



**UNIVERSITÀ
DI PAVIA**

Dipartimento di Scienze Economiche e Aziendali

Corso di Laurea Magistrale in Finance

**STABLECOIN STRESS AND CRYPTOCURRENCY
MARKETS: FORECASTING RETURNS, VOLATILITY
AND TAIL RISK WITH WHITE-BOX AND BLACK-BOX
MODELS**

Relatore:

Chiar.mo Prof. Paolo Giudici

**Tesi di Laurea
di Antonio Belardi
Matr. n. 555328**

Anno Accademico 2025–2026

Abstract

Stablecoins are the settlement layer of the cryptocurrency market: the two largest, Tether (USDT) and USD Coin (USDC), account for more than four fifths of a market worth roughly 315 billion dollars, and the overwhelming majority of crypto trades are priced and settled against them. Their one-dollar peg is the closest thing the crypto ecosystem has to a risk-free anchor — and when that anchor slips, the entire market feels it. This thesis asks whether *stablecoin stress* — measured through peg deviations, deviation volatility and abnormal trading volumes, and summarized in a daily Stablecoin Stress Index (SSI) — contains predictive information for Bitcoin returns, volatility and crash risk beyond standard market and macro-financial variables (gold, the US dollar index, the VIX).

The empirical design compares two families of models on three prediction tasks over the period July 2020 – March 2026: interpretable “white-box” models (OLS with Newey–West inference, ARIMAX, Ridge and Lasso, splines and generalized additive models, GARCH-family volatility models including a purpose-built GARCH-X with the SSI in the variance equation, logistic regression and discriminant analysis) against flexible “black-box” machine-learning models (decision trees, bagging, random forests, gradient boosting/XGBoost, support vector machines and feed-forward neural networks). All models follow the same protocol: estimation on data through January 2026, hyperparameter tuning by expanding-window time-series cross-validation, model selection on February 2026, and a single final evaluation on March 2026.

Three results stand out. First, next-day *returns* remain essentially unpredictable for every model — an honest confirmation of near-efficiency at the daily horizon. Second, *volatility* is where stablecoin stress earns its keep: the GARCH-X model augmented with the SSI significantly outperforms the standard GARCH(1,1) benchmark out of sample in both evaluation months (Diebold–Mariano $p = 0.013$ and $p = 0.035$), and gradient boosting delivers the lowest QLIKE loss in the validation month (in the calm test month a simple rolling 30-day variance is marginally better). Third, stablecoin stress variables rank among the most important predictors of *large-drawdown days*: the best classifier, a random forest, reaches an AUC of 0.71 on the validation month under

the main 10th-percentile crash definition — rising to 0.80 on the extreme 5th-percentile tail — and classifiers trained only on pre-event data produce sharply elevated real-time crash probabilities around the Terra/UST collapse (May 2022), the FTX failure (November 2022) and the USDC–SVB depeg (March 2023). The thesis concludes that the stablecoin layer operates as a barometer of crypto-market fragility: it does not tell us *where* prices go next, but it tells us a great deal about *how violently* they may move.

Sommario

Le stablecoin sono lo strato di regolamento del mercato delle criptovalute: le due maggiori, Tether (USDT) e USD Coin (USDC), rappresentano oltre quattro quinti di un mercato che vale circa 315 miliardi di dollari, e la grande maggioranza degli scambi in criptovalute è prezzata e regolata contro di esse. Il loro ancoraggio a un dollaro è quanto di più vicino a un'ancora priva di rischio esista nell'ecosistema crypto — e quando quell'ancora scivola, l'intero mercato lo avverte. Questa tesi si chiede se lo *stress delle stablecoin* — misurato attraverso le deviazioni dal peg, la loro volatilità e i volumi di scambio anomali, e sintetizzato in uno Stablecoin Stress Index (SSI) giornaliero — contenga informazione predittiva per i rendimenti, la volatilità e il rischio di crollo del Bitcoin, oltre alle variabili di mercato e macro-finanziarie standard (oro, indice del dollaro, VIX).

Il disegno empirico confronta due famiglie di modelli su tre compiti previsivi nel periodo luglio 2020 – marzo 2026: i modelli interpretabili “white-box” (OLS con inferenza Newey–West, ARIMAX, Ridge e Lasso, spline e modelli additivi generalizzati, modelli di volatilità della famiglia GARCH — incluso un GARCH-X costruito ad hoc con l'SSI nell'equazione della varianza —, regressione logistica e analisi discriminante) contro i modelli flessibili “black-box” di machine learning (alberi decisionali, bagging, random forest, gradient boosting/XGBoost, support vector machine e reti neurali feed-forward). Tutti i modelli seguono lo stesso protocollo: stima sui dati fino a gennaio 2026, scelta degli iperparametri con cross-validation a finestra espandibile per serie storiche, selezione dei modelli su febbraio 2026 e un'unica valutazione finale su marzo 2026.

Tre risultati emergono. Primo, i *rendimenti* a un giorno restano sostanzialmente imprevedibili per ogni modello — una conferma onesta della quasi-efficienza del mercato all'orizzonte giornaliero. Secondo, la *volatilità* è il terreno dove lo stress delle stablecoin si guadagna il posto: il modello GARCH-X aumentato con l'SSI batte significativamente il benchmark GARCH(1,1) fuori campione in entrambi i mesi di valutazione (Diebold–Mariano $p = 0,013$ e $p = 0,035$), e il gradient boosting ottiene la perdita QLIKE più bassa nel mese di validazione (nel mese di test, più calmo, una

semplice varianza mobile a 30 giorni risulta lievemente migliore). Terzo, le variabili di stress delle stablecoin figurano tra i predittori più importanti dei *giorni di forte ribasso*: il miglior classificatore, un random forest, raggiunge un'AUC di 0,71 nel mese di validazione con la definizione principale di crollo al decimo percentile — che sale a 0,80 sulla coda estrema del quinto percentile — e classificatori addestrati solo su dati pre-evento producono probabilità di crollo in tempo reale nettamente elevate attorno al collasso di Terra/UST (maggio 2022), al fallimento di FTX (novembre 2022) e al depeg di USDC–SVB (marzo 2023). La tesi conclude che lo strato delle stablecoin opera come un barometro della fragilità del mercato crypto: non ci dice *dove* andranno i prezzi, ma ci dice molto su *quanto violentemente* potranno muoversi.

Contents

Abstract.....	2
Sommario.....	4
Contents.....	6
List of Figures.....	6
List of Tables.....	6
List of Abbreviations.....	12
1. Introduction.....	14
1.1 Motivation.....	14
1.2 Research questions.....	15
1.3 What we find.....	16
1.4 Contributions.....	17
1.5 Outline.....	17
2. Stablecoins and crypto markets: background and literature.....	19
2.1 What stablecoins are and how the peg works.....	19
2.2 The two dominant stablecoins: USDT and USDC.....	20
2.3 A chronology of stress, 2020–2026.....	22
2.4 The regulatory turn.....	24
2.5 Related literature.....	25
3. Data and variable construction.....	27
3.1 Sources and sample.....	27
3.2 Alignment and cleaning.....	28
3.3 Targets: three definitions of “tomorrow”.....	28
3.4 Features.....	29
3.5 The Stablecoin Stress Index.....	30
3.6 Descriptive statistics.....	34
3.7 Training, validation and test windows.....	38
3.8 Beyond the first day: horizon and calendar patterns.....	38
4. Methodology.....	41
4.1 Tasks, protocol and metrics.....	41
4.2 The bias–variance trade-off.....	44
4.3 White-box models.....	45
4.4 Black-box models.....	51
4.5 Hyperparameters and implementation.....	55

5. Empirical results.....	57
5.1 Task A — predicting the next-day return.....	57
5.2 Task B — predicting the next-day variance.....	63
5.3 Task C — predicting large-drawdown days.....	68
5.4 The March 2026 test.....	72
5.5 Synthesis: white boxes, black boxes, and where each wins.....	75
6. Interpretation: opening the black boxes.....	77
6.1 What the ensembles look at.....	77
6.2 How the effects look.....	78
6.3 The stress-variable scorecard.....	79
6.4 Could the models have seen the crises coming?.....	81
6.5 The trade-off, reconsidered.....	85
6.6 Using the results: a stablecoin-stress monitor in practice.....	86
7. Conclusions.....	88
7.1 Answers to the research questions.....	88
7.2 Limitations.....	89
7.3 Future research.....	89
7.4 Policy postscript.....	90
7.5 Closing remark.....	90
References.....	91
Appendix A — The software pipeline.....	95
A.1 Reproducibility.....	95
A.2 Data and outputs.....	97
Appendix B — Complete results tables.....	98
B.1 Task A — full leaderboard, validation and test.....	98
B.2 Task B — full leaderboard, validation and test.....	98
B.3 Task C — full leaderboard, validation and test.....	99
B.4 Diebold–Mariano tests — complete results.....	99
B.5 Ljung–Box diagnostics — conditional-mean models.....	102
B.6 Ljung–Box diagnostics — volatility models.....	102
B.7 ARIMAX order selection — full criteria grid.....	105
B.8 The feature dictionary.....	105
B.9 Strongest pairwise feature correlations.....	106
B.10 OLS — complete coefficient table (Task A).....	107
B.11 Confusion matrices — complete counts.....	109

List of Figures

Figure 3.1 — Cryptocurrency prices, July 2020 – March 2026 (1 July 2020 = 100, log scale). Shad.....	28
Figure 3.2 — USDT and USDC prices with the three major stress episodes shaded. Daily closes; in.....	31
Figure 3.3 — The Stablecoin Stress Index (red, left scale) and the Bitcoin price (grey, right s.....	32
Figure 3.4 — Fat tails: the distribution of BTC daily returns against a normal with the same me.....	35
Figure 3.5 — BTC 30-day annualized volatility (blue, left) and the SSI (red, right).....	36
Figure 3.6 — The motivating plot: SSI at day t versus $ \text{BTC return} $ at $t+1$, with decile means. T.....	36
Figure 3.7 — Correlation matrix of the model features and the next-day BTC return. The stress b.....	37
Figure 5.1 — GAM partial effects on the next-day return (training window). Effects are small —	59
Figure 5.2 — Lasso coefficient paths, Task A. As the penalty relaxes (right to left), the first.....	59
Figure 5.3 — The champion regression tree for Task A (displayed to depth 3). A deliberately cru.....	60
Figure 5.4 — Task A out of sample: actual daily returns (bars) against model forecasts (lines),.....	62
Figure 5.5 — Task A forecast errors over time. The errors inherit essentially the full variabil.....	63
Figure 5.6 — GARCH(1,1) conditional volatility against absolute daily returns, full sample.....	67
Figure 5.7 — Out-of-sample one-step-ahead volatility forecasts, February–March 2026.....	68
Figure 5.8 — Confusion matrices on the validation month (28 days, 5 crash days) at each model's.....	70
Figure 5.9 — ROC curves on the validation month. The random forest dominates; the diagonal is c.....	72
Figure 5.10 — The March 2026 test: actual daily BTC returns (bars), the champion return forecas.....	75
Figure 6.1 — Permutation importance on the validation month, averaged over random forest and XG.....	77
Figure 6.2 — Partial dependence of the predicted next-day return (XGBoost). Effects are economi.....	78

Figure 6.3 — Partial dependence of the predicted crash probability (XGBoost). The USDT absolute.....	78
Figure 6.4 — Real-time crash probabilities (dark red, right scale) from classifiers trained onl.....	82
Figure 6.5 — Two lenses on the same question: OLS t-statistics (left, white box) against perm.....	85

List of Tables

Table 2.1 — Tether and USD Coin at a glance.....	21
Table 3.1 — The nine daily series and their roles.....	27
Table 3.2 — The feature set (33 candidates).....	29
Table 3.3 — Peg-deviation statistics, daily closes, July 2020 – March 2026.....	31
Table 3.4 — The fifteen highest readings of the Stablecoin Stress Index.....	33
Table 3.5 — Daily log returns, descriptive statistics (crypto on the 7-day calendar, macro on t.....	34
Table 3.6 — Sample splits, with crash-day counts under the main (10th-percentile) and robustnes.....	38
Table 3.7 — Horizon analysis of the SSI: correlation with future absolute BTC returns and avera.....	39
Table 3.8 — Tail risk (main 10th-percentile definition) and volatility by day of the week of th.....	39
Table 4.1 — Hyperparameter grids; all tuning by 5-fold expanding-window cross-validation on the.....	55
Table 5.1 — Task A, validation month (February 2026). Returns in percent; benchmarks in italics.....	57
Table 5.2 — ARIMAX(1,0,0) estimates, training window (N = 1,920; log-likelihood -4,868.8). Stan.....	61
Table 5.3 — Estimated variance-equation parameters, training window. The first three models are.....	63
Table 5.4 — GARCH-X full inferential output. Standard errors from the numerical observed-inform.....	64
Table 5.5 — Ljung–Box tests on standardized () and squared standardized () residuals, training	65
Table 5.6 — Task B, validation month (February 2026). QLIKE is the primary criterion; DM tests	66
Table 5.7 — Task C, validation month (February 2026): 5 crash days among 28 under the main 10th.....	68
Table 5.8 — Precision, recall and F1 on the validation month at the conventional 0.5 threshold	70
Table 5.8b — Robustness check on the 5th-percentile target, validation month (4 crash days amon.....	71
Table 5.9 — Champions selected on validation, evaluated once on test. “n.s.” = not significantl.....	73

Table 6.1 — Every stress variable under every lens. OLS: next-day return (%; HAC p-values); log.....	79
Table 6.2 — Terra/UST collapse, day by day (selected columns of the modelling dataset).....	82
Table 6.3 — FTX collapse, day by day. Note the sequence: on 8 November the volume anomaly spike.....	83
Table 6.4 — USDC–SVB depeg, day by day. The SSI’s sample maximum (9.32) coincides with USDC’s 2.....	83
Table A.1 — The pipeline. Model predictions are stored per model and split, so all comparisons	95
Table B.1 — Task A (next-day BTC return, %): all models, validation (Feb 2026) and test (Mar 20.....	98
Table B.2 — Task B (next-day realized-variance proxy, % ²): all models, both out-of-sample month.....	98
Table B.3 — Task C (large-drawdown days, main 10th-percentile target): all models, both out-of-.....	99
Table B.4 — All Diebold–Mariano tests. Task A: squared-error loss vs the zero forecast; Task B:.....	99
Table B.5 — Ljung–Box tests on conditional-mean residuals (training window). H0: no autocorrela.....	102
Table B.6 — Ljung–Box tests on standardized and squared standardized residuals of the volatilit.....	102
Table B.7 — Information criteria for every (p, q) candidate. AIC prefers (2,1); BIC and HQC pre.....	105
Table B.8 — The 32 model features: construction and interpretation. All quantities computed wit.....	105
Table B.9 — The ten largest absolute pairwise correlations among continuous features (training-.....	106
Table B.10 — Full OLS estimation output for the next-day BTC return (percent): all 32 features	107
Table B.11 — Confusion matrices on the validation month (28 days, 5 crash days, main target) at.....	109

List of Abbreviations

Abbreviation	Meaning
ADF	Augmented Dickey–Fuller (unit-root test)
AIC / BIC / HQC	Akaike / Bayesian / Hannan–Quinn information criterion
AR, ARMA, ARIMAX	Autoregressive (moving-average) model with exogenous regressors
AUC	Area under the ROC curve
BTC, ETH, SOL, XRP	Bitcoin, Ethereum, Solana, XRP
CV	Cross-validation
DM	Diebold–Mariano test
DXY	US Dollar Index
EGARCH / GJR	Exponential / Glosten–Jagannathan–Runkle GARCH
GAM	Generalized additive model
GARCH(-X)	Generalized autoregressive conditional heteroskedasticity (with exogenous variance regressor)
GLD	SPDR Gold Shares ETF
HAC	Heteroskedasticity- and autocorrelation-consistent (standard errors)
LB	Ljung–Box test
LDA	Linear discriminant analysis
MAE / MSE / RMSE	Mean absolute / mean squared / root mean squared error
MiCA	Markets in Crypto-Assets Regulation (EU)
MLP	Multilayer perceptron (feed-forward

	neural network)
OLS	Ordinary least squares
QLIKE	Quasi-likelihood volatility loss function
ROC	Receiver operating characteristic
RV	Realized variance
SSI	Stablecoin Stress Index
SVB	Silicon Valley Bank
SVM / SVR / SVC	Support vector machine / regression / classification
USDT / USDC	Tether / USD Coin
VIF	Variance inflation factor
VIX	CBOE Volatility Index
XGBoost	Extreme gradient boosting

1. Introduction

1.1 Motivation

Every market has an asset that participants treat as “cash”. In the cryptocurrency market that role is not played by the dollar itself — which most exchanges cannot hold or settle directly — but by *stablecoins*: blockchain tokens engineered to trade at exactly one dollar. The two largest, Tether (USDT) and USD Coin (USDC), together account for more than four fifths of a stablecoin market whose total capitalization has grown to roughly 315 billion dollars by mid-2026, nearly doubling in two years. The vast majority of cryptocurrency trading pairs are quoted against stablecoins rather than against fiat currency; stablecoins are the collateral posted in derivatives markets, the settlement asset of decentralized finance, and the vehicle through which capital enters and exits the crypto ecosystem. They are, in short, the plumbing of the entire market.

Plumbing is invisible until it breaks. A stablecoin whose price is exactly one dollar carries no information; a stablecoin trading at 99.5 or 100.5 cents reveals that someone, somewhere, is paying a premium to get in or out — that redemption pressure, collateral doubts or a flight to (or from) crypto liquidity is under way. The short history of the market offers dramatic examples. In May 2022 the algorithmic stablecoin TerraUSD collapsed from one dollar to essentially zero within a week, erasing tens of billions of dollars of value and dragging the entire crypto market down with it; USDT itself briefly traded below 99 cents amid record redemptions. In November 2022 the failure of the FTX exchange triggered another system-wide deleveraging. In March 2023 USDC — widely considered the most transparent of the major stablecoins — fell to around 88 cents intraday when Silicon Valley Bank, where part of its reserves were deposited, was closed by regulators; our daily data record a closing price of 0.9715 dollars, a deviation of 285 basis points from the peg. And in October 2025, a tariff announcement triggered the largest liquidation cascade in crypto history — roughly 19 billion dollars of leveraged positions wiped out in a day — during which the synthetic stablecoin USDe traded as low as 65 cents on the largest exchange while the majors wobbled on secondary venues.

Each of these episodes shares a common signature: stress in the stablecoin layer and violent moves in crypto prices arrived together. What is less obvious, and what this thesis investigates, is whether the stablecoin layer contains *predictive* information — whether measurable stress in the pegs of USDT and USDC today tells us something about Bitcoin’s return, its volatility, or its probability of a crash tomorrow, over and above what standard market and macro-financial variables already tell us.

1.2 Research questions

The thesis is organized around one economic question and one methodological question.

The economic question is: **does stablecoin stress predict cryptocurrency market behaviour?** To make “stablecoin stress” measurable, we construct a daily Stablecoin Stress Index (SSI) that aggregates six observable symptoms of strain — the absolute peg deviations of USDT and USDC, the rolling volatility of those deviations, and abnormal trading volumes in the two coins — into a single standardized series. The index, built with strictly backward-looking information, spikes precisely at the Terra/UST collapse, the FTX failure and the USDC–SVB depeg, which we take as *prima facie* evidence that it captures the phenomenon of interest. “Cryptocurrency market behaviour” is then given three concrete meanings, in increasing order of predictability: the *direction and size* of the next day’s Bitcoin return (Task A), the *magnitude* of the next day’s move — its realized variance (Task B) — and the occurrence of a *large drawdown*, defined as a daily return falling below its rolling tenth percentile (Task C), with the more extreme fifth percentile retained throughout as a robustness check.

The methodological question is: **how much do flexible machine-learning models add over classical interpretable models on these tasks, and at what cost?** Following the taxonomy of the statistical-learning literature (James, Witten, Hastie and Tibshirani, 2023), we contrast “white-box” models whose parameters carry direct economic interpretation — ordinary least squares with heteroskedasticity- and autocorrelation-consistent inference, ARIMAX, Ridge and Lasso regressions, spline-based generalized additive models, logistic regression and linear discriminant analysis, and the GARCH

family of volatility models including a purpose-built GARCH-X specification with the SSI in the variance equation — against “black-box” models that trade interpretability for flexibility: decision trees, bagging, random forests, gradient boosting (XGBoost), support vector machines and feed-forward neural networks. The comparison is disciplined by a strict out-of-sample protocol: all models are estimated on data from July 2020 through January 2026, hyperparameters are tuned by expanding-window time-series cross-validation within that sample, model selection uses February 2026 as a validation month, and the champions face a single, final test on March 2026 — a month never touched during development.

1.3 What we find

Three results structure the thesis. First, next-day returns are essentially unpredictable: no model, white or black, significantly beats the naïve zero-return forecast in Diebold–Mariano tests on either evaluation month. Out-of-sample R^2 values hover within a few percentage points of zero, in line with the asset-pricing evidence that daily return predictability, where it exists at all, is faint (Welch and Goyal, 2008; Gu, Kelly and Xiu, 2020). We regard this as a feature of the study rather than a failure: a design that could not detect unpredictability would not be credible when it detects predictability elsewhere.

Second, volatility is where stablecoin stress earns its keep. The GARCH-X model that admits the lagged SSI into the conditional-variance equation significantly outperforms the standard GARCH(1,1) benchmark out of sample in *both* evaluation months (Diebold–Mariano p-values of 0.013 and 0.035 under the QLIKE loss), and the machine-learning models — gradient boosting in particular — deliver the lowest variance-forecast losses of the validation month, beating the entire GARCH family; in the calm test month a simple rolling 30-day variance is marginally better still, a reminder that flexibility pays where there is turbulence to track. The message is consistent across model families: the magnitude of tomorrow’s Bitcoin move is forecastable, and stablecoin stress is part of the forecast.

Third, stablecoin stress variables rank among the strongest predictors of crash days. The best classifier, a random forest, attains an AUC of 0.71 on the stressed validation month

under the main 10th-percentile crash definition, rising to 0.80 when only the extreme 5th-percentile tail counts; permutation importance places the churn of the USDT peg — the 7-day volatility of its deviation — first among thirty-two candidate predictors, with the SSI sixth and abnormal USDC volume seventh. Most strikingly, classifiers re-trained using only information available *before* each historical stress episode produce sharply elevated real-time crash probabilities around Terra (May 2022), FTX (November 2022) and the USDC–SVB depeg (March 2023) — the model “sees” each of these three crises building in the stablecoin data before the worst of the price damage materializes — evidence that is striking but, with three events, necessarily anecdotal in sample size.

1.4 Contributions

The thesis makes three contributions. First, it proposes and validates a simple, fully replicable daily index of stablecoin stress built from public market data alone; the index requires no proprietary flow data and can be computed in real time. Second, it provides a systematic three-task comparison of ten model families under a common, leakage-free time-series protocol, bridging the econometric tradition (ARIMAX, GARCH) and the statistical-learning tradition (regularization, trees, ensembles, kernels, neural networks) that are usually evaluated in isolation. Third, it contributes a novel GARCH-X application: embedding a composite stablecoin-stress regressor in the conditional-variance equation of Bitcoin returns, estimated by direct maximum likelihood, and testing its out-of-sample value against both classical and machine-learning alternatives.

1.5 Outline

The remainder of the thesis proceeds as follows. Chapter 2 describes the institutional mechanics of stablecoins, reconstructs the major stress episodes of 2020–2026, and reviews the related literature. Chapter 3 documents the data, the construction of every variable including the SSI, and the exploratory evidence that motivates the modelling choices. Chapter 4 presents the methodology: the three prediction tasks, each model family with its essential mathematics, the bias–variance considerations that guide regularization, and the evaluation protocol. Chapter 5 reports the empirical results task by task, including the final March 2026 forecast. Chapter 6 opens the black boxes —

variable importance, partial-dependence analysis, and event backtests — and confronts what white-box and black-box models jointly say about the economic question. Chapter 7 concludes with limitations and directions for future research. Two appendices close the thesis: Appendix A documents the software pipeline, and Appendix B collects the complete results tables; every number, table and figure in the thesis is generated by the accompanying code and is exactly reproducible.

2. Stablecoins and crypto markets: background and literature

2.1 What stablecoins are and how the peg works

A stablecoin is a blockchain token whose issuer promises that one unit is worth, and can be exchanged for, one unit of a reference currency — in practice almost always the US dollar. The promise can be engineered in three broad ways. *Fiat-backed* stablecoins (USDT, USDC) hold a reserve of dollar assets — bank deposits, Treasury bills, repurchase agreements — and stand ready to mint new tokens against incoming dollars and to redeem tokens for dollars at par. *Crypto-collateralized* stablecoins (DAI) issue tokens against an over-collateralized pool of cryptoassets managed by a smart contract. *Algorithmic* and *synthetic* designs promise the peg without full backing: TerraUSD maintained it through an arbitrage relationship with a sister token, and more recent synthetic dollars such as USDe replicate it with a hedged derivatives position. The history examined in this thesis is, to a large extent, the history of the market learning which of these designs survive stress.

The peg of a fiat-backed stablecoin is maintained by a two-tier mechanism. In the *primary market*, authorized counterparties can create and redeem tokens with the issuer at exactly one dollar; in the *secondary market*, everyone else trades the token at whatever price clears. Whenever the secondary-market price moves away from one dollar, an arbitrage opens: at 99.5 cents an arbitrageur can buy tokens on the exchange and redeem them at par; at 100.5 cents she can mint at par and sell. If the primary market is frictionless, deviations should be closed within minutes. In reality access to the primary market is restricted — Tether, for instance, historically imposed minimum redemption sizes and fees, while Circle offered broader access — so deviations can persist, and their size is informative about the tightness of the arbitrage and the balance of order flow. Lyons and Viswanath-Natraj (2023) document precisely this: peg deviations behave like the currency-board discounts and premia of traditional monetary economics, widening when redemption is doubtful or costly and narrowing when arbitrage access improves.

This institutional detail is the foundation of the thesis: *because* the peg is maintained by arbitrage against a redemption promise, the observed deviation from one dollar is a market-priced signal of stress — a daily, publicly observable temperature reading of the crypto financial system. A premium signals a scramble into the safety or liquidity of the stablecoin (or demand to move dollars onto crypto rails); a discount signals doubts about the issuer or a rush to exit. Volume carries complementary information: redemptions and flight episodes leave footprints in trading activity even when the peg itself bends only slightly. Gorton and Zhang (2023) frame the same point in monetary-history terms: privately issued money that is not information-insensitive will trade at fluctuating discounts exactly when confidence wavers, as nineteenth-century wildcat banknotes did.

The run mechanics deserve one more turn of the screw, because they determine *which* observable moves first. A stablecoin run is a coordination problem among holders who cannot all redeem at par simultaneously: reserves are ample for ordinary redemptions but not for everyone at once, and the primary market processes redemptions with delays, minimums and fees. A holder who suspects trouble therefore faces a choice between the slow, safe exit (redeem with the issuer) and the fast, risky one (sell on the secondary market at whatever price stands). In the earliest phase of a run the informed money crowds into *both* channels at once — redemption queues lengthen and exchange turnover spikes — while the price still sits within a few basis points of par, because arbitrageurs are still willing to warehouse the risk. Only when their capacity or their nerve is exhausted does the deviation blow out. The observable sequence is thus volume first, price second: a prediction of the theory that Chapter 6 confirms day by day in the FTX episode, and the reason the SSI blends turnover anomalies with price deviations rather than relying on either alone.

2.2 The two dominant stablecoins: USDT and USDC

The empirical work focuses on Tether and USD Coin, which jointly represent over 80% of stablecoin capitalization throughout the sample and constitute the quoting and settlement layer of the exchanges from which our price data come.

Table 2.1 — Tether and USD Coin at a glance.

	USDT (Tether)	USDC (USD Coin)
Issuer	Tether Ltd.	Circle
Launched	2014	2018
Market cap, mid-2026	≈ 186 billion USD (≈ 59% share)	≈ 75 billion USD (≈ 24% share)
Reserves	Mainly short-term US Treasuries; quarterly attestations	Cash and short-term US Treasuries in a regulated fund; monthly attestations
Primary-market access	Restricted (verified customers, minimum sizes, fees)	Broader institutional access
Main historical controversy	Reserve disclosures (2021 settlements with NYAG and CFTC)	Reserve exposure to Silicon Valley Bank (March 2023)

The two coins embody different points on a transparency–reach frontier. Tether grew up alongside the offshore exchange ecosystem, has faced repeated questions about its reserves — settling with the New York Attorney General and the CFTC in 2021 over historical misstatements — and yet has proven remarkably resilient, with its market share *rising* after each industry crisis. USDC pursued the regulated-institution route, with monthly reserve attestations and US banking relationships; ironically, it was precisely a US bank — Silicon Valley Bank, holding 3.3 billion dollars of Circle’s reserves — whose failure produced USDC’s deepest depeg. The pair therefore gives us two imperfectly correlated stress sensors: USDT stress tends to reflect crypto-native and offshore funding conditions, while USDC stress in our sample is dominated by the March 2023 traditional-banking shock. This complementarity is one reason the Stablecoin Stress Index of Chapter 3 aggregates symptoms from both.

A word on market structure completes the picture. Stablecoins are not only traded assets but *units of account*: on most venues the liquid order books are BTC/USDT and ETH/USDT, not BTC/USD, so the effective dollar price of Bitcoin embeds the price of

the stablecoin leg. Stress in the numeraire therefore transmits to measured crypto prices mechanically, before any portfolio decision is taken. Stablecoins also dominate the collateral side of the perpetual-futures market, where the bulk of crypto leverage lives: a trader posting USDT margin against a long BTC perpetual is short the stablecoin's credit risk whether he knows it or not, and a depeg tightens his effective margin at exactly the moment prices fall — one channel through which stablecoin stress and liquidation cascades reinforce each other, visible at scale in the October 2025 episode.

2.3 A chronology of stress, 2020–2026

Six episodes punctuate the sample period and anchor the empirical analysis. They are visible — with different intensities — in the peg-deviation and volume data of Chapter 3, and three of them are used as event studies in Chapter 6.

April 2021 — the first leverage flush. After Bitcoin's run to what was then an all-time high near 65,000 dollars, the weekend of 17–18 April 2021 brought a sudden 15% drawdown amplified by derivatives liquidations. Both USDT and USDC briefly traded more than 100 basis points from par — the largest simultaneous deviation of the pre-Terra sample — and the episode ranks third in our Stablecoin Stress Index. It is a reminder that stablecoin stress accompanies not only issuer crises but also generic deleveraging.

May 2022 — Terra/UST. TerraUSD, an algorithmic stablecoin that had grown to roughly 18 billion dollars on the strength of a 20% deposit yield offered by the Anchor protocol, lost its peg during the week of 7–13 May 2022 and spiraled to zero as its stabilizing mechanism minted unlimited amounts of the collapsing sister token LUNA. The collapse vaporized several tens of billions of dollars, forced the liquidation of leveraged funds and lenders exposed to it, and pushed Bitcoin from about 36,000 to under 27,000 dollars in a week. Crucially for this thesis, the stress spilled into the *major* stablecoins: USDT traded as low as 95 cents on some venues on 12 May and processed record redemptions in the following weeks, while USDC traded at a premium as capital fled toward perceived quality. Our index registers the week as the second-strongest stress episode of the sample.

November 2022 — FTX. The failure of the FTX exchange — from the leak of its affiliated trading firm’s balance sheet in early November, through the halt of withdrawals on 8 November, to the bankruptcy filing on 11 November — was primarily an *exchange* crisis rather than a stablecoin crisis: pegs bent but did not break, with USDT dipping modestly below par amid redemption queues. The episode matters for our design precisely because of that contrast: it shows that not every crypto crash originates in, or fully transmits through, the stablecoin layer, and it gives the classifiers of Chapter 6 a harder test — predicting a crash whose stablecoin signature was comparatively faint.

March 2023 — USDC and Silicon Valley Bank. When Silicon Valley Bank was closed by regulators on Friday 10 March 2023, Circle disclosed that 3.3 billion dollars of USDC reserves were trapped in the bank. Over the weekend — with the primary redemption market closed — USDC traded down to roughly 88 cents intraday, and our daily closing data record 0.9715 dollars (a 285-basis-point discount), while USDT flipped to a *premium* above 1.01 as holders rotated between the two coins. The peg snapped back when US authorities guaranteed SVB deposits on Sunday evening. This is the strongest stablecoin-specific stress event in our sample — the SSI’s maximum — and a textbook illustration that stablecoin risk is inseparable from traditional banking risk.

October 2025 — the 10/10 liquidation cascade. On 10 October 2025 an announcement of 100% tariffs on Chinese imports triggered the largest liquidation event in crypto history: roughly 19 billion dollars of leveraged positions were forcibly closed within a day, Bitcoin fell about 14% intraday, several trading venues suffered outages, and the synthetic dollar USDe — outside our two-coin universe but emblematic of the newer synthetic designs — traded as low as 65 cents on the largest exchange. USDT deviated visibly from par in our daily data. The episode enters our sample as the largest post-2023 spike in the SSI and, in Chapter 3, as evidence that the index remains informative in the more regulated, ETF-era market structure.

February 2026 — the macro selloff. The validation month of our forecasting protocol was itself turbulent: the first weekend of February 2026 brought one of the ten largest liquidation events on record, and 5 February saw Bitcoin’s largest one-day fall since

November 2022, as tariff-driven inflation expectations removed hopes of rate cuts and spot-ETF flows turned negative; the month closed with a geopolitical shock. Notably, contemporaneous commentary emphasized that — unlike 2022 — *no major stablecoin depegged and no exchange failed*: the stress was macro-financial, not crypto-native. This composition matters for interpreting our results: February 2026 rewards models that read general market fragility, not only stablecoin-specific strain, and we return to this point in Chapters 5 and 6.

2.4 The regulatory turn

The sample spans the transition of stablecoins from regulatory grey zone to regulated instrument. In the European Union, the Markets in Crypto-Assets Regulation (MiCA) applied its stablecoin provisions from 30 June 2024, imposing authorization, reserve and redemption requirements on issuers of “e-money tokens” and pushing exchanges to delist non-compliant coins for EU customers. In the United States, the GENIUS Act — signed into law on 18 July 2025 — established the first federal framework for payment stablecoins: 100% reserve backing in cash and short-term Treasuries, monthly public reserve disclosures, federal or state licensing of issuers, and full application of anti-money-laundering law. The market read the framework as legitimization: stablecoin capitalization accelerated from roughly 160 billion dollars in mid-2024 to over 300 billion by early 2026. For our purposes the regulatory turn has two implications. First, it plausibly *dampens* peg-deviation volatility late in the sample — a regime change the expanding-window design of Chapter 4 is built to absorb. Second, it raises the systemic stakes: as stablecoins integrate with the Treasury market and the payment system, understanding what their stress signals predict is no longer a niche crypto question but a financial-stability one (Arner, Auer and Frost, 2020). A third implication cuts the other way and is worth stating honestly: regulation reshapes the *composition* of the stress signal rather than abolishing it. Reserve rules and disclosure make classic issuer-solvency runs on the majors less likely, but they do nothing about flight-to-quality rotations between coins (the SVB weekend pattern), about liquidity failures on individual venues, or about the newer synthetic designs that sit outside the licensed perimeter — and the sample’s late episodes are exactly of these kinds. If anything, the

migration of stress from the regulated core to the unregulated periphery strengthens the case for measuring it continuously rather than assuming it away.

2.5 Related literature

The thesis draws on, and contributes to, four strands of research.

Stablecoins and crypto-market dynamics. The seminal — and controversial — paper is Griffin and Shams (2020), who argue that unbacked Tether issuance was used to support Bitcoin prices during the 2017 boom; subsequent work with longer samples (Kristoufek, 2021; Ante, Fiedler and Strehle, 2021) finds that stablecoin issuances are followed by measurable but economically modest reactions in major cryptoassets, more consistent with demand-driven inflows than manipulation. Lyons and Viswanath-Natraj (2023) shift attention from issuance to *peg deviations*, showing they behave as economically meaningful prices of stress; Baur and Hoang (2021) document the mirror image, stablecoins acting as intra-crypto safe havens during Bitcoin drawdowns; Gorton and Zhang (2023) supply the monetary-theory frame. Our contribution to this strand is to move from contemporaneous description to out-of-sample *prediction*, and to compress the several stress symptoms this literature identifies into a single replicable index.

Cryptocurrency return and volatility modelling. Liu and Tsyvinski (2021) establish that cryptocurrency returns are driven by crypto-specific factors — momentum and attention — rather than by exposure to traditional asset classes; Makarov and Schoar (2020) document the market’s segmented arbitrage structure. On volatility, a large literature starting with Katsiampa (2017) fits GARCH-family models to Bitcoin, generally finding strong persistence and fat tails; Conrad, Custovic and Ghysels (2018) show that long-run Bitcoin volatility responds to macro-financial conditions — a finding our GARCH-X design extends by asking whether *stablecoin* conditions play an analogous role at the daily frequency. Borri (2019) documents conditional tail risk in crypto markets, motivating our third prediction task.

Machine learning in empirical asset pricing. Gu, Kelly and Xiu (2020) provide the canonical evidence that flexible methods — trees and neural networks — improve return forecasts relative to linear benchmarks, while warning that predictable variation

is a small fraction of total variance; Welch and Goyal (2008) and Campbell and Thompson (2008) supply the out-of-sample discipline (the R^2 relative to a naïve benchmark) that we adopt. Applications to crypto at the daily horizon (e.g., Jaquart, Dann and Weinhardt, 2021) typically find directional accuracy barely above chance, consistent with our Task A results.

Statistical-learning methodology. The modelling vocabulary of the thesis — the bias–variance trade-off, regularization, splines and additive models, trees and ensembles, support vector machines, feed-forward networks, and the discipline of validation — follows James, Witten, Hastie and Tibshirani (2023). The econometric components rest on Engle (1982) and Bollerslev (1986) for (G)ARCH, Nelson (1991) and Glosten, Jagannathan and Runkle (1993) for asymmetric extensions, Newey and West (1987) for HAC inference, Diebold and Mariano (1995) — with the small-sample correction of Harvey, Leybourne and Newbold (1997) — for forecast-accuracy testing, and Patton (2011) for the choice of the QLIKE loss when volatility is measured with a noisy proxy.

Relative to all four strands, the specific combination offered here — a composite stablecoin-stress index, three prediction targets of increasing predictability, a ten-family white-box/black-box horse race under a single leakage-free protocol, and an event-study validation of the resulting crash-probability model — is, to the best of our knowledge, new.

3. Data and variable construction

3.1 Sources and sample

All series are daily and come from Yahoo Finance, retrieved programmatically with the `yfinance` interface; the download, like every subsequent step, is scripted and reproducible (Appendix A). The sample runs from 1 July 2020 to 31 March 2026 — 2,100 calendar days chosen to start after the March 2020 pandemic dislocation, span a full boom–bust–boom cycle, and end with the two out-of-sample months of the forecasting protocol. Nine series are used, in three blocks.

Table 3.1 — The nine daily series and their roles.

Block	Series (ticker)	Role in the analysis
Crypto majors	Bitcoin (BTC-USD)	Target asset of all three prediction tasks
	Ethereum (ETH-USD), Solana (SOL-USD), XRP (XRP-USD)	Cross-asset momentum features
	Tether (USDT-USD), USD Coin (USDC-USD)	Stress measurement: prices and volumes
Stablecoins		
Macro-financial	SPDR Gold Shares (GLD)	Safe-haven flows
	US Dollar Index (DX-Y.NYB)	Dollar liquidity and strength
	CBOE VIX (^VIX)	Global risk aversion

Figure 3.1 plots the four cryptocurrencies over the sample, normalized to 100 at the start and drawn on a logarithmic scale; the shaded bands mark the three stress episodes studied in depth later (Terra, FTX, USDC–SVB). The sample contains two complete drawdowns of more than 60% in Bitcoin — 2022 and the 2025–26 correction — and even larger swings in the other coins, a first indication of the volatility scale the models must contend with.

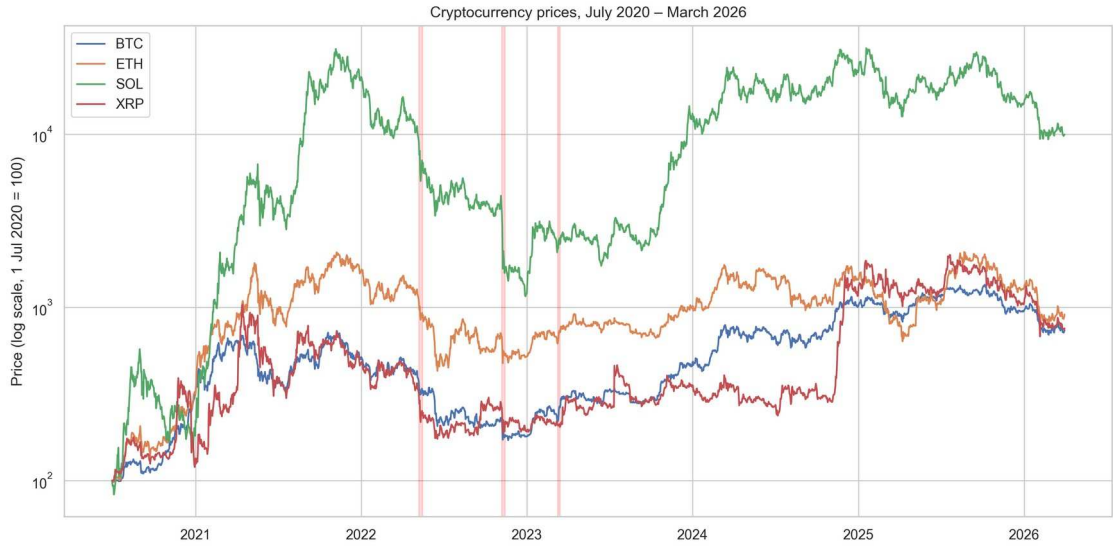


Figure 3.1 — Cryptocurrency prices, July 2020 – March 2026 (1 July 2020 = 100, log scale). Shaded bands: Terra/UST collapse, FTX failure, USDC–SVB depeg.

3.2 Alignment and cleaning

The two halves of the dataset live on different calendars: cryptocurrencies trade every day of the year, while GLD, the dollar index and the VIX exist only on exchange business days — about 69% of calendar days (roughly 655 of the 2,100 days of each macro series are non-trading days). We align everything on the full daily calendar and *forward-fill* the macro series: on a Saturday, the value of the VIX known to a crypto trader is Friday’s close, which is exactly what the forward fill encodes. No information from the future ever enters a row. The alternative — restricting the panel to business days — would discard weekends, which host some of crypto’s most dramatic moves (the February 2026 selloff began on a Sunday); we keep the crypto calendar and accept that macro variables are locally constant across weekends, with their weekend returns equal to zero. Cryptocurrency series have no missing closes in the sample.

3.3 Targets: three definitions of “tomorrow”

All returns are daily log returns, $r_t = \ln(P_t / P_{t-1})$, expressed in percent. Each observation of the modelling dataset is indexed by the *information date* t : features use only data available at the close of day t , and targets refer to day $t + 1$.

Task A — return. The continuous target is the next-day Bitcoin log return, $y_t^A = r_{t+1}$.

Task B — variance. The volatility target is the realized-variance proxy $y_t^B = r_{t+1}^2$. Squared daily returns are an unbiased but noisy proxy of the true conditional variance; Chapter 4 discusses why the QLIKE loss is the appropriate metric under such a proxy (Patton, 2011).

Task C — large drawdown. The binary target flags days whose return falls below a rolling lower tail: $y_t^C = 1\{r_{t+1} < Q_{0.10}(r_{t-249}, \dots, r_t)\}$, where $Q_{0.10}$ is the tenth percentile of the past 250 daily returns (minimum 120 observations). The threshold is computed strictly with information available at t , so the definition of a “crash day” is itself real-time: it adapts to the prevailing volatility regime, flagging roughly the worst 10% of days by construction. The tenth percentile is adopted as the *main* definition on statistical grounds: it guarantees enough positive days in a one-month evaluation window for confusion matrices, precision, recall and F1 to be informative, which a rarer event class cannot. In the training sample the base rate is 10.7% (205 of 1,920 days); in the February 2026 validation month it rises to 17.9% (5 of 28 days) — a genuinely stressed month — and in the March 2026 test month it is 16.1% (5 of 31 days). The more extreme *fifth*-percentile definition (base rate 5.4% in training, 4 validation and only 1 test positive) is retained throughout as a robustness check for the worst crashes: Section 5.3 reports the corresponding results.

3.4 Features

Thirty-three candidate predictors are constructed in four blocks; all are measurable at the close of day t .

Table 3.2 — The feature set (33 candidates).

Block	Variables	Rationale
Stablecoin stress (9)	Signed and absolute peg deviations of USDT and USDC (basis points); 7-day rolling std of each deviation; 30-day volume z-scores of each coin; the composite SSI (Section	The thesis variables of interest

3.5)

Crypto market (13)	BTC lagged returns (lags 1, 2, 3, 5, 7); BTC rolling volatility and mean return over 7 and 30 days; lag-1 returns of ETH, SOL, XRP; BTC volume z-score	Autoregressive structure, volatility clustering, cross-asset momentum
Macro-financial (4)	Lag-1 GLD return; lag-1 DXY return; VIX level; lag-1 VIX change	Safe-haven flows, dollar liquidity, global risk appetite
Calendar (7)	Day-of-week dummies for the target date (Monday as reference); depeg dummy = 1 if either coin deviates more than 50 bp	Crypto weekend effects; discrete stress regimes

Collinearity is screened with variance inflation factors on the continuous features. Only one variable exceeds the conventional threshold of 10 — the USDT volume z-score (VIF = 10.0), which is nearly collinear with the SSI that already absorbs it — and is dropped, with the SSI itself protected from removal as the designated variable of interest. The final feature set therefore contains 32 predictors. The full correlation structure is mapped in Section 3.6 (Figure 3.7), with the ten strongest pairs tabulated in Appendix B; the tightest links, unsurprisingly, sit inside the stress block and among the market’s own momentum and volatility windows.

3.5 The Stablecoin Stress Index

The six stress symptoms of the two coins — absolute deviations, deviation volatilities, volume anomalies — move together during crises but are measured in different units. The Stablecoin Stress Index compresses them into one series:

$$SSI_t = \frac{1}{6} \sum_{j=1}^6 \frac{x_{j,t} - \mu_{j,t}}{\sigma_{j,t}},$$

where $x_{j,t}$ are the six components and $\mu_{j,t}, \sigma_{j,t}$ are their *expanding-window* means and standard deviations computed with data up to t only (minimum 60 observations). The expanding standardization is the crucial design choice: a full-sample z-score would leak information about the future scale of stress into the past, inflating the apparent early-warning power of the index. Built this way, the SSI is a genuine real-time statistic: the value an analyst could have computed on each historical evening.

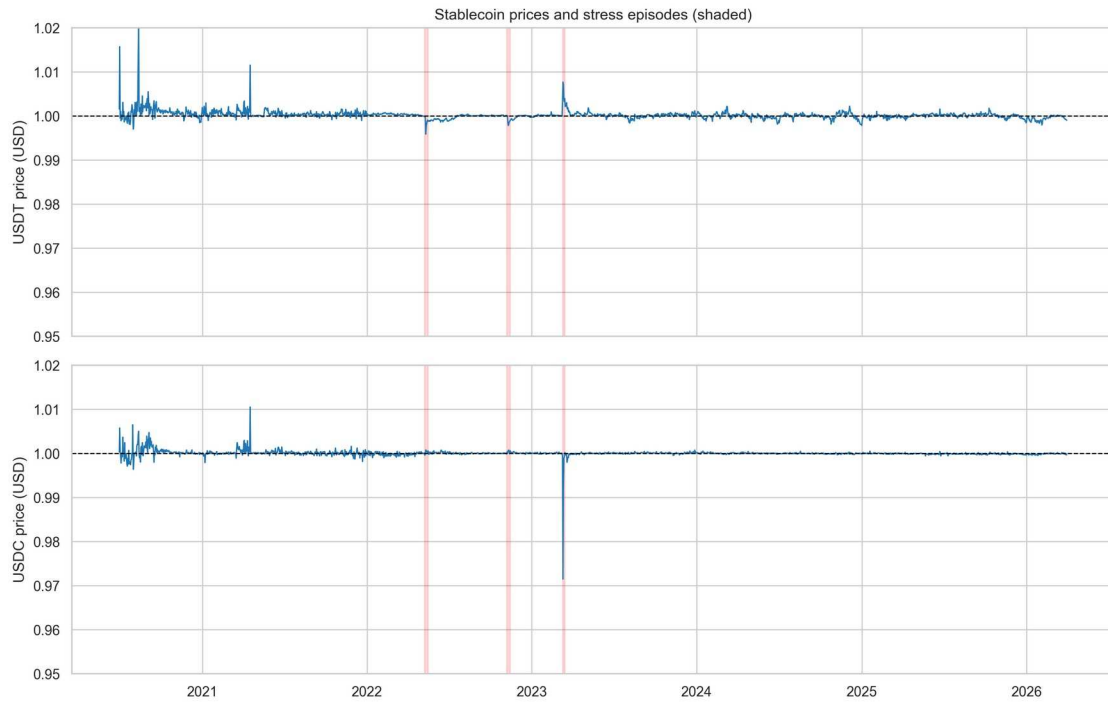


Figure 3.2 — USDT and USDC prices with the three major stress episodes shaded. Daily closes; intraday extremes were substantially wider (USDC touched ≈ 0.88 intraday on 11 March 2023).

Figure 3.2 shows the raw material: both coins hug the peg to within a few basis points in normal times — the *median* absolute deviation is 3.3 bp for USDT and 1.1 bp for USDC — and break from it violently but rarely. The tail statistics tell the story compactly.

Table 3.3 — Peg-deviation statistics, daily closes, July 2020 – March 2026.

	Mean		95th	99th		Days		
	dev	Media	pct	pct	Max	> 50	Min	Max
	(bp)	n (bp)	(bp)	(bp)	(bp)	bp	close	close
USDT	5.33	3.26	15.5	30.9	202.1	0.33%	0.995	1.020

							9	2
USDC	2.55	1.13	8.2	25.8	285.0	0.29%	0.971	1.010
							5	5

Two asymmetries deserve note. First, USDC keeps the tighter peg on ordinary days but owns the sample’s deepest discount (−285 bp at the SVB close), while USDT’s largest deviation is actually a *premium* (+202 bp) — recorded during the SVB weekend, when holders fled USDC *into* USDT. Signed deviations therefore carry information that absolute deviations alone would miss, which is why both enter the feature set. Second, daily closes understate intraday stress (USDC touched about 88 cents intraday on 11 March 2023); our index is thus a conservative, lower-bound measure of stress.

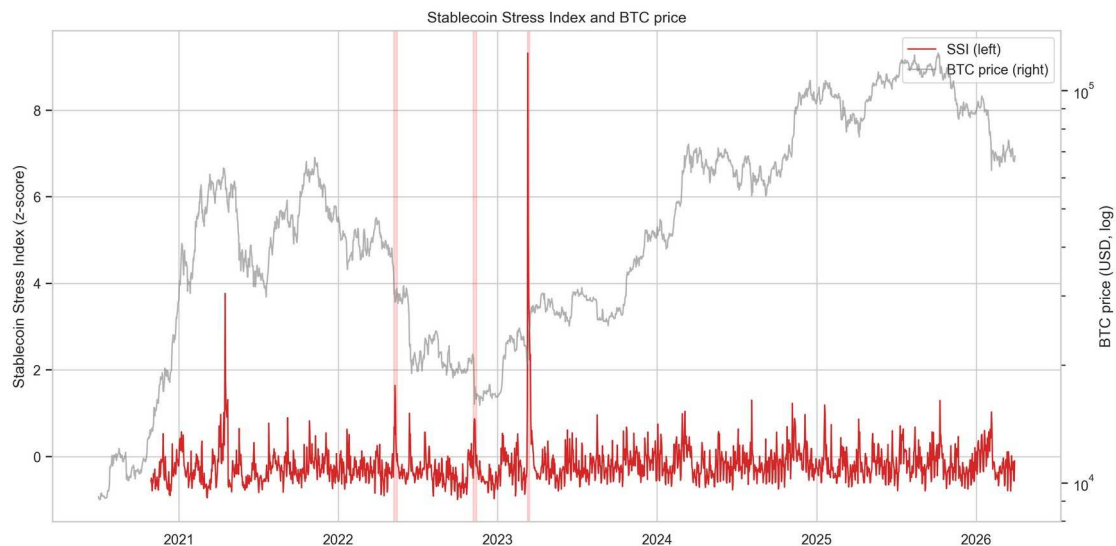


Figure 3.3 — The Stablecoin Stress Index (red, left scale) and the Bitcoin price (grey, right scale, log).
Shaded bands: Terra, FTX, USDC–SVB.

Figure 3.3 plots the resulting index. Its fifteen highest readings, listed below, provide an immediate validity check: the top of the ranking is dominated by the USDC–SVB weekend (SSI = 9.3 on 11 March 2023) and its aftermath, followed by the April 2021 leverage flush and the Terra week of May 2022. Post-2023 spikes appear exactly where the narrative of Chapter 2 predicts: the 5 August 2024 yen-carry unwind, the US election day of November 2024, and the 10 October 2025 liquidation cascade. The FTX week, consistent with its character as an exchange rather than stablecoin crisis, produces

only a moderate elevation — an observation we exploit when interpreting the classifiers in Chapter 6.

Table 3.4 — The fifteen highest readings of the Stablecoin Stress Index.

Date	SSI	dev USDT (bp)	dev USDC (bp)	BTC return same day	BTC return next day
2023-03-11	9.32	76.9	285.0	+2.2%	+7.2%
2023-03-12	5.21	71.5	79.3	+7.2%	+8.8%
2021-04-17	3.77	115.3	105.0	−1.5%	−7.7%
2023-03-13	3.71	38.7	10.5	+8.8%	+2.2%
2023-03-14	3.28	32.5	8.2	+2.2%	−1.5%
2023-03-15	2.91	39.5	1.8	−1.5%	+2.7%
2023-03-17	2.36	23.3	0.5	+9.0%	−1.7%
2023-03-16	2.23	21.3	0.4	+2.7%	+9.0%
2021-04-18	1.75	1.0	1.0	−7.7%	−0.9%
2022-05-11	1.64	41.3	7.7	−7.0%	+0.4%
2021-04-23	1.31	1.6	0.2	−1.3%	−2.1%
2024-08-05	1.30	3.7	0.2	−7.4%	+3.7%
2025-10-10	1.29	17.6	0.3	−7.2%	−2.2%
2024-11-06	1.23	8.3	0.2	+8.7%	+0.4%
2022-05-12	1.21	23.9	6.6	+0.4%	+0.8%

The table also previews a lesson that will recur throughout the thesis: extreme stablecoin stress predicts *turbulence*, not *direction*. The days following the highest SSI readings include both crashes (−7.7% after 17 April 2021) and violent rebounds (+7.2% and +8.8% after the SVB weekend, as the deposit guarantee reversed the panic). This is precisely why the empirical design separates the return task from the volatility and tail-risk tasks.

3.6 Descriptive statistics

Table 3.5 — Daily log returns, descriptive statistics (crypto on the 7-day calendar, macro on trading days). Jarque–Bera normality is rejected at any conventional level for every series.

Series	N	Mean (%)	Std (%)	Ann.	Skew	Exc.	Min (%)	Max (%)
				vol (%)		kurtos is		
BTC	2,099	0.095	3.04	58.1	−0.14	3.9	−17.4	+17.2
ETH	2,099	0.105	4.12	78.7	−0.23	4.8	−31.8	+23.1
SOL	2,099	0.219	6.29	120.2	−0.23	7.8	−55.0	+38.7
XRP	2,099	0.097	5.35	102.2	+0.72	19.6	−55.1	+54.9
GLD	1,128	0.064	1.10	17.5	−1.16	11.6	−10.8	+6.2
DXY	1,131	−0.00	0.44	7.0	−0.40	1.8	−2.1	+1.6

3

The table quantifies what Figure 3.1 suggests. Bitcoin’s annualized volatility over the sample is 58% — five times gold’s and eight times the dollar index’s — with daily moves reaching $\pm 17\%$; the altcoins are wilder still, with Solana annualizing above 120%. Every return series is leptokurtic (Bitcoin’s excess kurtosis of 3.9 implies far heavier tails than the normal, visualized in Figure 3.4), and Jarque–Bera tests reject normality overwhelmingly. The VIX enters the feature set in levels (sample range roughly 12 to 65) and in daily changes (mean -0.11 points, standard deviation 1.81 points). These facts drive three methodological choices in Chapter 4: heavy-tailed error distributions for the GARCH models, robust (HAC) inference for OLS, and the use of rank-based and tree methods that do not presume Gaussianity.

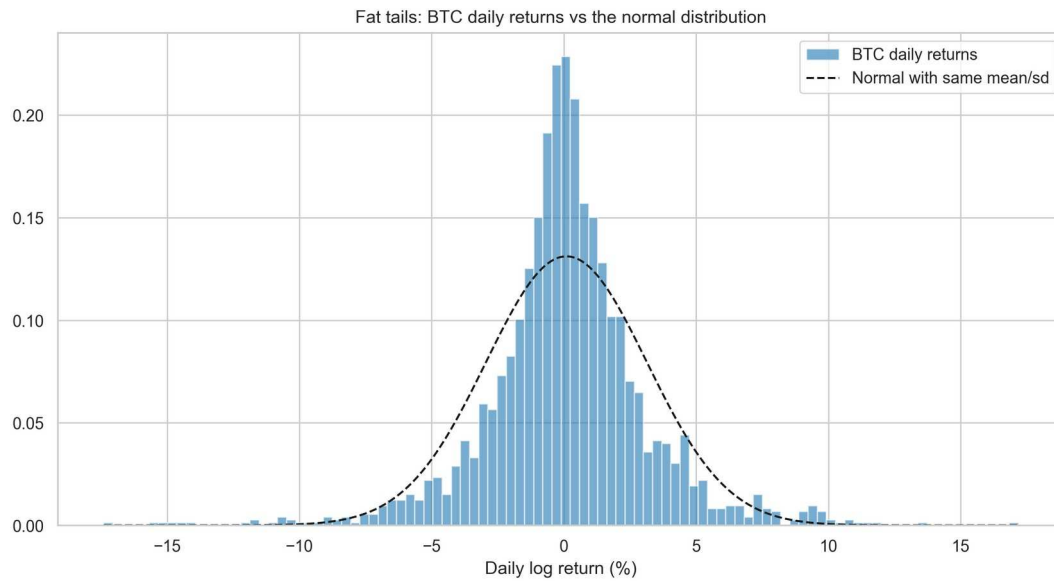


Figure 3.4 — Fat tails: the distribution of BTC daily returns against a normal with the same mean and variance.

Stationarity is confirmed where the modelling requires it: augmented Dickey–Fuller tests reject a unit root for Bitcoin returns (statistic -47.4), both peg-deviation series (-5.1 and -7.6), the SSI (-8.8) and the VIX level (-5.3), all with p-values below 0.01, while the Bitcoin log *price* is, as expected, non-stationary ($p = 0.18$). All models are therefore specified on returns, and the ARIMAX of Chapter 4 sets the integration order to zero.

Figure 3.5 provides the visual bridge from stress to volatility that motivates the whole exercise: Bitcoin’s 30-day realized volatility and the SSI trace the same crisis architecture. Figure 3.6 sharpens the point into the thesis’s motivating scatter: sorting days into SSI deciles, the average *absolute* next-day return rises visibly in the top deciles — stablecoin stress today anticipates larger moves tomorrow — while (as Chapter 5 confirms) the *signed* return remains unpredictable. The relationship is also visibly non-linear, flat over the central deciles and steep in the extremes, anticipating the value of splines, GAMs and tree-based models over purely linear specifications.

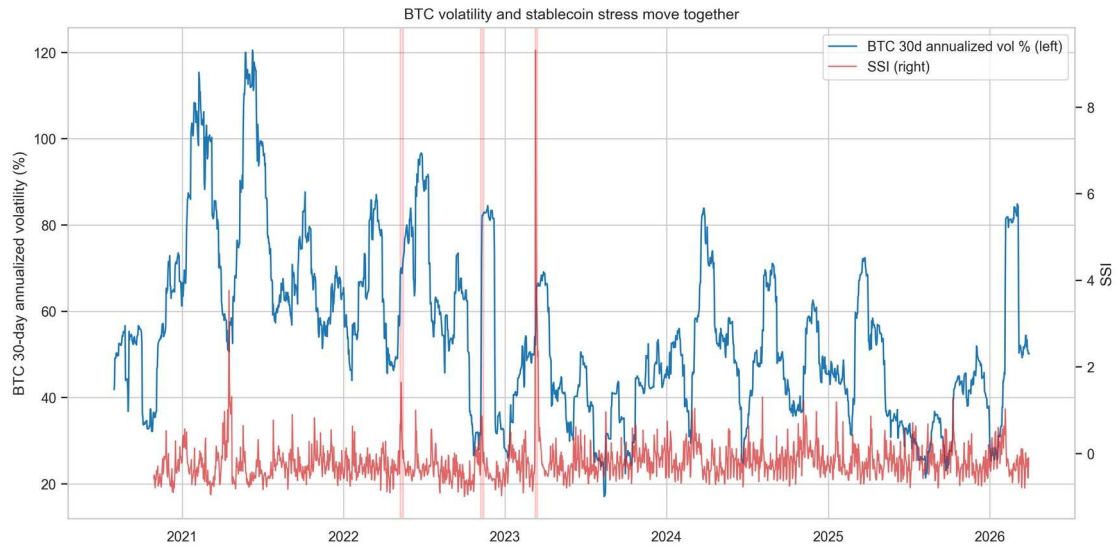


Figure 3.5 — BTC 30-day annualized volatility (blue, left) and the SSI (red, right).

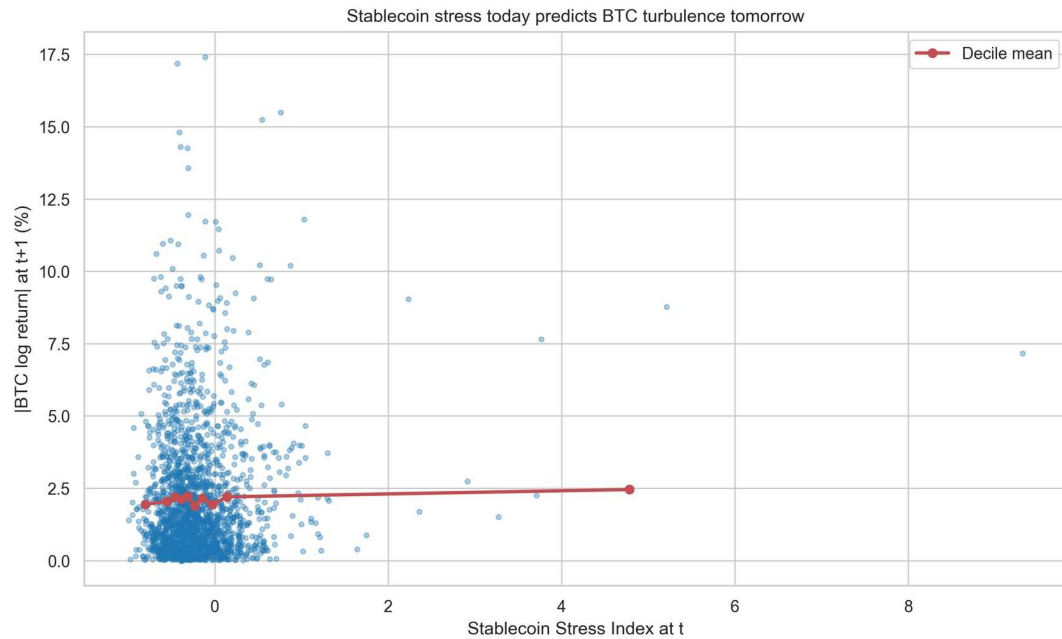


Figure 3.6 — The motivating plot: SSI at day t versus $|BTC\ return|$ at $t+1$, with decile means. Turbulence, not direction, is predictable.

The correlation structure of the full feature set, summarized in Figure 3.7, previews the collinearity landscape the models must navigate. Three clusters stand out. The strongest pairwise correlation in the entire panel is between the lagged BTC and ETH returns (0.81) — the two majors move as one at the daily frequency — closely followed by the pair of volume anomalies (USDC and BTC volume z-scores, 0.81): turbulence in the market and turnover in its settlement asset are nearly the same phenomenon. The stress

block is internally cohesive, exactly as its construction implies: the SSI correlates at 0.71 with the USDC volume z-score, at 0.64 with BTC volume, and at 0.60 with the USDC deviation volatility, while the two coins' deviation volatilities track each other at 0.68. The signed and absolute USDC deviations correlate at -0.64 , confirming that USDC's departures from the peg in this sample are predominantly *discounts* — while for USDT the same correlation is much weaker, the footprint of its 2023 premium episodes. These overlaps are precisely why Chapter 4 leans on regularization and why the VIF screen of Section 3.4 removed the single worst offender rather than attempting to orthogonalize a block whose internal correlations are economically meaningful. The ten strongest pairs are tabulated in Appendix B.

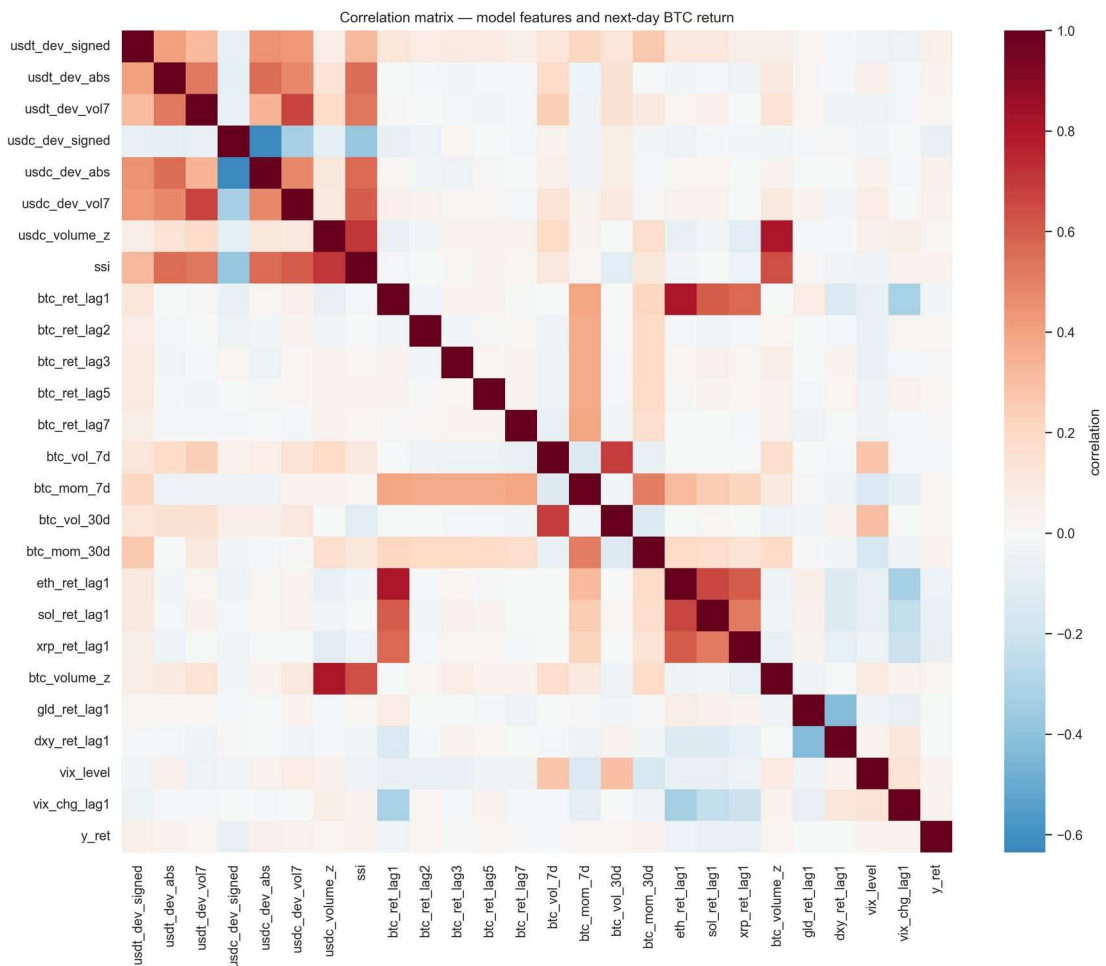


Figure 3.7 — Correlation matrix of the model features and the next-day BTC return. The stress block (top-left) and the momentum/volatility block are internally correlated; correlations with the target (last row) are uniformly close to zero.

3.7 Training, validation and test windows

After the burn-in imposed by the rolling and expanding windows (60 days for the SSI, 120 for the first tail threshold), the modelling dataset contains 1,979 daily observations with target dates from 30 October 2020 to 31 March 2026, partitioned as follows.

Table 3.6 — Sample splits, with crash-day counts under the main (10th-percentile) and robustness (5th-percentile) definitions. All model estimation uses training data only; hyperparameters are tuned by expanding-window cross-validation inside the training window; validation selects the champion models; the test month is evaluated once, in Chapter 5.

Split	Target dates	Obs.	Crash days (q10)	Base rate (q10)	Crash days (q05)
Training	30 Oct 2020 – 31 Jan 2026	1,920	205	10.7%	104
Validation	1 – 28 Feb 2026	28	5	17.9%	4
Test	1 – 31 Mar 2026	31	5	16.1%	1

The two out-of-sample months are usefully different. February 2026 was a stress month — one of the largest liquidation weekends on record, the sharpest one-day fall since November 2022, five crash days out of twenty-eight under the main definition — but of *macro* origin, with stablecoins holding their pegs. March 2026 was its calmer aftermath. A model-selection month this turbulent is demanding: models tuned to it must read general fragility, not merely recent momentum; and a calm test month then punishes any model that simply learned to predict chaos. Chapter 5 shows that this asymmetry is exactly where the white-box/black-box comparison becomes interesting.

3.8 Beyond the first day: horizon and calendar patterns

Two further descriptive questions sharpen the modelling choices of Chapter 4: for *how long* does stablecoin stress carry information, and *when* does crypto tail risk cluster in calendar time?

Table 3.7 — Horizon analysis of the SSI: correlation with future absolute BTC returns and average turbulence following high- and low-stress days.

Horizon h (days)	Corr($\text{SSI}_{t, r_{t+h} }$)	Mean $ r_{t+h} $, top SSI decile	Mean $ r_{t+h} $, bottom SSI decile	Ratio
1	0.093	2.45%	1.94%	1.27
2	0.032	2.33%	2.38%	0.98
3	0.003	2.10%	2.19%	0.96
4	0.015	2.13%	2.19%	0.97
5	0.011	2.35%	2.41%	0.98

The horizon analysis of Table 3.7 delivers a clean and somewhat surprising answer: the information in the index is *short-lived*. On the day after a top-decile stress reading, Bitcoin’s absolute return averages 2.45%, against 1.94% after a bottom-decile reading — a 27% difference in expected turbulence. By the second day the two groups are statistically indistinguishable, and the correlation with future absolute returns collapses from 0.093 to essentially zero. Stablecoin stress is a *one-day-ahead* barometer: it flags imminent turbulence, not persistent regimes. This is consistent with the arbitrage mechanics of Chapter 2 — peg deviations are either resolved or escalate within days, not weeks — and it justifies the one-step-ahead design of every prediction task in this thesis; nothing in the data suggests that stretching the forecast horizon would have found more signal.

Table 3.8 — Tail risk (main 10th-percentile definition) and volatility by day of the week of the target date, full modelling sample.

Day (target)	Days	Crash rate	Mean return	Std of return
Monday	283	14.5%	2.65%	3.82%
Tuesday	283	12.0%	2.21%	2.95%
Wednesday	282	12.4%	2.56%	3.54%
Thursday	282	13.8%	2.30%	3.27%

Friday	283	13.4%	2.35%	3.31%
Saturday	283	3.5%	1.11%	1.83%
Sunday	283	6.4%	1.64%	2.32%

The calendar analysis of Table 3.8 corrects a popular intuition. Crypto’s around-the-clock weekend trading is often described as its danger zone, yet *unconditionally* the weekend is the calmest part of the week: Saturday and Sunday have the lowest crash rates (3.5% and 6.4%, against a roughly 11% weekly average) and roughly half the weekday volatility, reflecting the absence of macro news and institutional flows. The dangerous day is **Monday** — a 14.5% crash rate and the highest return dispersion — when the week’s accumulated news and the weekend’s thin-liquidity price drift are marked to the reopening of traditional markets. This makes the prominence of the *Sunday* dummy in the February 2026 crash classification of Chapter 6 all the more interesting: it is not that weekends are ordinarily risky, but that when systemic deleveraging does strike, it exploits exactly the thin weekend order books that are harmless on an average Saturday — the “Black Sunday” liquidation of early February 2026 being the case in point. Unconditional calm and conditional fragility coexist, and the day-of-week dummies let the models express both.

4. Methodology

This chapter sets out the common evaluation framework (Section 4.1), the statistical principle that organizes the entire model menu — the bias–variance trade-off (Section 4.2) — and then each model family: the white boxes in Section 4.3 and the black boxes in Section 4.4. Notation and treatment follow James, Witten, Hastie and Tibshirani (2023) for the statistical-learning material and the original econometric sources for the time-series material. Section 4.5 collects the hyperparameter grids and implementation details.

4.1 Tasks, protocol and metrics

Setup. For each information date t we observe the feature vector $x_t \in \mathbb{R}^{32}$ of Chapter 3 and the three targets: $y_t^A = r_{t+1}$ (next-day BTC log return, %), $y_t^B = r_{t+1}^2$ (realized-variance proxy, %²) and $y_t^C \in \{0, 1\}$ (large-drawdown indicator). Task A and B are regressions, Task C a binary classification whose main target — the rolling 10th-percentile drawdown of Section 3.3 — has a training base rate of about 11%; the more extreme 5th-percentile target (base rate about 5%) is carried throughout as a robustness check.

Protocol. The data are split by target date into training (through 31 January 2026, 1,920 observations), validation (February 2026, 28) and test (March 2026, 31). Three rules discipline every model identically. First, *estimation uses training data only*; the validation and test months are never seen during fitting. Second, *hyperparameters are chosen by expanding-window cross-validation inside the training window*: the training sample is split into five consecutive folds, and each candidate configuration is repeatedly estimated on an initial segment and scored on the following fold, mimicking real-time forecasting. Ordinary k -fold cross-validation, which shuffles observations, is inadmissible here — it would train on the future to predict the past, and with volatility clustering this leaks substantial information. Third, *model selection uses the validation month and the test month is evaluated exactly once*, in Chapter 5, after all modelling decisions are frozen. The champion of each task is the model with the best validation score (MSE for Task A, QLIKE for Task B, AUC for Task C) among non-benchmark models.

Concretely, the expanding-window scheme splits the 1,920 training observations into five consecutive blocks of roughly 320 days each: the first fold trains on the initial block and predicts the second; the second trains on the first two blocks and predicts the third; and so on, so that the last fold trains on about 1,600 observations. Averaging a candidate configuration’s loss over the five folds yields an estimate of its true out-of-sample risk that respects the arrow of time and spans several market regimes — the 2021 bull market, the 2022 collapse, the 2023 recovery and the 2024–25 cycle — which protects the tuning from overfitting any single regime. Two further disciplines are worth making explicit because they are the most common sources of silent look-ahead bias in published backtests: all standardization statistics (means and variances of the features) are estimated on the training portion of each fold only and then applied to its validation portion; and every feature was constructed in Chapter 3 with rolling or expanding windows, so no quantity anywhere in the pipeline aggregates information across the train/validation boundary.

Benchmarks. Every task includes naïve benchmarks that any useful model must beat: for Task A the zero forecast (the random-walk prediction, optimal under martingale prices), the training-mean forecast and an AR(1); for Task B the training-sample variance and a rolling 30-day variance; for Task C the constant base-rate probability. The benchmarks are not straw men — a large literature (Welch and Goyal, 2008) documents how often published predictors fail to beat exactly these.

Regression metrics. Reported for Task A are the mean squared error and mean absolute error, the out-of-sample R^2 relative to the zero forecast,

$$R_{\text{zero}}^2 = 1 - \frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t y_t^2},$$

its analogue relative to the training mean (Campbell and Thompson, 2008), and directional accuracy — the share of days on which the forecast has the correct sign.

Volatility loss. For Task B, alongside the MSE we report the QLIKE loss

$$QLIKE = \frac{1}{T} \sum_t \left(\frac{r v_t}{\hat{\sigma}_t^2} - \ln \frac{r v_t}{\hat{\sigma}_t^2} - 1 \right),$$

which equals zero for a perfect forecast and penalizes under-prediction of variance more than over-prediction. Patton (2011) shows that QLIKE (and the MSE) are *robust* losses: rankings based on them are consistent even when the true conditional variance is observed only through a noisy proxy such as r_{t+1}^2 — precisely our situation. QLIKE is the primary Task B criterion; MSE on the squared-return proxy is dominated by a few extreme days and is reported for completeness.

Classification metrics. Even with the 10th-percentile crash definition — adopted precisely so that the one-month evaluation windows contain enough positives (5 in validation, 5 in test) for count-based statistics to be informative — threshold-based measures remain fragile; the primary criterion is therefore the area under the ROC curve (AUC), a threshold-free measure of ranking quality equal to the probability that a randomly chosen crash day receives a higher predicted probability than a randomly chosen normal day. Error rate, balanced accuracy and the F1 score at the conventional 0.5 threshold are reported alongside, mainly to illustrate the pathologies of imbalanced classification (a classifier that never signals a crash attains an 82.1% accuracy in February 2026 while being useless). Because a 0.5 threshold under a low base rate produces degenerate operating points for well-calibrated models, each classifier is additionally given an *F1-optimal alert threshold*, chosen by maximizing F1 on out-of-fold predicted probabilities within the training window — the threshold is thus treated exactly like a hyperparameter and never sees validation or test data. For consistency, all classifier hyperparameter grids are tuned with ROC-AUC as the cross-validation scoring metric (regression grids use mean squared error), so the tuning criterion always matches the task’s primary evaluation metric.

Residual diagnostics. Every estimated time-series model is subjected to Ljung–Box portmanteau tests (Ljung and Box, 1978) at lags 7, 14 and 21: on the residuals of the conditional-mean models (OLS, ARIMAX, with degrees of freedom adjusted for estimated ARMA terms) and, for the volatility models, on both the standardized residuals $\hat{z}_t = \hat{\varepsilon}_t / \hat{\sigma}_t$ and their squares. The null hypothesis is the absence of residual autocorrelation: a p-value above 5% indicates that the model has captured the relevant time dependence, while a rejection signals dynamics the specification has missed and disqualifies it as a well-specified description of the data.

Forecast-accuracy testing. Whether one forecast beats another is itself a statistical question. For loss series $L_t^{(1)}, L_t^{(2)}$ we compute the Diebold–Mariano (1995) statistic on the loss differential $d_t = L_t^{(1)} - L_t^{(2)}$,

$$DM = \frac{\bar{d}}{\sqrt{\widehat{V}(\bar{d})/T}},$$

with a Newey–West estimate of the long-run variance of d_t (three lags), the small-sample correction of Harvey, Leybourne and Newbold (1997), and Student- t critical values with $T - 1$ degrees of freedom. Task A models are tested against the zero forecast under squared-error loss; Task B models against GARCH(1,1) under QLIKE.

4.2 The bias–variance trade-off

Every modelling choice in this thesis is an instance of one decision: how much flexibility to allow. For a regression target with irreducible noise variance σ_ε^2 , the expected prediction error of a fitted function $\hat{f}$ at a point x_0 decomposes as

$$E[(y_0 - \hat{f}(x_0))^2] = \text{Var}(\hat{f}(x_0)) + [E\hat{f}(x_0) - f(x_0)]^2 + \sigma_\varepsilon^2.$$

Rigid models (a linear regression with few variables) have high bias and low variance; flexible models (a deep tree, an unregularized network) have low bias and high variance. The optimum lies between, and *where* it lies depends on the signal-to-noise ratio of the task. This is the deep reason the thesis studies three tasks rather than one. Daily returns have a signal-to-noise ratio close to zero — the irreducible term dominates — so the optimal flexibility is very low, and we should expect heavily regularized models to converge on near-constant forecasts (Chapter 5 confirms this: the cross-validated Lasso selects zero variables, and the cross-validated Ridge shrinks with $\lambda = 10^4$). Realized variance, by contrast, is strongly autocorrelated and predictable, so flexibility pays. Crash classification sits in between. The cross-validation of Section 4.1 is the empirical instrument that locates the optimum: each model family exposes one or two complexity parameters — λ for penalized regressions, depth and leaf size for trees, the number of boosting rounds, the SVM’s C , the network’s weight decay — and the expanding-window CV picks the value minimizing estimated out-of-sample loss.

It is worth dwelling on what the trade-off implies for a noise-dominated target, because it inverts the usual machine-learning instincts. When the predictable share of variance is large — image recognition, say, or realized volatility — richer models reduce bias faster than they accumulate variance, and complexity pays. When the predictable share is a percent or two — daily returns — the bias term is bounded above by that same small number, so there is almost nothing for flexibility to buy, while the variance cost of estimating hundreds of effective parameters on two thousand noisy observations remains fully priced. The rational response of a well-tuned learner is to converge toward the sample mean, and an honest empirical design must therefore treat “the best model predicts almost nothing” as an admissible, indeed expected, outcome rather than a failure to be engineered away. The three tasks of this thesis are chosen to sit at three points of this spectrum, so that the same model menu can be observed behaving rationally under three different signal-to-noise ratios.

4.3 White-box models

4.3.1 Linear regression with robust inference

The baseline model for Task A is ordinary least squares,

$$y_t^A = \beta_0 + \beta' x_t + \varepsilon_t,$$

estimated on the training window. OLS coefficients remain unbiased under heteroskedasticity and autocorrelation of ε_t , but their conventional standard errors do not; since volatility clustering guarantees heteroskedasticity, all inference uses Newey–West (1987) HAC standard errors with seven lags (one trading week). OLS plays two roles: it is a forecasting contender, and it is the interpretable reference whose coefficient table (Chapter 6) tells us which variables carry *marginal* predictive information conditional on all others.

The Newey–West estimator replaces the homoskedastic covariance $\hat{\sigma}^2 (X'X)^{-1}$ with a sandwich form whose middle matrix accumulates lagged cross-products of the residual-weighted regressors, down-weighted by Bartlett kernel weights $w_k = 1 - k/(L+1)$ up to the truncation lag $L=7$; the choice of one trading week balances capturing the autocorrelation left by overlapping rolling-window features against the loss of precision

from estimating higher-order autocovariances. Coefficients retain their natural units — with the target in percent and the deviation in basis points, a coefficient of +0.025 on the signed USDT deviation means that a ten-basis-point Tether premium is associated with about a quarter of a *percentage point* (25 basis points) of additional expected BTC return the next day, everything else equal: roughly a twelfth of a typical daily move, small but not negligible — which is the readability that makes OLS the reference white box even when it does not win the forecasting race.

4.3.2 Ridge and Lasso

With 32 correlated predictors and a weak signal, unregularized OLS coefficients are noisy. Shrinkage trades a little bias for a large variance reduction. Ridge regression (Hoerl and Kennard, 1970) solves

$$\min_{\beta} \sum_t (y_t - \beta_0 - \beta' x_t)^2 + \lambda \sum_j \beta_j^2,$$

shrinking all coefficients smoothly toward zero; the Lasso (Tibshirani, 1996) replaces the penalty with $\lambda \sum_j |\beta_j|$, whose geometry sets some coefficients exactly to zero and thus performs variable *selection*. Predictors are standardized inside each CV fold (the scaler is fit on the fold's training part only), and λ is chosen on grids of 30 values spanning 10^{-1} – 10^4 (Ridge) and 10^{-4} – $10^{0.5}$ (Lasso) by expanding-window CV. The selected λ values are themselves results: they measure how much shrinkage the data demand. For the classification task the same two penalties are applied to the logistic likelihood (Section 4.3.5).

The two penalties differ in geometry, and the geometry is the economics. The ridge solution can be written in terms of the singular values d_j of the feature matrix: each principal direction of the data is shrunk by the factor $d_j^2/(d_j^2 + \lambda)$, so directions with little variance — exactly those along which collinear features contradict each other — are suppressed most, which is why ridge is the canonical answer to the correlated stress block of Chapter 3. The lasso's diamond-shaped constraint set instead has corners on the coordinate axes, and the solution lands on a corner with positive probability: coefficients are not merely shrunk but set exactly to zero, turning estimation into variable selection. Standardization before penalization is not cosmetic — the penalty is

measured in coefficient units, so unscaled features would be penalized in proportion to their arbitrary units of measurement. Finally, λ traces a continuous path from OLS ($\lambda=0$) to the constant forecast ($\lambda \rightarrow \infty$): the cross-validated choice along this path is a direct, data-driven measurement of how much signal the penalized likelihood believes the features contain, and Chapter 5 reads it as such.

4.3.3 ARIMAX

The time-series counterpart of OLS models the return as an ARMA process with exogenous regressors:

$$y_t^A = c + \sum_{i=1}^p \phi_i y_{t-i}^A + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \gamma' z_t + \varepsilon_t,$$

where z_t contains the six exogenous stress and macro variables (SSI, the two absolute deviations, the VIX level, gold and dollar returns). Returns being stationary (Chapter 3), the integration order is zero. The orders (p, q) are selected on the training window over the grid $p, q \in \{0, 1, 2\}$ using three information criteria jointly — AIC (Akaike, 1974), BIC (Schwarz, 1978) and HQC (Hannan and Quinn, 1979) — with the order preferred by the majority of the criteria estimated; since BIC and HQC penalize parameters more heavily than AIC, this rule leans deliberately toward parsimony, and Chapter 5 reports the full criteria grid so that the sensitivity of the choice can be inspected. Estimation is by maximum likelihood in state-space form. The state-space representation deserves a word: the ARMA process is cast as a linear Gaussian system whose latent state carries the moving-average memory, and the Kalman filter delivers both the exact likelihood for estimation and, as a by-product, the one-step-ahead predictions used out of sample. The exogenous block is restricted to six regressors — the SSI, the two absolute deviations, the VIX level and the gold and dollar returns — rather than the full feature set, for two reasons: the ARMAX likelihood is estimated jointly and parsimony protects it from the collinearity documented in Section 3.4, and the specification isolates the thesis question (does *stress* add to an otherwise standard time-series model?) from the generic autoregressive content already carried by the ARMA terms. Out-of-sample predictions for the validation and test months are genuine one-step-ahead forecasts computed with *parameters frozen at their training-window estimates*: the filter is run

forward over the out-of-sample period, so each prediction uses only information available the previous evening.

4.3.4 Splines and the GAM

Figure 3.6 suggested that stress relates to future market behaviour non-linearly. The generalized additive model (Hastie and Tibshirani, 1986) captures smooth non-linearity while remaining additive — hence still interpretable, one curve per variable:

$$y_t^A = \beta_0 + \sum_{j \in S} f_j(x_{j,t}) + \sum_{k \notin S} \beta_k x_{k,t} + \varepsilon_t.$$

Six variables with the strongest non-linear candidacy (the SSI, both absolute deviations, the VIX level, 30-day volatility, and the lag-1 return) receive smooth terms f_j , represented as cubic B-spline expansions with six knots at training-sample quantiles; the remaining variables enter linearly. The model is estimated as a penalized least-squares regression on the expanded basis with a ridge penalty selected by expanding-window CV — the classical “splines in a linear model” implementation of the GAM. The fitted partial curves $\hat{f}_j$ are reported in Chapter 5.

The construction rests on the classical basis-expansion idea: a cubic spline with K interior knots is a piecewise cubic polynomial constrained to be continuous with continuous first and second derivatives at every knot, and the set of such functions is a linear space of dimension $K+4$ spanned by a B-spline basis whose members are non-zero over at most five adjacent knot intervals — a locality that keeps the design matrix well-conditioned where a global polynomial basis would explode. Placing the knots at training-sample quantiles allocates flexibility where the data actually live, which matters for stress variables whose distributions are extremely right-skewed: equally spaced knots would waste most of the budget on a nearly empty right tail. The ridge penalty on the basis coefficients then plays the role of the smoothing parameter of a classical smoothing spline, and its cross-validated choice determines the effective degrees of freedom of each curve. The result occupies the exact midpoint of the interpretability spectrum this thesis explores: each variable’s contribution is still a single drawable curve — Chapter 5 draws them — but the curves are free to bend where linearity is wrong.

4.3.5 Logistic regression and discriminant analysis

For Task C the canonical white box is logistic regression,

$$\log \frac{p_t}{1-p_t} = \beta_0 + \beta' x_t, p_t = \Pr(y_t^C = 1 \mid x_t),$$

estimated by maximum likelihood. Class imbalance is handled by weighting observations inversely to class frequency, which prevents the fit from collapsing onto the majority class. Its L1-penalized version performs variable selection exactly as the Lasso does in regression. Linear discriminant analysis (Fisher, 1936) approaches the same problem generatively: it assumes the features are Gaussian within each class with a common covariance matrix, and classifies by the resulting linear posterior log-odds. LDA is included both because it is part of the classical canon and because its comparison with logistic regression is informative: when the Gaussian assumption is grossly violated — as fat-tailed financial features guarantee — LDA's probability estimates suffer in characteristic ways.

Formally, logistic regression maximizes the Bernoulli log-likelihood $\sum_t [y_t \log p_t + (1 - y_t) \log (1 - p_t)]$, a smooth concave problem solved by iteratively reweighted least squares; class weighting multiplies each term by the inverse frequency of its class, which is equivalent to re-balancing the sample and shifts the fitted probabilities upward for the rare class without changing the ranking. Its coefficients are log-odds effects: a standardized coefficient of +0.40 on the 7-day USDT deviation volatility (Chapter 6) means one standard deviation of extra peg churn multiplies the odds of a next-day crash by $e^{0.40} \approx 1.5$. LDA reaches a linear score of the same form from the opposite, generative direction: assuming $x \mid y=k \sim N(\mu_k, \Sigma)$ with a shared covariance, Bayes' rule yields the discriminant $\delta_k(x) = x' \Sigma^{-1} \mu_k - 1/2 \mu_k' \Sigma^{-1} \mu_k + \log \pi_k$. The two models therefore share a functional form but disagree about what is being estimated: logistic regression fits the conditional probability directly and is robust to the feature distribution, while LDA borrows strength from the (here badly violated) Gaussian assumption — efficient when true, mis-calibrated when not, which is precisely the pattern the validation month exposes.

4.3.6 The GARCH family and the GARCH-X extension

Volatility clustering — visible to the naked eye in Figure 3.5 — is the most robust stylized fact of financial returns (Engle, 1982; Bollerslev, 1986). The GARCH(1,1) model specifies the conditional variance recursively:

$$r_t = \mu + \varepsilon_t, \varepsilon_t = \sigma_t z_t, \sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2,$$

with z_t i.i.d. Student- t to accommodate the heavy tails documented in Chapter 3. The sum $\alpha + \beta$ measures persistence. Two asymmetric extensions are estimated: GJR-GARCH (Glosten, Jagannathan and Runkle, 1993), which adds a term $\gamma \varepsilon_{t-1}^2 1\{\varepsilon_{t-1} < 0\}$ allowing negative shocks to raise variance more than positive ones, and EGARCH (Nelson, 1991), which models $\ln \sigma_t^2$ and hence needs no positivity constraints.

The thesis's central specification is the **GARCH-X**, which admits the lagged stablecoin stress index into the variance equation:

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 + \pi S S I_{t-1}^+, S S I_t^+ = \max(S S I_t, 0),$$

with $\pi \geq 0$ so that positive stress can only add variance and positivity of σ_t^2 is preserved. Because standard software does not accommodate exogenous variance regressors, the model is estimated by direct numerical maximum likelihood (Gaussian innovations; L-BFGS-B with parameter bounds), and an identically implemented plain GARCH(1,1) provides the nested null so that the in-sample contribution of stress can be assessed by a likelihood-ratio test, $LR = 2(\ell_x - \ell_0) \sim \chi_1^2$. The estimation routine is part of the accompanying code, documented in Appendix A.

Out-of-sample variance forecasts for all GARCH-family models follow the same fixed-parameter logic as the ARIMAX: parameters are frozen at training-window estimates and the variance recursion is filtered forward through February and March 2026, producing one-step-ahead forecasts $\hat{\sigma}_{t+1|t}^2$ that use only past information. The adequacy of each variance specification is verified with the Ljung–Box tests of Section 4.1 applied to standardized and squared standardized residuals: a specification that leaves significant autocorrelation in $\hat{z}_t^2$ has not fully captured the volatility clustering it exists to model.

Three properties of the estimated processes matter for interpretation. First, stationarity of the variance recursion requires $\alpha + \beta < 1$, and the *unconditional* variance is then $\omega / (1 - \alpha - \beta)$; our estimate sits on the boundary ($\alpha + \beta \approx 1.000$), the integrated-GARCH configuration typical of crypto samples, under which shocks to variance decay with a half-life of $\ln 0.5 / \ln(\alpha + \beta)$ — effectively months — and the unconditional variance is undefined: volatility wanders rather than mean-reverts, which is why fixed-parameter filtered forecasts remain sensible while long-horizon extrapolation would not be. Second, Student- t innovations with $\hat{\nu} \approx 3.3$ imply standardized shocks whose fourth moment barely exists; a Gaussian-innovation GARCH would misread every tail day as a regime change, inflating α . Third, the asymmetry extensions nest clean hypotheses: GJR’s γ measures the extra variance impact of negative shocks (the equity “leverage effect”), and EGARCH’s log formulation makes variance respond multiplicatively, removing positivity constraints but — as the diagnostics of Chapter 5 show — at the price of a fragile fit on this sample. The GARCH-X design keeps the stress term inside the positive region by construction ($\pi \geq 0$, $SS I^+ \geq 0$), so the likelihood-ratio test of $\pi = 0$ is a one-parameter boundary test whose χ_1^2 reference is conservative. It bears emphasis that in-sample significance and out-of-sample value are different questions — a regressor can fail a likelihood-ratio test yet improve forecasts, and vice versa (Hansen and Lunde, 2005). Chapter 5 shows the GARCH-X is precisely such a case.

4.4 Black-box models

4.4.1 Decision trees

A regression tree (CART) partitions the feature space by recursive binary splitting: at each step it chooses the variable j and threshold s whose split most reduces the residual sum of squares, and predicts within each terminal leaf by the leaf mean (for classification, by leaf class frequencies, with splits chosen by Gini impurity). Trees are the bridge model of the thesis: they are drawn and read like a white box (the fitted tree is displayed in Chapter 5) but their instability — small data perturbations produce different trees — marks them as the high-variance building block that the ensemble methods below exist to tame. Complexity is controlled by maximum depth and minimum leaf size, tuned by expanding-window CV. The estimation is greedy: at each

node the algorithm scans every feature and every observed threshold, chooses the split minimizing the immediate loss, and never revisits the choice — a stagewise strategy that makes trees fast (each level costs order $p \times n \log n$) but path-dependent, since an early split reshapes every later one. This path dependence is the microscopic source of the tree’s famous instability: perturbing a handful of observations can flip the first split from one momentum window to another and produce a structurally different tree with similar fit. Depth and leaf-size limits act as pre-pruning, bounding the number of leaves — a depth-2 tree has at most four — and hence the model’s effective degrees of freedom; the classical alternative, growing deep and pruning back by cost-complexity, arrives at similar places at greater cost and is omitted here in favour of direct CV over the two limits.

4.4.2 Bagging and random forests

Bagging (Breiman, 1996) reduces the variance of trees by averaging: B trees are grown on bootstrap resamples of the training set and their predictions averaged, $\hat{f}_{bag}(x) = B^{-1} \sum_b \hat{f}^{(b)}(x)$. Averaging helps most when the averaged objects are uncorrelated; the random forest (Breiman, 2001) strengthens the effect by also drawing, at every split of every tree, a random subset of $m < p$ candidate variables, which decorrelates the trees (otherwise every tree would lead with the same dominant predictors). We grow 400 trees and tune m (between $\sqrt{p}$ and $p/3$) and the leaf size; bagging is implemented as the forest with $m = p$. Forests also yield impurity-based and permutation-based variable importances, the interpretive tools of Chapter 6.

The variance arithmetic explains why the recipe works. Averaging B identically distributed predictors, each with variance σ^2 and pairwise correlation ρ , yields variance $\rho\sigma^2 + (1-\rho)\sigma^2/B$: growing more trees kills the second term for free, but the first survives however large B becomes — the ensemble can never be less variable than its trees are correlated. Bootstrap resampling alone leaves ρ high because dominant predictors (trailing momentum, trailing volatility) head almost every tree; restricting each split to a random subset of m candidates forces trees to explore secondary structure — including the stress block — and drives ρ down, which is the entire content of Breiman’s improvement. Each tree also leaves out roughly a third of the sample (the

out-of-bag observations), giving the forest an internal validation device; we rely on the expanding-window CV instead, since out-of-bag error ignores the time ordering that our protocol treats as sacred.

4.4.3 Gradient boosting (XGBoost)

Where bagging grows deep trees in parallel and averages away their variance, boosting grows *shallow* trees sequentially, each one fit to the errors of the current ensemble.

Formally, gradient boosting builds $F_M(x) = \sum_{m=1}^M v h_m(x)$, where each h_m is a small tree fit

to the negative gradient of the loss at the current predictions and v is a learning rate that deliberately slows the fit. The bias–variance logic is the mirror image of the forest’s: each weak tree is high-bias/low-variance, and complexity is added gradually, dose by dose. XGBoost (Chen and Guestrin, 2016) is the reference implementation, adding second-order gradient information, L2 regularization on leaf weights, and stochastic subsampling of rows and columns. We tune depth (2–3), learning rate (0.01–0.05) and the number of rounds (200–600), with 80% row and column subsampling; for classification, the minority class is up-weighted through the scale-pos-weight parameter.

Formally, boosting performs gradient descent in function space: at stage m the current ensemble’s predictions are differentiated through the loss, the negative gradient (for squared error, simply the residuals) becomes the working response, a shallow tree h_m is fitted to it, and the ensemble takes a damped step $F_m = F_{m-1} + v h_m$. The learning rate and the number of rounds trade off along a one-dimensional path — halving v roughly doubles the M needed — and small v with more rounds almost always generalizes better, because each step corrects a smaller fraction of the residual and the trajectory can stop earlier when validation loss turns. XGBoost sharpens the classic recipe by expanding the loss to second order, so each leaf weight has the closed form $-G/(H + \lambda_{L2})$ with G and H the sums of first and second derivatives in the leaf, and by evaluating candidate splits directly on the regularized objective; row and column subsampling add a stochastic decorrelation reminiscent of the forest. Depth two to three is not a compromise but a statement about interaction order: a depth- d tree can represent interactions of at most d variables, and the CV’s preference for shallow trees says the predictable structure in these data involves pairs of variables at most.

4.4.4 Support vector machines

The support vector machine (Cortes and Vapnik, 1995) fits the flattest function within an ε -tube of the data (regression, SVR) or the maximum-margin separating boundary (classification, SVC), in both cases solving a convex program in which only observations on or outside the margin — the support vectors — matter. Non-linearity enters through the kernel trick: replacing inner products with the radial basis function kernel $K(x, x') = \exp(-\gamma \|x - x'\|^2)$ implicitly fits in an infinite-dimensional feature space while computing only pairwise similarities. The cost parameter C (and ε for regression) governs the bias–variance position and is tuned by CV on standardized features; class weights handle the imbalance in Task C.

Two features of the machinery deserve note. The ε -insensitive loss $\max(0, |y - f(x)| - \varepsilon)$ creates a tube of indifference around the fitted function: observations inside it exert no pull at all, so the solution depends only on the minority of points at or outside the boundary — the support vectors — which is what makes the method robust to the mass of unremarkable days and sensitive to the informative extremes. And the kernel trick works because the dual optimization touches the data exclusively through inner products: replacing $x'x'$ with $K(x, x') = \exp(-\gamma \|x - x'\|^2)$ implicitly fits an infinite-dimensional expansion while computing only an $n \times n$ similarity matrix, with γ setting the bandwidth at which two market days count as “similar”. The cost of this elegance is opacity of a specific kind: the fitted function is a weighted sum of kernel bumps centred on historical days, so the model is best interrogated by asking *which* days anchor it — a question the interpretation toolkit of Chapter 6 is not designed for, and one reason the SVM contributes accuracy but little narrative in what follows.

4.4.5 Feed-forward neural network

The deep-learning representative is a fully connected multilayer perceptron: for hidden layers $h^{(1)}, h^{(2)}$,

$$h^{(1)} = g(W^{(1)}x + b^{(1)}), h^{(2)} = g(W^{(2)}h^{(1)} + b^{(2)}), \hat{y} = W^{(3)}h^{(2)} + b^{(3)},$$

with ReLU activations $g(u) = \max(0, u)$, trained by stochastic gradient descent (Adam) on standardized inputs. Architectures of one or two hidden layers (32, or 64–32 units) are compared, regularized by an L2 weight-decay penalty and by early stopping on a held-out fraction of the training data. With under two thousand training observations and a weak signal, the appropriate network is deliberately small: the flexibility that lets deep networks shine on large datasets is exactly what must be restrained here.

The universal-approximation property — a single hidden layer can approximate any continuous function arbitrarily well — explains both the promise and the danger: expressive capacity is unlimited, so on a noise-dominated target the network’s capacity must be starved deliberately. Three brakes are applied. The L2 weight decay $\alpha \sum_l \|W^{(l)}\|^2$ shrinks the network toward the linear regime (small weights keep every ReLU near its linear region, making the whole network approximately linear); early stopping halts gradient descent when held-out loss deteriorates, which is itself an implicit regularizer equivalent to bounding the distance travelled from the small random initialization; and the architecture search is confined to one or two narrow layers. Adam’s adaptive step sizes make the optimization reliable despite the loss surface’s non-convexity, but non-convexity has a reproducibility consequence — different seeds reach different minima — which is why the seed is fixed and reported. The validation-month failure of the classifier variant (Chapter 5) is the pedagogical payoff: with 205 positive examples, even a starved network finds spurious structure, a textbook illustration of why sample size, not algorithmic sophistication, is the binding constraint for deep learning in low-frequency finance.

4.5 Hyperparameters and implementation

Table 4.1 — Hyperparameter grids; all tuning by 5-fold expanding-window cross-validation on the training window.

Model	Tuned parameters (grid)	Fixed settings
Ridge	λ : 30 values, 10^{-1} – 10^4	standardized features
Lasso	λ : 30 values, 10^{-4} – $10^{0.5}$	standardized features
ARIMAX	$(p, q) \in \{0, 1, 2\}^2$ by AIC	$d=0$; 6 exogenous

		regressors
GAM (splines + ridge)	ridge λ : 20 values, 10^{-1} – 10^3	cubic B-splines, 6 knots, 6 smooth terms
Logistic / L1-logistic	$C \in \{0.01, 0.1, 1, 10\} / \{0.001, \dots, 1\}$	balanced class weights
GARCH family	—	Student-t; GARCH-X: Gaussian MLE, $\pi \geq 0$
Tree	depth $\{2, 3, 4, 6\}$; min leaf $\{10, 30, 100\}$	—
Bagging	min leaf $\{5, 20, 50\}$	400 trees, $m = p$
Random forest	$m \in \{\sqrt{p}, 0.33p\}$; min leaf $\{5, 20, 50\}$	400 trees
XGBoost	depth $\{2, 3\}$; $v \in \{0.01, 0.05\}$; rounds $\{200, 600\}$	subsample 0.8, colsample 0.8, L2 = 1
SVR / SVC	$C \in \{0.1, 1, 10\}$; SVR $\varepsilon \in \{0.01, 0.1, 0.5\}$	RBF kernel, standardized features
MLP	layers $\{(32), (64, 32)\}$; decay $\{10^{-3}, 10^{-1}\}$	ReLU, Adam, early stopping

The pipeline is implemented in Python on pandas and NumPy, with statsmodels for OLS/HAC, ARIMAX and the diagnostic tests, the arch package for the GARCH family, scikit-learn for the regularized, kernel, tree, ensemble and neural models, and XGBoost for gradient boosting; the GARCH-X likelihood is purpose-written with SciPy optimization. All stochastic components run under a fixed random seed, library versions are pinned, and one command regenerates every table and figure of the thesis (Appendix A). The complete pipeline — download, preprocessing, all grid searches, all diagnostics and every figure — executes in a few minutes on a consumer laptop; determinism was verified by re-running the pipeline end-to-end and checking that every stored table reproduces to the last reported digit. Three Jupyter notebooks accompany the scripts, one per prediction task, presenting the same analyses in sequential, executable form.

5. Empirical results

The chapter reports the three horse races in sequence, always in the same order: validation-month leaderboard, statistical tests, and only then — in Section 5.4 — the single look at the March 2026 test month. Every number is generated by the accompanying pipeline and traceable to a CSV file in the code repository.

5.1 Task A — predicting the next-day return

Table 5.1 — Task A, validation month (February 2026). Returns in percent; benchmarks in italics.

Model	MSE	RMSE	R ² (zero)	Dir. acc.	Chosen hyperpara meters
<i>Zero forecast</i>	18.58	4.31	0.000	—	—
<i>Training mean</i>	18.69	4.32	−0.006	0.357	—
<i>AR(1)</i>	18.46	4.30	0.006	0.393	—
OLS	18.36	4.29	0.012	0.571	—
Ridge	18.45	4.30	0.007	0.500	$\lambda = 10^4$
Lasso	18.69	4.32	−0.006	0.357	$\lambda = 0.76$, 0 variables
ARIMAX	18.75	4.33	−0.009	0.429	(1,0,0), 6 exog.
GAM (splines)	18.04	4.25	0.029	0.536	$\lambda = 10^3$
Tree	17.53	4.19	0.056	0.464	depth 2, leaf ≥ 100
Bagging	18.63	4.32	−0.003	0.500	400 trees
Random forest	18.91	4.35	−0.018	0.571	$m = \sqrt{p}$, leaf ≥ 50
XGBoost	18.00	4.24	0.031	0.536	depth 2, $\nu = 0.01$,

					200 rounds
SVR	18.88	4.35	-0.016	0.393	$C = 0.1$, $\varepsilon = 0.5$
MLP	18.52	4.30	0.003	0.536	(64, 32), decay 0.1

Three features of the leaderboard deserve attention before any ranking is taken seriously. First, the *spread is tiny*: the gap between the best model and the zero forecast is about 5% of the MSE, and the out-of-sample R^2 of every model lies within six percentage points of zero. Second, the regularized models deliver a verdict of their own: the cross-validated Lasso selects **zero** of the 32 predictors — its optimal forecast is a constant — and the cross-validated Ridge shrinks so aggressively ($\lambda = 10^4$) that its forecasts are nearly constant too. When the two methods whose entire purpose is to find the optimal bias–variance point both choose “predict almost nothing”, the message about the signal-to-noise ratio of daily returns is difficult to miss. Third, the nominal winner — a depth-2 regression tree that essentially sorts days into two or three coarse regimes — beats the zero forecast by 5.6% of MSE, but the Diebold–Mariano test cannot distinguish the two (DM = -0.90, $p = 0.38$); indeed *no* model differs significantly from the random walk in February (all $p \geq 0.26$). The white boxes are not humiliated: OLS attains the best directional accuracy of the month (57%), and the GAM is the best non-tree model, its fitted curves (Figure 5.1) showing exactly the mild nonlinearities in volatility and stress that Figure 3.6 anticipated. But honesty requires the summary to be blunt: on returns, nobody reliably beats “predict zero”.

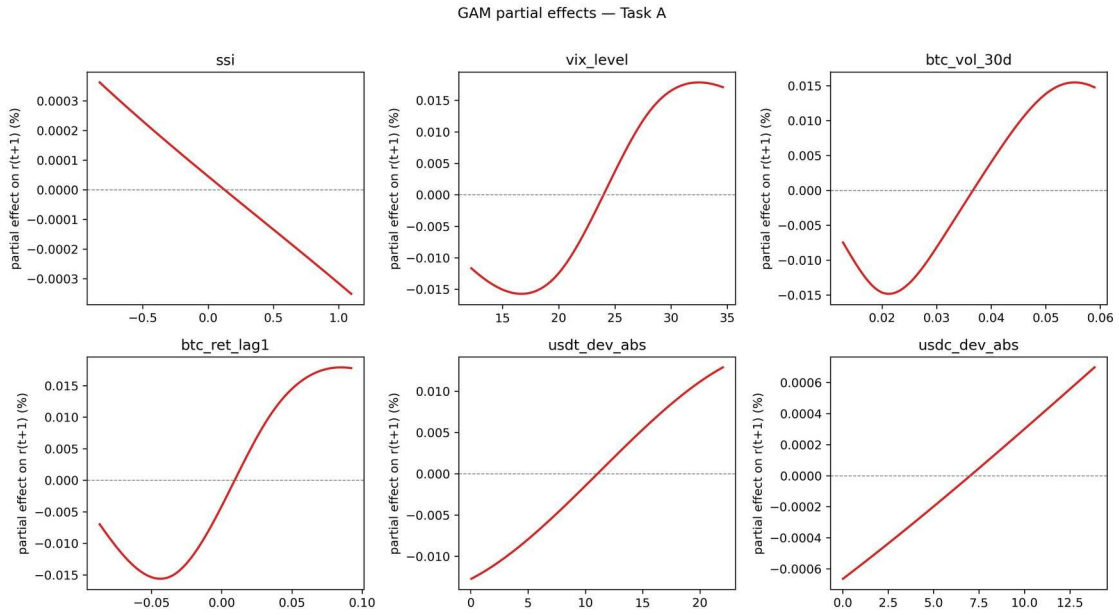


Figure 5.1 — GAM partial effects on the next-day return (training window). Effects are small — a fraction of a percentage point — and concentrated in the volatility and stress variables.

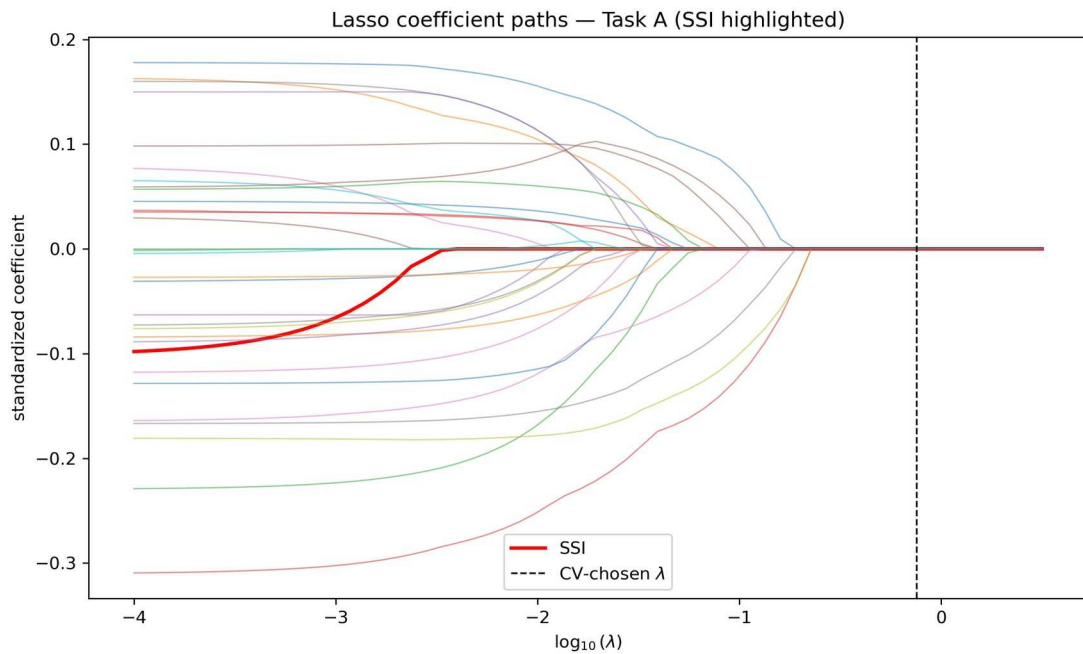


Figure 5.2 — Lasso coefficient paths, Task A. As the penalty relaxes (right to left), the first variables to enter carry stress and momentum information, but the cross-validated penalty (dashed line) excludes them all.

The Lasso path (Figure 5.2) refines the null result usefully: the *order* in which variables would enter — signed stablecoin deviations and short-horizon momentum first — is

economically sensible, but the cross-validated penalty sits to the right of every entry point. The signal is not absent; it is merely too faint, relative to daily noise, to survive honest regularization. The fitted regression tree (Figure 5.3) tells the same story in tree language: its first split is on trailing 30-day momentum and its leaves differ by fractions of a percent per day.

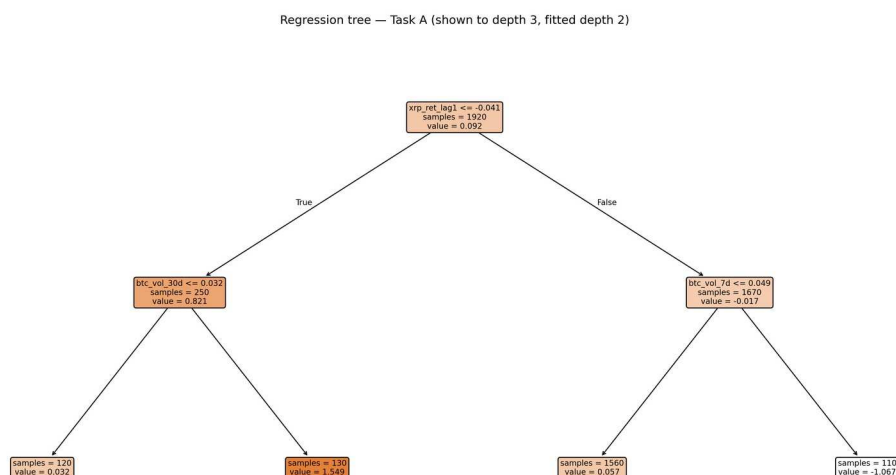


Figure 5.3 — The champion regression tree for Task A (displayed to depth 3). A deliberately crude model: two splits, three effective regimes.

Two specification checks complete the picture. First, the ARIMAX order is selected jointly by three information criteria: AIC alone would prefer the richer ARMA(2,1), but BIC and HQC — which penalize parameters more heavily — both select the parsimonious AR(1), and the majority rule adopts it; the performance difference between the two candidates is economically trivial (validation MSE 18.75 against 18.45, both statistically indistinguishable from the zero forecast), which is itself informative about how little ARMA structure daily Bitcoin returns contain. Second, Ljung–Box tests find no significant residual autocorrelation in either conditional-mean model — OLS residuals ($p = 0.35$ – 1.00 at lags 7–21) and ARIMAX residuals ($p = 0.10$ – 0.67) are both clean — so the unpredictability result cannot be attributed to a mis-specified mean equation: the models are adequate; the signal is simply not there.

The full estimation output of the selected ARIMAX(1,0,0) makes the point in coefficients rather than in fit statistics.

Table 5.2 — ARIMAX(1,0,0) estimates, training window ($N = 1,920$; log-likelihood $-4,868.8$). Standard errors from the outer product of gradients.

Parameter	Coef.	Std. err.	z	p-value	95% conf. interval
intercept	-0.166	0.279	-0.60	0.551	[-0.713, 0.381]
SSI	+0.121	0.167	+0.72	0.472	[-0.208, 0.448]
USDT dev., abs.	+0.006	0.013	+0.45	0.653	[-0.020, 0.032]
USDC dev., abs.	+0.015	0.015	+1.03	0.305	[-0.014, 0.044]
VIX level	+0.011	0.012	+0.91	0.362	[-0.013, 0.035]
DXY return (lag 1)	+2.667	22.332	+0.12	0.905	[-41.10, 46.44]
GLD return (lag 1)	+6.039	8.095	+0.75	0.456	[-9.83, 21.91]
AR(1)	-0.028	0.019	-1.47	0.141	[-0.066, 0.009]
σ^2	9.335	0.182	51.38	0.000	[8.979, 9.691]

The table reads as a formal restatement of the chapter's theme: the only parameter estimated with any precision is the innovation variance. The autoregressive coefficient is a statistically indistinguishable -0.03 — daily Bitcoin returns carry essentially no linear memory — and every exogenous regressor, the SSI included, has a z-statistic below 1.5. The confidence intervals translate the imprecision into economic units. The SSI point estimate says that one unit of stress raises the expected next-day return by 0.12 percentage points, but the 95% interval runs from -0.21 to $+0.45$: the data cannot even settle the *sign* of the effect, let alone its size. The AR(1) interval $[-0.066, +0.009]$

is economically as informative as it is statistically null: even at its most negative plausible value, yesterday's return would explain well under half a percent of today's variance. Against an innovation standard deviation of $\sqrt{9.34} \approx 3.1\%$ per day, every conditional-mean effect in the table is an order of magnitude too small to matter for a trader — the intervals quantify just how flat the predictable component is. Under the loose 10% criterion of the HAC-robust OLS (whose complete coefficient table with confidence intervals is reported in Appendix B), only the *signed* deviations and short-window volatility graze significance; in the jointly estimated ARIMAX, which uses absolute deviations and a single stress index, nothing survives. Both mean models pass their Ljung–Box diagnostics, so this is not mis-specification: it is the estimated shape of an (almost) efficient market.

Figures 5.4 and 5.5 make the null result visible day by day. The forecasts of every model hug the zero line while realized returns swing by percentage points around them, and the forecast-error series is almost indistinguishable from the return series itself — the graphical signature of forecasts that explain almost none of the variance. February's errors dwarf March's, reflecting the volatility regimes documented in Task B.

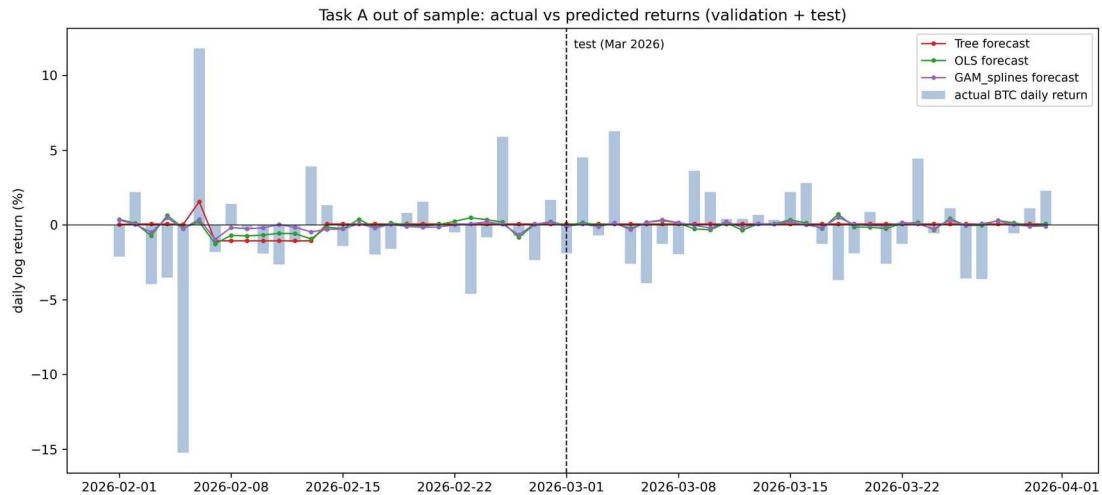


Figure 5.4 — Task A out of sample: actual daily returns (bars) against model forecasts (lines), validation and test months.

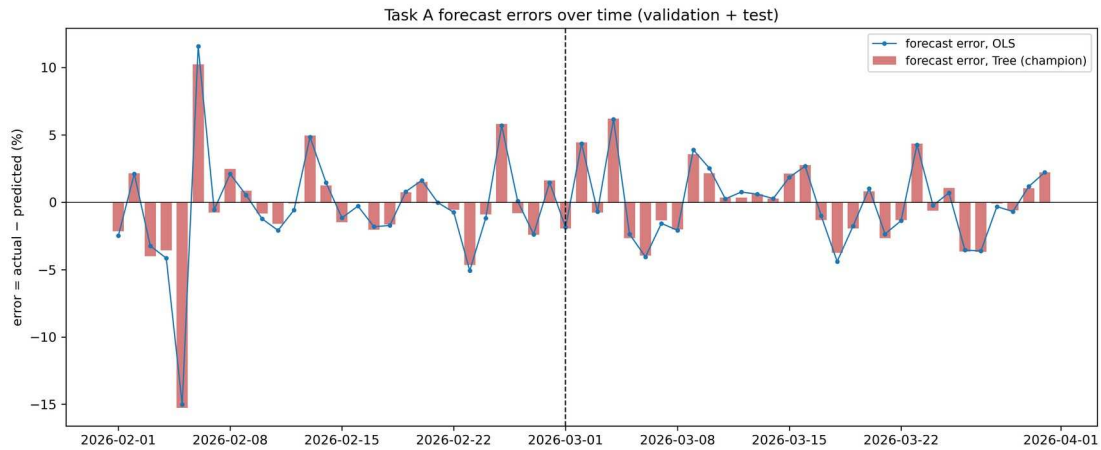


Figure 5.5 — Task A forecast errors over time. The errors inherit essentially the full variability of the returns.

5.2 Task B — predicting the next-day variance

The estimated GARCH-family parameters (training window) set the stage.

Table 5.3 — Estimated variance-equation parameters, training window. The first three models are estimated with Student- t innovations; the GARCH-X by Gaussian MLE — log-likelihoods across the two groups are therefore not directly comparable. Full inference for the GARCH-X is reported in the next table; the likelihood-ratio test of $\pi = 0$ against the Gaussian-restricted nested model gives $LR = 0.93$.

Model	ω	α	β	Asym.	ν	Pers.
GARCH(1,1)	0.088	0.058	0.942	—	3.34	1.000
GJR-GARCH	0.100	0.054	0.939	0.014	3.34	0.993
EGARCH	0.038	0.131	0.992	−0.004	3.20	—
GARCH-X (MLE)	0.168	0.059	0.923	—	Gauss.	0.982

Bitcoin’s conditional variance is extraordinarily persistent — $\alpha + \beta$ equals 1.000 to three decimals, the integrated-GARCH boundary — and extraordinarily fat-tailed, with a Student- t degrees-of-freedom estimate of 3.3, implying tails so heavy that the fourth moment barely exists. The asymmetry terms, by contrast, are economically negligible: unlike equity indices, Bitcoin in this sample shows no robust leverage effect, consistent with the mixed findings of the crypto-GARCH literature (Katsiampa, 2017). The stress coefficient of the GARCH-X is positive ($\pi = 0.153$: a one-standard-deviation stress

reading adds about 0.15%² to the next day's variance) but *not significant in sample*. The LR test deserves a technical note: because the constraint $\pi \geq 0$ places the null hypothesis on the boundary of the parameter space, the LR statistic is not asymptotically χ^2_1 but a 50:50 mixture $1/2 \chi^2_0 + 1/2 \chi^2_1$ (Self and Liang, 1987; Andrews, 2001). The nominal χ^2_1 p-value of 0.34 is therefore conservative; the mixture p-value is half of it, roughly 0.17 — kinder to the SSI, but still short of conventional significance. Two comparability caveats complete the in-sample picture. First, the LR test is internally consistent because it compares the GARCH-X with its Gaussian-restricted nested version; second, since the three classical models use Student-*t* likelihoods while the GARCH-X is Gaussian, parameter p-values and information criteria are not homogeneous across the two groups, and the only distribution-free cross-model comparison is the out-of-sample one under QLIKE. If the analysis stopped at in-sample inference — as much of the empirical literature does — the SSI would be dismissed. The residual diagnostics and the out-of-sample leaderboard suggest a more nuanced verdict.

Table 5.4 — *GARCH-X full inferential output. Standard errors from the numerical observed-information (Hessian) matrix; $\pi \geq 0$ is a boundary constraint, so its two-sided p-value (and the interval's negative portion) is conservative, consistent with the boundary-corrected LR test above.*

GARCH-X parameter	Coef.	Std. err.	z	p-value	95% conf. interval
μ	0.100	0.061	1.65	0.099	[−0.019, 0.220]
ω	0.168	0.057	2.94	0.003	[0.056, 0.280]
α	0.059	0.012	4.77	< 0.001	[0.035, 0.083]
β	0.923	0.016	57.12	< 0.001	[0.891, 0.954]
π_{SSI}	0.153	0.173	0.88	0.377	[−0.186, 0.491]

The inferential detail sharpens the persistence discussion. The GARCH core of the X-specification is estimated with textbook precision — α and β carry z-statistics of 5 and 57, and the β interval [0.891, 0.954] excludes the integrated boundary — implying persistence $\alpha + \beta = 0.982$ and a shock half-life of roughly 37 days: high, but visibly *below* the exact unit persistence of the plain GARCH(1,1). The stress loading absorbs part of what the pure recursion would otherwise force into β , pulling the model off the integrated boundary. The loading itself is where the economic and the statistical readings part ways. At the point estimate, a two-unit stress day (a Terra- or FTX-scale reading) adds $2 \times 0.153 \approx 0.31\%^2$ to the next day's variance — lifting a baseline daily volatility of 2% to about 2.08%, a modest but not trivial repricing of risk; at the upper bound of the interval the same day would add nearly $1\%^2$, pushing volatility above 2.2%. But the interval also contains zero: in sample, stress days are too few for π to be pinned down, and any conclusion about the SSI's role in the variance equation must lean on out-of-sample evidence rather than on this coefficient — which is precisely why the model's case rests on the Diebold–Mariano tests below and the residual diagnostics next, and why even those should be read as supportive rather than conclusive given the one-month evaluation windows.

Table 5.5 — Ljung–Box tests on standardized ($\hat{z}_t$) and squared standardized ($\hat{z}_t^2$) residuals, training window. H_0 : no residual autocorrelation; bold p-values reject at 5%.

Model	LB on $\hat{z}_t$, lags	LB on $\hat{z}_t^2$, lags	Adequate?
	7/14/21 (p)	7/14/21 (p)	
GARCH(1,1)	0.87 / 0.45 / 0.59	0.052 / 0.064 / 0.030	borderline
GJR-GARCH	0.84 / 0.44 / 0.59	0.080 / 0.098 / 0.052	yes
EGARCH	0.82 / 0.46 / 0.63	0.004 / 0.004 / 0.002	no
GARCH-X (SSI)	0.89 / 0.45 / 0.57	0.21 / 0.17 / 0.066	yes

The diagnostics of Table 5.5 deliver a first verdict before any forecast is scored. All four models whiten the *level* of the standardized residuals, but they differ on the squares — the direct test of whether the volatility dynamics have been captured. The EGARCH is rejected at every lag ($p \leq 0.004$): its log-variance recursion leaves systematic volatility dependence unexplained, and it should be discounted regardless of any in-sample fit. The plain GARCH(1,1) is borderline, slipping below 5% at the longest lag ($p = 0.030$). **Both the GJR-GARCH and the GARCH-X pass at every lag at the 5% level:** the GJR with its longest-lag p-value essentially on the threshold ($p = 0.052$), the GARCH-X more comfortably ($p = 0.066$ – 0.21). The stress term thus appears to absorb much of the residual volatility dependence that the pure recursion misses, but the diagnostics warrant a measured reading: the GJR achieves a comparable whitening through a different channel (asymmetry), and — given the insignificant π — the tests are consistent with a useful role for stablecoin stress in the variance equation rather than proof that it is indispensable.

Table 5.6 — Task B, validation month (February 2026). QLIKE is the primary criterion; DM tests use QLIKE loss. Benchmarks in italics.

Model	Val. MSE	Val. QLIKE	DM vs GARCH(1,1)
<i>Rolling 30-day variance</i>	2600.1	3.376	$p = 0.29$
<i>Constant (train variance)</i>	2443.3	2.589	$p = 0.62$
GARCH(1,1)	2509.2	2.713	—
GJR-GARCH	2498.8	2.632	$p = 0.44$
EGARCH	2465.5	2.617	$p = 0.033$
GARCH-X (SSI)	2467.9	2.620	$p = 0.013$
Random forest	2262.2	2.061	$p = 0.058$
XGBoost	2168.6	2.046	$p = 0.039$

Two results stand out. First, the **GARCH-X significantly outperforms the plain GARCH(1,1)** under QLIKE (DM $p = 0.013$) — the stablecoin stress term, statistically invisible in sample, earns its place exactly when it matters, in a stressed out-of-sample month. The contrast between the failed likelihood-ratio test and the successful forecast

comparison is a textbook illustration of Hansen and Lunde’s (2005) warning that in-sample fit and predictive value are different currencies. Second, in this stressed month the **machine-learning models dominate the entire GARCH family**: XGBoost improves QLIKE over GARCH(1,1) by 25% (2.05 against 2.71, DM $p = 0.039$), with the random forest close behind — a lead that Section 5.4 shows does not survive into the calm test month, where the rolling 30-day benchmark edges every model. Both consume the full 32-variable information set — trailing volatilities, stress measures, the VIX — where GARCH sees only its own past squared residuals plus, in the X variant, one stress index. Figure 5.6 shows the classical fit at its best (conditional volatility tracking every crisis of the sample); Figure 5.7 shows the out-of-sample months, where the boosted trees adapt to February’s turbulence faster than the near-integrated GARCH recursion.

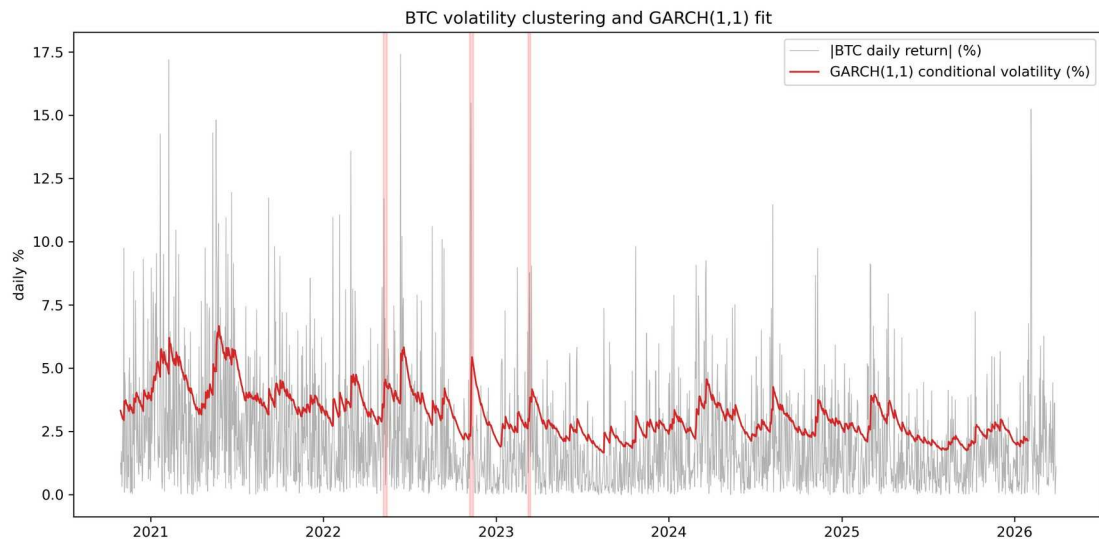


Figure 5.6 — GARCH(1,1) conditional volatility against absolute daily returns, full sample.

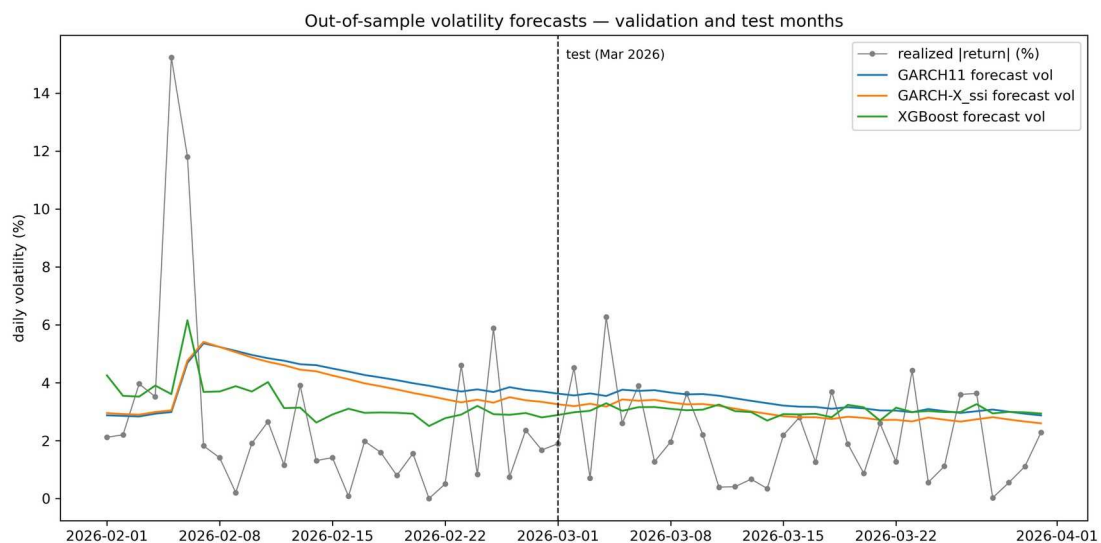


Figure 5.7 — Out-of-sample one-step-ahead volatility forecasts, February–March 2026.

5.3 Task C — predicting large-drawdown days

Table 5.7 — Task C, validation month (February 2026): 5 crash days among 28 under the main 10th-percentile definition. Alerts = days with predicted probability above the threshold; thr^* = the F1-optimal threshold chosen on out-of-fold training probabilities (Section 4.1).

Model	CV AUC	Val. AUC	Val. F1 @0.5	Alerts @0.5	thr^*	Val. F1 @ thr^*	Alerts @ thr^*
Base rate	—	0.500	0.00	0	—	—	—
Logistic (balanced)	0.589	0.644	0.29	16	0.50	0.29	16
LDA	—	0.670	0.00	0	0.07	0.36	23
L1-logistic	0.596	0.626	0.32	14	0.48	0.27	17
Tree	0.556	0.613	0.32	14	0.62	0.00	0
Random forest	0.559	0.713	0.00	3	0.38	0.42	14
XGBoost	0.573	0.617	0.38	11	0.39	0.38	21
SVC	0.581	0.678	0.00	0	0.12	0.32	20
MLP	0.503	0.313	0.00	0	0.30	0.00	0

The classification task separates the field, though the broader crash definition narrows the margins. The random forest remains the best ranker of the month (AUC 0.713), but

the gap to the rest of the field shrinks relative to the extreme-tail robustness results below: the SVC (0.678) and the LDA (0.670) sit close behind, the logistic family clusters around 0.63, gradient boosting slips to 0.617, and the neural network fails outright (AUC 0.313, a reminder that small samples and noisy targets are hostile territory for flexible optimizers). The cross-validated AUCs on the training window (first column) are modest for everyone, 0.50–0.60: ranking ordinary bad days — the 10th percentile includes many merely poor sessions alongside the catastrophes — is genuinely hard on average. The table also documents the pathology of the conventional 0.5 threshold, which persists even at a 10.7% base rate: the best-ranked models are exactly the ones that degenerate there (the forest issues 3 alerts, all false; SVC, LDA and MLP issue none), not because they miss the risk but because their calibrated probabilities rarely exceed one half; XGBoost, whose probabilities run hotter, operates respectably at 0.5 ($F1 = 0.38$). Re-thresholding repairs the degeneracy without touching validation or test data: with the F1-optimal cut-off learned on out-of-fold training probabilities, the random forest issues 14 alerts and converts its ranking advantage into the month’s best F1 (0.42), while the LDA climbs from a degenerate zero to 0.36. The exercise also shows the tool’s limits: the tuned threshold is itself an estimated quantity, and for the decision tree the out-of-fold choice (0.62) lands *above* the model’s validation probabilities, silencing it entirely — threshold tuning inherits the noise of the probabilities it is tuned on. Under class imbalance the threshold is a decision parameter to be chosen, like any hyperparameter, on training data; which operating point is “right” ultimately depends on the user’s loss function, which is why the threshold-free AUC remains the primary criterion. The ROC curves (Figure 5.9) make the ranking picture visible across all thresholds.

The confusion matrices, shown in Figure 5.8 at each model’s tuned threshold, spell the trade-off out in raw counts; Table 5.8 summarizes precision, recall and F1 at both operating points, and the complete count-level table is reported in Appendix B.

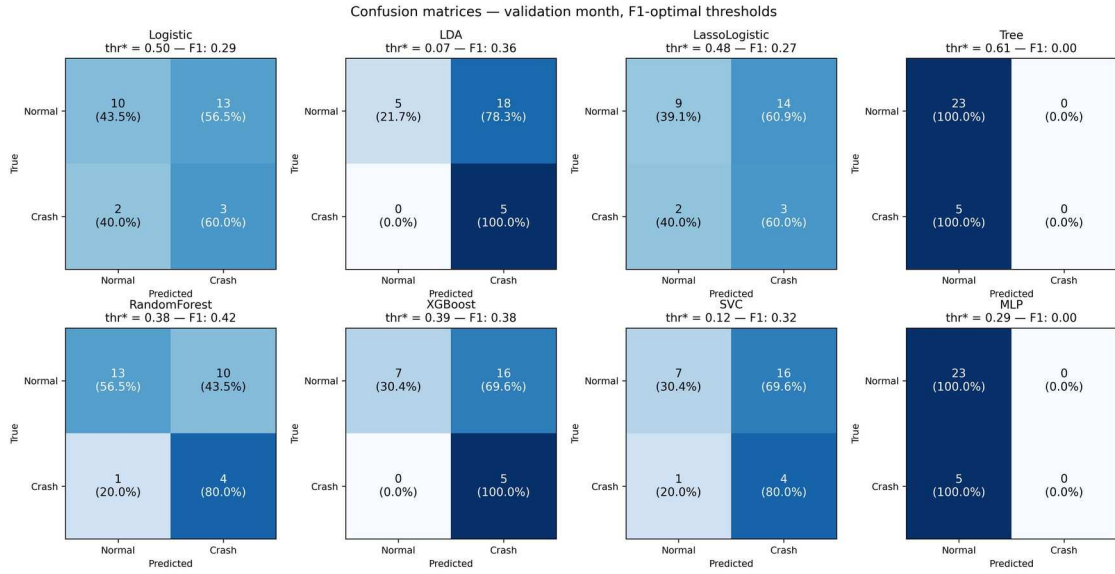


Figure 5.8 — Confusion matrices on the validation month (28 days, 5 crash days) at each model’s F1-optimal threshold thr^* . Cell shading is row-normalized; each cell reports the count and the row percentage.

Table 5.8 — Precision, recall and F1 on the validation month at the conventional 0.5 threshold and at the F1-optimal threshold thr^* (alerts = predicted crash days out of 28). Full confusion-matrix counts in Appendix B.

Model	Prec.	Rec.	F1	thr^*	Prec.	Rec.	F1	Alerts
	@0.5	@0.5	@0.5		@ thr^*	@ thr^*	@ thr^*	@ thr^*
Logistic	0.19	0.60	0.29	0.50	0.19	0.60	0.29	16
LDA	0.00	0.00	0.00	0.07	0.22	1.00	0.36	23
L1-logistic	0.21	0.60	0.32	0.48	0.18	0.60	0.27	17
Tree	0.21	0.60	0.32	0.62	0.00	0.00	0.00	0
Random forest	0.00	0.00	0.00	0.38	0.29	0.80	0.42	14
XGBoost	0.27	0.60	0.38	0.39	0.24	1.00	0.38	21
SVC	0.00	0.00	0.00	0.12	0.20	0.80	0.32	20
MLP	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0

Read together, figure and table make three points that the aggregate metrics compress. First, at the conventional threshold four of the eight classifiers — including the three best rankers — degenerate to “never alarm” (or nearly so): zero true positives, all five

crashes missed, and a deceptively low error rate. Second, re-thresholding moves each model along its own precision–recall frontier, and where it lands reveals the quality of the underlying ranking: the random forest reaches the month’s best balance (4 of 5 crashes caught, precision 0.29), XGBoost buys perfect recall at the price of alarming on 21 of 28 days, and the LDA does the same on 23 — fire alarms that ring most mornings. Third, thresholding is no cure-all: the MLP’s probabilities are too poorly ordered to rescue, and the decision tree is silenced by its own tuned cut-off, two different ways for the same tool to fail. For a risk desk the relevant row is the forest’s second: in the stressed month the tuned monitor would have caught the entire early-February crash cluster — the three consecutive crash days of 3–5 February, including the –15.2% collapse, and the mid-month tail day of 11 February — and missed only the isolated late-month drop (23 February, –4.6%, whose predicted probability, 0.31, sits just below the 0.38 threshold), at the cost of ten false alarms concentrated in the fragile first half of the month, roughly 2–3 per week. That is an operating point a human committee can reason about, which is what the exercise was for. Two cautions temper the enthusiasm: with five positives, validation AUCs carry wide sampling bands; and February 2026 — a month whose fragility built up visibly in the data — was favourable terrain for stress-sensitive models.

Robustness: the extreme 5th-percentile target. Re-running the key models on the rarer definition — only the worst ~5% of days count as crashes: 104 training positives, 4 in validation — with identical grids, cross-validation and out-of-fold threshold tuning gives the following validation results.

Table 5.8b — Robustness check on the 5th-percentile target, validation month (4 crash days among 28). Same protocol as the main analysis.

Model (q05 target)	CV AUC	Val. AUC	thr*	Val. F1 @thr*	TP / FP / FN
Logistic (balanced)	0.570	0.646	0.38	0.22	2 / 12 / 2
Random forest	0.558	0.802	0.38	0.55	3 / 4 / 1
XGBoost	0.554	0.771	0.24	0.33	3 / 11 / 1

The robustness check strengthens, rather than weakens, the main conclusions. When only the worst crashes count, the ensembles' advantage *widens*: the forest's AUC rises to 0.80 against 0.65 for the logistic — extreme days are where the non-linear stress signature is most distinctive, and where the 10th-percentile target dilutes it with ordinary bad days. The operating signature is unchanged: at its tuned threshold the forest catches 3 of the 4 extreme crashes (again including 5 February) with only 4 false alarms, and again misses only 23 February — the same isolated drop, under both definitions, whose stress fingerprint simply was not there (a point Section 6.4 returns to). The main-target results above are therefore, if anything, conservative about the ensembles' edge on the events that matter most.

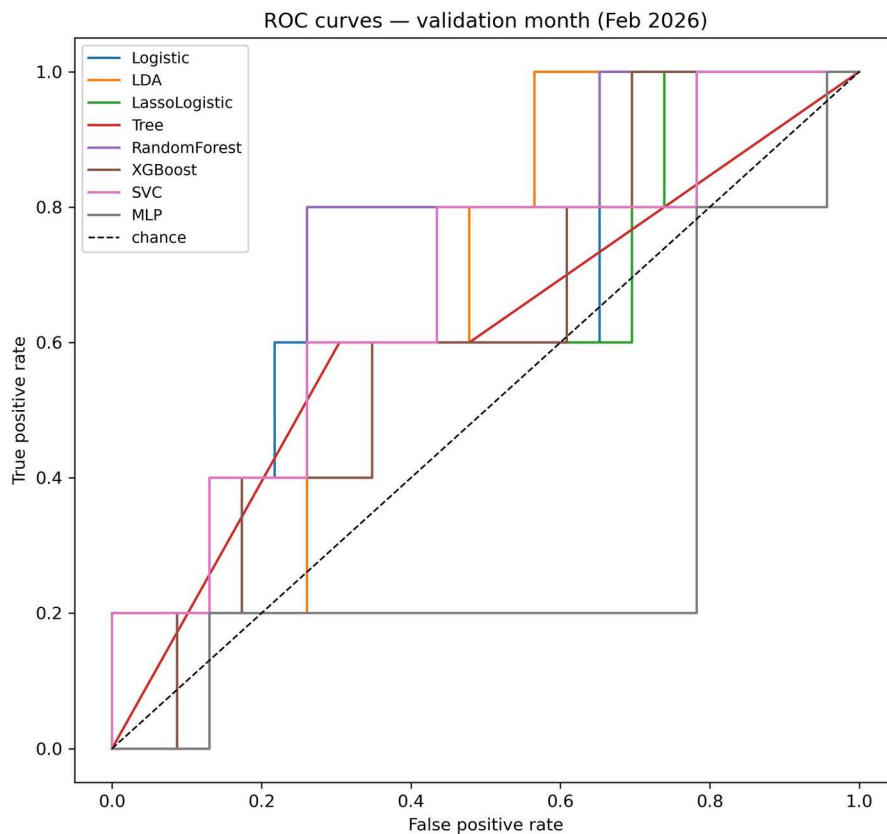


Figure 5.9 — ROC curves on the validation month. The random forest dominates; the diagonal is chance.

5.4 The March 2026 test

With all choices frozen — champion return model: the regression tree; champion variance model: XGBoost; champion classifier: the random forest — the test month is unveiled once.

Table 5.9 — Champions selected on validation, evaluated once on test. “n.s.” = not significantly different from the zero benchmark.

Task	Champion	Validation	Test (March 2026)	Best test performer (post hoc)
A — return MSE / R ² (zero)	Tree	17.53 / 0.056	6.60 / 0.0005	SVR (6.58 / 0.003) — n.s.
B — variance QLIKE	XGBoost	2.046	1.204	Rolling-30 benchmark (1.169)
C — crash AUC	Random forest	0.713	0.592	SVC (0.808)

March 2026 was a very different month from February: realized daily volatility halved (the average squared return fell from about 18.6%² to 6.6%²), and although five days breached the rolling 10th-percentile threshold (5, 6, 18, 26 and 27 March), all were shallow drops between −2.6% and −3.9% — routine bad days of a calming market, not crashes of February’s kind; only 6 March breached the extreme 5th-percentile threshold. The test-month verdicts, model by model, are instructive.

Task A. The champion tree’s February advantage evaporates completely: its test R²(zero) is 0.0005, statistically and economically zero, and its forecast path (Figure 5.10) is revealingly *flat* — the depth-2 tree, faced with March’s unremarkable features, outputs an almost constant +0.05% per day. No model beats the zero forecast significantly in March either (the only marginal DM result runs in the *opposite* direction: XGBoost is nearly significantly worse, $p = 0.08$). OLS again posts the best directional accuracy (61%), an intriguing but — over 31 days — statistically fragile consolation. The February ranking, in other words, did not transfer; in a signal-starved task, one good month is selection noise, and the honest conclusion is the one the Lasso drew from the start.

Task B. The volatility story *does* transfer, with one twist. The GARCH-X again significantly beats the GARCH(1,1) (QLIKE 1.201 vs 1.268, DM $p = 0.035$) — the stress term helps in a calm month too, chiefly by *not* inflating variance when stress is absent — and gradient boosting again edges the GARCH family ($p = 0.051$). The twist is the winner: the humble rolling 30-day variance posts the month’s best QLIKE (1.169), because in a placid month the best forecast is a short memory. No sophisticated model beats it significantly, and EGARCH — the most flexible classical specification — is significantly *worse* than plain GARCH ($p = 0.0003$), having carried February’s alarm too far into March, exactly the fragility its failed residual diagnostics (Table 5.5) had foreshadowed. The pattern is a clean bias–variance parable: models tuned amid stress over-predict calm.

Task C. The forest’s AUC declines from 0.71 to 0.59: March’s five crash days were mild macro drips, not stress events, and the forest — trained to smell stablecoin-mediated fragility — ranks them only weakly. At its tuned threshold it issues four alerts in the month (catching 5 March, missing the other four, whose predicted probabilities sit between 0.29 and 0.34), and none of its probabilities reaches one half: a stress monitor staying appropriately quiet in a month without stress, at the cost of missing days that were “crashes” only by the widened definition. On the extreme 5th-percentile robustness target the picture is cleaner: the forest’s test AUC is 0.70, ranking the one genuine tail day (6 March) above roughly 70% of the month. The post-hoc column shows the SVC as March’s best ranker (AUC 0.81) after placing mid-field in February — with the Task A lesson fresh, one month of superiority is selection noise, not a verdict. The instability of the podium across two adjacent months is itself the finding: in short evaluation windows, model selection is a noisy estimator, which is why Chapter 6 leans on evidence that aggregates across the full sample.

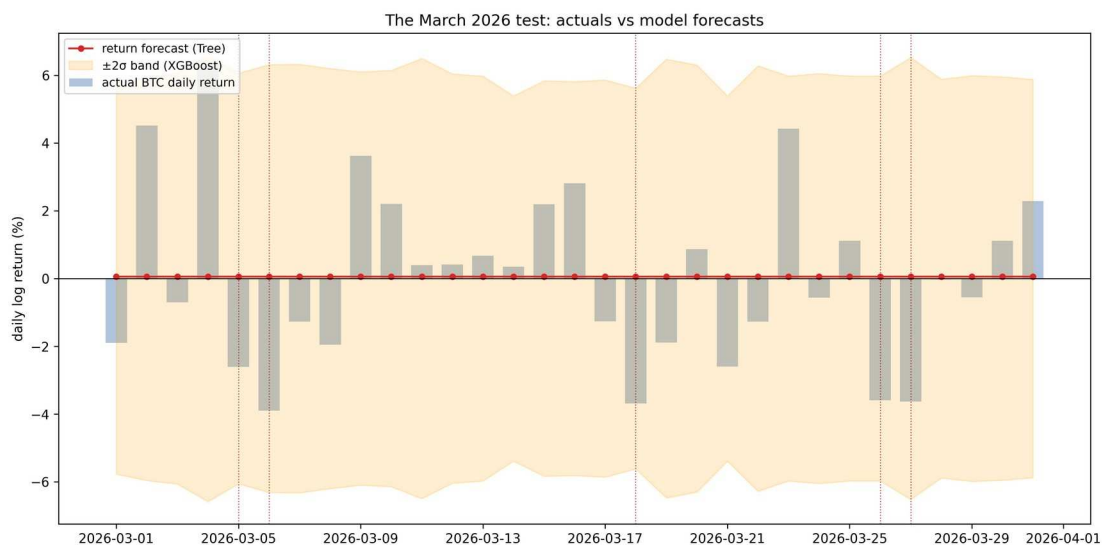


Figure 5.10 — The March 2026 test: actual daily BTC returns (bars), the champion return forecast (red), the $\pm 2\sigma$ band from the champion variance model (shaded), and the five realized crash days under the main definition (dotted lines). The champion classifier issues no high-confidence ($p \geq 0.5$) alert in the calm month. All 31 realized returns fall inside the band.

Figure 5.10 assembles the final forecast product. Its most defensible component is the band: the XGBoost variance forecasts convert into a $\pm 2\sigma$ envelope that contains all 31 March returns while being far from trivially wide — early-March widths near $\pm 6\%$ narrow visibly as the month calms. The return line through the middle is, appropriately, almost flat. If one image summarizes the thesis, it is this one: the market’s *level* tomorrow is anyone’s guess; its *temperature* is not.

5.5 Synthesis: white boxes, black boxes, and where each wins

Across the three tasks the comparison resolves into a pattern rather than a winner. On returns (Task A), regularization theory predicts and the data confirm a draw at zero: flexibility has nothing to buy. On variance (Task B), the black boxes win clearly in the stressed month — 25% QLIKE improvement over GARCH(1,1), with the direction confirmed in the calm month, where however the humble rolling 30-day variance posts the lowest loss of all — because volatility is genuinely predictable and non-linear, and because they ingest thirty-two conditioning variables instead of one. Yet the white box closes half the gap at a fraction of the opacity: adding a *single* interpretable stress regressor to GARCH buys a significant improvement in both months, with parameters an examiner can read (point estimate $\pi = 0.15$ variance points per stress unit, itself not

significant in sample). On crash risk (Task C), the random forest ranks risk best under both crash definitions — narrowly on the broad 10th-percentile target, decisively on the extreme tail (AUC 0.80 against 0.65 for the logistic) — while the logistic family remains the only model class that also *explains* its alerts in coefficient form. The interpretability–accuracy trade-off, in short, is real but narrower than advertised — and Chapter 6 shows that with the right tools, much of the black boxes’ reasoning can be read after the fact.

6. Interpretation: opening the black boxes

Predictive accuracy answers *whether* stablecoin stress helps; this chapter asks *how*. Four instruments are used: permutation importance (Section 6.1), partial-dependence curves (Section 6.2), a “scorecard” that confronts every stress variable with the verdict of every model family (Section 6.3), and real-time event backtests around the three historical crises, cross-checked against their day-by-day anatomy (Section 6.4). Section 6.5 draws the methodological moral, and Section 6.6 translates the findings into an operational monitoring design.

6.1 What the ensembles look at

Permutation importance measures how much a model’s out-of-sample performance deteriorates when one feature’s values are randomly shuffled — destroying its information while preserving its distribution. Unlike impurity-based importances, it is computed on the validation month and is comparable across models. Figure 6.1 reports the average over the two tree ensembles for Tasks A and C.

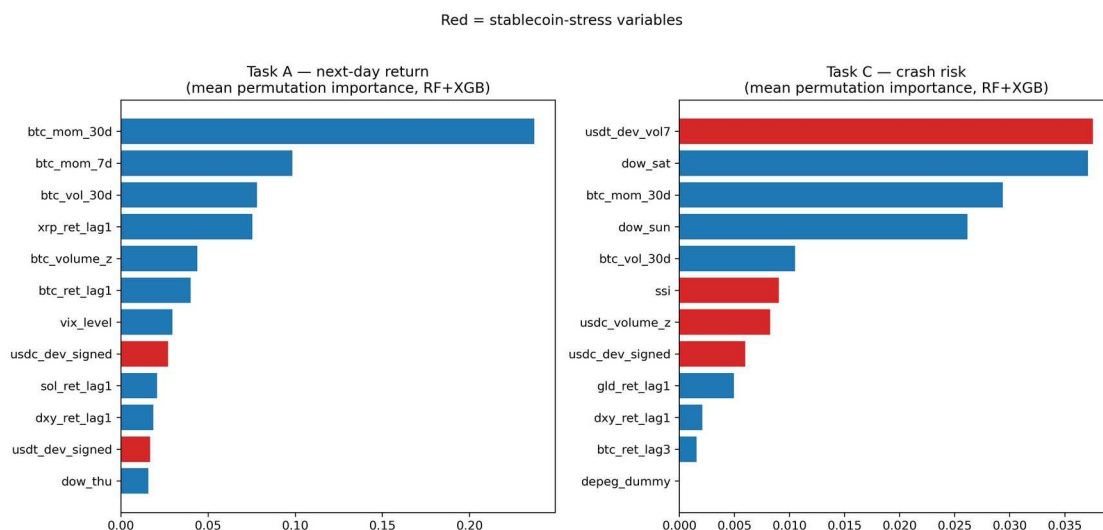


Figure 6.1 — Permutation importance on the validation month, averaged over random forest and XGBoost. Left: return task; right: crash-day task. Stablecoin-stress variables in red.

For the return task the ranking is dominated by the market’s own state: 30-day and 7-day momentum, 30-day volatility, cross-asset momentum (the lagged XRP return), and volume anomalies, with the first stress variable — the signed USDC deviation —

eighth. This is exactly the hierarchy one expects where predictable structure is weak: the ensembles cling to slow-moving conditioning variables. For the *crash* task the picture inverts: the single most important variable of all thirty-two is the **7-day volatility of the USDT peg deviation** — not the level of the peg but its churn: a peg that trembles, even within a basis point of par, is a peg under strain — with the SSI sixth, the USDC volume z-score seventh and the signed USDC deviation eighth. The two weekend dummies rank second and fourth, the econometric echo of crypto’s weekend liquidity holes, of which February 2026’s “Black Sunday” was only the latest example. Stress variables, marginal for direction, are central for danger — and what matters about them is their *instability*, not their level.

6.2 How the effects look

Partial-dependence curves trace the model’s average prediction as one feature is varied with the joint distribution of the others held fixed. Figure 6.2 (return task) and Figure 6.3 (crash task) show the four most instructive variables.

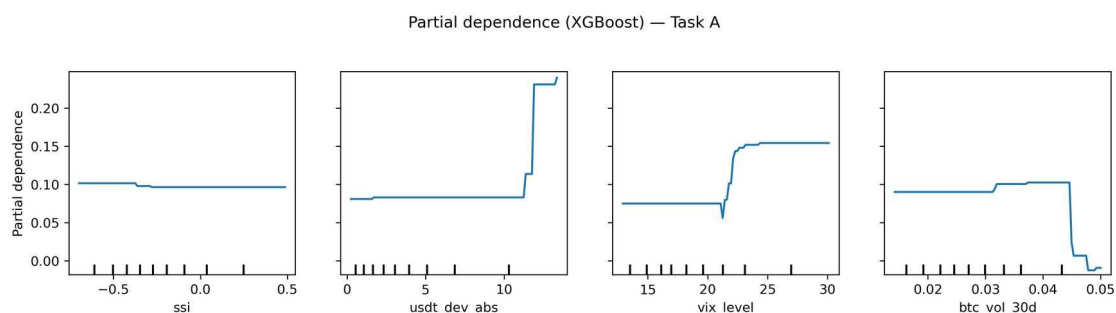


Figure 6.2 — Partial dependence of the predicted next-day return (XGBoost). Effects are economically tiny; note the flat SSI panel.

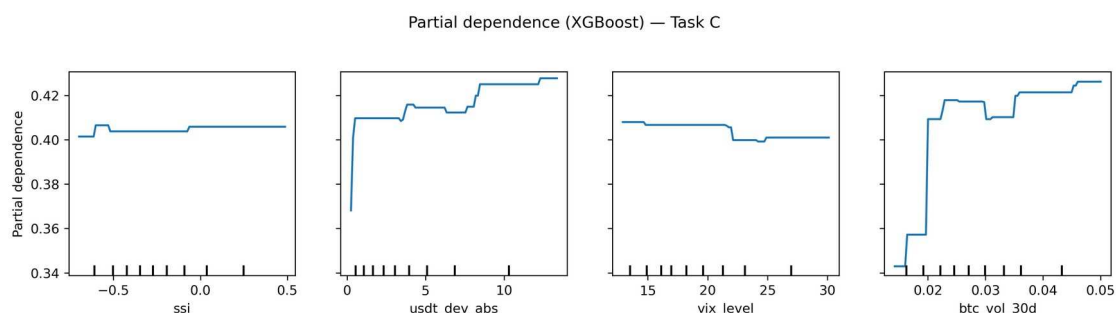


Figure 6.3 — Partial dependence of the predicted crash probability (XGBoost). The USDT absolute deviation shows the cleanest monotone risk effect.

On returns, the SSI panel is a horizontal line — the boosted trees, free to use it, choose not to: stress simply does not move the *expected* return. The mild effects that do exist are consistent with rebound dynamics (expected returns tick up when the USDT deviation or the VIX is extreme) and with momentum crashes (expected returns turn negative at the top of the trailing-volatility range). On crash probability the structure is stronger and economically legible: the probability rises monotonically in the USDT absolute deviation, with a sharp step around 4 basis points — the model has effectively learned a depeg threshold — and rises in trailing volatility. The composite SSI’s own partial curve is non-monotonic, a statistical artefact worth understanding: since its components (deviations, volumes) sit beside it in the feature set, the curve shows the effect of the *residual* index variation with components fixed, and the ensembles prefer the raw components. The index earns its keep as a parsimonious summary — one variable instead of six in the GARCH-X, where it improves out-of-sample forecasts — not as the dominant feature in models rich enough to consume the ingredients directly.

6.3 The stress-variable scorecard

Table 6.1 — Every stress variable under every lens. OLS: next-day return (% , HAC p-values); logit: crash probability (standardized coefficients, unweighted inferential fit); permutation ranks out of 32 features.

Variable	OLS coef. (p)	In Lasso?	Logit coef. (p)	In L1- logit?	Perm. rank A	Perm. rank C
SSI	−0.22 (0.75)	no	−0.24 (0.47)	no	14	6
USDT dev., signed	+0.025 (0.07)	no	+0.005 (0.96)	yes	11	25
USDT dev., abs.	+0.029 (0.27)	no	+0.07 (0.61)	yes	30	26
USDC dev., signed	−0.040 (0.03)	no	−0.07 (0.66)	no	8	8
USDC dev., abs.	−0.008 (0.79)	no	−0.05 (0.83)	no	28	20
USDT dev.	+0.000	no	+0.40	yes	25	1

vol. (7d)	(1.00)		(0.004)			
USDC dev.	+0.005	no	−0.45	yes	15	13
vol. (7d)	(0.84)		(0.12)			
USDC volume z	+0.073	no	+0.10	yes	27	7
	(0.70)		(0.60)			
Depeg dummy (> 50 bp)	−3.25	no	+0.18	no	20	12
	(0.55)		(0.24)			

Read row by row, the scorecard converges on three findings. First, the *signed* deviations are the only stress variables with (marginal) significance for returns, and their signs tell a coherent flow story: a USDC premium — capital paying up for the regulated safe asset — predicts a *lower* Bitcoin return tomorrow (-0.040 , $p = 0.03$), while a USDT premium — capital queuing to enter crypto through the offshore on-ramp — predicts a *higher* one ($+0.025$, $p = 0.07$). The two coins are not interchangeable sensors; they sit on opposite sides of the flight-to-safety corridor. Second, for crash risk the *second moments* of the plumbing beat its levels: the only stress variable significant in the crash logit is the 7-day volatility of the USDT deviation ($+0.40$, $p = 0.004$) — one standardized unit of peg churn raises the crash log-odds by 0.4, roughly a fifty-percent increase in the odds — and the same variable tops the ensembles’ permutation ranking, while the USDC volume anomaly sits just behind (rank 7; its starring role in the FTX anatomy of Section 6.4 shows *when* it leads). The deviation *levels*, by contrast, are insignificant everywhere: peg prices move last — churn and turnover move first. Third, the Lasso column is uniformly “no” while the cross-validated L1-logit retains 22 of the 32 variables, including the deviation-volatility pair: regularization itself distinguishes the task with signal from the task without, and within the crash task it keeps precisely the second-moment stress variables. One instructive exclusion: the L1-logit drops the SSI while keeping its most informative components — the same partialling logic as Figure 6.3, where the index is redundant *given* its ingredients; in the GARCH-X, where it enters alone, the index is exactly what earns its keep.

Two clarifications discipline the reading of Table 6.1. The first concerns *magnitudes*. With returns in percent and deviations in basis points, the USDC coefficient of -0.040 means that a ten-basis-point USDC premium maps into about -0.4 percentage points of

expected next-day BTC return, and the USDT coefficient into +0.25 points for the same premium — against a residual daily standard deviation of roughly 3%, effects of about a tenth of a typical move: economically visible for a levered position, far too small to trade on after costs, and estimated with HAC confidence intervals (Appendix B) wide enough to admit effects half or twice as large. The second concerns the *logit column*. The coefficients and p-values reported here come from the unweighted maximum-likelihood Logit estimated for inference; the predictive classifier of Chapter 5 uses balanced class weights, which rescale the intercept and tilt the fit toward the rare class. The two objects share the feature set but not the estimation objective, so the table’s p-values describe the statistical association between each variable and crash days — they are not inference on the internal parameters of the deployed alert engine.

6.4 Could the models have seen the crises coming?

The sharpest test of an early-warning claim is counterfactual deployment. For each of the three major episodes we re-estimate the crash classifier *using only data ending before the event window*, then let it produce daily crash probabilities through the crisis — exactly what a risk desk running the model would have seen.

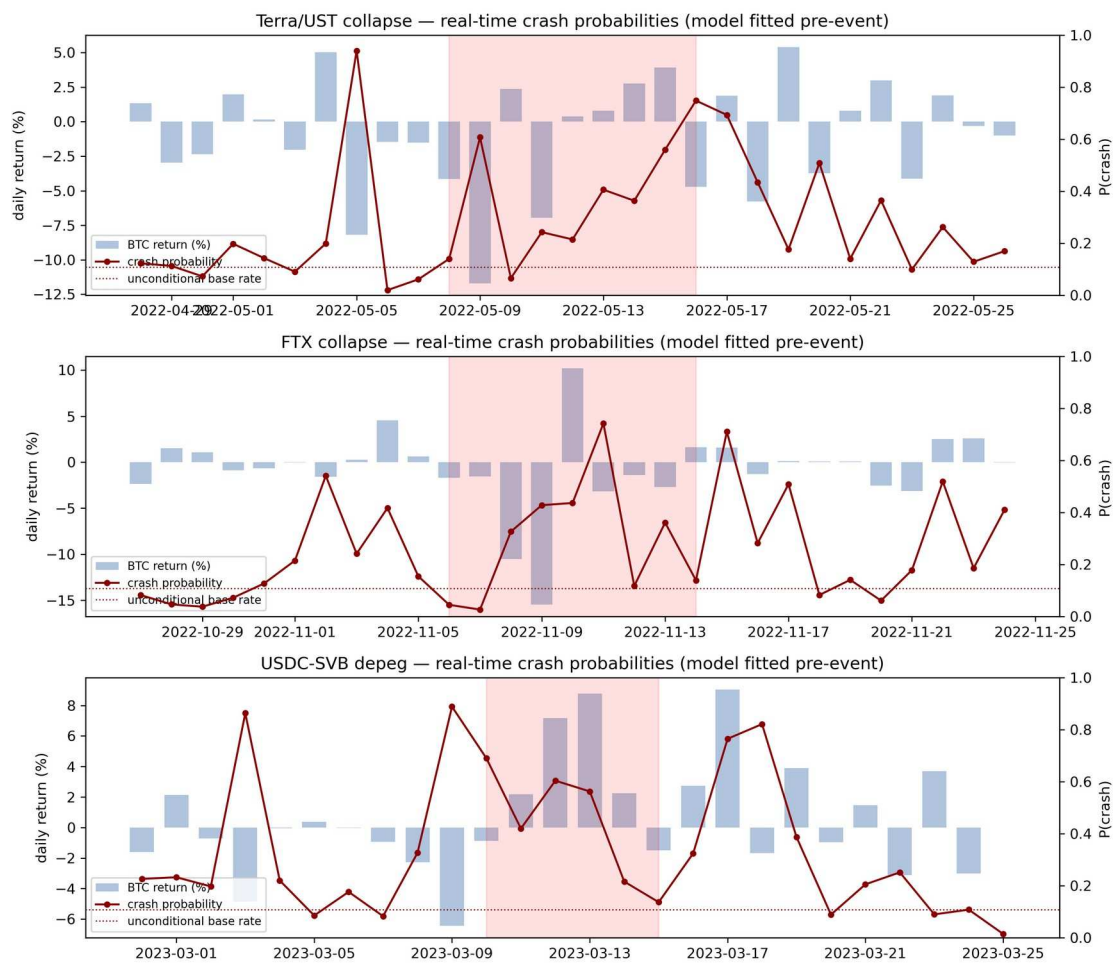


Figure 6.4 — Real-time crash probabilities (dark red, right scale) from classifiers trained only on pre-event data, with daily BTC returns (bars, left scale). Top: Terra/UST, May 2022; middle: FTX, November 2022; bottom: USDC–SVB, March 2023. Dotted line: unconditional base rate.

The day-by-day anatomy of the three windows, reconstructed in Tables 6.2–6.4 from the modelling dataset, lets the narrative be checked against the raw numbers.

Table 6.2 — Terra/UST collapse, day by day (selected columns of the modelling dataset).

Date	SSI	dev USDT (bp)	dev USDC (bp)	USDC volume z	BTC ret. next day
2022-05-05	0.05	0.1	0.4	+1.5	–1.5%
2022-05-06	0.07	0.2	0.7	+1.8	–1.5%
2022-05-07	–0.47	0.8	0.9	–0.2	–4.2%
2022-05-08	0.01	1.5	0.9	+1.7	–11.7%

2022-05-09	0.69	0.9	4.1	+3.7	+2.4%
2022-05-10	0.52	2.1	1.7	+3.1	−7.0%
2022-05-11	1.64	41.3	7.7	+4.3	+0.4%
2022-05-12	1.21	23.9	6.6	+3.8	+0.8%
2022-05-13	0.36	17.6	3.5	+0.8	+2.8%

Table 6.3 — FTX collapse, day by day. Note the sequence: on 8 November the volume anomaly spikes (+3.9 σ) while both pegs still sit within one basis point of par; the deviations blow out only on 9–10 November.

Date	SSI	dev USDT (bp)	dev USDC (bp)	USDC volume z	BTC ret. next day
2022-11-04	0.02	0.8	0.4	+0.6	+0.6%
2022-11-05	−0.40	1.2	1.0	−0.7	−1.7%
2022-11-06	−0.50	0.9	1.3	−1.0	−1.6%
2022-11-07	−0.13	0.9	3.2	−0.0	−10.5%
2022-11-08	0.76	0.3	0.2	+3.9	−15.5%
2022-11-09	0.87	14.0	3.4	+3.5	+10.2%
2022-11-10	0.84	21.7	6.9	+2.4	−3.2%
2022-11-11	0.33	17.8	7.3	+0.2	−1.4%

Table 6.4 — USDC–SVB depeg, day by day. The SSI’s sample maximum (9.32) coincides with USDC’s 285 bp discount and USDT’s simultaneous premium; the days after the peak deliver large positive returns as the deposit guarantee reverses the panic.

Date	SSI	dev USDT (bp)	dev USDC (bp)	USDC volume z	BTC ret. next day
2023-03-08	−0.47	0.1	1.3	−0.4	−6.4%
2023-03-09	−0.27	0.1	0.1	+0.2	−0.9%
2023-03-10	1.19	30.5	5.2	+3.5	+2.2%

2023-03-11	9.32	76.9	285.0	+5.1	+7.2%
2023-03-12	5.21	71.5	79.3	+1.2	+8.8%
2023-03-13	3.71	38.7	10.5	+1.0	+2.2%
2023-03-14	3.28	32.5	8.2	+0.6	−1.5%
2023-03-15	2.91	39.5	1.8	+0.5	+2.7%

Read together, the tables document the theoretical sequence of Chapter 2 with unusual clarity. In the FTX window the single most informative row is 8 November: BTC has just fallen 10.5% and is about to fall 15.5% more, both pegs are still intact to within a basis point — and USDC turnover is running almost four standard deviations above normal. Volume moves first; price moves later; and a classifier trained on the pre-event sample, in which that same volume variable already ranked highly, raises its alarm on exactly that day. The Terra window shows the slower variant: volume anomalies build from 8–9 May while the USDT deviation grows from one basis point to 41 over four days — a run in slow motion. And the SVB window shows the limiting case in which the stress is so acute that every symptom fires at once, followed by the sharpest *positive* returns in the tables: confirmation that extreme stablecoin stress forecasts turbulence rather than direction, and that the second leg of a depeg episode is as likely to be a violent relief rally as a continued collapse.

The backtests, shown in Figure 6.4, are the thesis’s most striking evidence. Around **Terra**, the model trained through April 2022 pushes the crash probability to roughly 0.95 on 5 May — as stablecoin volumes and deviation volatility build, three days before the collapse window — lets it fall back as the first leg plays out, and climbs again to about 0.75 as the run reaches its climax in mid-May. Around **FTX**, whose stablecoin signature Chapter 3 showed to be faint, the pre-event model is appropriately quiet in early November (probabilities near 0.03 on 5–6 November), then lifts its alarm roughly tenfold on 8 November — the volume-anomaly day dissected above — and peaks near 0.75 on 11 November, the day of the bankruptcy filing: late relative to Terra, exactly as the weaker stablecoin footprint would predict, and a useful reminder that this sensor detects stablecoin-mediated stress, not all stress. Around **SVB**, the model’s probability reaches roughly 0.9 on 9 March 2023 — *the eve of the depeg window* — driven by the very churn-and-turnover variables that Section 6.3 identified, and relaxes through the

window as the deposit guarantee restores the peg. Three episodes are three anecdotes, not a rejection region; but the anecdotes are the right ones, they are out of sample by construction, and they are produced by the same fixed pipeline as every other number in the thesis.

6.5 The trade-off, reconsidered

The chapter’s tools blur the boundary the thesis started from. The black boxes turned out to be *auditable*: permutation importance revealed a coherent variable hierarchy, partial dependence revealed threshold effects an economist can name, and the event backtests turned the classifier into a narrative any risk committee could follow. The white boxes turned out to be *upgradable*: one stress regressor inside GARCH captured a significant share of what the ensembles exploit, at zero cost in transparency.

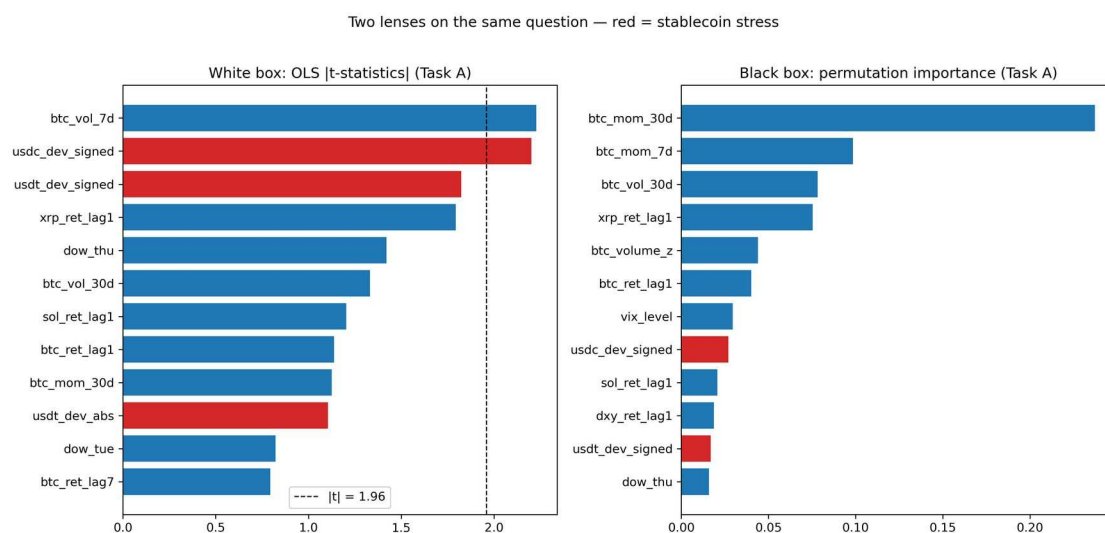


Figure 6.5 — Two lenses on the same question: OLS $|t\text{-statistics}|$ (left, white box) against permutation importance (right, black box) for the return task; stablecoin-stress variables in red. The two rankings disagree exactly where inference and prediction measure different things.

Figure 6.5 closes the loop by juxtaposing the two families’ native notions of “what matters” for the same task. The white box ranks variables by inferential significance — signed deviations and short-window volatility clear the $|t| = 1.96$ bar — while the black box ranks them by predictive leverage, crowning slow momentum and volatility windows that the OLS t -statistics treat as unremarkable. The disagreement is not a contradiction: a variable can be marginally significant yet contribute nothing recursively

partitionable, and a variable can drive predictions while its linear projection is buried by collinearity. Reading both columns together is more informative than either alone, which is perhaps the most practical lesson of the comparison.

What remains genuinely black is not the fitted models but the *selection* process — the recognition, quantified in Chapter 5, that a validation month can crown a lucky champion in a signal-poor task. The practical synthesis for a market-risk user is therefore layered: a GARCH-X for the volatility core, because it is defensible line by line; an ensemble overlay for tail-risk ranking, because it demonstrably ranks danger better; and the SSI dashboard underneath both, because the stablecoin-mediated crises of the sample — Terra, FTX, SVB — announced themselves there first (the macro-driven selloffs did not, and a three-event sample warrants humility about how general the pattern is).

6.6 Using the results: a stablecoin-stress monitor in practice

The thesis’s findings assemble naturally into an operational monitoring layer, and spelling it out makes the economic content concrete. The base of the stack is the SSI itself: six public inputs, updated daily at negligible cost, with the one-day-ahead reading of Section 3.8 defining its use — a top-decile print is a same-week warning, not a forecast of a persistent regime. The middle layer is the GARCH-X: a risk desk sizing positions or setting margins needs tomorrow’s variance, the model delivers it with a single interpretable stress loading ($\pi=0.15$ variance points per unit of stress), and its clean residual diagnostics mean the number can be defended before a model-validation committee — the institutional constraint that still excludes fully opaque models from regulatory capital pipelines. The top layer is the ensemble classifier run at the F1-optimal threshold of Section 5.3, which in the stressed validation month would have caught four of the five crash days — all of the early-February cluster — at the best F1 of any model, concentrating its false alarms in the days when the market was in fact fragile; its alerts are auditable after the fact through the permutation and partial-dependence toolkit, with peg-deviation churn and the USDC volume anomaly as the headline triggers. Two honest caveats bound the deployment. First, the monitor detects *stablecoin-mediated* fragility: a crash arriving through channels that leave the settlement layer untouched — the macro-driven February 2026 selloff being the archetype — will

be flagged late or not at all, so the stack complements rather than replaces macro risk management. Second, all thresholds inherit the estimation sample's crisis mix; a structural change in stablecoin regulation or market microstructure (Section 2.4) warrants re-estimation before the alerts are trusted. Within those limits, the monitor operationalizes the thesis in one sentence: watch the plumbing's turnover, not just its prices, and size risk — not direction — accordingly.

7. Conclusions

7.1 Answers to the research questions

Does stablecoin stress predict cryptocurrency market behaviour? Yes — but selectively, and the selectivity is the finding. Stress in the USDT and USDC pegs carries no exploitable information about the *direction* of tomorrow’s Bitcoin return: in a market efficient enough at the daily horizon, neither peg deviations nor anything else in our thirty-two-variable information set moves the conditional mean by more than noise. It does carry information about *risk*. A single interpretable stress index added to the variance equation of a GARCH model improves volatility forecasts significantly in two consecutive out-of-sample months — one stressed, one calm — and stablecoin variables — above all the churn of the USDT peg and abnormal USDC volume — rank at the top of what the best crash-day classifiers use. Classifiers deployed counterfactually before Terra, FTX and SVB produce elevated real-time alert probabilities in all three windows — earliest and loudest for the two crises that were genuinely stablecoin crises — though three episodes are a basis for confidence, not proof, that the pattern generalizes. The economic reading is coherent with the institutional analysis of Chapter 2: because stablecoins are the market’s settlement and safe-haven layer, strain in that layer — priced in deviations, but signalled first in volumes — is a barometer of systemic fragility, not a directional signal.

What do black-box models add over white-box models? Where there is no signal (returns), nothing — and the regularized white boxes are the most honest about it, the Lasso electing to predict a constant. Where there is signal (volatility, tail risk), the ensembles add real accuracy: a 25% QLIKE improvement over GARCH(1,1) in the stressed validation month and the best crash-day ranking on every available measure. Yet the trade-off proved narrower than its reputation. Half of the ensembles’ volatility advantage was recovered by one transparent modification of the classical model (the GARCH-X), and the interpretation toolkit of Chapter 6 — permutation importance, partial dependence, event backtests — rendered the ensembles’ behaviour auditable after the fact. The residual, irreducible cost of the black boxes is not opacity of mechanism but fragility of *selection*: our validation-month champion for returns was a

lucky tree whose advantage vanished in the test month, a caution that applies to any model horse race run on short evaluation windows.

7.2 Limitations

Honesty about scope is part of the method. First, the stress measures are built from *daily closing* prices of two stablecoins on one composite data source; intraday extremes (USDC’s 88 cents) and newer synthetic designs (USDe, whose 65-cent print on 10 October 2025 sits outside our universe) are invisible to the index, which is therefore a conservative sensor. Second, the out-of-sample evidence, however cleanly protocolled, spans two months and three event studies: effect *directions* replicate across them, but magnitudes carry wide bands, and the Task C test month contains only five positive observations under the main definition (one under the extreme one). Third, the analysis is predictive, not causal: stablecoin volume anomalies may anticipate crashes because they cause them, because they reflect informed flight, or because both respond to a common shock — our design cannot separate these channels. Fourth, all results concern Bitcoin; the framework transfers to other cryptoassets, but the estimates do not automatically. Fifth, forecast value is measured in statistical loss functions; a trading or risk-budgeting evaluation net of costs is a different exercise, deliberately left outside the thesis.

7.3 Future research

Five extensions follow naturally. Intraday data would let the stress index capture the within-day depegs that daily closes smooth away, and would multiply the effective sample of stress episodes. Broadening the stablecoin universe — euro stablecoins, DAI, and the synthetic-dollar designs whose fragility the October 2025 episode exposed — would test whether the periphery of the stablecoin system is an even earlier warning site than its core. On-chain data (redemption queues, mint/burn flows, bridge activity) would replace our volume proxies with the primary-market quantities they approximate. Methodologically, the GARCH-X invites a multivariate treatment — stablecoin stress in the covariance dynamics of several cryptoassets — and the classification task invites proper scoring-rule evaluation and utility-based thresholds tied to a risk manager’s loss function. Finally, as the GENIUS Act and MiCA reshape the industry, the structural

stability of every relationship estimated here becomes a research question in its own right: a regulated stablecoin system may be one in which the barometer, at last, goes quiet.

7.4 Policy postscript

The results also carry a modest message for the regulatory debate. Supervisors designing stablecoin oversight have concentrated, reasonably, on reserve quality and redemption rights — the determinants of whether a peg *should* hold. The evidence here suggests a complementary, cheaply implementable monitoring dimension: the joint behaviour of secondary-market turnover and peg deviations is an early symptom of system-wide deleveraging, informative one day ahead of realized turbulence and, historically, in advance of the worst days of every stablecoin-mediated crisis in the sample. A supervisor needs nothing beyond publicly available data to compute it — the inputs are exchange prices and volumes — and the GARCH-X evidence shows the signal survives after conditioning on the market’s own volatility history. At the same time, the February 2026 episode is the cautionary counterweight: a fully macro-driven crash left the stablecoin layer quiet, so stablecoin-stress monitoring identifies one fragility channel among several, not a general crisis thermometer. The policy reading, like the statistical one, is about the variance equation rather than the mean: stablecoin stress tells the system how hard the next shock will hit, not when or from where it will come.

7.5 Closing remark

The thesis set out to measure whether the crypto market’s plumbing carries predictive information. The answer fits in one sentence: stablecoin stress does not tell us where the market is going, but it tells us — significantly, out of sample, and through every major crisis of the last six years — how dangerous the journey is about to be.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69(3), 683–734.
- Ante, L., Fiedler, I., & Strehle, E. (2021). The influence of stablecoin issuances on cryptocurrency markets. *Finance Research Letters*, 41, 101867.
- Arner, D. W., Auer, R., & Frost, J. (2020). Stablecoins: risks, potential and regulation. *BIS Working Papers*, No. 905, Bank for International Settlements.
- Baur, D. G., & Hoang, L. T. (2021). A crypto safe haven against Bitcoin. *Finance Research Letters*, 38, 101431.
- Belardi, A. (2026). *Stablecoin stress and cryptocurrency markets: replication archive* [Python code and data]. Università di Pavia. Available at: <https://github.com/antoniobelardi7-lang/stablecoin-stress-thesis>
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.
- Borri, N. (2019). Conditional tail-risk in cryptocurrency markets. *Journal of Empirical Finance*, 50, 1–19.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4), 1509–1531.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

- Conrad, C., Custovic, A., & Ghysels, E. (2018). Long- and short-term cryptocurrency volatility components: A GARCH-MIDAS analysis. *Journal of Risk and Financial Management*, 11(2), 23.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
- Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5), 1779–1801.
- Gorton, G. B., & Zhang, J. (2023). Taming wildcat stablecoins. *University of Chicago Law Review*, 90(3), 909–971.
- Griffin, J. M., & Shams, A. (2020). Is Bitcoin really untethered? *Journal of Finance*, 75(4), 1913–1964.
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.
- Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B*, 41(2), 190–195.
- Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7), 873–889.
- Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291.

- Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3), 297–310.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in Python*. Springer.
- Jaquart, P., Dann, D., & Weinhardt, C. (2021). Short-term bitcoin market prediction via machine learning. *Journal of Finance and Data Science*, 7, 45–66.
- Katsiampa, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters*, 158, 3–6.
- Kristoufek, L. (2021). Tethered, or untethered? On the interplay between stablecoins and major cryptoassets. *Finance Research Letters*, 43, 101991.
- Liu, Y., & Tsyvinski, A. (2021). Risks and returns of cryptocurrency. *Review of Financial Studies*, 34(6), 2689–2727.
- Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297–303.
- Lyons, R. K., & Viswanath-Natraj, G. (2023). What keeps stablecoins stable? *Journal of International Money and Finance*, 131, 102777.
- Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2), 293–319.
- Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2), 347–370.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398), 605–610.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1), 267–288.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4), 1455–1508.

Institutional and market sources. Congress.gov, S.1582 — GENIUS Act (119th Congress, signed 18 July 2025); European Union, Markets in Crypto-Assets Regulation (EU) 2023/1114 (stablecoin provisions applicable 30 June 2024); DefiLlama and CoinGecko stablecoin market-capitalization statistics (accessed June 2026); CoinGecko, “What is October 10th? Crypto’s 10/10 mass market liquidation event” (2025); Bloomberg, “Bitcoin selloff explained” (6 February 2026); VanEck Digital Assets Research, “What triggered Bitcoin’s major selloff in February 2026?” (2026). Market data: Yahoo Finance.

Appendix A — The software pipeline

A.1 Reproducibility

Every table and figure in the thesis is generated by a Python pipeline of ten numbered scripts. From a clean machine, the entire thesis reproduces by installing the pinned dependencies (`pip install -r requirements.txt`) and running the single driver script `run_all.py`, which executes the ten steps in order.

Randomness (bootstrap resampling, forest construction, network initialization) is controlled by a single global seed (42) set in `config.py`, which also fixes every date, ticker, window length and threshold used in the analysis; library versions are pinned in `requirements.txt` (pandas, NumPy, statsmodels, arch, scikit-learn, XGBoost, SciPy, matplotlib).

Table A.1 — The pipeline. Model predictions are stored per model and split, so all comparisons and tests in Chapters 5–6 are computed from the same prediction files.

Script	Role	Thesis material produced
00_download_data.py	Yahoo Finance download, integrity checks	—
01_preprocessing.py	Cleaning, alignment, features, SSI, VIF screening, splits	Tables 3.2 and 3.6
02_eda.py	Descriptive statistics, stationarity tests, exploratory figures	Figures 3.1–3.7; Tables 3.3–3.5
03_task_A_whitebox.py	Benchmarks, OLS-HAC, Ridge, Lasso, ARIMAX (AIC/BIC/HQC), GAM, Ljung–Box tests	Figures 5.1–5.2; Tables 5.1 (part), 5.2
04_task_A_blackbox.py	Tree, bagging, forest, XGBoost, SVR, MLP	Figure 5.3; Table 5.1 (part)

05_task_B_volatility.py	GARCH family, own GARCH-X MLE, ML on variance, residual diagnostics	Figures 5.6–5.7; Tables 5.3–5.6
06_task_C_classification.py	Logistic, LDA, L1-logit, tree ensembles, SVC, MLP, F1-optimal thresholds, 5th-percentile robustness	Figures 5.8–5.9; Tables 5.7, 5.8, 5.8b
07_comparison_forecasts.py	Leaderboards, Diebold– Mariano tests, OOS plots, March 2026 run	Figures 5.4–5.5, 5.10; Table 5.9, DM results
08_interpretation.py	Permutation importance, partial dependence, scorecard, event backtests	Figures 6.1–6.5; Table 6.1
09_extra_descriptives.py	SSI horizon analysis, weekday tail risk, event anatomies, correlation ranking	Tables 3.7–3.8, 6.2–6.4, B.9

In addition to the numbered scripts, three Jupyter notebooks — `notebook_task_A_returns.ipynb`, `notebook_task_B_volatility.ipynb` and `notebook_task_C_classification.ipynb` — present the complete sequential workflow of each prediction task (code, diagnostics, tables, metrics and figures inline) and reproduce the headline results of Chapter 5 interactively.

The two components written specifically for this thesis are documented in the main text rather than reproduced here: the Stablecoin Stress Index, whose construction from six expanding-window z-scores is defined in Section 3.5, and the GARCH-X likelihood, whose filter, Gaussian objective, parameter bounds and out-of-sample filtering rule are specified in Section 4.3.6. Both are implemented in `01_preprocessing.py` and `05_task_B_volatility.py` respectively, in the accompanying code.

A.2 Data and outputs

Raw data (nine CSV series), the processed modelling dataset (1,979 rows $\times$ 32 features + 4 targets), all prediction files, and every generated table (CSV) and figure (PNG, 300 dpi) are organized under data/ and outputs/ in the code archive accompanying the thesis. The mapping from each thesis exhibit to its generating script in the table above makes every reported number auditable.

The complete archive — the ten scripts, the three notebooks, the pinned requirements.txt, and the full data/ and outputs/ trees — is deposited with this thesis and publicly available in the repository (Belardi, 2026):

<https://github.com/antoniobelardi7-lang/stablecoin-stress-thesis>

Appendix B — Complete results tables

This appendix collects, unabridged, the result tables that Chapter 5 reports in condensed form, plus the feature dictionary and the collinearity map referenced in Chapter 3. Every table is written verbatim by the pipeline of Appendix A.

B.1 Task A — full leaderboard, validation and test

Table B.1 — Task A (next-day BTC return, %): all models, validation (Feb 2026) and test (Mar 2026).

model	val MSE	val R2	val	test	test R2	test
		zero	DirAcc	MSE	zero	DirAcc
Tree	17.5312	0.0563	0.4643	6.5965	0.0005	0.4839
XGBoost	18.0028	0.0309	0.5357	6.6713	-0.0108	0.4194
GAM_splines	18.0417	0.0288	0.5357	6.674	-0.0112	0.4839
OLS	18.3617	0.0116	0.5714	6.6531	-0.008	0.6129
Ridge	18.4513	0.0068	0.5	6.6237	-0.0036	0.5484
bench_AR1	18.46	0.0063	0.3929	6.6044	-0.0007	0.5484
MLP	18.5217	0.003	0.5357	6.8487	-0.0377	0.3548
bench_zero	18.5771	0.0	0.0	6.6	0.0	0.0
Bagging	18.633	-0.003	0.5	6.8264	-0.0343	0.4194
Lasso	18.6907	-0.0061	0.3571	6.5976	0.0004	0.4839
bench_mean	18.6907	-0.0061	0.3571	6.5976	0.0004	0.4839
ARIMAX	18.751	-0.0094	0.4286	6.7002	-0.0152	0.5484
SVR	18.8825	-0.0164	0.3929	6.5822	0.0027	0.4194
RandomForest	18.9098	-0.0179	0.5714	6.7268	-0.0192	0.4839

B.2 Task B — full leaderboard, validation and test

Table B.2 — Task B (next-day realized-variance proxy, %²): all models, both out-of-sample months.

model	val MSE	val QLIKE	test MSE	test QLIKE
XGBoost	2168.5852	2.0462	74.5343	1.2036

RandomForest	2262.1931	2.0613	74.5471	1.1951
bench_const	2443.314	2.5885	78.7398	1.2288
EGARCH	2465.4763	2.6167	103.1905	1.3239
GARCH-X_ssi	2467.9328	2.6202	75.2754	1.2014
GJR-GARCH	2498.7628	2.6322	88.1935	1.2648
GARCH11	2509.1975	2.713	88.8439	1.2676
bench_roll30	2600.1375	3.3756	78.3131	1.1694

B.3 Task C — full leaderboard, validation and test

Table B.3 — Task C (large-drawdown days, main 10th-percentile target): all models, both out-of-sample months. The 5th-percentile robustness results are in Table 5.8b.

model	val AUC	val error	val F1	test		
				AUC	test error	test F1
RandomForest	0.713	0.2857	0	0.5923	0.1613	0
SVC	0.6783	0.1786	0	0.8077	0.1613	0
LDA	0.6696	0.1786	0	0.5769	0.1613	0
Logistic	0.6435	0.5357	0.2857	0.6231	0.2903	0.3077
LassoLogistic	0.6261	0.4643	0.3158	0.6154	0.3548	0.2667
XGBoost	0.6174	0.3571	0.375	0.5462	0.3548	0
Tree	0.613	0.4643	0.3158	0.6308	0.4839	0.2857
bench_baserate	0.5	0.1786	0	0.5	0.1613	0
MLP	0.313	0.1786	0	0.5692	0.1613	0

B.4 Diebold–Mariano tests — complete results

Table B.4 — All Diebold–Mariano tests. Task A: squared-error loss vs the zero forecast; Task B: QLIKE loss vs GARCH(1,1). Negative statistics favour the model.

Task	Model	Ref.	Split	Loss	DM	p
A	ARIMA	zero	val	squared	0.434	0.6676
	X			error		
A	ARIMA	zero	test	squared	0.763	0.4511

	X			error		
A	Bagging	zero	val	squared error	0.075	0.9411
A	Bagging	zero	test	squared error	0.832	0.4121
A	GAM_s plines	zero	val	squared error	-0.904	0.3739
A	GAM_s plines	zero	test	squared error	0.434	0.6677
A	Lasso	zero	val	squared error	1.142	0.2636
A	Lasso	zero	test	squared error	-0.029	0.9767
A	MLP	zero	val	squared error	-0.227	0.822
A	MLP	zero	test	squared error	1.484	0.1481
A	OLS	zero	val	squared error	-0.44	0.6637
A	OLS	zero	test	squared error	0.219	0.828
A	Random Forest	zero	val	squared error	0.713	0.482
A	Random Forest	zero	test	squared error	1.075	0.2911
A	Ridge	zero	val	squared error	-0.531	0.5998
A	Ridge	zero	test	squared error	0.23	0.8196
A	SVR	zero	val	squared	1.09	0.2852

				error		
A	SVR	zero	test	squared error	-0.141	0.8889
A	Tree	zero	val	squared error	-0.901	0.3757
A	Tree	zero	test	squared error	-0.069	0.9455
A	XGBoos t	zero	val	squared error	-0.737	0.4672
A	XGBoos t	zero	test	squared error	1.814	0.0797
B	EGARCH H	GARCH 11	val	QLIKE	-2.254	0.0325
B	EGARCH H	GARCH 11	test	QLIKE	4.094	0.0003
B	GARCH -X_ssi	GARCH 11	val	QLIKE	-2.656	0.0131
B	GARCH -X_ssi	GARCH 11	test	QLIKE	-2.212	0.0347
B	GJR- GARCH	GARCH 11	val	QLIKE	-0.785	0.4395
B	GJR- GARCH	GARCH 11	test	QLIKE	-1.491	0.1463
B	Random Forest	GARCH 11	val	QLIKE	-1.98	0.058
B	Random Forest	GARCH 11	test	QLIKE	-1.621	0.1154
B	XGBoos t	GARCH 11	val	QLIKE	-2.171	0.0389
B	XGBoos	GARCH	test	QLIKE	-2.036	0.0506

	t	11				
B	bench_c	GARCH	val	QLIKE	-0.502	0.6201
	onst	11				
B	bench_c	GARCH	test	QLIKE	-1.135	0.2652
	onst	11				
B	bench_r	GARCH	val	QLIKE	1.086	0.2871
	oll30	11				
B	bench_r	GARCH	test	QLIKE	-1.901	0.067
	oll30	11				

B.5 Ljung–Box diagnostics — conditional-mean models

Table B.5 — Ljung–Box tests on conditional-mean residuals (training window). H_0 : no autocorrelation.

model	lag	LB stat	p value	no autocorr 5pct
OLS residuals	7	0.68	0.9985	yes
OLS residuals	14	14.29	0.4282	yes
OLS residuals	21	22.89	0.3499	yes
ARIMAX residuals	7	4.09	0.665	yes
ARIMAX residuals	14	19.96	0.0961	yes
ARIMAX residuals	21	25.6	0.1796	yes

B.6 Ljung–Box diagnostics — volatility models

Table B.6 — Ljung–Box tests on standardized and squared standardized residuals of the volatility models (training window).

model	series	lag	LB stat	p value	no autocorr 5pct
GARCH11	std. residuals	7	3.12	0.8733	yes
GARCH11	std.	14	14.01	0.4489	yes

	residuals				
GARCH11	std.	21	19.0	0.5853	yes
	residuals				
GARCH11	squared	7	13.97	0.0517	yes
	std.				
	residuals				
GARCH11	squared	14	22.77	0.0642	yes
	std.				
	residuals				
GARCH11	squared	21	34.81	0.0297	no
	std.				
	residuals				
GJR-GARCH	std.	7	3.45	0.84	yes
	residuals				
GJR-GARCH	std.	14	14.07	0.4443	yes
	residuals				
GJR-GARCH	std.	21	18.94	0.5888	yes
	residuals				
GJR-GARCH	squared	7	12.7	0.0797	yes
	std.				
	residuals				
GJR-GARCH	squared	14	21.16	0.0977	yes
	std.				
	residuals				
GJR-GARCH	squared	21	32.53	0.0517	yes
	std.				
	residuals				
EGARCH	std.	7	3.66	0.8184	yes
	residuals				
EGARCH	std.	14	13.83	0.4628	yes

	residuals				
EGARCH	std.	21	18.31	0.6295	yes
	residuals				
EGARCH	squared	7	21.1	0.0036	no
	std.				
	residuals				
EGARCH	squared	14	31.87	0.0042	no
	std.				
	residuals				
EGARCH	squared	21	44.13	0.0022	no
	std.				
	residuals				
GARCH-X (own MLE)	std.	7	2.93	0.8912	yes
	residuals				
GARCH-X (own MLE)	std.	14	14.06	0.4454	yes
	residuals				
GARCH-X (own MLE)	std.	21	19.29	0.5667	yes
	residuals				
GARCH-X (own MLE)	squared	7	9.68	0.2075	yes
	std.				
	residuals				
GARCH-X (own MLE)	squared	14	18.98	0.1655	yes
	std.				
	residuals				
GARCH-X (own MLE)	squared	21	31.47	0.0661	yes
	std.				
	residuals				

B.7 ARIMAX order selection — full criteria grid

Table B.7 — Information criteria for every (p, q) candidate. AIC prefers (2,1); BIC and HQC prefer (1,0); the majority rule selects (1,0).

p	q	AIC	BIC	HQC	chosen
0	1	9755.6	9805.7	9774.1	
0	2	9756.2	9811.8	9776.7	
1	0	9755.6	9805.6	9774.0	<— selected
1	1	9756.8	9812.4	9777.3	
1	2	9753.5	9814.7	9776.0	
2	0	9756.3	9811.9	9776.7	
2	1	9753.4	9814.6	9775.9	
2	2	9755.1	9821.8	9779.6	

B.8 The feature dictionary

Table B.8 — The 32 model features: construction and interpretation. All quantities computed with information available at day t .

Variable(s)	Definition	Meaning
usdt_dev_signed / usdc_dev_signed	$(P - 1) \times 10,000$	Signed peg deviation, basis points
usdt_dev_abs / usdc_dev_abs		$P - 1$
usdt_dev_vol7 / usdc_dev_vol7	rolling std, 7 days	Peg-deviation volatility
usdc_volume_z	$(V - \text{mean}_{30})/\text{std}_{30}$	USDC volume anomaly (USDT twin dropped by VIF)
ssi	mean of 6 expanding z-	Stablecoin Stress Index

	scores	(Section 3.5)
depeg_dummy	1{max	dev
btc_ret_lag1 ... lag7	r(t), r(t-1), r(t-2), r(t-4), r(t-6)	Lagged BTC log returns (lags 1,2,3,5,7 to target)
btc_vol_7d / btc_vol_30d	rolling std of r, 7/30 days	Realized volatility, short and medium window
btc_mom_7d / btc_mom_30d	rolling mean of r, 7/30 days	Momentum, short and medium window
eth/sol/xrp_ret_lag1	r(t) of ETH, SOL, XRP	Cross-asset momentum
btc_volume_z	(V - mean30)/std30	BTC volume anomaly
gld_ret_lag1 / dxy_ret_lag1	log return at t	Gold and dollar-index returns (ffilled calendar)
vix_level / vix_chg_lag1	level; first difference	Risk aversion, level and change
dow_tue ... dow_sun	1{target date weekday}	Day-of-week dummies (Monday reference)

B.9 Strongest pairwise feature correlations

Table B.9 — The ten largest absolute pairwise correlations among continuous features (training-relevant collinearity map).

feature 1	feature 2	corr
btc_ret_lag1	eth_ret_lag1	0.814
usdc_volume_z	btc_volume_z	0.811
usdc_volume_z	ssi	0.711
btc_vol_7d	btc_vol_30d	0.691
usdt_dev_vol7	usdc_dev_vol7	0.676
eth_ret_lag1	sol_ret_lag1	0.664
ssi	btc_volume_z	0.644
usdc_dev_signed	usdc_dev_abs	-0.636

eth_ret_lag1	xrp_ret_lag1	0.604
usdc_dev_vol7	ssi	0.603

B.10 OLS — complete coefficient table (Task A)

Table B.10 — Full OLS estimation output for the next-day BTC return (percent): all 32 features plus intercept, with Newey–West (HAC, 7 lags) standard errors, t-statistics, p-values and 95% confidence intervals. Training window, $N = 1,920$.

Variable	coef	HAC se	t stat	p value	95% CI
const	-0.1521	0.4011	-0.3792	0.7046	[-0.938, 0.634]
usdt_dev_signed	0.0248	0.0136	1.824	0.0682	[-0.002, 0.051]
usdt_dev_abs	0.0286	0.0259	1.106	0.2687	[-0.022, 0.079]
usdt_dev_vol7	0.0001	0.0391	0.0028	0.9978	[-0.077, 0.077]
usdc_dev_signed	-0.0399	0.0181	-2.202	0.0277	[-0.075, -0.004]
usdc_dev_abs	-0.0084	0.0317	-0.2644	0.7915	[-0.071, 0.054]
usdc_dev_vol7	0.0045	0.0224	0.1986	0.8426	[-0.039, 0.048]
usdc_volume_z	0.0726	0.1861	0.3901	0.6964	[-0.292, 0.437]
ssi	-0.2157	0.6886	-0.3132	0.7541	[-1.565, 1.134]
btc_ret_lag1	5.236	4.601	1.138	0.2552	[-3.783, 14.254]
btc_ret_lag2	-0.0473	3.086	-0.0153	0.9878	[-6.096, 6.002]

btc_ret_lag3	-0.0638	3.326	-0.0192	0.9847	[-6.583, 6.455]
btc_ret_lag5	-1.024	3.233	-0.3167	0.7514	[-7.360, 5.312]
btc_ret_lag7	-2.763	3.478	-0.7945	0.4269	[-9.579, 4.053]
btc_vol_7d	-16.39	7.354	-2.228	0.0258	[-30.800, -1.974]
btc_mom_7d	3.193	14.31	0.2231	0.8234	[-24.855, 31.241]
btc_vol_30d	14.23	10.69	1.331	0.183	[-6.717, 35.181]
btc_mom_30d	16.31	14.49	1.125	0.2606	[-12.101, 44.714]
eth_ret_lag1	-2.873	3.73	-0.7702	0.4412	[-10.184, 4.438]
sol_ret_lag1	-2.773	2.303	-1.204	0.2284	[-7.287, 1.740]
xrp_ret_lag1	-3.302	1.84	-1.794	0.0727	[-6.908, 0.305]
btc_volume_z	0.0619	0.128	0.484	0.6284	[-0.189, 0.313]
gld_ret_lag1	5.23	8.956	0.5839	0.5593	[-12.324, 22.783]
dxy_ret_lag1	-7.391	21.51	-0.3436	0.7312	[-49.556, 34.773]
vix_level	0.0107	0.0144	0.7399	0.4594	[-0.018, 0.039]
vix_chg_lag1	0.0233	0.0543	0.4287	0.6681	[-0.083,

					0.130]
dow_tue	-0.2552	0.31	-0.8233	0.4103	[-0.863, 0.352]
dow_wed	0.1679	0.3388	0.4955	0.6203	[-0.496, 0.832]
dow_thu	-0.4704	0.3311	-1.421	0.1554	[-1.119, 0.179]
dow_fri	-0.2092	0.3253	-0.6432	0.5201	[-0.847, 0.428]
dow_sat	-0.2194	0.2826	-0.7763	0.4376	[-0.773, 0.334]
dow_sun	-0.014	0.2672	-0.0522	0.9584	[-0.538, 0.510]
depeg_dummy	-3.254	5.385	-0.6044	0.5456	[-13.808, 7.300]

B.11 Confusion matrices — complete counts

Table B.11 — Confusion matrices on the validation month (28 days, 5 crash days, main target) at the conventional 0.5 threshold and at the F1-optimal threshold thr^* tuned on out-of-fold training probabilities. $TN/FP/FN/TP$ = true negatives, false positives, false negatives, true positives.

Model	Thres					Precis		
	hold	TN	FP	FN	TP	ion	Recall	F1
Logistic	0.5	10	13	2	3	0.19	0.60	0.29
Logistic	$thr^* =$ 0.50	10	13	2	3	0.19	0.60	0.29
LDA	0.5	23	0	5	0	0.00	0.00	0.00
LDA	$thr^* =$ 0.07	5	18	0	5	0.22	1.00	0.36
L1-logistic	0.5	12	11	2	3	0.21	0.60	0.32
L1-logistic	$thr^* =$	9	14	2	3	0.18	0.60	0.27

	0.48							
Tree	0.5	12	11	2	3	0.21	0.60	0.32
Tree	thr* =	23	0	5	0	0.00	0.00	0.00
	0.61							
Random forest	0.5	20	3	5	0	0.00	0.00	0.00
Random forest	thr* =	13	10	1	4	0.29	0.80	0.42
	0.38							
XGBoost	0.5	15	8	2	3	0.27	0.60	0.38
XGBoost	thr* =	7	16	0	5	0.24	1.00	0.39
	0.39							
SVC	0.5	23	0	5	0	0.00	0.00	0.00
SVC	thr* =	7	16	1	4	0.20	0.80	0.32
	0.12							
MLP	0.5	23	0	5	0	0.00	0.00	0.00
MLP	thr* =	23	0	5	0	0.00	0.00	0.00
	0.29							