Realized Volatility Forecasting with Machine Learning Methods

Ella Veysel · Felipe Jaramillo · Vicky Xu · Chenyao An · Aditya Raju

Prof. Alessio Brini - FINTECH540
Duke University

Executive Summary

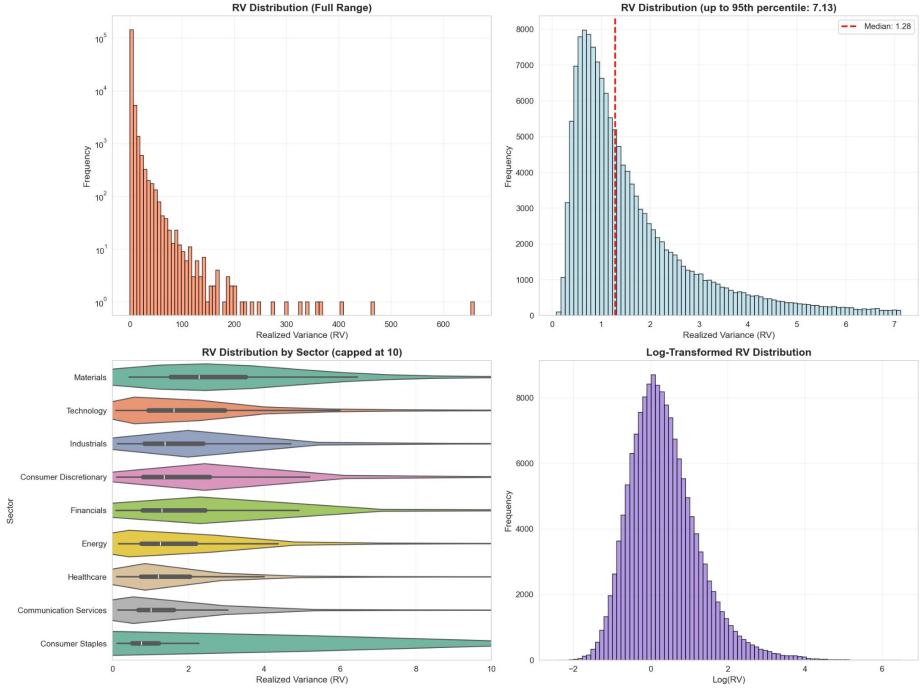
1. Goal
Test if ML improves **next-day RV forecasts** for 30 DJIA stocks.

2. Data
Use **2003–2024** high-frequency data with **70+ engineered features**.

3. Models
Compare **Random Forest, XGBoost, and LightGBM**; choosing the best.

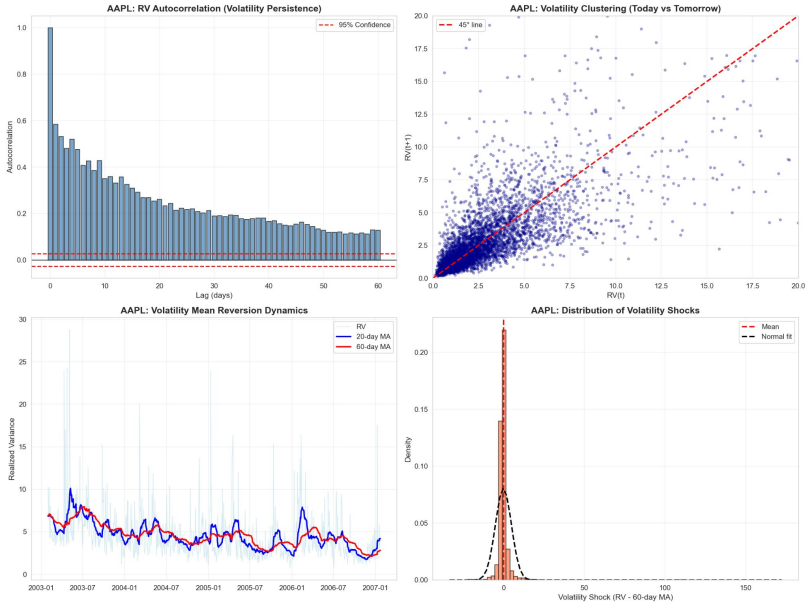
4. Result
Tuned LightGBM achieves **test $R^2 \approx 0.63$**.

5. Key Drivers
SHAP: **recent volatility, jumps, market risk** matter most.



Motivation & Introduction

- Volatility forecasting is important for **risk management, portfolio decisions, and derivatives pricing**.
- Realized volatility from high-frequency data gives a **very accurate measure of true market volatility**.
- RV is **nonlinear, clustered, and influenced by jumps**, which makes prediction difficult.
- ML models—especially tree-based ones—capture **nonlinear patterns and interactions** better than classical models.



EDA - Data Description

We use high-frequency realized measures for all **30 DJIA stocks**

from **2003 to 2024**.

For each stock-day we observe:

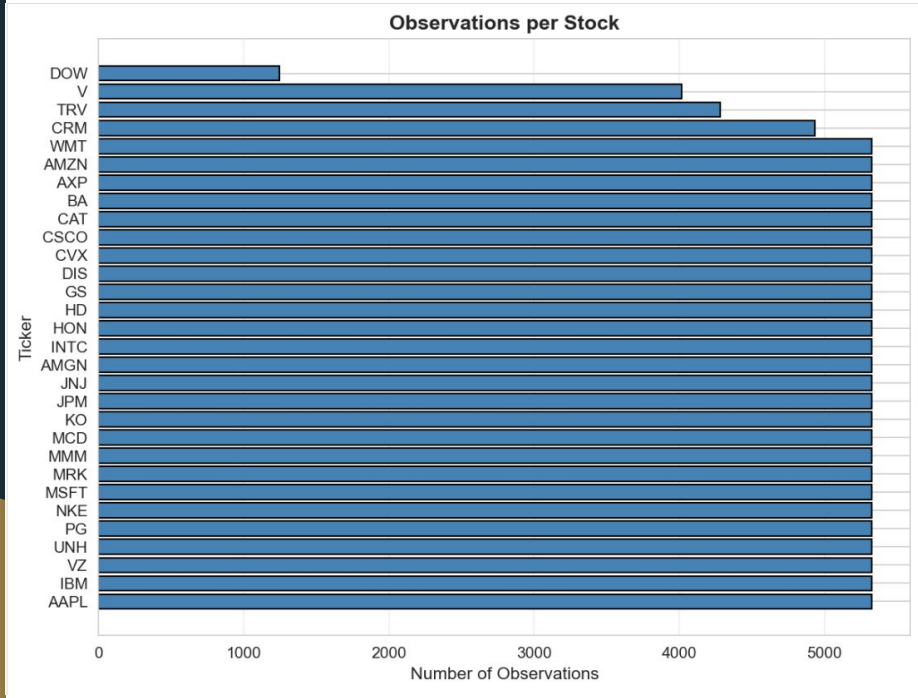
- **RV** (Realized Variance)
- **BPV** (Bipower Variation — continuous variation)
- **RQ** (Realized Quarticity)
- **Good / Bad semivariances** (upside vs downside moves)
- Both **1-minute** and **5-minute** versions

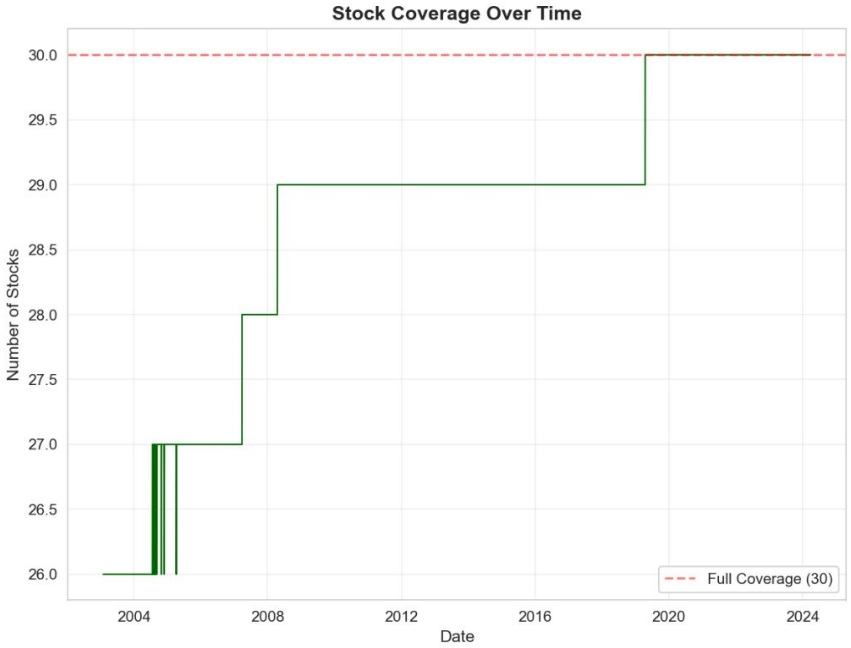
The panel is unbalanced early on but very complete after 2010.

```
Loading 10 data sheets...
  ✓ RV       : (5346, 30)
  ✓ BPV      : (5346, 30)
  ✓ Good     : (5346, 30)
  ✓ Bad      : (5346, 30)
  ✓ RQ       : (5346, 30)
  ✓ RV_5     : (5346, 30)
  ✓ BPV_5    : (5346, 30)
  ✓ Good_5   : (5346, 30)
  ✓ Bad_5    : (5346, 30)
  ✓ RQ_5     : (5346, 30)
```

```
------------------------------
DATA PREPROCESSING
------------------------------
RV       : 4.24% missing data
BPV      : 4.26% missing data
Good     : 4.24% missing data
Bad      : 4.24% missing data
RQ       : 4.24% missing data
RV_5     : 4.24% missing data
BPV_5    : 4.24% missing data
Good_5   : 4.24% missing data
Bad_5    : 4.24% missing data
RQ_5     : 4.24% missing data
```

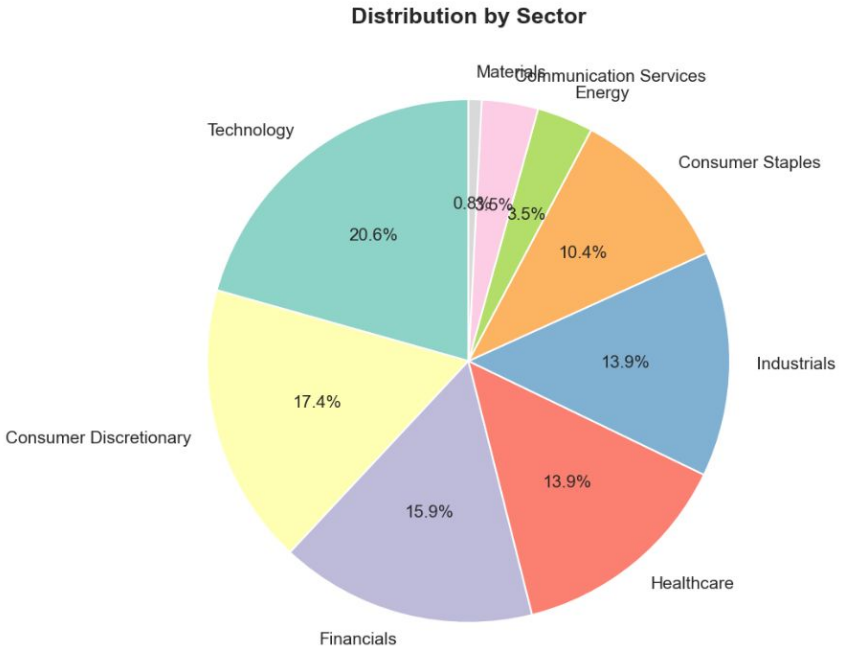
	Date	Ticker	RV	BPV	Good	Bad	RQ	RV_5	BPV_5	Good_5	Bad_5	RQ_5
0	2003-01-02	AAPL	8.3082	6.2018	5.4820	2.8262	263.0252	6.4939	3.7720	5.1023	1.3916	152.7294
1	2003-01-03	AAPL	6.5682	5.3314	3.3633	3.2048	98.4764	6.5745	5.8933	3.6380	2.9365	62.0543
2	2003-01-06	AAPL	7.3444	6.1792	3.8547	3.4897	301.8829	5.9923	4.4432	3.8791	2.1132	180.2133
3	2003-01-07	AAPL	10.0133	9.1303	5.5418	4.4715	215.2682	9.5007	7.1200	5.7827	3.7179	197.3819
4	2003-01-08	AAPL	6.0982	4.9211	3.2482	2.8500	97.9670	4.9405	5.4333	2.4732	2.4674	35.7264
...	...	...	...	...	...	...	...	...	...	...	...	...
160375	2024-03-22	WMT	0.5350	0.4092	0.2549	0.2800	1.3144	0.4101	0.3899	0.1622	0.2478	0.2477
160376	2024-03-25	WMT	0.6223	0.5255	0.2983	0.3240	0.7816	0.5458	0.5431	0.2363	0.3095	0.2871
160377	2024-03-26	WMT	0.6235	0.4772	0.3023	0.3213	1.8720	0.5238	0.5012	0.2955	0.2283	0.5659
160378	2024-03-27	WMT	0.4281	0.3431	0.2193	0.2088	0.3398	0.4057	0.4784	0.1935	0.2123	0.1592
160379	2024-03-28	WMT	0.4626	0.4684	0.2023	0.2603	0.2801	0.3594	0.3361	0.1065	0.2530	0.1510

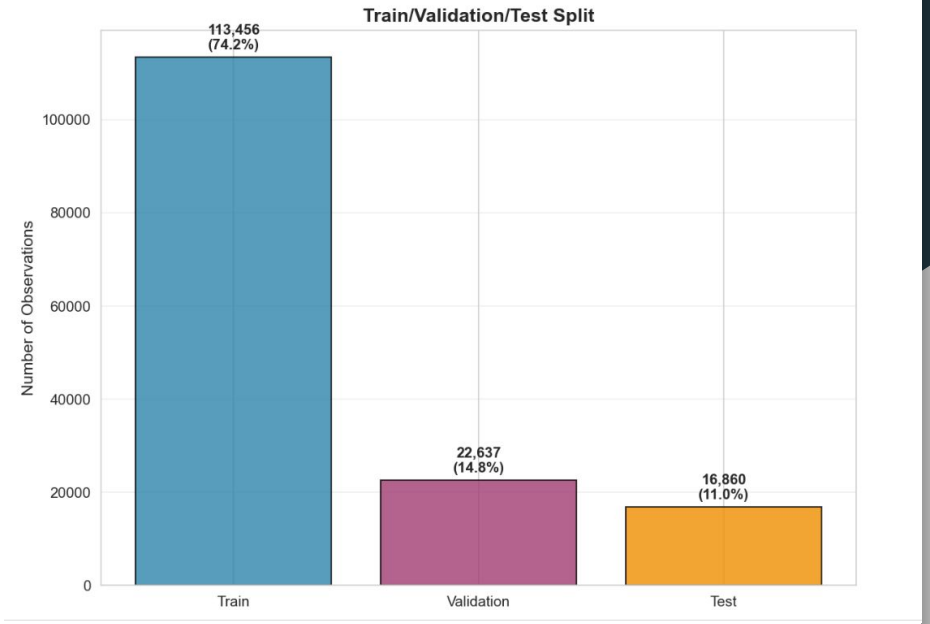




Most Dow Jones stocks have around 5,000 daily observations, giving us a complete and balanced dataset.

The dataset reaches full 30-stock coverage after 2010, making the panel stable and reliable.





The dataset covers diverse sectors, with technology being the largest group.

We use a chronological split with 74% train, 15% validation, and 11% test for realistic forecasting.

Jumps & Cross-Sectional Patterns

Jumps

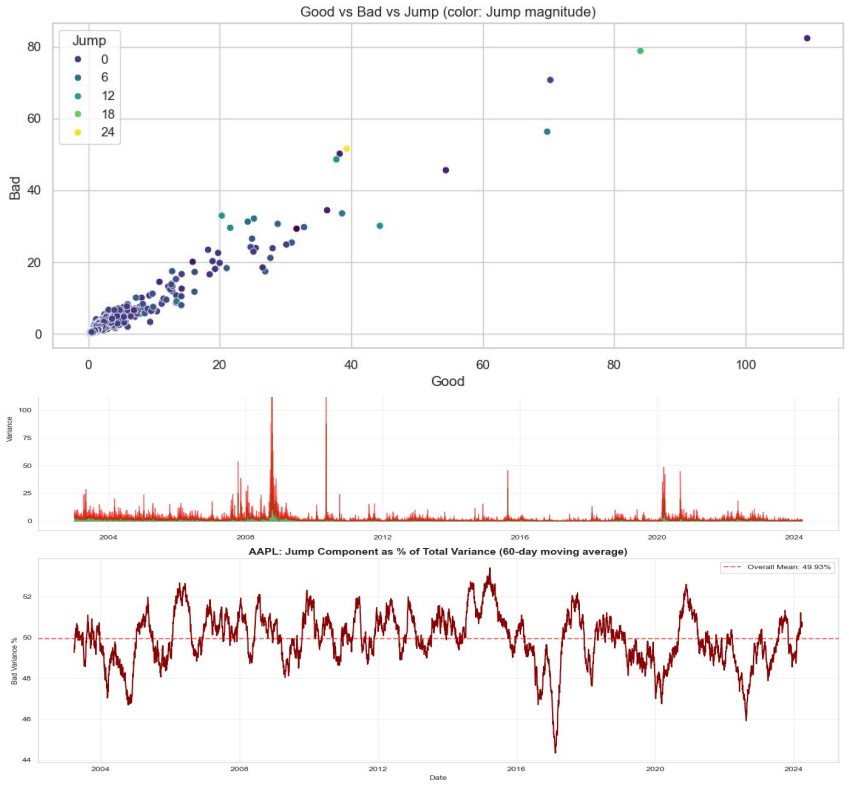
- Rare but large → strong predictors of volatility spikes
- Bad variance (downside semivariance) drives crisis periods

Cross-Sectional Patterns

- Tech/Financials most volatile
- Staples/Healthcare most stable
- Market dispersion widens in stress regimes

Variance Decomposition

- **Variance decomposition:** (RV = Good + Bad) separates continuous diffusion from discontinuous jumps—enables distinct modeling of each component
- **Jump detection threshold:** Bad_pct > 20% flags abnormal days; AAPL averages 49.85% jump contribution across full sample
- **Rolling jump metrics** capture time-varying intensity—60-day windows show regime shifts from ~30% (calm) to >60% (crisis)
- **Crisis events validate features:** 2008 and 2020 spikes demonstrate that engineered jump indicators successfully identify systemic shocks



Feature Engineering

- **Temporal:** RV lags (1, 5, 10, 20), rolling means (5, 20, 60)
- **Jump features:** Bad share, Good/Bad ratios, jump indicator
- **Market features:** Cross-sectional mean, dispersion, volatility rank
- **Frequency:** 1-min vs 5-min RV ratios
- **Calendar/data quality:** DOW, month, missingness flags

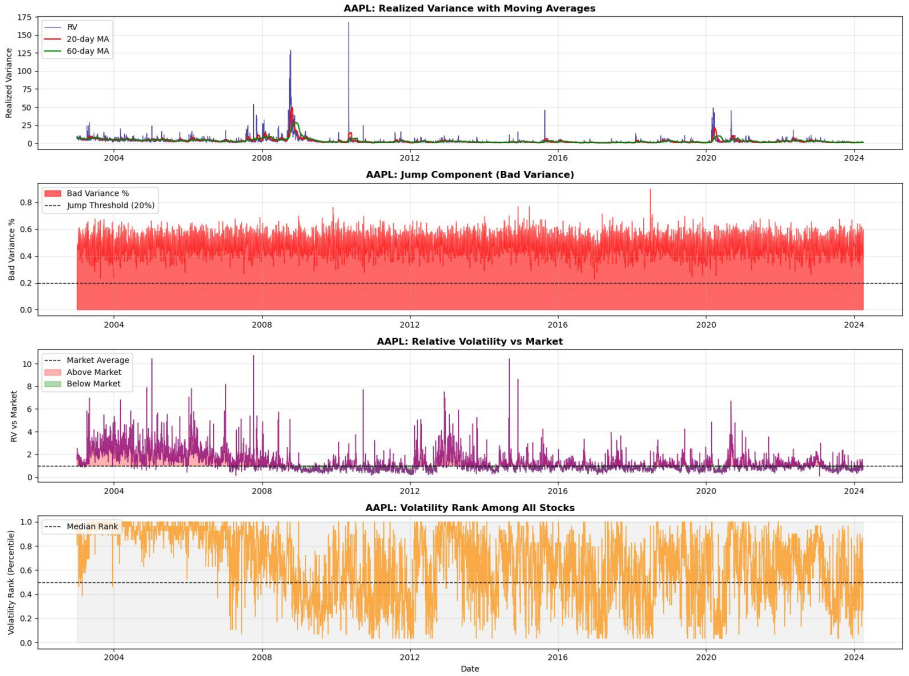
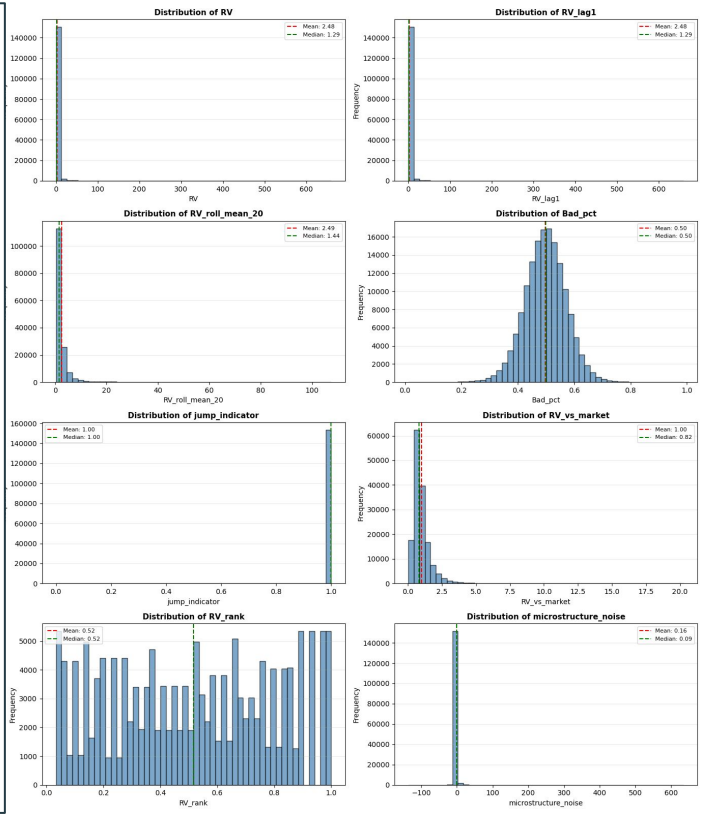


Table 5: Feature Engineering Summary

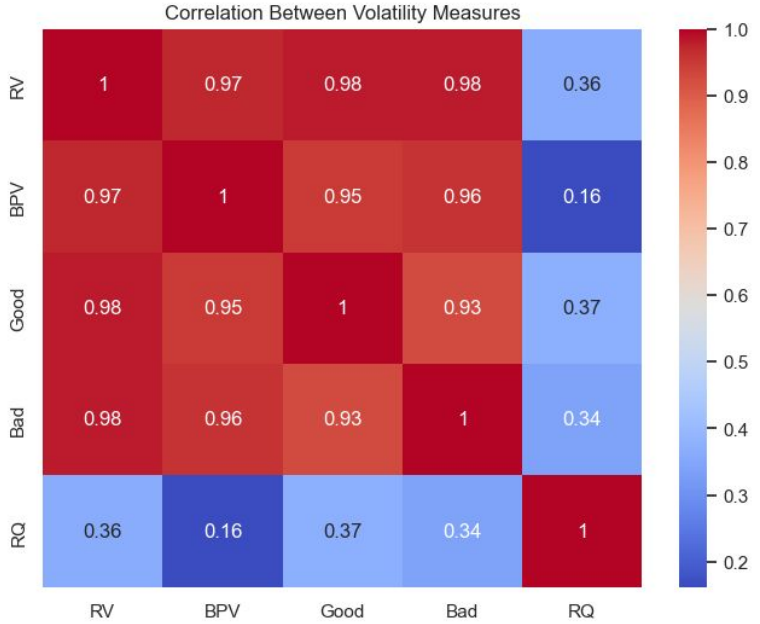
#	Category	Features	Total	Key Components
1	Temporal Lags	4		RV lag-1, lag-5, lag-10, lag-20
2	Temporal Rolling	12	18	5/20/60-day mean, std, min, max
3	Temporal Momentum	2		5-day and 20-day rate of change
4	Variance Ratios	4		Good/Bad ratios, Bad percentage
5	Jump Detection	4	12	Binary jump flags, rolling frequency
6	Jump Intensity	4		Magnitude and severity metrics
7	Market Aggregates	3		Cross-sectional mean, median, std
8	Market Relative	3	8	RV vs market (ratio, z-score, rank)
9	Market Dispersion	2		Range and coefficient of variation
10	Sector Aggregates	3	6	Industry-level mean, median, std
11	Sector Relative	3		Stock vs sector (ratio, z-score, rank)
12	VIX Levels	3		VIX, 20-day MA, log(VIX)
13	VIX Dynamics	3	9	Lag, change, percentage change
14	VIX Regimes	3		Low/medium/high flags, z-score
15	Frequency Ratios	4	6	1-min/5-min (RV, BPV, Good, Bad)
16	Microstructure Noise	2		Excess noise, consistency
17	Calendar	9	9	Month, quarter, weekday, month-end flags
18	Log Transforms	11	11	log(RV), log(BPV), log(Good/Bad), log(VIX)
	TOTAL		**79**	*Plus 10 original measures = 89 features*

Dataset: 5,346 trading days (2003–2024) × 30 Dow Jones stocks ≈ 160,000 observations
Missing Data: 4.24% treated via forward-fill (3 days), interpolation (4–10 days), and deletion.



Modeling Framework

- **Target Variable**: $\log(1 + RV_{t+1})$
- **Strong justification**: reduces *tail* impact
- **Dataset Split**:
  - Train: 2003-2018 (72.2%)
  - Validation: 2019-2021 (14.8%)
  - Test: 2022-2024 (11.0%)
- **Models Evaluated**:
  - Random Forest
  - XGBoost
  - LightGBM



Validation Results

Model	Train RMSE	Train R^2	Train MAE	Val RMSE	Val R^2	Val MAE
Random Forest	0.1934	0.8957	0.1350	0.3140	0.7319	0.2136
XGBoost	0.2268	0.8566	0.1609	0.3147	0.7307	0.2142
LightGBM	0.2557	0.8177	0.1806	0.3037	0.7492	0.2119

- **Key Insight**: LightGBM achieved the highest validation
- **Reasoning**: Smallest gap between Training R^2 and Validation R^2, showing better generalization
- Based on this, our final evaluation was conducted by tuning the **LightGBM model** and evaluating against the test data

Final Test Results

Model	Train R^2	Train RMSE	Val R^2	Val RMSE	Test R^2	Test RMSE
Baseline LightGBM	0.8177	0.2557	0.7492	0.3037	—	—
Tuned LightGBM	0.8777	0.2095	0.7331	0.3133	0.6269	0.2739

Final Performance (2022–2024 Test Set)

- Tuned LightGBM achieves **$R^2 \approx 0.63$**
- Low MAE and stable performance across market regimes
- Captures inflation shocks, 2022–2023 volatility spikes
- After completing the hyperparameter tuning of the model, we evaluated the performance against the test set

Why LightGBM Wins

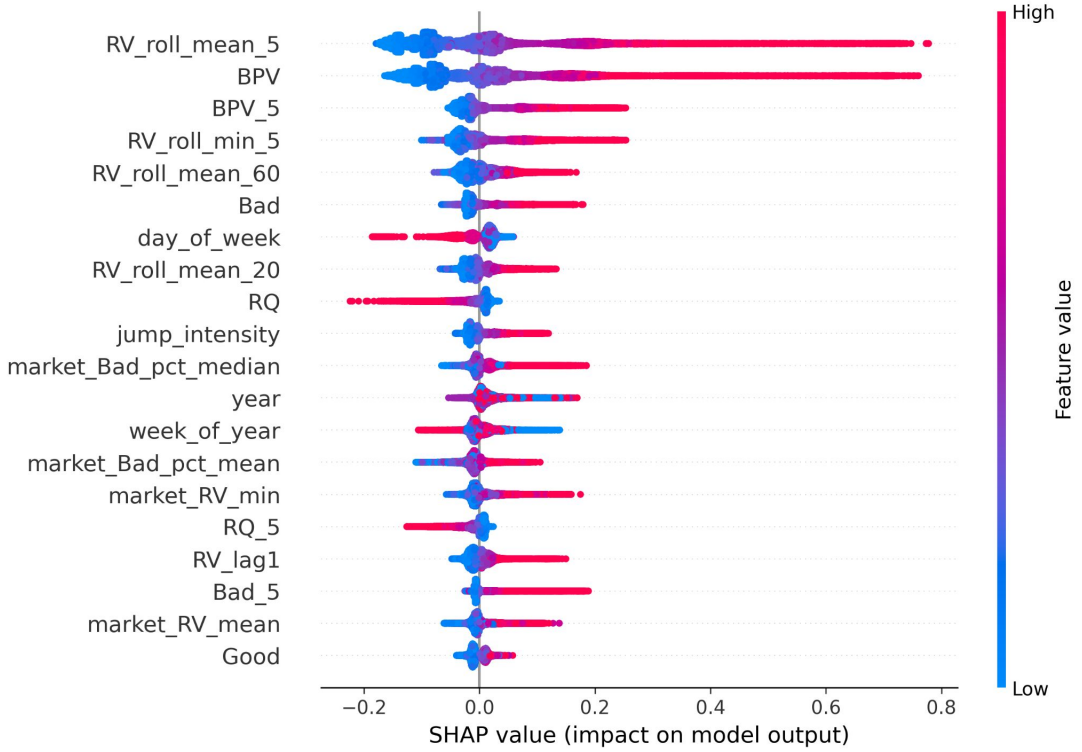
- **Superior Generalization**
- **Effective Handling of Volatility Characteristics**
    - **Nonlinear Thresholds**
        - Captures sudden regime shifts (ex. When RV spikes)
    - **Jump Awareness**
        - Utilizes Good/Bad and BPV measures to predict persistence after shocks
    - **Cross-sectional Interactions**
        - Incorporates market conditions such as dispersion and semivariance to inform firm-level predictions
- **Most stable generalization vs RF/XGB**

SHAP Analysis

Why Do We Use SHAP for Model Interpretation?

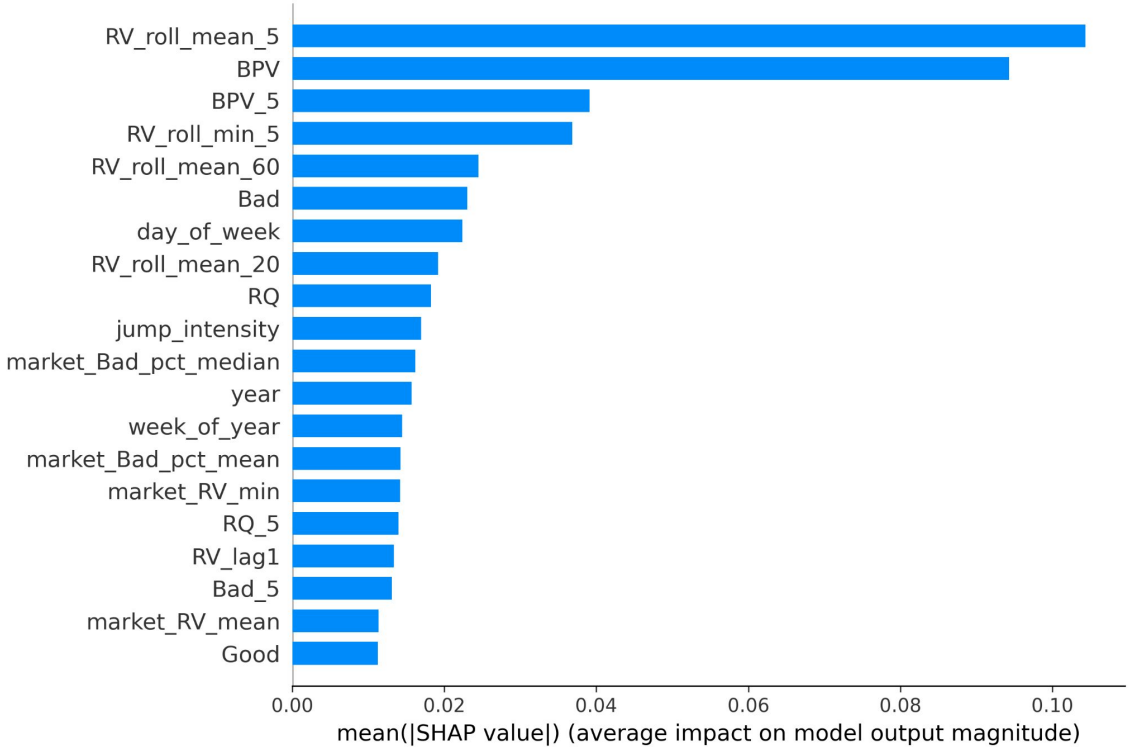
- LightGBM is a complex nonlinear model → hard to interpret directly
- SHAP provides *consistent, game-theoretic* feature attribution
- **Helps us answer:**
  - *Which features drive next-day RV forecast the most?*
  - *How do feature values influence the prediction direction?*
  - *Are there nonlinear effects or interactions?*

SHAP Summary Plot (Overall Feature Importance)



- Top predictors: RV_roll_mean_5, BPV, BPV_5, RV_roll_min_5
- Strong dominance of short-horizon RV and bipower variation
- Color = feature value (blue = low, red = high)
- Long right tails → higher feature values strongly increase predicted RV

SHAP Bar Plot (Overall Feature Importance)



- RV_roll_mean_5 and BPV contribute the largest average impact
- Market-level downside measures also matter (Bad_pct_mean, Bad_pct_median)
- Confirms model reliance on both firm-level and market-level risk

Project Significance

1. LightGBM clearly improves next-day RV forecasting

- Tuned model achieved **test $R^2 \approx 0.63$** — competitive for volatility forecasting.

2. Model reveals key economic drivers of risk

- Short-horizon volatility dominates (RV_roll_mean_5, RV_roll_min_5, BPV).
- Downside risk and jumps matter (Bad, BPV_5, jump_intensity).
- Market-level downside semivariance also influences firm-level RV.

3. Provides interpretable, actionable signals

- Higher short-term volatility → **forecasted RV rises sharply**.
- Joint spikes in market-level and firm-level risk strongly amplify predictions.
- Useful for **risk management, liquidity planning, portfolio hedging**.

Future Improvements

- Experiment with **alternative ML models** (CatBoost, neural networks, temporal transformers)
- Incorporate more **macroeconomic variables** or sentiment data for regime detection
- Explore **probabilistic forecasting** (e.g., quantile loss, distributional models) rather than point predictions
- Reduce tail-risk error by modeling extreme events separately
- Apply **rolling-window** or online learning to adapt to structural market shifts

Thank you for your Time!

Ella Veysel · Felipe Jaramillo · Vicky Xu · Chenyao An · Aditya Raju

Prof. Alessio Brini - FINTECH540
Duke University