Realized Volatility Forecasting with Machine Learning Methods

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

`Github link`

Professor Alessio Brini
Duke University Pratt School of Engineering

December 1, 2025

Executive Summary

This project investigates whether modern machine learning methods can improve short–horizon volatility forecasting relative to traditional econometric benchmarks. Using high–frequency realized measures for the thirty Dow Jones Industrial Average (DJIA) stocks from 2003 through 2024, we construct next–day realized variance forecasts based on a set of engineered features that capture time–series persistence, jump behavior, cross–sectional market conditions, frequency differences, calendar, and data–quality indicators.

Topic. The exploratory analysis confirms several well–known characteristics of financial volatility. Realized variance is highly right–skewed with heavy tails, strong clustering, and large jumps during crisis periods such as 2008 and 2020. Volatility comoves across stocks, and downside semivariance contributes disproportionately during high–stress regimes. Missing data are concentrated in early years due to index entry timing and sparse intraday coverage, but coverage becomes nearly complete after 2010. These patterns motivate the construction of lagged features, rolling statistics, jump indicators, market–wide dispersion measures, and cross–frequency consistency metrics.

We evaluate three tree–based ensemble models under a unified forecasting framework: Random Forest, XGBoost, and LightGBM. All models are trained on 2003–2018, tuned on 2019–2021, and tested on 2022–2024, using the same engineered feature matrix. Performance is measured using RMSE, MAE, and out–of–sample R^2 on the original realized variance scale. Volatility persistence allows all three models to extract substantial predictive signal, but LightGBM generalizes best: a tuned LightGBM model achieves a test R^2 of about 0.63 on next–day RV when predicting $t + 1$ volatility.

SHAP (SHapley Additive exPlanations) analysis shows that the most informative predictors are short–horizon rolling averages of realized variance, bipower variation, downside–risk and jump–intensity measures, and market–level indicators summarizing downside semivariance and cross–sectional volatility dispersion. Higher values of these variables push the LightGBM forecast sharply upward, particularly when both firm–level and market–level risk are elevated.

Overall, the project shows that machine learning methods meaningfully improve next–day realized volatility forecasts when paired with a thoughtful feature–engineering pipeline. Tree–based ensembles, and especially a tuned LightGBM model, provide robust gains over simple benchmarks and remain stable during recent high–volatility regimes. These findings have direct implications for risk management, derivatives trading, and dynamic portfolio allocation.

1   Introduction

Volatility forecasting is central to financial risk management, portfolio construction, and derivatives pricing. Even though realized volatility can be measured accurately using high–frequency intraday returns, predicting it is difficult because volatility is nonlinear, persistent, and strongly state–dependent. These features make volatility a natural setting for evaluating whether modern machine learning methods can extract more predictive structure than traditional econometric models.

We study one–day–ahead realized variance (RV) forecasts for the thirty DJIA constituents from 2003 through early 2024. The underlying data set consists of daily realized measures constructed from one–minute and five–minute returns, including RV, bipower variation (BPV), realized quarticity (RQ), and upside and downside semivariances (Good and Bad). Figure 1 in Appendix A summarizes the coverage of these measures across firms, sectors, and time.

The empirical properties of RV motivate the modeling choices. The unconditional distribution is highly right–skewed with a long tail driven by crisis episodes (Figure 2), and statistical diagnostics in Figure 3 show strong deviations from normality even after log transforms. Volatility clusters in time and remains elevated after shocks (Figure 4), and it reacts asymmetrically to market stress, with higher VIX levels associated with disproportionately higher RV (Figure 5). Volatility also exhibits strong cross–sectional comovement and sector differences (Figure 6). These patterns suggest that useful predictors should include not only own lags of RV but also jump components and market–level information.

Our contribution is to (i) engineer a rich feature set based on high–frequency realized measures, jump decompositions, and cross–sectional summaries; (ii) compare several tree–based machine learning models under a consistent forecasting framework focused on next–day RV; and (iii) interpret which features drive forecast performance, particularly during high–volatility regimes, using SHAP explanations.

2   Data and Feature Engineering

The data set contains daily observations for thirty DJIA stocks from January 2003 to early 2024. For each stock and trading day we observe RV, BPV, Good and Bad semivariances, and RQ at both one–minute and five–minute sampling frequencies. These measures allow us to separate continuous and jump components of price variation and to compare how volatility behaves across frequencies. The panel is unbalanced because some firms enter the index late, but most components have long histories.

The realized measures are strictly positive and vary over several orders of magnitude. Figure 2 shows the full distribution of RV and a version truncated at the 95th percentile. Together with the diagnostics in Figure 3, these plots confirm heavy tails, excess kurtosis, and the presence of extreme observations during events such as the 2008 financial crisis and the 2020 COVID shock. Sector and cross–sectional dispersion patterns (Figure 6) indicate that technology and financial firms tend to be more volatile than consumer staples and healthcare.

We exploit the variance decomposition into Good and Bad semivariances to capture asymmetric jump behavior. For AAPL, Figure 7 shows that Bad variance spikes sharply during stress periods and often dominates total RV. The joint behavior of Good, Bad, and jump intensity in Figure 8 highlights that large jumps tend to occur in already high–volatility regimes. These empirical regularities motivate features based on jump shares, jump frequencies, and jump intensity.

From the raw realized measures we construct a feature matrix with more than 70 variables for

each firm–day, grouped into:

- **Temporal features:** lags of RV at 1, 5, 10, and 20 days; rolling means, standard deviations, minima, and maxima over 5, 20, and 60 days; and simple momentum ratios such as $RV_t/RV_{t-5} - 1$.
- **Variance–decomposition and jump features:** Good and Bad shares of RV, Good/Bad ratios, a jump indicator when Bad share exceeds a threshold, and a 20–day jump frequency.
- **Cross–sectional market features:** daily cross–sectional mean, median, standard deviation, and dispersion of RV across the 30 stocks; relative volatility RV_t divided by the market mean; and each stock's percentile rank of RV.
- **Frequency and microstructure features:** ratios of one–minute to five–minute RV and related measures, along with a simple microstructure–noise proxy given by the difference between the two RV estimates.
- **Calendar and data–quality features:** year, month, day–of–week indicators, month–end flags, and indicators for originally missing realized measures.

Most strictly positive measures are log–transformed to stabilize variance, and all continuous features (except simple integer calendar variables) are standardized using training–set statistics. Figure 10 (Appendix) shows examples of the resulting feature distributions, and Figure 9 ranks features by their absolute correlation with next–day RV. Lagged RV, rolling means, jump shares, and market–level volatility measures appear among the most informative predictors.

3   Modeling Approach

For each stock i and day t we aim to forecast next–day realized variance,

$$\widehat{RV}_{i,t+1} = f(X_{i,t}),$$

where $X_{i,t}$ contains only information available up to time t. We model the log–transformed target, $y_{i,t+1} = \log(1 + RV_{i,t+1})$, for numerical stability as the RV is heavily skewed with extreme values.

The data are split chronologically into a training set (2003–2018), validation set (2019–2021), and test set (2022–2024). All models use the same engineered feature matrix described in Section 2. We evaluate performance using root mean squared error (RMSE), mean absolute error (MAE), and out–of–sample R^2 on the original RV scale.

We estimate three baseline tree–based models and a tuned version of the best one:

- **Random Forest:** an ensemble of decorrelated decision trees, which reduces variance by averaging over multiple bootstrap samples. Random Forests are robust to noisy features and useful for uncovering nonlinear effects in lagged and jump–related variables.
- **XGBoost:** a gradient–boosted tree ensemble where each tree fits the residuals of the previous ones, capturing complex additive nonlinearities.
- **LightGBM:** a gradient–boosted tree ensemble that uses histogram–based splits and leaf–wise growth. Like XGBoost, it handles nonlinear interactions but trains faster and with lower memory usage.
- **Tuned LightGBM:** a LightGBM model with hyperparameters (tree depth, learning rate, number of estimators, subsampling, and regularization) selected via randomized search on the validation set. This tuned model is used for final out–of–sample evaluation on 2022–2024.

All models are trained on the same train/validation split and use identical feature preprocessing, ensuring that performance differences reflect model choice rather than data handling.

4   Results

Table 1 summarizes training and validation performance for the three baseline tree–based models, evaluated on the original RV scale.

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

Table 1: Training and validation performance (2003–2021) of baseline Random Forest, XGBoost, and LightGBM models. Metrics are computed on the original RV scale.

All three models extract substantial predictive signal, but the tradeoff between in–sample and validation performance differs. Random Forest fits the training data most closely yet has the weakest validation R^2, indicating overfitting. LightGBM yields the best validation RMSE and highest validation R^2, despite a looser in–sample fit. The smaller train–validation gap suggests that LightGBM generalizes more effectively across different volatility regimes, motivating its selection for further tuning.

4.1   Final Tuned LightGBM: Out–of–Sample Test Performance

A separate randomized hyperparameter search over LightGBM settings produces a tuned model that generalizes well to the 2022–2024 test period. Table 2 reports its out–of–sample metrics.

Model	Test RMSE	Test MAE	Test R^2
LightGBM (Tuned)	0.2739	0.1918	0.6269

Table 2: Out–of–sample performance (2022–2024) of the tuned LightGBM model.

The tuned LightGBM explains roughly 63% of the variation in next–day RV on the test set and maintains a relatively low MAE despite high–volatility episodes such as the 2022 inflation and rate–hike cycle and the 2023 regional banking stress. The RMSE remains elevated by occasional crisis spikes, which is expected given the heavy–tailed nature of realized volatility.

4.2   Baseline vs. Tuned LightGBM

To quantify the effect of tuning, Table 3 compares the baseline and tuned LightGBM models across train, validation, and test samples.

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

Table 3: Comparison of baseline and tuned LightGBM models. Test results are available only for the tuned model.

The baseline model attains slightly higher validation R^2 than the tuned version, but the tuned LightGBM improves in–sample fit and achieves strong out–of–sample performance on the 2022–2024 test set. This suggests that tuning helped stabilize the model across shifting volatility regimes rather than simply optimizing for the particular 2019–2021 validation window.

4.3   SHAP–Based Feature Interpretation

We use SHAP values to interpret the tuned LightGBM model at the feature level. The SHAP summary plots (Figures 11 and 12) show that the most important predictors are:

- short–horizon rolling averages of realized variance, especially $RV\_roll\_mean\_5$,
- bipower variation measures (BPV and BPV_5),
- short–horizon minimum volatility $RV\_roll\_min\_5$,
- longer–horizon rolling means such as $RV\_roll\_mean\_60$,
- downside–risk and jump–intensity variables (Bad, Bad_5, jump_intensity),
- market–level downside semivariance statistics (e.g., `market_Bad_pct_mean` and `market_Bad_pct_median`).

For the top features, higher values are associated with positive SHAP values, meaning they push predicted next–day volatility upward. Dependence plots for $RV\_roll\_mean\_5$, $RV\_roll\_min\_5$, BPV, and BPV_5 (Figures 13–16) reveal approximately monotonic, nonlinear relationships: as recent volatility or jump–related measures increase, the model's forecast rises steeply. Color gradients in these plots indicate that interactions with other risk variables (such as BPV or recent market–wide downside semivariance) further amplify predicted volatility when both firm–level and market–level risk are high.

5   Discussion and Conclusion

The empirical results align with the volatility patterns observed in the exploratory analysis. Persistence and clustering, visible in Figure 4, make lagged RV and rolling averages the strongest predictors of next-day volatility. Embedding these directly into the feature set allows the tree-based models to approximate heterogeneous autoregressive behavior without imposing a rigid functional form.

Jump-related features also contribute meaningfully. The variance decomposition in Figure 7 and the joint structure in Figure 8 show that downside jumps correspond to persistent high-volatility regimes. SHAP dependence plots confirm that large BPV, Bad variance, and jump intensity sharply increase predicted volatility, especially when recent RV is already elevated.

Cross-sectional features further enhance accuracy. Figure 6 highlights that volatility co-moves across DJIA stocks, and SHAP results show that market-level variables—such as cross-sectional mean RV and dispersion—rank among the top predictors, indicating that index-wide risk conditions materially shape firm-level volatility dynamics.

Some limitations remain. Heavy-tailed residuals show that extreme shocks remain difficult to forecast, and shifting macro regimes may require periodic retraining. While SHAP improves interpretability, the underlying boosted-tree model is still a complex ensemble, so explanations should be used cautiously.

Overall, combining high-frequency realized measures with engineered temporal, jump-based, and cross-sectional features provides substantial gains in forecasting next-day volatility. Among the models considered, the tuned LightGBM model offers the best balance of flexibility and generalization, achieving test R^2 above 0.6 with economically intuitive SHAP insights, reinforcing the value of machine-learning methods in risk management and portfolio applications.

A Figures and Diagnostics
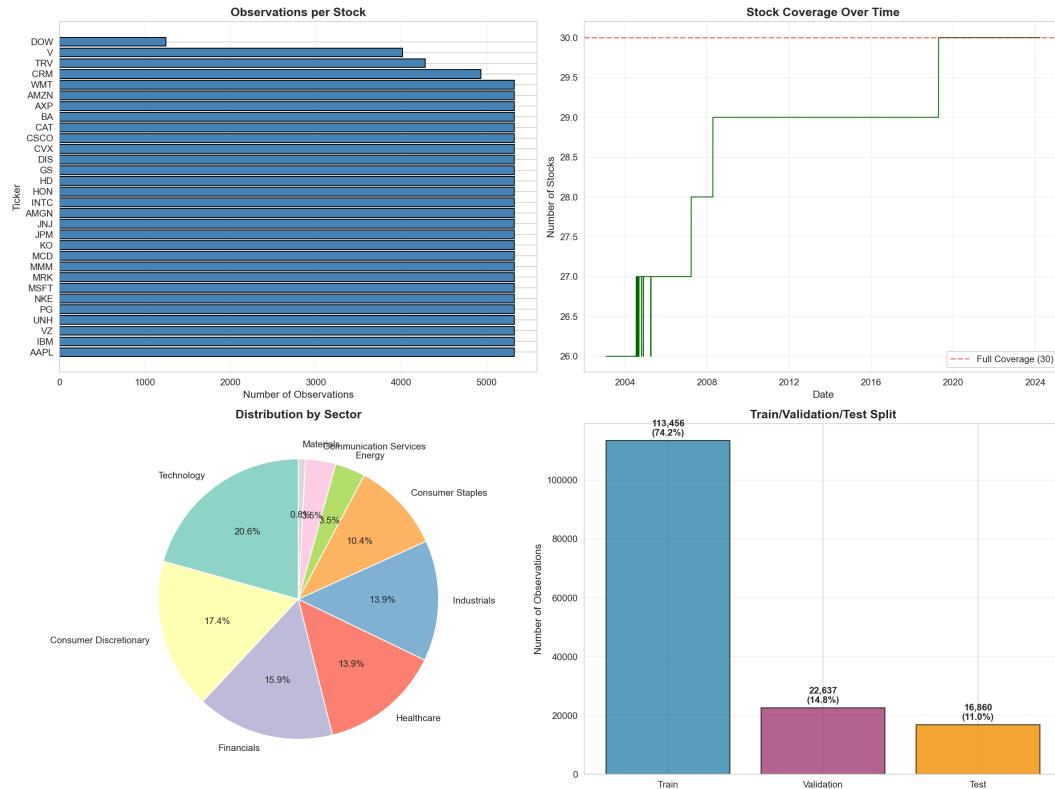
Figure 1: Overview of dataset composition: observations per stock, sector distribution, and coverage over time.
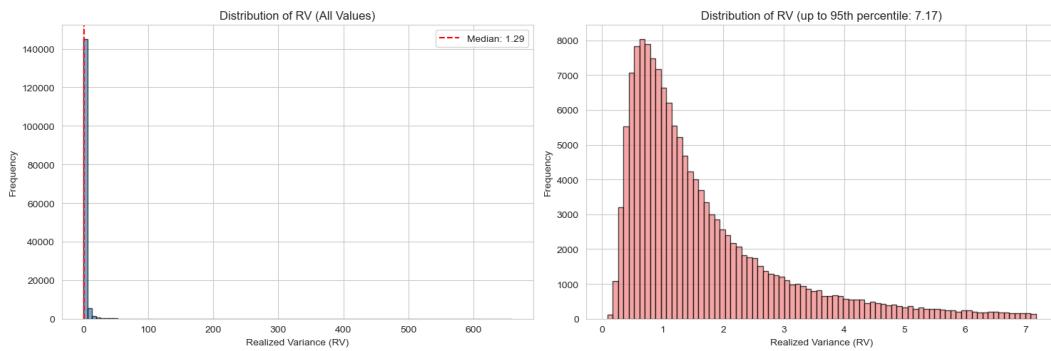

Figure 2: Full distribution of realized variance (left) and distribution truncated at the 95th percentile (right).
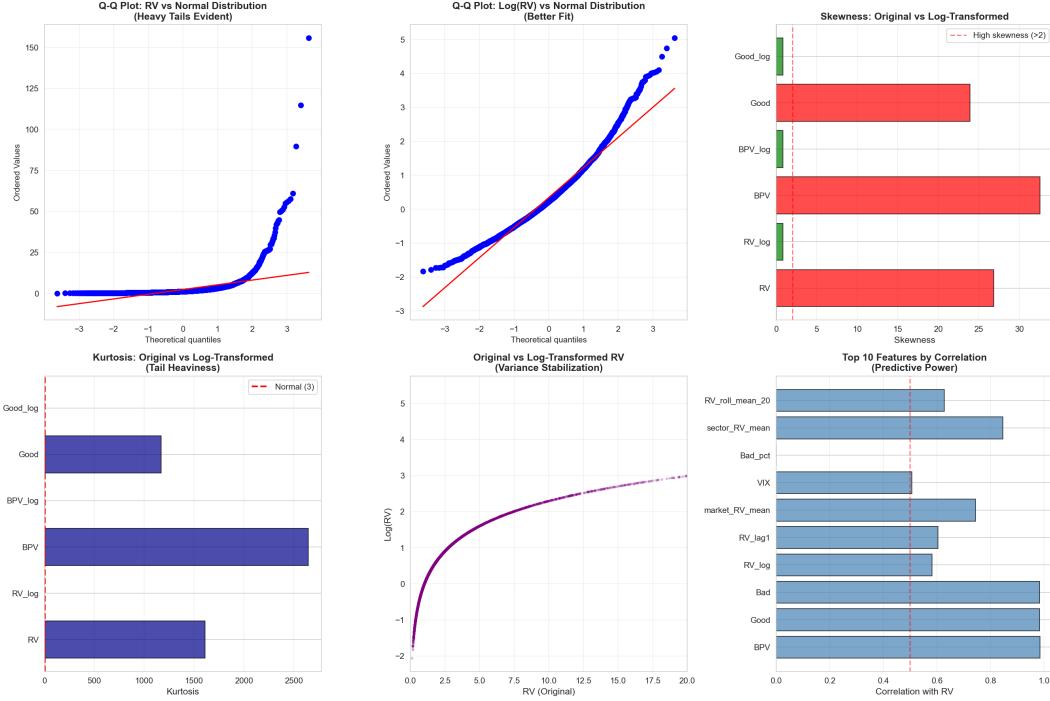
Figure 3: Statistical diagnostics for realized measures, including Q–Q plots, skewness, kurtosis, and log–scale comparisons.
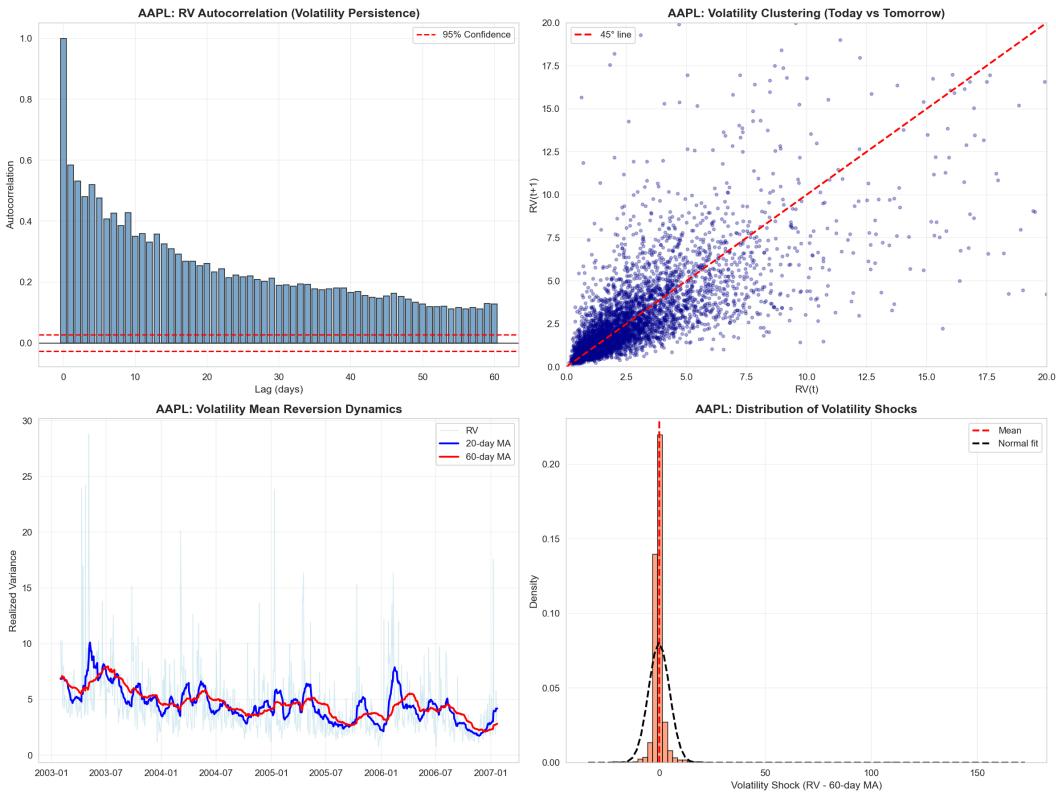

Figure 4: Evidence of volatility persistence and clustering in the realized variance series.
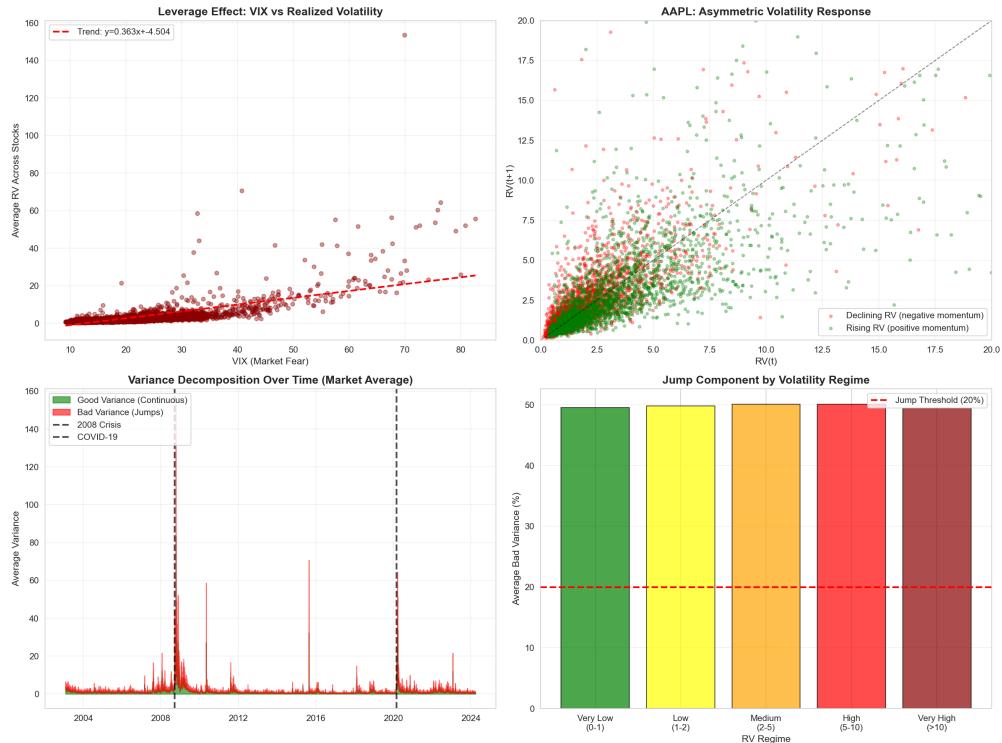
Figure 5: Relationship between VIX and realized volatility. Higher VIX levels are associated with sharply higher RV.
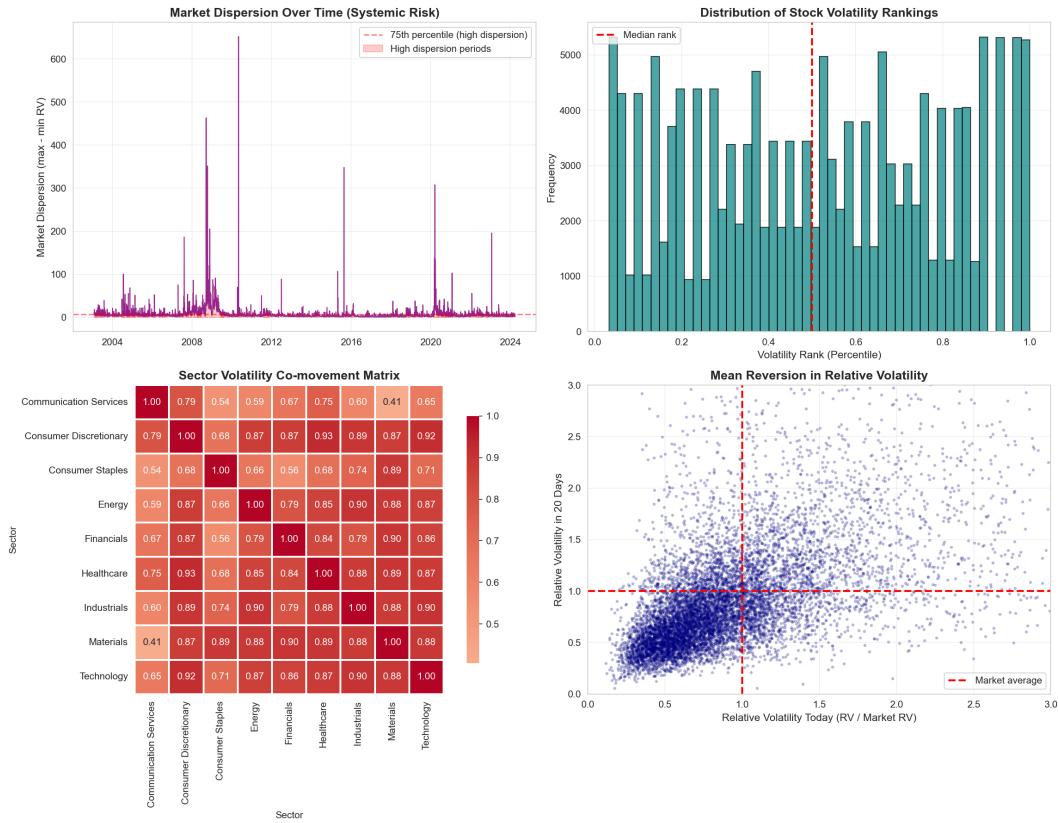
Figure 6: Cross–sectional characteristics of realized volatility, including sector patterns and dispersion across DJIA stocks.
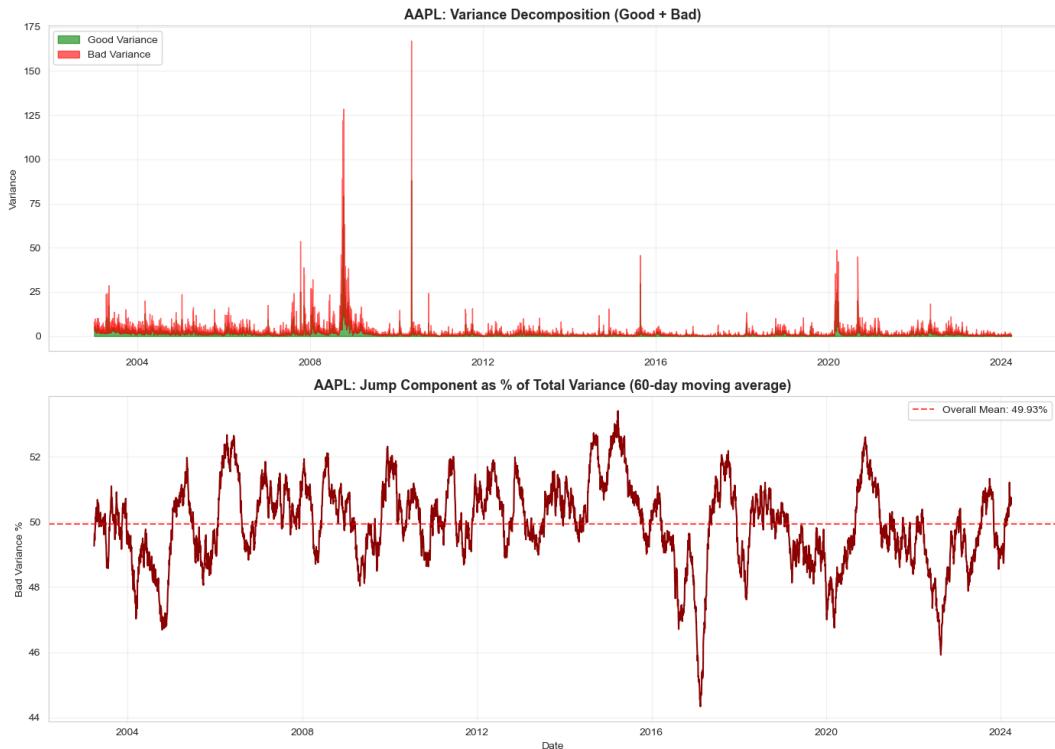
Figure 7: Variance decomposition for AAPL into Good and Bad semivariances, highlighting downside risk during stress periods.
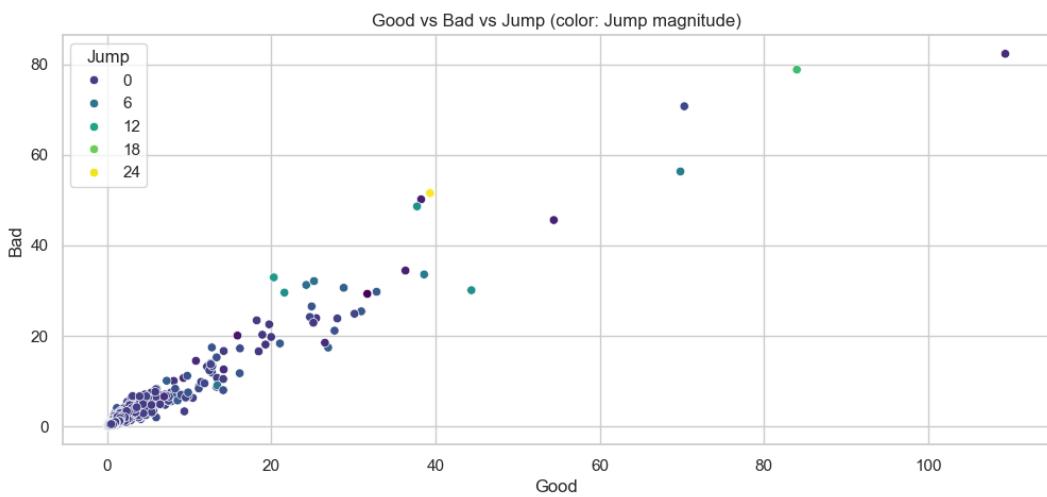
Figure 8: Scatterplot of Good vs. Bad variance, colored by jump magnitude. Large jumps occur disproportionately on high–volatility days.
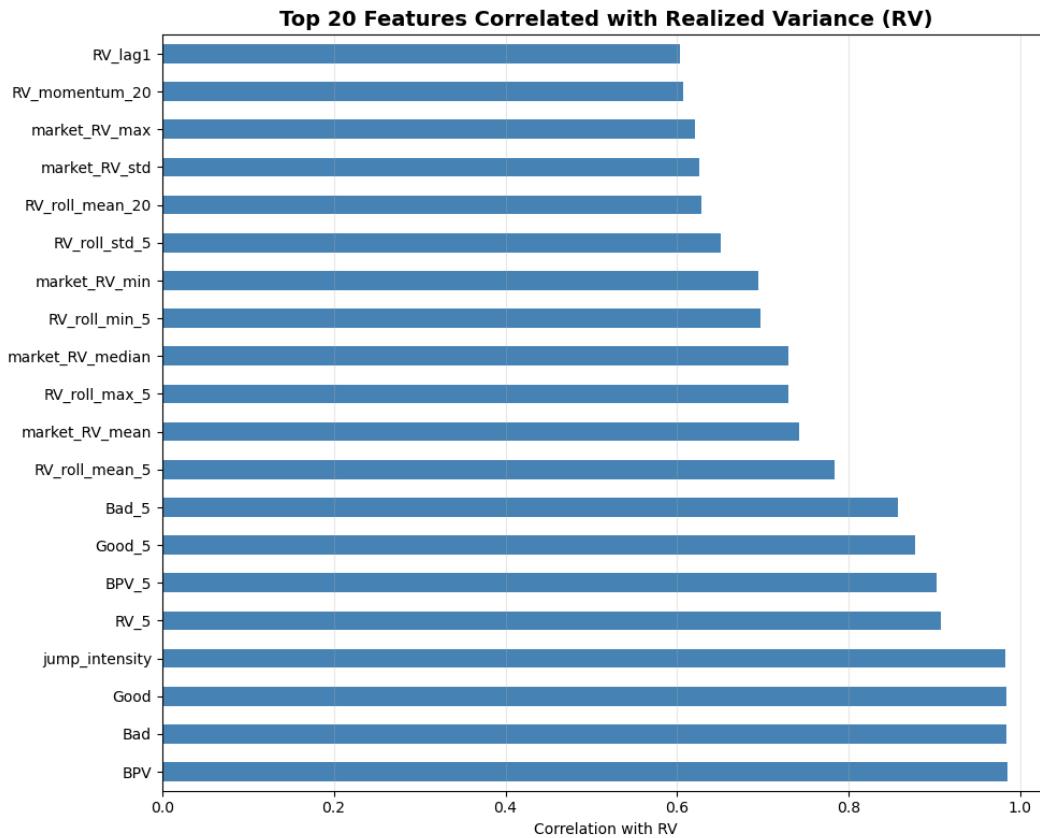
Figure 9: Top engineered features ranked by absolute correlation with next–day realized variance. Temporal and jump–related features dominate, along with market–level volatility measures.
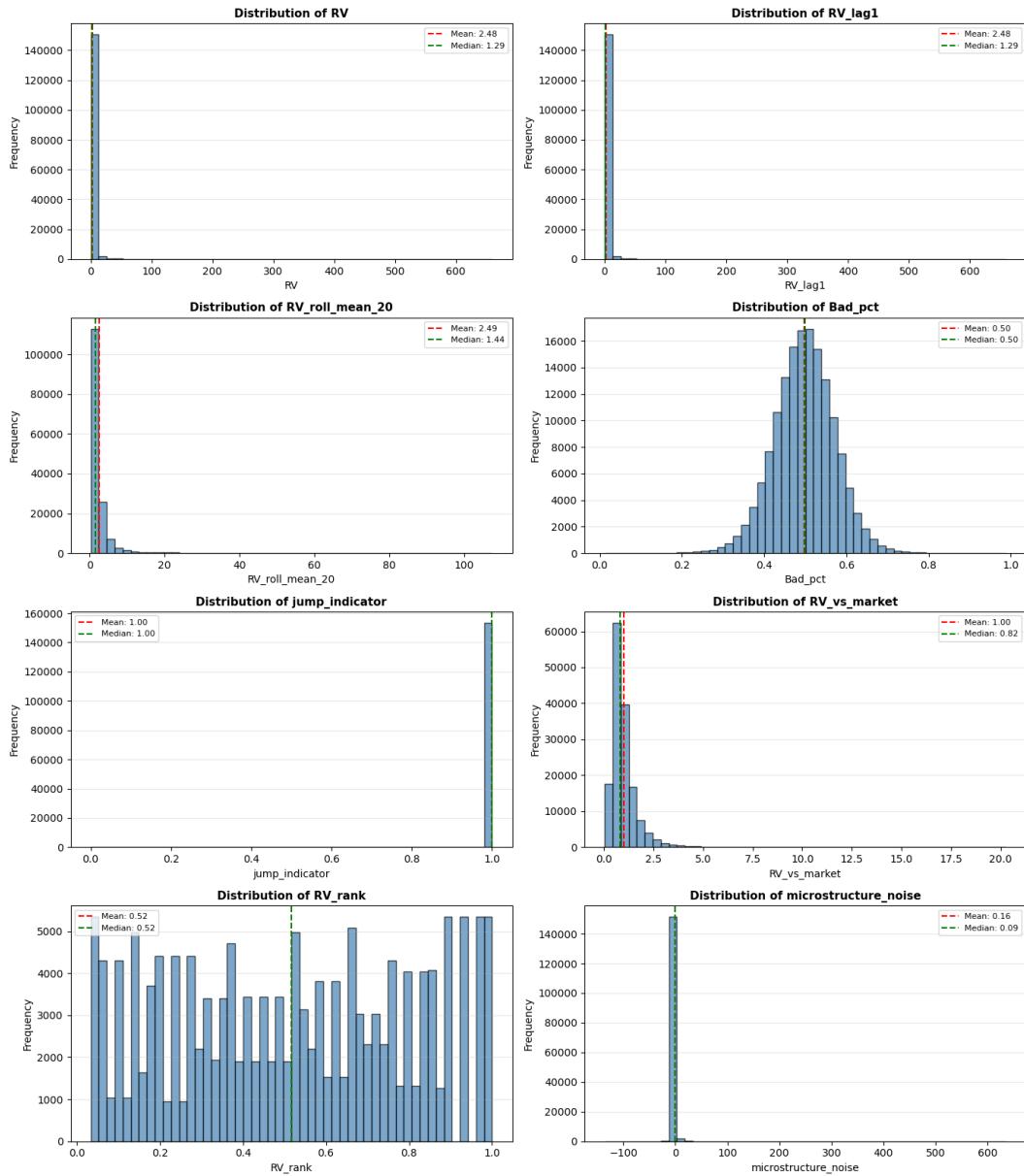
Figure 10: Distributions of selected engineered features, including lagged RV, rolling means, jump shares, relative volatility, and volatility rank.
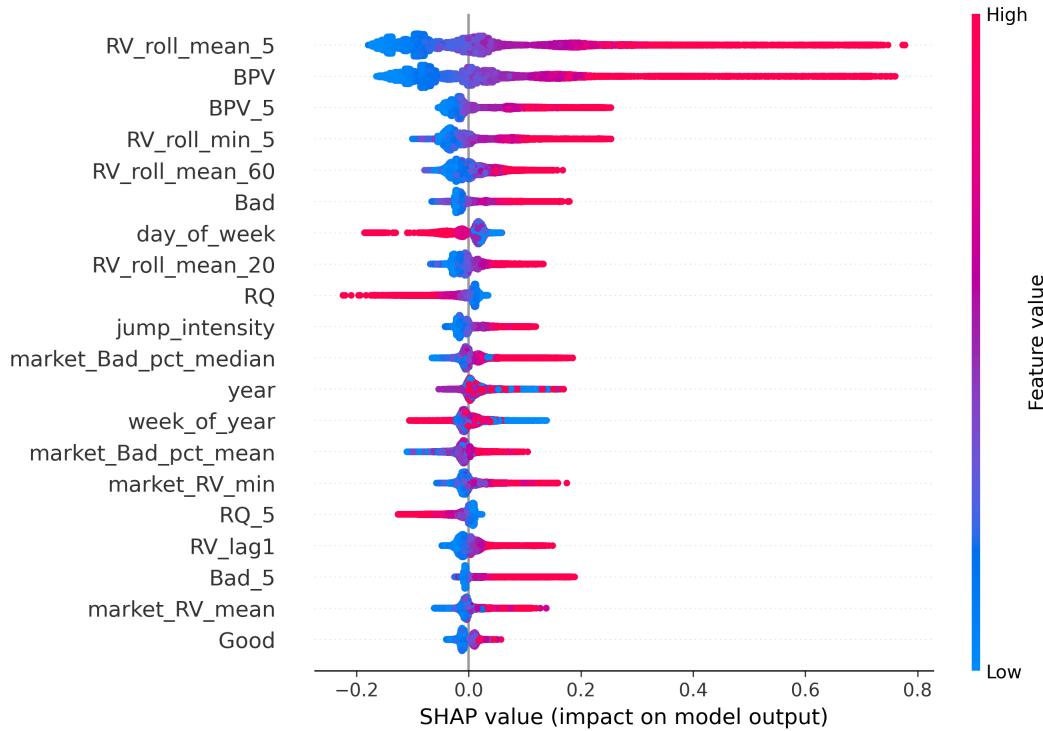
Figure 11: SHAP summary plot for the tuned LightGBM model. Each point represents a stock–day observation; color encodes feature value and horizontal position indicates its contribution to predicted log RV.
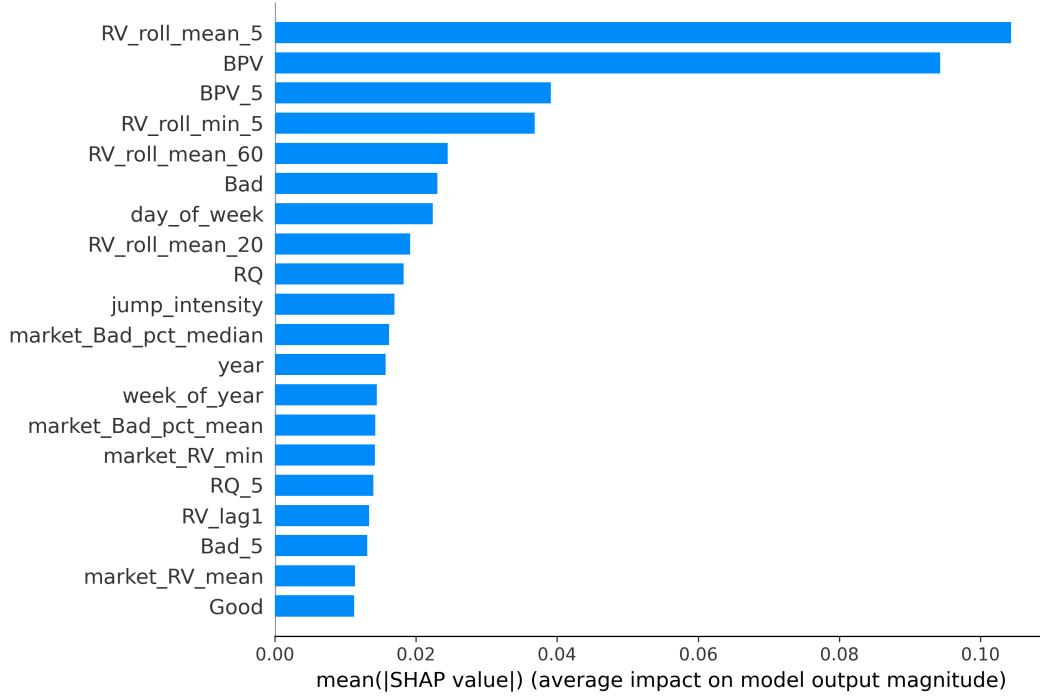
Figure 12: Mean absolute SHAP values for the top features. Short–horizon rolling RV, BPV measures, downside–risk variables, and market–level downside semivariance dominate the model's predictions.
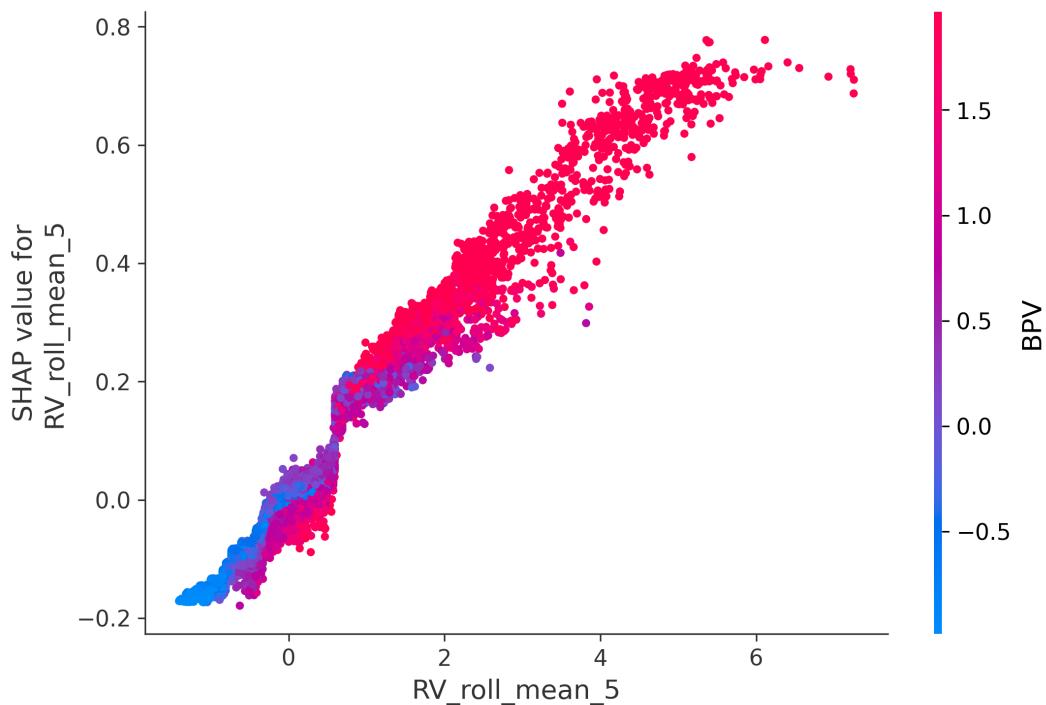

Figure 13: SHAP dependence plot for $RV\_roll\_mean\_5$. Higher recent volatility strongly increases predicted next–day RV, with color showing interaction with BPV.
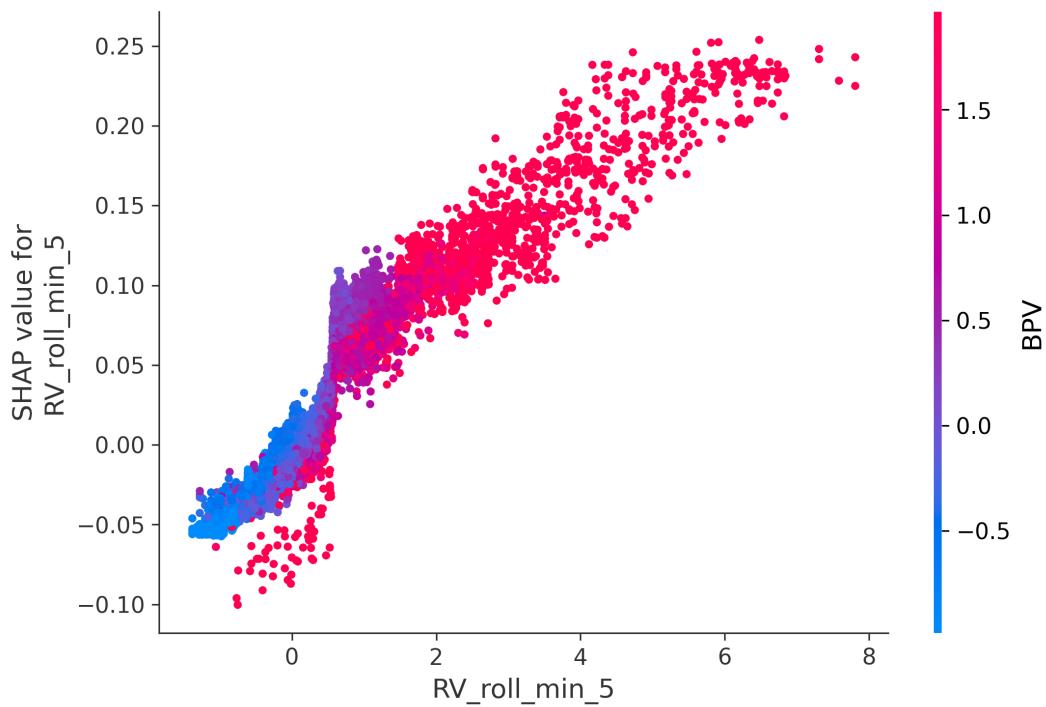
Figure 14: SHAP dependence plot for $RV\_roll\_min\_5$. Low recent minima correspond to negative contributions, while high minima shift forecasts upward.
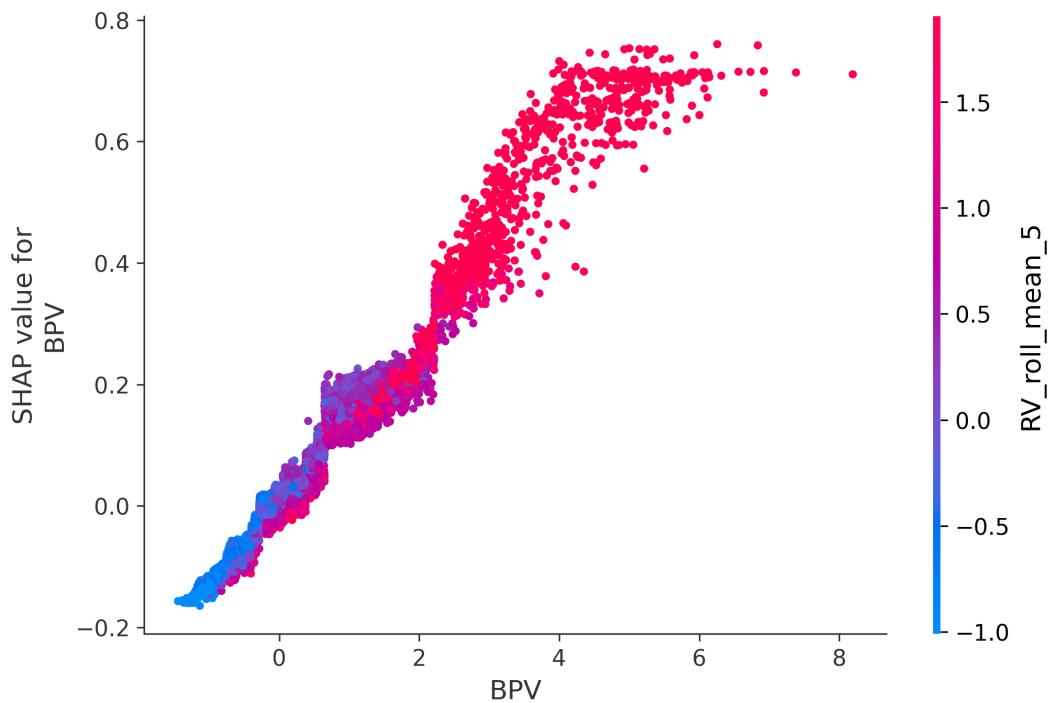
Figure 15: SHAP dependence plot for BPV. Large bipower variation values are associated with substantially higher predicted volatility, especially when combined with high short–run RV rolling means (color).
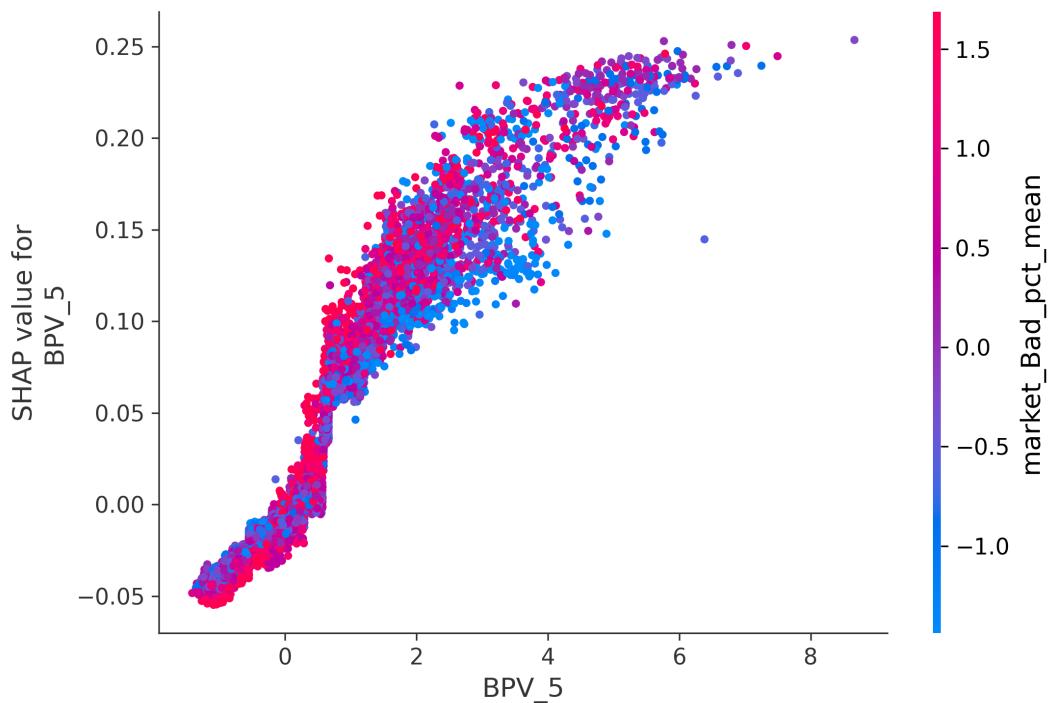
Figure 16: SHAP dependence plot for BPV_5. The pattern mirrors BPV, confirming the strong role of jump–related and continuous–variation measures in the LightGBM forecasts.