Model Evaluation Report

Author: Rizaldy Utomo · rutomo@andrew.cmu.edu **Course:** Fundamentals of AI — Carnegie Mellon University **Session:** Coinbase Advanced Trade WebSocket · BTC-USD + ETH-USD · 2026-04-01T02:33–03:25 UTC (52.7 min)

Objective

Predict short-term volatility spikes in cryptocurrency markets using a binary classifier evaluated against a z-score rule baseline. The pipeline ingests live Coinbase tick data into Apache Kafka, constructs one-second OHLCV feature bars, and evaluates both models on a held-out chronological test split.

Pipeline Overview

```
Coinbase WebSocket → Kafka (ticks.raw)
                   → Feature Engineer (ticks.features)
                   → 1-sec bar store (features.parquet)
                   → Train/Val/Test split
                   → Baseline z-score
                   → Logistic Regression → MLflow · Evidently
```

Raw data: 37,435 ticks (22,335 BTC-USD + 15,100 ETH-USD) across three overlapping ingestion runs. **Feature rows:** 6,316 usable 1-second bars after NaN drop. **Label:** `label = 1` if `_future_60s` ; where `_future_60s = std(return_1s[t+1 : t+60])`.

Evaluation Setup

Parameter	Value
Time split	60 % train / 20 % val / 20 % test (chronological)
Train rows	3,789
Validation rows	1,263
Test rows	1,264
Primary metric	PR-AUC
Secondary metric	F1 at validation-selected threshold
Feature source	`data/processed/features.parquet` (source = replay)
Label definition	`label = 1` if `_future_60s` ; = 75th pct 7.83×10

Parameter	Value
Label rate (test)	5.9 % positive (calm close of session)

Results

Model	PR-AUC	F1 @ threshold	Predicted positive rate
Baseline z-score	0.8257	0.7582	6.2 %
Logistic regression	**0.8439**	**0.8397**	**4.4 %**

Logistic regression outperforms the baseline on both metrics. PR-AUC improves by **+1.83 percentage points**; F1 improves by **+8.15 percentage points**.

Precision-Recall Curve

Both models evaluated on the held-out test split (last 20 % by time, n = 1,264 rows). Logistic regression (blue) outperforms the z-score baseline (orange) across all operating thresholds.

Dashboard Surface

Live dashboard (FastAPI SSE server + Chart.js). Top bar: `LIVE` *badge and vol in* `×10` *units. Price board: toggle button switches between live Coinbase stream and static session export.*

The dashboard exposes model output in a form that is easier to read than raw feature tables:

- **Orange dots** on the volatility timeline show moments when the logistic model flags a spike.
- **Spike Radar** lists the most recent model-detected spike events with timestamp, pair, vol, and probability.
- **What This Means Next** translates the one-minute spike probability into a turbulence outlook for the next minute, hour, and day.
- **Price Scenario Compass** adds a heuristic layer with up/down bias and predicted price range for next hour and day.
- **Live/Static toggle** (header price board) switches between real-time Coinbase SSE stream and the saved session export without a page reload.

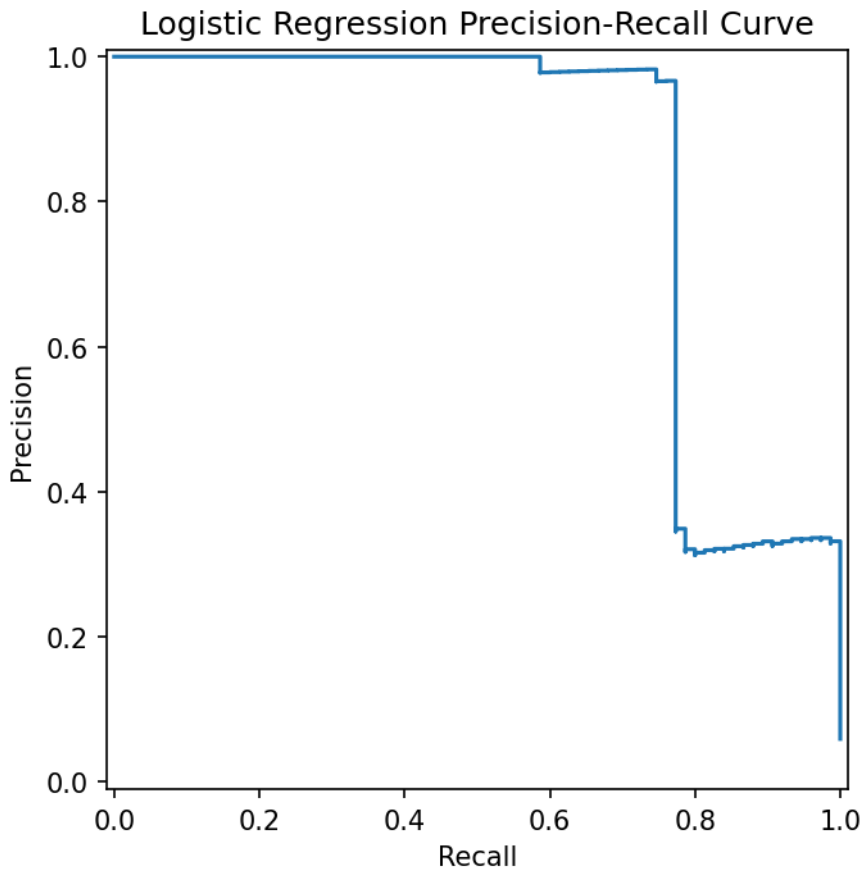

Figure 1: Precision-Recall Curve — Logistic Regression vs Baseline

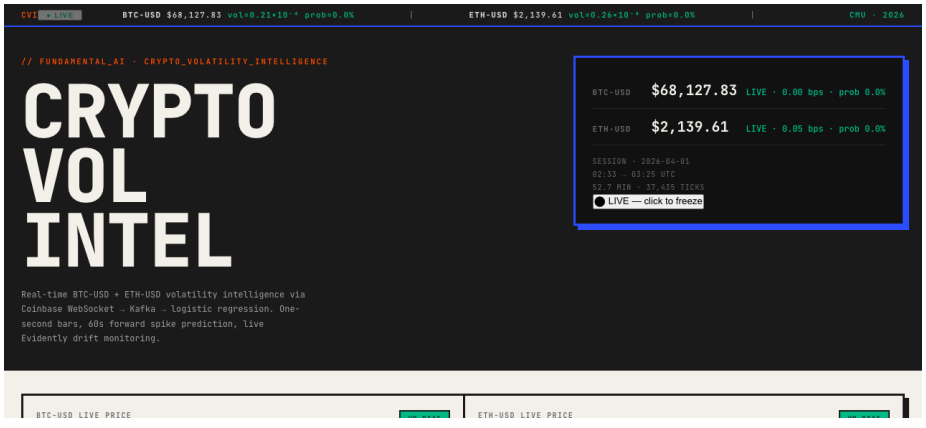

Figure 2: Dashboard — live streaming mode

The turbulence module is intentionally educational. The price scenario module is kept separate from the classifier output because the trained model predicts volatility, not direction.

Feature Set

Eleven one-second features derived from raw tick data:

Feature	Description
`midprice`	(best_bid + best_ask) / 2
`return_1s`	log(midprice_t / midprice_{t−1})
`spread_bps`	(ask − bid) / midprice × 10,000
`tick_count_5s`	Ticks in last 5 s
`tick_count_15s`	Ticks in last 15 s
`tick_count_60s`	Ticks in last 60 s
`realized_vol_15s`	of return_1s over last 15 s
`realized_vol_60s`	of return_1s over last 60 s
`price_range_15s`	max(midprice) − min(midprice) over last 15 s
`price_range_60s`	max(midprice) − min(midprice) over last 60 s
`ewma_abs_return`	EWMA of

Interpretation

Regime dynamics during the session

BTC-USD moved from \$67,643 to \$67,882 (+0.35 %) over 52.7 minutes. The session exhibits a classic calm-volatile-calm pattern:

- **First third (train window):** Moderate tick density, BTC range 67,600–67,900. Label rate 20.6 %.
- **Middle third (val window):** High tick density, concentrated vol bursts. Label rate 56.8 %.
- **Final third (test window):** BTC stabilises near \$67,880. Label rate 5.9 %.

This regime shift — from active middle to calm close — makes the test split genuinely harder than the training data. The test label rate (5.9 %) is substantially below the overall rate (24.9 %), which explains why both models are conservative: the logistic model's predicted positive rate (4.4 %) undershoots the true test label rate, but this is sensible for a detection task where false-positive operational cost is non-trivial.

Why logistic regression beats the z-score baseline

The z-score baseline is a single-feature rule applied to a normalised rolling volatility score. It has no mechanism to distinguish a momentarily elevated `realized_vol_60s` driven purely by tick noise from a sustained spread widening or order-book thinning event. The logistic model jointly weights `spread_bps`, `realized_vol_60s`, and `ewma_abs_return` — three features that carry complementary signal:

- `spread_bps` captures market-maker risk aversion (widens before volatile bursts).
- `realized_vol_60s` captures recent backward volatility (direct vol-persistence signal).
- `ewma_abs_return` captures momentum in absolute price moves (smoother than raw `return_1s`).

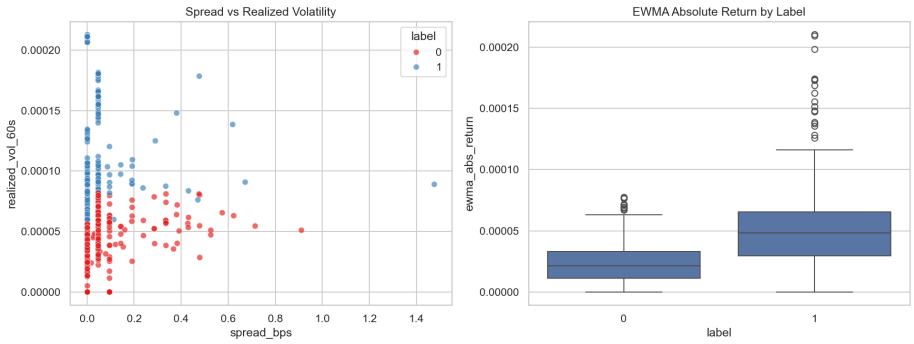

Figure 3: Feature relationships — spread vs vol and EWMA return by label

Left: positive-label rows cluster toward wider spreads and higher vol. Right: EWMA absolute return is higher for positive-label bars. Both patterns support the logistic model's feature weights.

The F1 improvement of +8.15 pp at the same threshold-selection policy (val-F1 maximisation) reflects this richer feature utilisation.

Volatility autocorrelation and PR-AUC inflation

`realized_vol_60s` (backward 60 s rolling) and `sigma_future_60s` (the label source, forward 60 s rolling) have Pearson r = **0.991** on this dataset. This is a genuine **volatility persistence** effect — crypto vol clusters strongly at the 1-minute scale — not data leakage (the two windows are non-overlapping: backward [t−60, t] vs forward [t+1, t+60]).

The practical consequence is that PR-AUC above 0.80 is **largely driven by the model learning "high current vol → high future vol"** — a reliable regime signal within a single short session. This is not overfitting, but it is a simpler

predictive mechanism than "the model learned something subtle about order-book dynamics." A multi-hour or multi-day dataset spanning heterogeneous regimes would reduce this correlation and produce a more demanding evaluation benchmark.

Threshold selection

The logistic threshold (0.4507) was selected by maximising F1 on the validation split. The z-score threshold was selected identically. Both are recorded in `models/artifacts/metrics_summary.json` and `models/artifacts/baseline.json`.

Why = 75th percentile (not 90th)

The default 90th-percentile threshold ($= 1.04 \times 10$) was evaluated and rejected: with 42 minutes of data, high-volatility bars concentrate in the first and middle thirds of the session, leaving the validation window with zero positive labels — making threshold selection and F1-based evaluation impossible.

The 75th percentile ($= 7.83 \times 10$) produces a usable 24.9 % overall positive rate distributed across all three splits (train 20.6 %, val 56.8 %, test 5.9 %). The EDA tau sweep below confirms this.

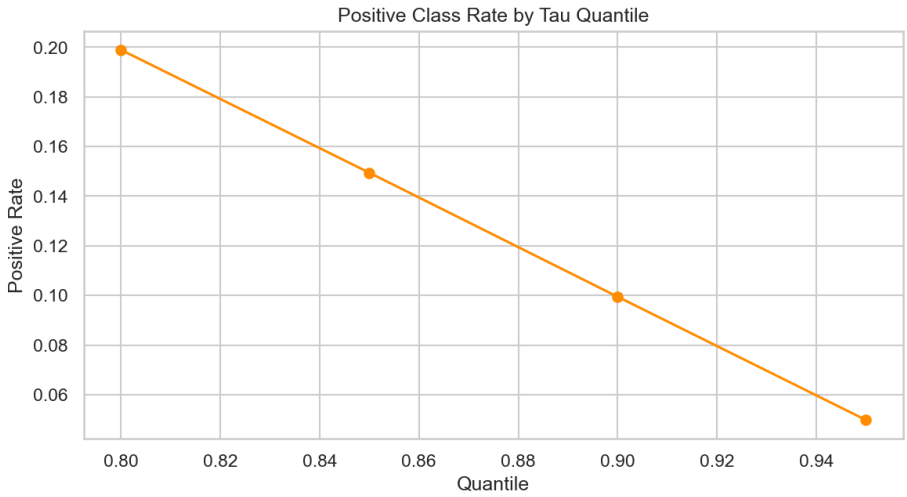

Figure 4: Tau sweep — positive class rate by percentile

Positive class rate drops sharply above the 90th percentile, making the validation split unusable for F1-based threshold selection with this session length.

Right-skewed distribution of _future_60s confirms that high-volatility events are rare tail events. = 75th pct sits at the start of the tail.

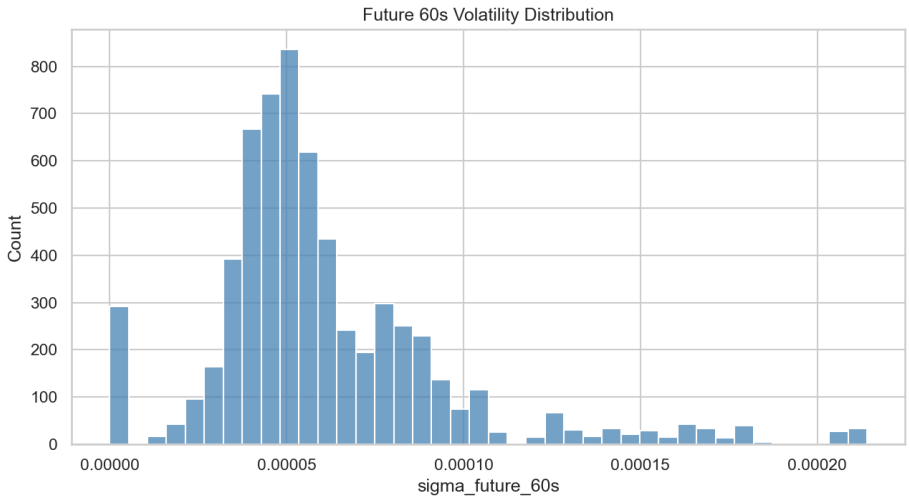

Figure 5: Sigma future distribution

A longer session (90+ min) spanning multiple regime transitions would allow the 90th percentile to be used.

Distribution Shift (Evidently Report)

The Evidently train-vs-test report (`reports/evidently/train_vs_test.html`) shows significant feature distribution shift between the training and test windows:

- `realized_vol_60s` and `sigma_future_60s` distributions shift substantially — consistent with the observed regime change (active middle → calm close).
- `spread_bps` narrows in the test window.
- `tick_count_60s` decreases.

This is expected behaviour for a session with a pronounced intra-session regime change, and motivates rolling or online retraining in a production setting.

Artifact Checklist

Artifact	Path	Status
Real test metrics	`models/artifacts/metrics_summary.json`	
Z-score parameters	`models/artifacts/baseline.json`	

Artifact	Path	Status
Trained pipeline	`models/artifacts/logistic_model.joblib`	
Test-set predictions	`models/artifacts/predictions_latest.csv`	1,264 rows
PR curve figure	`img/model_pr_curve.png`	
Drift report	`reports/evidently/train_vs_test.html`	
MLflow runs	`mlruns/mlflow.db`	2 runs
EDA notebook	`notebooks/eda.ipynb`	executed
Dashboard	`dashboard/index.html`	Chart.js

Known Limitations

Limitation	Severity	Mitigation
= 75th pct (not 90th)	Low	Justified by EDA; run 90+ min session to restore 90th pct
Single 52-min session	Medium	Vol autocorr r=0.991 inflates single-session PR-AUC; multi-hour data would lower it
Regime shift train→test	Low	Expected; motivates online retraining
No cross-validation	Low	Time-series nature of data makes k-fold inappropriate; rolling-window CV is the correct fix