1.9	99.9	99.9	100.0	1.00	-0.00	111-
PacWest	2022Q3	41.4	62.3	41.2	5.0	0.52	0.28	0100
PacWest	2022Q4	41.2	48.0	23.9	6.5	0.47	0.24	1000
PacWest	2023Q1	44.3	94.1	77.9	31.2	0.24	0.31	000-
Western Alliance	2022Q3	69.1	45.6	23.4	2.4	0.74	0.22	0100
Western Alliance	2022Q4	67.7	51.7	23.3	2.3	0.71	0.18	1000
Western Alliance	2023Q1	71.0	75.3	57.8	3.1	0.52	0.16	000-
Heartland Tri-State	2022Q3	0.1	85.1	76.1	89.9	0.23	1.15	00-1
Heartland Tri-State	2022Q4	0.1	87.7	72.3	91.7	0.28	0.83	0-1-
Heartland Tri-State	2023Q1	0.1	85.9	83.7	93.5	0.27	0.80	-1--
Citizens (Sac City)	2022Q3	0.1	99.8	99.7	92.8	0.39	1.41	1010
Citizens (Sac City)	2022Q4	0.1	99.6	99.3	91.3	0.31	1.21	0101
Citizens (Sac City)	2023Q1	0.1	99.2	98.6	88.7	0.33	1.02	101-

Note: Pct: percentile of the bank's four-quarter risk among all banks at the same origin. Unins. share: uninsured deposits/deposits; UL/equity: unrealized securities losses/equity. LSE $t+1..t+4$: realized events ("-": no report after failure or outside the sample).

Table 9 Failed Banks in the Test Window: Percentile of One-Quarter Risk at the Last Call Report Before Failure

Failed bank	Last report	SAFE	LightGBM + TASF	LightGBM	CAMELS-Logit
Citizens Bank	2023Q3	99.6	99.6	99.3	99.9
Silicon Valley Bank	2022Q4	76.0	76.0	73.5	45.5
Signature Bank	2022Q4	79.7	79.7	80.2	13.2
First Republic Bank	2023Q1	66.1	66.1	64.5	2.2

6.7 What Drives the Forecasts

Figure 6 reports mean absolute Shapley values of the one-quarter and four-quarter hazards on the test set [36]. At $k=1$ the most important inputs are the volatility of deposit growth, the bank's cash buffer and its peer rank, the empirical-Bayes frailty and the same-season deposit changes one and two years earlier, followed by the change in capital and the rank of uninsured leverage of mark-to-market equity. At $k=4$ the frailty becomes the single most important input, ahead of deposit volatility, the same-season deposit history two years earlier and the length of the bank's history, which shows that persistent bank-level propensity dominates the longer horizon while current liquidity conditions dominate the next quarter. For SVB at 2022Q4 the largest positive contributions to the one-quarter hazard come from deposit volatility, low mark-to-market capital, the high uninsured leverage of mark-to-market equity, the rising cost of deposits and the uninsured share. These are the vulnerabilities emphasized by the supervisory review [3] and by [2]; the same-season deposit history, which showed no seasonal outflows in previous years, lowered its risk.

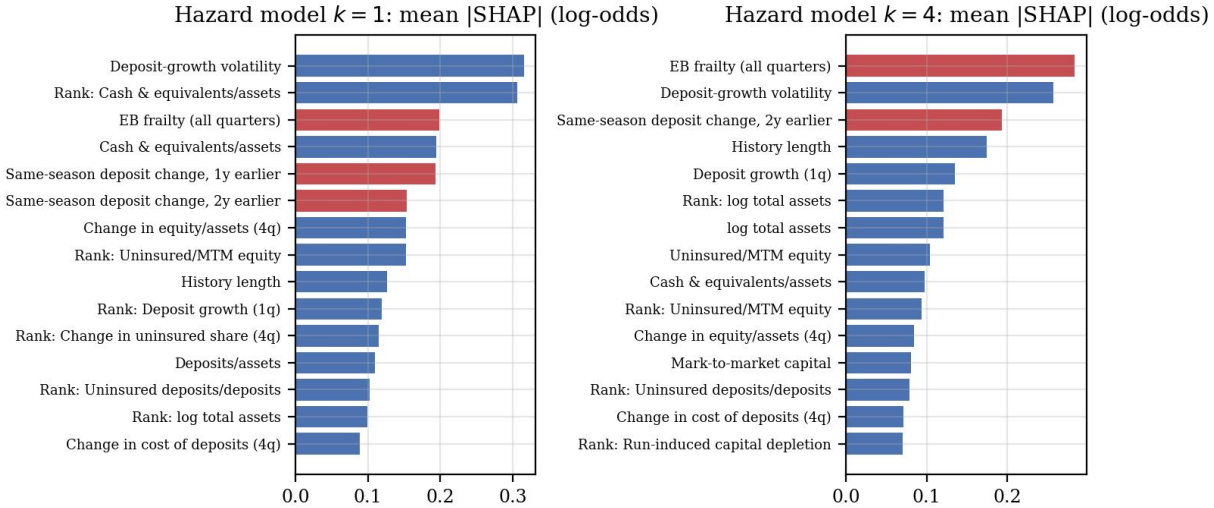


Figure 6 Mean Absolute Shapley Values of the SAFE Hazards for $k=1$ (Left) and $k=4$ (Right) on the Test Set
Note: TASF inputs are shown in red.

6.8 Robustness

6.8.1 Statistical uncertainty

Table 10 reports bank-level cluster-bootstrap confidence intervals for the differences in AUROC and AUPRC between SAFE and each baseline. At $h = 1$ the AUPRC gain over LightGBM is 0.027 with a 95% interval of [0.010, 0.041], and over XGBoost 0.024 [0.007, 0.039]; against the CAMELS, linear and random-forest baselines all intervals exclude zero by a wide margin. At $h = 1$ SAFE coincides with LightGBM + TASF and is statistically indistinguishable from XGBoost + TASF, as expected from Section 6.1. At $h = 4$ the AUPRC gain is significant against every baseline, including the direct classifiers with identical information: +0.012 [0.003, 0.020] against LightGBM + TASF and +0.018 [0.009, 0.026] against XGBoost + TASF. The same ranking holds when AUPRC is computed separately for each origin quarter and then averaged, which removes the influence of the time variation in event rates: SAFE attains 0.253 at $h = 1$ against 0.226 for LightGBM, and 0.402 at $h = 4$ against 0.390 for LightGBM + TASF and 0.384 for LightGBM.

Table 10 Bank-Level Cluster Bootstrap (300 Replications) of the Differences Between SAFE and Each Baseline on the Rate-Shock Test Set

	ΔAUPRC [95% CI]		ΔAUROC	
SAFE minus	$h = 1$	$h = 4$	$h = 1$	$h = 4$
CAMELS-Logit	+0.176 [0.146, 0.208]	+0.223 [0.187, 0.257]	+0.197	+0.166
Hazard-Logit	+0.108 [0.084, 0.133]	+0.078 [0.055, 0.100]	+0.041	+0.036
RF	+0.079 [0.060, 0.101]	+0.040 [0.025, 0.055]	+0.021	+0.011
XGBoost	+0.024 [0.007, 0.039]	+0.025 [0.012, 0.037]	+0.008	+0.006
LightGBM	+0.027 [0.010, 0.041]	+0.018 [0.007, 0.030]	+0.008	+0.002
XGBoost + TASF	+0.001 [-0.007, 0.011]	+0.018 [0.009, 0.026]	+0.005	+0.008
LightGBM + TASF	+0.000 [0.000, 0.000]	+0.012 [0.003, 0.020]	+0.000	+0.003

Note: Mean difference and 95% percentile interval.

6.8.2 Bank size

Table 11 splits the test set at $1 billion of total assets (30,064 small-bank and 7,859 large-bank forecasts at $h = 1$). SAFE has the highest AUPRC in both groups and at both horizons; the one-quarter gain over LightGBM is 0.029 for small banks and 0.019 for large banks, and the four-quarter gain over LightGBM + TASF is 0.015 and 0.004.

Table 11 Out-of-Time Performance by Bank Size (Rate-Shock Test Set)

Size group	Model	AUROC $_1$	AUPRC $_1$	AUROC $_4$	AUPRC $_4$
Assets < $1bn	CAMELS-Logit	0.667	0.054	0.653	0.181
	XGBoost	0.865	0.203	0.818	0.368
	LightGBM	0.864	0.197	0.822	0.376
	LightGBM + TASF	0.873	0.226	0.822	0.380
	SAFE (ours)	0.873	0.226	0.824	0.395
Assets ≥ $1bn	CAMELS-Logit	0.729	0.101	0.691	0.200
	XGBoost	0.885	0.240	0.847	0.413
	LightGBM	0.889	0.250	0.848	0.415
	LightGBM + TASF	0.890	0.269	0.839	0.423
	SAFE (ours)	0.890	0.269	0.855	0.427

6.8.3 Random seeds

Table 12 re-estimates SAFE and its strongest competitor, LightGBM + TASF, with three random seeds. The four-quarter AUPRC of SAFE is 0.401 ± 0.001 against 0.390 ± 0.002, and SAFE is ahead in 3 of the 3 seeds; the seed-to-seed standard deviations are an order of magnitude smaller than the gap between the two models. At $h = 1$ the two models coincide in every seed (0.235 ± 0.005), as expected.

Table 12 Robustness to Random Seeds (Mean ± Standard Deviation over Three Seeds; Rate-Shock Test Set; Full Data)

Model	AUROC $_1$	AUPRC $_1$	AUROC $_4$	AUPRC $_4$
LightGBM + TASF	0.876 ± 0.001	0.235 ± 0.005	0.826 ± 0.002	0.390 ± 0.002
SAFE (ours)	0.876 ± 0.001	0.235 ± 0.005	0.832 ± 0.001	0.401 ± 0.001

6.9 Sensitivity Analysis

6.9.1 Event definition

Table 13 re-estimates SAFE and LightGBM under alternative event definitions. Lowering the outflow threshold to $\tau = 5\%$ almost triples the event rate (3.9% per quarter, 14.4% per year), raising it to $\tau = 10\%$ halves it (0.7% per quarter), and requiring that three quarters of the outflow be funded by borrowing or asset sales ($\kappa = 0.75$) yields an intermediate rate. SAFE improves the one-quarter AUPRC under every definition, by 0.014 to 0.029, and the four-quarter AUPRC under the stricter definitions (0.354 against 0.323 for $\tau = 10\%$ and 0.378 against 0.354 for $\kappa = 0.75$). Under the broad definition $\tau = 5\%$, for which one bank in seven experiences an event within a year, the two models are on par at four quarters (0.531 against 0.534).

Table 13 Sensitivity to the Event Definition (Rate-Shock Test Set; Both Models Trained with $q = 0.3$)

	Event rate		AUPRC $_1$		AUPRC $_4$	
Label definition	$h = 1$	$h = 4$	LightGBM	SAFE	LightGBM	SAFE
$\tau = 5.0\%,\ \kappa = 0.5$	3.91%	14.35%	0.243	0.272	0.534	0.531
$\tau = 10.0\%,\ \kappa = 0.5$	0.73%	2.76%	0.194	0.208	0.323	0.354
$\tau = 7.5\%,\ \kappa = 0.75$	1.37%	5.21%	0.193	0.214	0.354	0.378

6.9.2 Hyperparameters and frailty prior

Table 14 varies the main hyperparameters of SAFE. The model is insensitive to the number of leaves (one-quarter AUPRC between 0.226 and 0.234, four-quarter AUPRC between 0.391 and 0.395) and to the strength of the empirical-Bayes prior over a sixteen-fold range ($a = 2$ to $a = 32$: 0.227–0.228 and 0.392–0.393), which shows that the frailty features do not depend on a finely tuned degree of shrinkage. A larger learning rate with fewer trees lowers the accuracy to 0.205 and 0.380, the level that LightGBM attains on the standard representation, so the default configuration is not at a boundary of favourable settings.

Table 14 Sensitivity of SAFE to Hyperparameters (Rate-Shock Test Set; $q = 0.3$)

SAFE configuration ($q = 0.3$)	AUROC $_1$	AUPRC $_1$	AUROC $_4$	AUPRC $_4$
Baseline (31 leaves, lr 0.05, 500 trees, $a = 8$)	0.872	0.227	0.828	0.397
15 leaves	0.873	0.234	0.829	0.395
63 leaves	0.872	0.226	0.825	0.391
Learning rate 0.10, 250 trees	0.867	0.205	0.825	0.380
Frailty prior strength $a = 2$	0.871	0.227	0.828	0.392
Frailty prior strength $a = 32$	0.873	0.228	0.829	0.393

7 DISCUSSION

7.1 Why Alignment Works

The central empirical regularity exploited by SAFE is that funding runs are recurrent and calendar-specific at the bank level. Municipal and other public deposits, agricultural cycles and tax payments move deposits in the same quarters every year, and some funding models are persistently more run-prone than others. A direct classifier for "an event within the next h quarters" can only learn this structure indirectly, by discovering which of several seasonal blocks is relevant for which part of its forecast window. The landmark hazard decomposition removes that burden: each hazard forecasts one quarter, receives the history of that quarter, and the product-limit identity assembles the term structure. The ablations confirm that the improvement does not come from additional information alone, because direct classifiers with identical inputs are less accurate at multi-quarter horizons and produce incoherent term structures for a quarter to a third of the banks.

7.2 Implications for Supervision

Three practical implications follow. First, run-type liquidity events are predictable from public regulatory data with an accuracy that is high enough to be useful: at a 10% false-alarm rate SAFE identifies almost two thirds of next-quarter liquidity stress events in the 2022–2023 episode, and the banks in the top 5% of four-quarter risk experience an LSE in about half of the cases, against an unconditional rate below 10%. Second, supervisory dashboards should report coherent multi-horizon risk; incoherent term structures are not a cosmetic problem, because they produce contradictory rankings across horizons. Third, alarm thresholds should be treated as an adaptive control problem. DF-ACA is cheap, model-agnostic and comes with a guarantee that does not rely on stationarity, which makes it suitable for the transitions between regimes in which early warning matters most.

7.3 Scope and Extensions

The event definition uses quarterly reports, so intra-quarter runs that are fully reversed before the reporting date are not recorded, and the models use only public Call Report data. Supervisors with access to daily deposit flows, confidential examination ratings or market data can add these sources to the representation without changing the architecture. The same design extends naturally to other recurrent, seasonal events in banking, such as liquidity facility draws or covenant breaches, and to other forecast frequencies.

8 CONCLUSION

We have proposed SAFE, an early-warning system for liquidity stress in U.S. banks built entirely from public Call Reports. SAFE combines a censoring-aware landmark hazard model, whose horizon-specific gradient-boosted hazards yield coherent term structures, with a target-aligned seasonal frailty representation that lets each hazard use the bank's history in the calendar quarter it forecasts. A delayed-feedback adaptive conformal alarm layer converts scores into alarms whose long-run miss rate provably tracks a supervisory target under arbitrary distribution shift. On the 2021Q4–2023Q3 rate-shock regime SAFE improves the one-quarter AUPRC of the strongest baseline by 13% and the four-quarter AUPRC by 5%, outperforms direct classifiers that receive the same features at every multi-quarter horizon, and is coherent for every bank. The improvements are stable across eight years of rolling-origin backtests, event definitions, hyperparameters, random seeds and bank size groups. The results show that the structure of the forecasting problem, and not only the choice of learner, determines how much of the information in regulatory data can be turned into timely warnings.

COMPETING INTERESTS

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

DATA AVAILABILITY

The Call Report bulk files are published by the FFIEC Central Data Repository; the copies used in this study were obtained from the public GitHub repositories poc1673/Call_Report_Analysis (2011Q2–2020Q3) and mattrwyrick/bank-comparison-toolkit (2020Q4–2023Q4). The failure list is the FDIC Failed Bank List. All code needed to rebuild the panel, the features, the experiments, the tables and the figures is provided with the paper.

REFERENCES

[1] Drechsler I, Savov A, Schnabl P, et al. Banking on uninsured deposits. NBER Working Paper 31138, National Bureau of Economic Research, 2023.
[2] Jiang E X, Matvos G, Piskorski T, et al. Monetary tightening and U.S. bank fragility in 2023: Mark-to-market losses and uninsured depositor runs? NBER Working Paper 31048, National Bureau of Economic Research, 2023.
[3] Board of Governors of the Federal Reserve System. Review of the Federal Reserve's supervision and regulation of Silicon Valley Bank. Board of Governors of the Federal Reserve System, Washington, DC, 2023.
[4] Cole R A, White L J. Déjà vu all over again: The causes of U.S. commercial bank failures this time around. Journal of Financial Services Research, 2012, 42: 5-29.
[5] Betz F, Oprică S, Peltonen T A, et al. Predicting distress in European banks. Journal of Banking & Finance, 2014, 45: 225-241.
[6] Carmona P, Climent F, Momparler A. Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 2019, 61: 304-323.
[7] Petropoulos A, Siakoulis V, Stavroulakis E, et al. Predicting bank insolvencies using machine learning techniques. International Journal of Forecasting, 2020, 36: 1092-1113.
[8] Gama J, Žliobaitė I, Bifet A, et al. A survey on concept drift adaptation. ACM Computing Surveys, 2014, 46: 44:1-44:37.
[9] Berger A N, Bouwman C H S. How does capital affect bank performance during financial crises? Journal of Financial Economics, 2013, 109: 146-176.
[10] Constantin A, Peltonen T A, Sarlin P. Network linkages to predict bank distress. Journal of Financial Stability, 2018, 35: 226-241.
[11] Climent F, Momparler A, Carmona P. Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. Journal of Business Research, 2019, 101: 885-896.
[12] Gogas P, Papadimitriou T, Agrapetidou A. Forecasting bank failures and stress testing: A machine learning approach. International Journal of Forecasting, 2018, 34: 440-455.
[13] Le H H, Viviani J-L. Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. Research in International Business and Finance, 2018, 44: 16-25.
[14] Sarlin P. On policymakers' loss functions and the evaluation of early warning systems. Economics Letters, 2013, 119: 1-7.
[15] Holopainen M, Sarlin P. Toward robust early-warning models: A horse race, ensembles and model uncertainty. Quantitative Finance, 2017, 17: 1933-1963.
[16] Lang J H, Peltonen T A, Sarlin P. A framework for early-warning modeling with an application to banks. ECB Working Paper No. 2182, European Central Bank, 2018.
[17] Beutel J, List S, Schweinitz G von. Does machine learning help us predict banking crises? Journal of Financial Stability, 2019, 45: 100693.
[18] Bluwstein K, Buckmann M, Joseph A, et al. Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. Journal of International Economics, 2023, 145: 103773.
[19] Diamond D W, Dybvig P H. Bank runs, deposit insurance, and liquidity. Journal of Political Economy, 1983, 91: 401-419.

[20] Iyer R, Puri M. Understanding bank runs: The importance of depositor-bank relationships and networks. American Economic Review, 2012, 102: 1414-1445.
[21] Iyer R, Puri M, Ryan N. A tale of two runs: Depositor responses to bank solvency risk. Journal of Finance, 2016, 71: 2687-2726.
[22] Egan M, Hortaçsu A, Matvos G. Deposit competition and financial fragility: Evidence from the US banking sector. American Economic Review, 2017, 107: 169-216.
[23] Acharya V V, Mora N. A crisis of banks as liquidity providers. Journal of Finance, 2015, 70: 1-43.
[24] Bai J, Krishnamurthy A, Weymuller C-H. Measuring liquidity mismatch in the banking sector. Journal of Finance, 2018, 73: 51-93.
[25] Hong H, Huang JZ, Wu D. The information content of Basel III liquidity risk measures. Journal of Financial Stability, 2014, 15: 91-111.
[26] Drechsler I, Savov A, Schnabl P. The deposits channel of monetary policy. Quarterly Journal of Economics, 2017, 132: 1819-1876.
[27] Choi D B, Goldsmith-Pinkham P, Yorulmazer T. Contagion effects of the Silicon Valley Bank run. NBER Working Paper 31772, National Bureau of Economic Research, 2023.
[28] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? In: Advances in neural information processing systems 35, datasets and benchmarks track, 2022.
[29] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016: 785-794.
[30] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems 30, 2017: 3146-3154.
[31] Angelopoulos A N, Bates S. Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 2023, 16: 494-591.
[32] Tibshirani R J, Barber R F, Candès E J, et al. Conformal prediction under covariate shift. In: Advances in neural information processing systems 32, 2019.
[33] Barber R F, Candès E J, Ramdas A, et al. Conformal prediction beyond exchangeability. Annals of Statistics, 2023, 51: 816-845.
[34] Gibbs I, Candès E. Adaptive conformal inference under distribution shift. In: Advances in neural information processing systems 34, 2021.
[35] Tong X, Feng Y, Li J J. Neyman-Pearson classification algorithms and NP receiver operating characteristics. Science Advances, 2018, 4: eaao1659.
[36] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. In: Advances in neural information processing systems 30, 2017: 4765-4774.
[37] Basel Committee on Banking Supervision. Basel III: The liquidity coverage ratio and liquidity risk monitoring tools. Bank for International Settlements, Basel, 2013.