Crypto Volatility Spike Detection

Hard Dataset – Model Comparison Report

Generated: 2025-11-19 11:55:34
HARD Train file: features_hard_train_20251119_114901.parquet
HARD Test file : features_hard_test_20251119_114901.parquet

1. Dataset Summary (Hard, Time-Based Split)

Train rows : 788
Test rows : 198
Number of features : 42
Train spike rate : 2.54%
Test spike rate : 15.15%
Train time range : 2025-11-07 22:09:00.382094+00:00 → 2025-11-08 15:20:17.136652+00:00
Test time range : 2025-11-08 15:20:17.433996+00:00 → 2025-11-08 15:31:16.274429+00:00
Construction notes:
- Label is a FUTURE volatility spike (proxy computed from future_volatility window).
- Train/test split is strictly time-based (no shuffling) to simulate live deployment.
- All volatility-like and future-looking columns are removed from features to avoid leakage.
- Simple median imputation is used for missing numeric values.

2. Modeling & Evaluation Setup

Four models are trained on the HARD TRAIN split and evaluated on the HARD TEST split:
• Logistic Regression
• Random Forest
• Gradient Boosting
• XGBoost
All models:
• Use the same engineered numeric feature set (after leakage removal).
• Are trained only on past information (no access to future volatility).
• Are evaluated on the hard test set with ~15% spike rate.
Metrics reported:
• Accuracy – overall correctness (dominated by the majority class).
• AUC – ranking quality for spike vs non-spike.
• Precision – fraction of predicted spikes that are real spikes.
• Recall – fraction of true spikes that the model catches.
• F1 – harmonic mean of precision and recall (primary selection metric).

3. Model Comparison (Hard Test Set)

Model	Thr	Acc	AUC	Prec	Rec	F1
★ random_forest_hard_20251119_	0.260	0.848	0.835	0.000	0.000	0.000
gradient_boosting_hard_20251	0.010	0.838	0.654	0.000	0.000	0.000
xgboost_hard_20251119_114951	0.260	0.848	0.564	0.000	0.000	0.000
logistic_regression_hard_202	0.880	0.848	0.500	0.000	0.000	0.000

4. Best Performing Model (Winner)

Winner model (by F1, then AUC):
• Model name : random_forest_hard_20251119_114950
• Accuracy : 0.848
• AUC : 0.835
• Precision : 0.000
• Recall : 0.000
• F1 score : 0.000
• Decision threshold: 0.260
Interpretation:
• The winner is selected to maximize F1, which balances precision and recall.
• A higher AUC indicates better ranking quality, even if classification
at a fixed threshold is conservative.
• In this hard, time-based setup with rare spikes, even modest recall