International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

A Methodology for Identifying Crisis Mechanisms in Equity Markets: Externally Grounded Labeling, Gradient Boosting, and Synthetic Validation

Arnav Chakole

Abstract: ***This paper proposes a methodology for classifying financial market stress into four market regimes comprising normal conditions, liquidity shocks, credit events, and systemic contagion using gradient-boosted tree classifiers. The principal contribution is an externally grounded labeling protocol in which crisis labels are defined exclusively from non-equity financial stress indicators before any equity return features are examined, thereby eliminating circular identification between labels and predictors. The study presents the complete mathematical formulation of the gradient boosting objective, SHAP interpretation framework, and walk-forward cross-validation procedure. The methodology is validated using synthetic data generated from label-conditional distributions calibrated to established financial crisis theories (Brunnermeier and Pedersen; Lo). The synthetic experiments demonstrate strong classification performance (macro-averaged F1 of 0.93) and recover theoretically expected feature-importance patterns through SHAP analysis in all three crisis regimes. The proposed framework, including labeling, feature construction, model training, walk-forward validation, and SHAP analysis, is provided as a self-contained, reproducible Python script and is presented as ready for empirical application to real market data.***

Keywords: financial crisis typology, labeling protocol, gradient boosting, SHAP, liquidity spirals, synthetic validation

JEL Codes: G01, G11, G17, C53, C45

1. Introduction

Financial crises are not a homogeneous class of events, and treating them as one is a mistake with real consequences. The Federal Reserve's response to the Bear Stearns collapse in March 2008 was a repo facility emergency short-term cash for an institution facing a funding run. Six months later, facing a collapse of interbank trust and cascading institutional failures, the response shifted to capital injections and blanket guarantees. In March 2020, the mechanism was again a liquidity freeze. Treasury markets seized, not because anyone doubted the US government's solvency, but because dealers lacked balance sheet capacity to intermediate and the Fed again used repo facilities and money market support. Same institution, same mandate, three different diagnoses, and three different toolkits.

This paper addresses the gap with a methodology, not an empirical result. I do not claim to have demonstrated that a machine learning classifier correctly identifies crisis types on historical data. That demonstration requires real CRSP and FRED data which I have not yet run. What I do claim is the following: I have designed a labeling protocol that is scientifically valid by construction, specified the mathematical framework completely, and demonstrated that the pipeline produces theoretically consistent results on synthetic data calibrated to theoretical predictions. The methodology is ready for empirical application.

The paper's contributions are as follows. First and this is the main one I introduce a labeling protocol that defines crisis types from external, non-equity stress indicators before any equity return data is examined. The circularity problem in crisis detection research is this: if you look at equity returns to decide which weeks are liquidity crises, and then train a model on equity returns to predict liquidity crises, you have not learned anything. You have constructed a tautology. My labeling protocol is immune to this by design. Second, I provide the complete mathematical specification of the gradient boosting objective, regularization, SHAP decomposition, and walk-forward cross-validation not as a literature summary but as the exact formulation implemented in my pipeline. Third, I validate the methodology on synthetic data and show that the classifier correctly recovers theoretically predicted feature-crisis associations under controlled conditions, which is a necessary (though not sufficient) condition for the methodology to work on real data.

2. Related Work and the Circularity Problem

2.1 Binary Crisis Detection

The mainstream approach to quantitative crisis detection identifies a binary *crisis* or *stress state*. Hamilton's Markov-switching model defines states by the statistical properties of the return process; the crisis state is simply the regime with lower mean and higher variance. Ang and Bekaert extend this to international equity markets. Bianchi, Büchner, and Tamoni use random forests to identify crisis periods in bond markets. These papers are valuable contributions. None of them attempt to distinguish among crisis mechanisms, and none of them need to their research questions do not require it. My research question does.

2.2 The Circularity Problem

The circularity problem arises when crisis labels are defined using the same data that is later used to train a classifier. Suppose a researcher labels a week as a "liquidity crisis" partly because of high equity volatility, then trains a model to predict "liquidity crises" from equity features including realized volatility. The classifier will find high realized volatility predictive of the liquidity crisis label but this is not a discovery about the world. It is a consequence of the labeling definition. The model has not learned that liquidity

crises generate volatility spikes; it has learned that the researcher's definition of liquidity crisis included volatility spikes.

This problem appears in various forms across the crisis-detection literature. It is not always intentional that researchers often use the best available crisis chronologies, which are themselves constructed from asset market data. The NBER business cycle dates, for example, are defined partly by asset market conditions. The BIS crisis database uses equity price declines as part of its crisis identification criteria. Using these as labels and then predicting them from equity features is formally circular.

My solution is straightforward: use only stress indicators that are logically and temporally prior to the equity return data used as features, and define the labeling criteria completely before opening any equity return files. LIBOR-OIS spreads, NFCI sub-indices, and the VIX term structure slope are measures of interbank funding markets, credit conditions, and implied volatility, not equity returns. A labeling protocol built from these indicators does not pre-encode any equity market signal.

2.3 Crisis Typology in Theory

The theoretical finance literature offers a rich crisis typology that motivates my three-way classification. Brunnermeier and Pedersen model liquidity crises as arising from the interaction of margin constraints and market maker balance sheet capacity: forced deleveraging generates a self-amplifying spiral in which asset prices fall, margins tighten, and selling begets more selling. The distinctive empirical signature is a sudden, large spike in short-term realized volatility relative to long run levels forced selling moves prices faster than the long-run volatility baseline.

Credit crises operate through the financial accelerator channel of Bernanke, Gertler, and Gilchrist (1999): deteriorating borrower net worth tightens external financing constraints, reducing investment and amplifying downturns. This mechanism is gradual and sectoral; the distressed sector (technology in 2001, European banks in 2011) diverges sharply from the rest of the market, generating high cross-sectional beta dispersion.

Systemic contagion episodes involve a third mechanism: uncertainty about the structure of counterparty networks, formalized in Caballero and Simsek (2013).

These three mechanisms generate distinct feature signatures, volatility term structure spikes for liquidity, beta dispersion for credit, correlation collapse for systemic crises, which is what motivates a multi-class classification approach and provides a theoretical basis for the feature design in Section 5.

3. Mathematical Framework

3.1 Gradient Boosting Objective

Let the training data be $(\mathbf{x}_i, y_i)$ for (i = 1,...., N) where $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector for week i and $y_i \in N, L, C, S$ is the crisis label. Gradient boosting builds an additive ensemble of M shallow decision trees. After m rounds:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(\mathbf{x}_i)$$

where $\eta \in (0, 1]$ is the learning rate and f_m is the tree added at round m. The XGBoost objective at round m is:

$$\mathcal{L}^{(m)} = \sum_{i=1}^{N} w_i, \ell! \left(y_i, , \hat{y}_i^{(m)}\right) + \Omega(f_m)$$

where the softmax cross-entropy loss for K = 4 classes is:

$$\ell(y_i, \hat{y}_i) = -\log! \left(\frac{e^{\hat{y}i^{(k^*)}}}{\sum k = 1^K e^{\hat{y}_i^{(k)}}}\right), \quad k^* = y_i$$

and the regularization term penalizes tree complexity:

$$\Omega(f_m) = \gamma T + \frac{\lambda}{2}|\mathbf{w}|^2$$

Here T is the number of leaves, w the leaf weight vector, γ an Lo leaf-count penalty, and λ an L2 weight penalty. Taking a second-order Taylor expansion around $\hat{y}_i^{(m-1)}$:

$$\mathcal{L}^{(m)} \approx \sum_{i=1}^{N} w_i \left[g_i f_m(\mathbf{x}_i) + \tfrac{1}{2} h_i f_m(\mathbf{x}_i)^2\right] + \Omega(f_m)$$

where gi and hi are the first and second derivatives of ℓ with respect to $\hat{y}_i^{(m-1)}$.

For leaf j with instance set Ij, the closed-form optimal weight is:

$$w_j^* = -\frac{\sum_{i \in I_j} w_i g_i}{\sum_{i \in I_j} w_i h_i + \lambda}$$

How XGBoost decides where to split a tree: the algorithm tries candidate splits and picks the one that most reduces the loss. The gain from a split is computed from the gi and hi values on the left and right child nodes, minus a penalty for adding a leaf γ. The important point is that the penalty stops the tree from growing too deep, which is critical for my small crisis subsets where overfitting is a real risk.

3.2 Class Imbalance Weighting

The Normal class constitutes roughly 80% of observations in any realistic labeling. Unweighted training causes the model to treat rare crisis classes as noise. I assign observation i in class k an inverse-frequency weight:

$$w_i = \frac{n}{K \cdot n_k}$$

where Nk is the class count and n the total. This amplifies the gradient and Hessian contributions of minority-class observations, forcing the model to classify crisis weeks correctly rather than defaulting to Normal.

3.3 SHAP Value Decomposition

I use SHAP values both for interpretability and as a validation check: if the model is identifying the right patterns, the features that dominate each crisis class should match the theoretical predictions from Section 2.3. The SHAP value

$\phi_j(\mathbf{x})$ feature j is defined as the average marginal contribution of that feature across all possible subsets of the other features. SHAP values have the useful property that they sum to the difference between the prediction and the average prediction. For tree ensembles this computation is exact and fast Lundberg 2020. Class-conditional feature importance is:

$$\bar{\phi}_j^{(k)} = \frac{1}{|\mathcal{I}k|} \sum i \in \mathcal{I}_k |\phi_j^{(k)}(\mathbf{x}_i)|$$

3.4 Walk-Forward Cross-Validation

Standard K fold cross-validation is invalid for time-series data: random fold assignment allows the model to train on observations that postdate its test set, a look-ahead bias that can produce arbitrarily optimistic performance estimates (López de Prado 2018). Walk-forward CV enforces strict time order.

With initial training window To and step size h, fold s uses:
$\text{Train}(s) = [1, ; T_0 + (s-1)h], \qquad \text{Test}(s) = [T_0 + (s-1)h + 1, ; T_0 + sh]$

The guarantee max (Train (s) < min (Test(s)) holds by construction for all s. Performance is reported as macro-averaged F1:

$$F1_{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} \frac{2 \cdot P_k \cdot R_k}{P_k + R_k}$$

Macro-averaging weights all K classes equally regardless of frequency, penalizing the model for failing on minority crisis classes.

4. The Labeling Protocol

This section describes the main methodological contribution. The labeling protocol assigns each week to one of four regimes, Normal (N), Liquidity (L), Credit (C), Systemic (S), using only external, non-equity stress indicators. The complete specification below constitutes the protocol; no additional judgment or equity market data enters the labeling step.

4.1 External Stress Indicators

National Financial Conditions Index: Published weekly by the Federal Reserve Bank of Chicago. I use the overall index and three sub-indices- risk, credit, leverage. The risk sub-index captures market-based volatility and credit measures. The credit sub-index captures traditional lending conditions. The separation between risk and credit sub-indices is what operationalizes the Brunnermeier-Pedersen distinction: a liquidity crisis is characterized by market-based risk stress in excess of fundamental credit deterioration.

LIBOR-OIS spread: The 3-month USD LIBOR rate minus the overnight indexed swap rate of the same tenor, from the Federal Reserve H.15 statistical release. This spread measures interbank funding stress directly: it reflects the premium banks charge each other for unsecured short-term lending above the risk-free rate implied by the OIS. In September 2008 this exceeded 360 basis points; in normal conditions it is near zero.

ICE BofA OAS indices: Option-adjusted spreads for investment grade and high yield corporate bond indices, from FRED. These measure the credit risk premium demanded by bond investors, capturing deterioration in corporate credit conditions independently of equity markets.

VIX term structure slope: Defined as VXV (3-month implied volatility) / VIX (1-month implied volatility). When this ratio falls below 1.0, near-term implied volatility exceeds medium-term markets expect the next few weeks to be more uncertain than the next few months. This term structure inversion is a signature of acute, concentrated stress rather than elevated but diffuse uncertainty.

4.2 Classification Rules

The following rules are applied in order of priority: S > L > C > N. All thresholds are defined relative to the empirical distribution of each indicator over the pre-sample period or, where unavailable, the full sample. Critically, they are fixed before any equity return data is examined.

Systemic (S) All three conditions must hold simultaneously:
(S1) NFCIoverall > 0.87 (95th percentile of indicator distribution)
(S2) VXV/VIX < 0.90
(S3) LIBOR-OIS > 50 bps OR HY OAS > 700 bps

The conjunction of all three conditions is intentional. Systemic episodes are defined by simultaneous broad-based stress (NFCI), acute near-term uncertainty (VIX slope), and either funding market seizure or extreme credit pricing. This corresponds to the Caballero-Simsek mechanism where network uncertainty triggers wholesale risk withdrawal across asset classes.

Liquidity (L). If not Systemic, both conditions must hold:
(L1) LIBOR-OIS ≥ 30bps [material interbank funding stress]
(L2) NFCI risk > 90th percentile and NFCI credit ≤ 90th percentile

The asymmetric NFCI condition is the key design choice. It operationalizes the Brunnermeier-Pedersen distinction between liquidity and credit: a pure liquidity crisis exhibits elevated market-based risk stress (NFCI risk) without a corresponding deterioration in fundamental credit conditions (NFCI credit). This prevents mis-labeling late-stage credit crises (when LIBOR-OIS also widens) as liquidity crises.

Credit (C). If not S or L:
(C1) NFCI credit > 90th percentile
(C2) IG OAS > 200bps OR HY OAS > 500 bps

Requiring both NFCI credit deterioration and spread widening prevents labeling episodes of idiosyncratic spread volatility as credit crises when broader credit conditions remain stable.

Normal (N). All weeks not meeting any of the above criteria.

Table 1: Label Distribution in Synthetic Validation Sample Constructed from the crisis episode chronology in crisis_classification.py. Not a claim about the real distribution.

Label	Count	Percentage
Normal(N)	935	80.10%
Credit(C)	124	10.60%
Systemic(S)	63	5.40%
Liquidity (L)	45	3.90%
Total	1167	100%

5. Feature Design

Features are computed from equity return data after the labeling step is complete. Each feature is motivated by a specific theoretical prediction about how a crisis mechanism manifests in equity market microstructure.

5.1 Volatility Term Structure Ratio (VR)

$$VR_t = \underset{i}{median} \left\{ \frac{\hat{\sigma}_{i,t}^{(21)}}{\hat{\sigma}_{i,t}^{(252)}} \right\}$$

where $\hat{\sigma}^{(21)}$ and $\hat{\sigma}^{(252)}$ are 21-day and 252-day realized volatility for constituent i.

Theoretical prediction: $VR \gg 1$ during liquidity crises and near 1 during credit crises. The forced-selling spiral of a liquidity crisis produces sudden large contemporaneous returns across many stocks, spiking 21-day vol while 252-day vol barely moves. VR should therefore be the primary Liquidity discriminator and a weak Credit discriminator.

5.2 Cross-Sectional Return Dispersion (CSD)

$$CSD_t = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (r_{i,t} - \bar{r}_t)^2}$$

Theoretical prediction: CSD is high in Credit crises (sectors diverge) and low relative to market-level stress in Systemic crises (correlation collapse means everything falls together). CSD does not cleanly separate Liquidity from Normal because both involve market-wide stress.

5.3 Mean Pairwise Correlation ($\bar{\rho}$)

$$\bar{\rho}t = \frac{1}{M} \sum (i,j) \in \mathcal{P} \hat{\rho}_{ij,t}^{(21)}$$

$M = 50$ randomly sampled constituent pairs, resampled monthly.

Theoretical prediction: $\bar{\rho} \to 1$ during Systemic contagion . This should be the dominant Systemic feature and a weak discriminator for the other classes.

5.4 Momentum Reversal (MomRev)

$$MomRev_t = \text{Spearman\_rank\_corr}_i \{\pi(i, t-4), r_{i,t}\}$$

where $\pi(i, t-4)$ is the 252-day momentum decile rank as of four weeks prior.

Theoretical prediction: $MomRev \ll 0$ during Liquidity crises. Momentum strategies are crowded precisely when liquidity is abundant; margin calls during a liquidity crisis force funds to sell winners first, generating sharp reversals (Shleifer and Vishny 1997).

5.5 Beta Interquartile Range (BetaIQR)

$$BetaIQR_t = \hat{\beta}t, 75\% - \hat{\beta}t, 25\%$$

Theoretical prediction: BetaIQR rises in Credit crises (sectoral divergence) and falls in Systemic episodes (beta heterogeneity collapses as everything co-moves). This opposite behavior under Credit vs. Systemic is the primary discriminator between these two regimes.

Table 2: Theoretical Feature Predictions by Crisis Mechanism Expected direction of each feature relative to Normal baseline.

Feature	Liquidity	Credit	Systemic
volatility ratio	↑↑ (spike)	→ (flat)	↑ (moderate)
CSD	→ (flat)	↑ (high)	↓ (low)
pairwise correlation	→ (flat)	→ (flat)	↑↑ (very high)
MomRev	↓↓ (negative)	→ (flat)	↓ (negative)
BetaIQR	→ (flat)	↑ (high)	↓ (low)

6. Synthetic Validation

I validate the methodology on synthetic data rather than claiming empirical results I have not produced. This is not a limitation to apologize for- it is the honest and scientifically appropriate choice. A methodology paper's job is to demonstrate that the framework is internally consistent, correctly implemented, and recovers known ground truth under controlled conditions.

6.1 Synthetic Data Design

I generate N = 1167 weekly observations spanning January 2000 to December 2023. Crisis labels are assigned using historically-grounded episode chronologies: the 2001 till 2002 technology credit crisis, the 2008 Bear Stearns liquidity episode, the 2008 to 2009 GFC systemic phase, the 2011 to 2012 European sovereign credit crisis, and the 2020 COVID 19 liquidity and systemic phases.

Features are drawn from label-conditional Gaussian distributions with parameters calibrated to the theoretical predictions in Table 2 and the empirical measurements in Brunnermeier and Pedersen (2009), Lo (2008), and Gu, Kelly, and Xiu (2020). For example, the Liquidity mean of volatility ratio = 2.2 is chosen to reflect a roughly 2X spike in short-term volatility relative to long run levels, consistent with the forced selling dynamics documented in the 2008 Bear Stearns period. Critically, features are generated after labels are assigned the same design principle that governs the real pipeline.

6.2 Methodology Validation Results

Table 3 reports walk-forward CV performance on the synthetic sample (To = 400, h = 52 weeks, 13 folds). These are real numbers produced by running crisis_classification.py not projections, and not claims about real-world performance.

Table 3: Walk-Forward CV Performance Synthetic Validation Sample All numbers produced by crisis_classification.py on synthetic data. Not a claim about performance on real financial data.

Model	Macro F1	Std Dev
Naive majority	0.222	-
Logistic regression (full)	0.979	0.075
XGBoost	0.946	0.131

Both models substantially outperform the naive baseline, confirming that the feature distributions carry sufficient class-discriminating information. The standard deviation of 0.131 across XGBoost folds reflects genuine variation folds covering crisis transitions (fold 2: GFC onset; fold 13: COVID) have lower F1 than folds covering stable regimes, which is expected and economically meaningful. The logistic regression performing comparably to XGBoost on synthetic data is also expected: the synthetic features are drawn from Gaussian distributions, which are well-separated by linear boundaries. On real data the XGBoost advantage should be more pronounced.

Label	Precision	Recall	F1 Score
Credit (C)	0.97	0.99	0.98
Liquidity (L)	0.83	0.75	0.79
Normal (N)	1	1	1
Systemic (S)	0.94	0.94	0.94
Macro avg	0.94	0.92	0.93

Liquidity has the lowest F1 (0.79) despite having the most theoretically distinct feature signature. This is because the synthetic Liquidity class has only 20 out-of-sample observations statistical noise dominates at this sample size. On real data the Liquidity class would have roughly 45 –103 observations (Table 1), which should improve per-class performance.

6.3 SHAP Validation

The SHAP validation is the methodology's most important test. If the pipeline is correctly recovering the feature-class associations designed into the synthetic data, then the feature that dominates each class in Table 2 should also dominate in the SHAP output. Table 5 reports the normalized mean SHAP per class.

Table 5: Normalized SHAP Importance. Theory predictions from Table 2 in parentheses

Feature	Credit (C)	Liquidity (L)	Normal (N)	Systemic (S)
rho_bar	0.001	0.125	0.001	0.597
volatility ratio p90	0.054	0.273	0.001	0.038
volatility ratio	0.146	0.251	0.002	0.03
MomRev	0.056	0.153	0.003	0.046
FinVolDiff	0.212	0.001	0.217	0.013
VolDiff_Financials	0.178	0.003	0.138	0.016
CSD_fin	0.131	0.015	0.333	0.13
VolDiff_IT	0.091	0	0.238	0.012

All three theory checks pass:

- **Liquidity:** VR/VR_p90 is the dominant feature (0.273), consistent with the Brunnermeier-Pedersen volatility spike prediction.
- **Systemic:** rho_bar is overwhelmingly dominant (0.597), consistent with Lo's (2008) correlation collapse signature.
- **Credit:** FinVolDiff (financial sector volatility differential) leads at 0.212, consistent with the sectoral divergence prediction from Bernanke-Gertler-Gilchrist.

This is a validation of correct implementation, not a claim about real-world crisis behavior. The synthetic data was designed so these features would discriminate between these classes. The fact that the SHAP analysis recovers the design confirms that the pipeline code is implemented correctly and is ready for application to real data.

7. Limitations and Future Work

What this paper establishes. The labeling protocol is internally consistent and immune to the circularity problem by construction. The mathematical framework is fully specified and correctly implemented; synthetic validation demonstrates this. The feature design is theoretically motivated, and the SHAP analysis correctly recovers theoretically predicted associations under controlled conditions.

What this paper does not establish. I make no claim about the classifier's ability to correctly identify crisis types on real historical data. I do not know whether the equity market feature signatures that theory predicts actually appear in the data at the magnitudes and timings that would make classification feasible. I do not know whether the labeling protocol, when applied to real NFCI/LIBOR-OIS/OAS/VIX data, produces a crisis chronology that is economically coherent. These are empirical questions that require running the pipeline on real data.

1) **Threshold sensitivity:** The labeling protocol uses fixed thresholds (0.87 for NFCI, 30 bps for LIBOR-OIS, etc.). These thresholds should be validated against the existing crisis chronology literature (Reinhart and Rogoff 2009; BIS crisis database) when applied to real data. Robustness to ±20% threshold perturbations is a minimum requirement before publication of empirical results.
2) **Sample size:** Crisis observations are rare by definition. Even over a 24-year sample, the three crisis classes together represent only roughly 20% of weeks, and some crisis types are rarer than others. This limits the statistical power of any performance estimate and makes a walk-forward CV, which further reduces the effective training sample, a genuine challenge. Addressing this may require multi-country or longer historical data.
3) **Feature set scope:** The current feature set is restricted to equity market data, which preserves the clean separation from the external stress indicators used in labeling. A real implementation could incorporate options-implied correlation, CDS spreads, or repo market haircuts as additional features but doing so requires care to ensure these features do not overlap with the labeling indicators.

8. Conclusion

I have proposed a methodology for multi-class financial crisis classification and demonstrated its internal consistency through synthetic validation. The central contribution is the labeling protocol: crisis type labels are defined exclusively from external, non-equity stress indicators before any equity return data is examined. This design eliminates the circular identification problem that affects much of the existing crisis-detection literature, where crisis labels are defined using the same asset market signals later used to predict them.

The gradient boosting framework, SHAP attribution, and walk-forward cross-validation are fully mathematically specified and reproducibly implemented in the accompanying Python pipeline. On synthetic data calibrated to theoretical feature-crisis associations, the SHAP analysis recovers all three theoretically predicted crisis signatures: volatility term structure dominance for Liquidity (Brunnermeier-Pedersen), correlation collapse dominance for Systemic (Lo, Caballero-Simsek), and financial sector divergence dominance for Credit (Bernanke-Gertler-Gilchrist). This confirms that the pipeline is correctly implemented and ready for empirical application.

I deliberately make no claims about real-world classification performance. The empirical paper applying this methodology to actual CRSP or the FRED data and producing a genuine crisis chronology with out-of-sample performance statistics is identified as the primary direction for future work. The methodology this paper establishes is the necessary foundation for that work.

References

[1] Ang, A., & Bekaert, G. (2002). International Asset Allocation with Regime Shifts. *Review of Financial Studies*, 15(4), 1137–1187. https://doi.org/10.1093/rfs/15.4.1137

[2] Bernanke, B. S., Gertler, M., & Gilchrist, S. (1999). The Financial Accelerator in a Quantitative Business Cycle Framework. In J. B. Taylor & M. Woodford (Eds.), *Handbook of Macroeconomics*, Vol. 1C (pp. 1341–1393).

[3] Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond Risk Premiums with Machine Learning. *Review of Financial Studies*, 34(2), 1046–1089. https://doi.org/10.1093/rfs/hhaa062

[4] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market Liquidity and Funding Liquidity. *Review of Financial Studies*, 22(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[5] Caballero, R. J., & Simsek, A. (2013). Fire Sales in a Model of Complexity. *Journal of Finance*, 68(6), 2549–2587. https://doi.org/10.1111/jofi.12088

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[7] Duffie, D. (2020). *Still the World's Safe Haven? Redesigning the U.S. Treasury Market After the COVID-19 Crisis*. Hutchins Center Working Paper No. 62, Brookings Institution. [verify]

[8] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

[9] Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2), 357–384. https://doi.org/10.2307/1912559

[10] He, Z., Nagel, S., & Song, Z. (2022). Treasury Inconvenience Yields During the COVID-19 Crisis. *Journal of Financial Economics*, 143(1), 57–79. https://doi.org/10.1016/j.jfineco.2021.06.014

[11] Lo, A. W. (2008). *Hedge Funds: An Analytic Perspective*. Princeton University Press.

[12] López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.

[13] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 4765–4774). arXiv:1705.07874.

[14] Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9

[15] Reinhart, C. M., & Rogoff, K. S. (2009). *This Time Is Different: Eight Centuries of Financial Folly*. Princeton University Press.

[16] Shleifer, A., & Vishny, R. W. (1997). The Limits of Arbitrage. *Journal of Finance*, 52(1), 35–55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

Appendix

The complete reproducible pipeline including synthetic data generation, walk-forward cross-validation (Eqs. 11–13), XGBoost training, logistic regression baselines, and SHAP analysis (Eqs. 9–10) is provided as a single Python script. Running it with python crisis_classification.py reproduces all tables and numbers in Section 6 (synthetic validation).

Link:crisis_classification.ipynb

To run with real data, replace the synthetic data block with CRSP/FRED downloads as described in the script's comments.