SELECTION REASONING

Group 3, Fundamentals of Operationalizing AI, Fall 2025

EXECUTIVE SUMMARY

This project builds and evaluates real-time models for predicting short-horizon volatility spikes in BTC/USD markets. To ensure realism, our approach uses strict time-based train/test splitting, constructs labels from future volatility, and removes all leakage-prone features, forcing models to rely only on information that would be available at prediction time. The resulting dataset is intentionally “hard,” exhibiting low spike frequency and noisy microstructure dynamics—conditions consistent with genuine trading environments.

Multiple models were trained on identical data and preprocessing pipelines, including Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. Evaluation uses both default and tuned probability thresholds and emphasizes AUC/ROC performance, which is more informative than accuracy under rare-event imbalance. While several models achieved similar accuracy (approximately 0.85) due to majority behavior, Random Forest significantly outperformed others in AUC (approximately 0.83), demonstrating superior ability to rank spike risk. This makes it the most reliable foundation for cost-sensitive triggering, threshold calibration, or further improvements with additional data.

All models use the same carefully curated feature set composed exclusively of lag-based market signals—including spreads, returns, rolling volatility context, and order-flow intensity—organized to reflect microstructure mechanisms tied to liquidity stress and volatility emergence. Leakage-related features (for example, future volatility metrics) are explicitly excluded. This design reflects realistic forecasting constraints and preserves causal directionality between features and the target.

Together, these design choices and results support our selection of Random Forest as the primary predictive model and validate the microstructure-driven feature framework as a principled approach to short-horizon volatility forecasting.

---

1. WHY WE SELECTED THE RANDOM FOREST MODEL

1.1 Problem and Evaluation Setup

The objective is to forecast future volatility spikes in BTC/USD using only information available at prediction time. To accomplish this, we constructed a hard, production-like dataset defined by:

- Label: Spike based on a future volatility proxy
- Train/Test Split: Strictly time-based (no shuffle)
- Leakage Handling: Removal of any forward-looking features
- Rare-Event Nature: Spikes represent less than 20% of the test window
- Identical Preprocessing Across Models: Ensures a fair comparison

All models use the exact same training pipeline:

- numeric-only features
- median imputation (no class rebalancing)
- identical train/test time windows

Models compared:

- Logistic Regression
- Random Forest
- Gradient Boosting
- XGBoost

Final evaluation results are provided in the PDF leaderboard:
“hard_leaderboard_report_20251119_115534.pdf”

1.2 Why Accuracy Alone Is Misleading

At threshold 0.5, every model appears “good” with approximately 0.85 accuracy. However, this is misleading because models are simply predicting the majority class (no spike).

Important metrics for rare-event detection:

- Accuracy: Not meaningful; dominated by majority class
- F1 / Recall: Measures spike detection effectiveness
- AUC-ROC: Measures how well the model separates spike from non-spike cases

Therefore, AUC is the correct primary indicator because it measures ranking quality even if a threshold has not been tuned yet.

1.3 Threshold Tuning and ROC Interpretation

After scanning multiple probability thresholds, the following insights emerge:

- Random Forest: AUC ≈ 0.83 (best), clearly superior ROC shape
- Gradient Boosting: AUC ≈ 0.65, limited improvement under tuning
- XGBoost: AUC ≈ 0.56, weak performance
- Logistic Regression: AUC ≈ 0.50, equivalent to random guessing

Even if no model yields a high F1 score on such a small test window, the best AUC model remains the best foundation for future improvements.

1.4 Final Choice: Random Forest

Random Forest is selected as the final model because it provides:

- The highest AUC (best ranking of spike risk)
- Ability to model non-linear microstructure interactions
- Robustness to noise and missing values
- Interpretability via feature importance methods
- A strong foundation for threshold tuning and cost-sensitive decisions

Random Forest is the best base model for future operational tuning, such as:

- asymmetric loss strategies
- confidence calibration for alerts or triggers
- real-time trading and risk decision systems

---

2. FEATURES USED BY THE FINAL MODELS — AND WHY

All final models share the exact same curated feature set, designed for real-time microstructure forecasting and free from leakage.

2.1 Price and Spread Features

Examples: price, bid, ask, spread, spread_bps, midprice

Why they matter: Widening spreads and thin order books often signal instability or directional risk.

2.2 Short-Horizon Returns

Examples: return_10s, return_30s, return_60s

Why they matter: Short bursts in returns can indicate continuation, reversal pressure, or volatility clustering.

2.3 Rolling Price and Return Statistics

Examples: price_std_30s, return_std_300s, return_max_60s, return_min_30s

Why they matter: They capture recent volatility regimes without referencing future values.

2.4 Order-Flow Intensity Features

Examples: tick_count_, *intensity*_, trade_intensity, last_size

Why they matter: Volatility spikes follow activity imbalances and bursts in trading pressure.

2.5 Weak Positional Fields (Kept but Low Importance)

Examples: sequence, trade_id

Notes: Useful for ordering in data processing but not expected to contain economic meaning; tree models down-weight them naturally.

2.6 Explicitly Excluded Features

Removed to avoid leakage:

- future_volatility
- volatility_30s, volatility_60s, volatility_120s
- volatility_spike
- future_volatility_proxy
- volatility_spike_future (label itself)

These features artificially boosted performance in earlier experiments because they reveal information from the future.

2.7 Summary of Feature Philosophy

Our feature design is:
• Microstructure-driven (spreads, returns, order flow)
• Operationally realistic (observable at prediction time)
• Label-safe (no target-derived contamination)
• Extensible (allows future expansion such as deeper order book layers)

FINAL CONCLUSION

Random Forest, combined with a leak-free and microstructure-aware feature design, provides a robust, realistic, and extensible framework for real-time volatility spike forecasting. The chosen model is not simply the best on current metrics; it establishes a trustworthy and scalable foundation for further research, additional data, and cost-sensitive deployment in real-world trading systems.