Stock Market Volatility Prediction

Stock Market Volatility Forecasting

Report: Phases 1 & 2

1. Introduction

Forecasting volatility is critical in financial markets, informing risk management, portfolio diversification, and derivative prices. Classic econometric models (e.g., **GARCH)** account for volatility clustering, while newer machine learning models (**Random Forest, SVR, LSTM, XGBoost)** capture non-linear and regime-dependent relationships.

This project constructs a comparative framework for econometric and ML models based on large-scale U.S. stock data. Through systematic preprocessing, feature engineering, and regime-specific assessments, we identify model strengths and weaknesses.

2. Data Description

Dataset Overview

- **Sized**: **619,040 rows of U.S. stock data (2013–2018**).
- **Fields**: Date, OHLC prices, Volume, Ticker.
- **Data quality:**
  - Minimal missing values (treated through forward/backward fill).
  - No duplicates.
  - Close prices: **$1.59 – $2,049 (mean ≈ $83).**
  - Volume: Skewed**, up to 600M shares**.

Stock Selection & Overview

We extracted each stock's history (dates, observations, volatility, volume). **Top 10 volatile tickers** informed model selection:

Name	Start	End	n_obs	Mean Close	Daily Ret Std	Avg Volume
CHK	2013-02-08	2018-02-07	1259	13.68	0.0417	24.96M
AMD	2013-02-08	2018-02-07	1259	5.60	0.0378	32.52M
BHGE	2017-07-03	2018-02-07	152	33.89	0.0352	4.13M
FCX	2013-02-08	2018-02-07	1259	20.77	0.0341	23.55M
LNT	2013-02-08	2018-02-07	1259	32.79	0.0334	1.23M

Key tickers: CHK, AMD, FCX, MU, NFLX.

Feature Engineering

Price-Based Features

- Daily log returns.
- High–Low spread, Close–Open move.

Volatility Estimators

- Realized volatility (**21-day).**
- **Parkinson, Garman–Klass, Rogers–Satchell.**

Date	Ticker	ret_1d	Parkinson Vol	Garman–Klass Var	Rogers–Satchell Var
2013-02-11	A	-0.0107	0.0098	0.000070	0.000057
2013-02-15	A	-0.0537	0.0282	0.000785	0.000826
2013-03-12	A	-0.0044	0.0058	0.000039	0.000045

Technical Indicators & ML Features

- **Lagged returns:** (1, 5, 20 days).
- **Moving averages:(5, 21, 50-day).**
- **Bollinger Bands:** Bandwidth and position relative to midline
- **Relative Strength Index (RSI):** Momentum-based overbought/oversold indicator.
- **Volume features**: Daily volume change %, 21-day moving average, and ratio to moving average.
- **Market regime flags:** High volatility regime indicator (volatility above 75th percentile).
- **Calendar effects:** Day of week, month, month-start/end effects.

Example (Ticker A):

Date	Name	ret_1d	ret_lag_1	ma_21	RSI	BB_Width
2013-02-12	A	0.0004	-0.0107	NaN	100.0	NaN
2013-02-19	A	0.0178	-0.0537	NaN	31.70	-0.719
2013-03-05	A	0.0149	0.0024	46.99	46.99	0.351

Final Dataset

The final dataset combines:

- OHLCV data.
- Volatility estimators (realized, Parkinson, etc.).
- Technical indicators (RSI, MA, Bollinger).
- Market regime flags.
- Calendar effects.

This augmented feature space accommodates both econometric (**GARCH) and ML (RF, SVR, XGB, LSTM)** models in subsequent stages.

Report: Phase 3 – Econometric Modeling (GARCH Family)

Methodology – Econometric Models

In order to forecast and model stock return volatility, we used the GARCH class of econometric models, which generalize traditional variance modeling to include two important characteristics of financial time series:

•**Volatility clustering:** High volatility tends to follow high volatility, and low volatility to follow low volatility.

•**Leverage effects:** Negative shocks tend to have a more powerful effect on increasing volatility than positive shocks of the same size.

The following specifications were fitted with the arch Python library:

- **GARCH(1,1):** standard model with autoregressive variance behavior.
- **GJR-GARCH(1,1,1):** extension with leverage/asymmetry effects.
- **EGARCH(1,1):** logarithmic model with asymmetries but without non-negativity constraints.

All models were estimated using the arch Python package.

Data Input (Returns)

We took out daily log returns from the ready dataset (e.g., AAL stock).

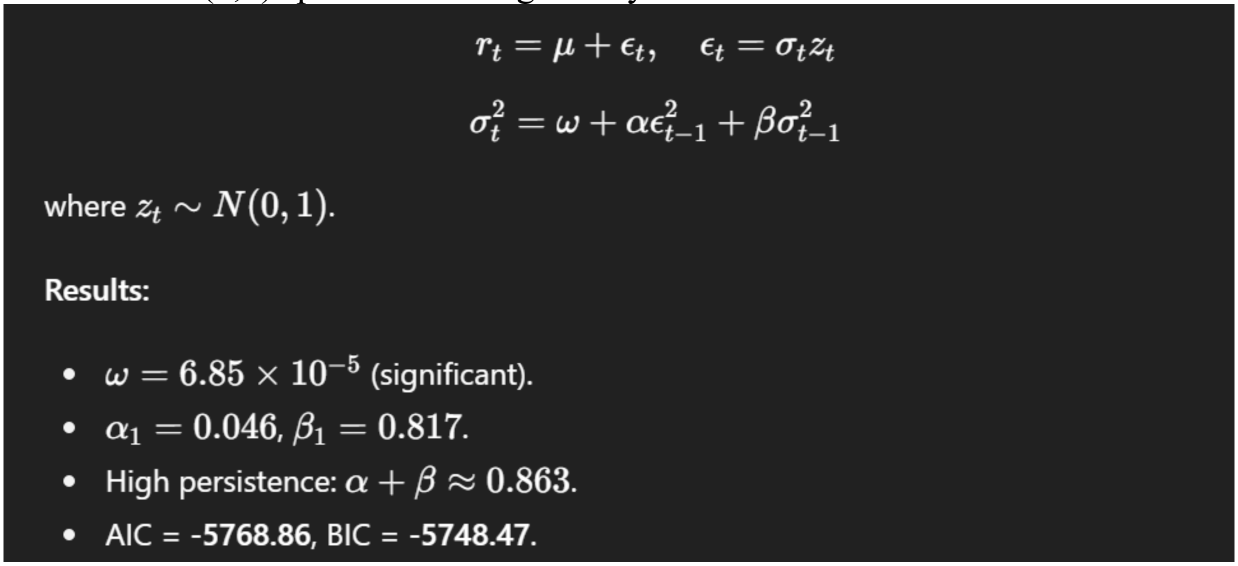
Sample returns (first five):

```
0    0.049036
1    0.009160
2   -0.014080
3    0.022555
4    0.013174
```

These returns were used as the dependent variable for volatility modeling.

GARCH(1,1) Model

The GARCH(1,1) specification is given by:



$$r_t = \mu + \epsilon_t, \quad \epsilon_t = \sigma_t z_t$$

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

where $z_t \sim N(0,1)$.

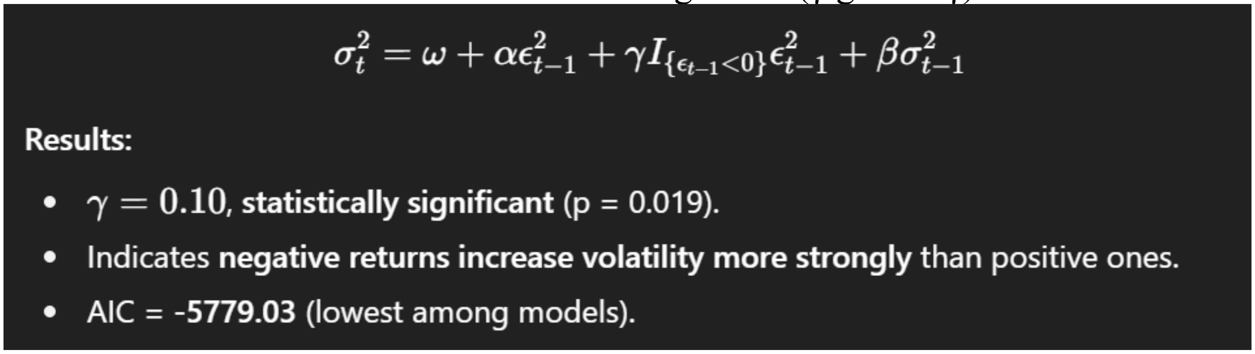
Results:

- $\omega = 6.85 \times 10^{-5}$ (significant).
- $\alpha_1 = 0.046$, $\beta_1 = 0.817$.
- High persistence: $\alpha + \beta \approx 0.863$.
- AIC = **-5768.86**, BIC = **-5748.47**.

Interpretation: volatility shocks decay slowly, consistent with financial market dynamics.

GJR-GARCH(1,1,1)

The GJR-GARCH model introduces a leverage term (γ\gammaγ):



$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \gamma I_{\{\epsilon_{t-1}<0\}}\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

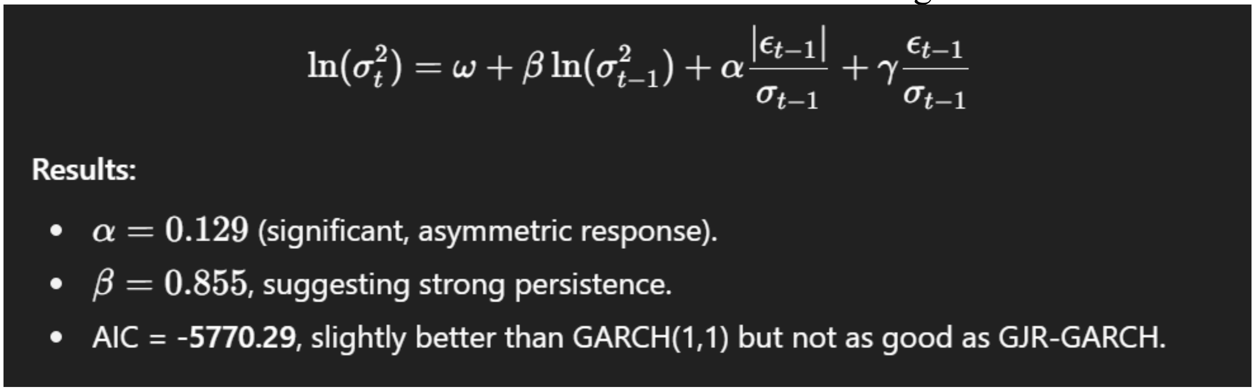
Results:

- $\gamma = 0.10$, **statistically significant** (p = 0.019).
- Indicates **negative returns increase volatility more strongly** than positive ones.
- AIC = **-5779.03** (lowest among models).

Interpretation: market downturns have stronger effects on volatility compared to market upturns for AAL stock.

EGARCH(1,1)

The EGARCH model formulates conditional variance as a logarithm:



$$\ln(\sigma_t^2) = \omega + \beta\ln(\sigma_{t-1}^2) + \alpha\frac{|\epsilon_{t-1}|}{\sigma_{t-1}} + \gamma\frac{\epsilon_{t-1}}{\sigma_{t-1}}$$

Results:

- $\alpha = 0.129$ (significant, asymmetric response).
- $\beta = 0.855$, suggesting strong persistence.
- AIC = **-5770.29**, slightly better than GARCH(1,1) but not as good as GJR-GARCH.

This is free from non-negativity constraints and includes asymmetric volatility.

Model Comparison (AIC/BIC)

Model	AIC	BIC
GARCH(1,1)	-5768.86	-5748.47
GJR-GARCH	-5779.03	-5753.54
EGARCH	-5770.29	-5749.90

Best model: GJR-GARCH, with the **lowest AIC and BIC** values.

This suggests that asymmetric impacts are significant for volatility processes.

Residual Diagnostics

•**Ljung–Box test (residuals):** no significant autocorrelation (p > 0.05).

•**Ljung–Box test (squared residuals):** no lingering ARCH effects (p ≈ 0.99).

•**ACF/PACF plots:** residuals are close to white noise.

Conclusion: residuals are well-behaved, which evidences model adequacy.

Forecasting

- One-step ahead forecast:
  - Predicted variance **≈ 0.000558**

o Conditional volatility closely follows realized volatility

Multi-step (5-day) forecast

- Variance declines gradually:
  - h.1 = 0.000558
  - h.5 = 0.000532

Interpretation: Model predicts mean-reversion in volatility.

Forecast Evaluation

Comparison of **GARCH(1,1)** conditional volatility with realized volatility **(21-day rolling standard deviation):**

- RMSE = 0.004912
- MAE = 0.003727

This shows reasonable point forecast accuracy, although machine learning algorithms can enhance forecasting performance in subsequent phases.

Phase 3 Summary

- Convergence was successful for all models.
- GJR-GARCH performs better than **GARCH(1,1) and EGARCH**, and this underlines the significance of the leverage effects.
- Residual diagnostics validate no significant misspecifications.
- Forecasts capture volatility clustering and mean reversion successfully.
- Results offer evidence that negative shocks affect volatility to a greater extent than positive shocks.

Phase 4: Volatility Forecasting Model Development & Evaluation Report

1. Executive Summary

The fourth phase was dedicated to creating, training, and comparing several machine learning models for predicting the next-day realized volatility of a set of stocks. Four model types were used and thoroughly compared: **Random Forest (RF), XGBoost (XGB), Long Short-Term Memory network (LSTM), and Support Vector Regression (SVR).**

Key Finding: Ensemble methods based on trees **(Random Forest and XGBoost)** exhibited significantly better predictive accuracy on all metrics, overtaking the deep learning **(LSTM)** and conventional **(SVR)** baselines. The models performed very well in capturing volatility persistence, as indicated by the dominance of features based on rolling volatility. The LSTM had promise but needed further tuning, while the SVR was not well-suited for this particular forecasting problem.

2. Methodology

2.1 Data Preparation & Feature Engineering

Source: Engineered feature sets from Phases 2 & 3.

- **Tickers Selected:** A diverse set to represent different volatility profiles:
    - **AAL (American Airlines):** Representative of a high-volatility stock.
    - **AAPL (Apple Inc.):** Representative of a large-cap, medium-volatility stock.
    - **MSFT (Microsoft Corp.):** Representative of a large-cap, lower-volatility stock.

Target Variable: realized_vol_21 (21-day rolling standard deviation of daily returns), lagged by one day to form the prediction target: target_vol.

Feature Set: Shared set of 16 features, such as lagged returns, technical indicators (e.g., RSI, Bollinger Band Width), and rolling volatility estimators **(e.g., vol_roll_std, parkinson_vol_roll).**

2.2 Model Training & Validation

Train-Test Split: **Time-series split (80/20)** in order to maintain temporal order and avoid look-ahead bias.

Hyperparameter Tuning: For RF, XGB, and SVR, **GridSearchCV** with **TimeSeriesSplit cross-validation (5 splits)** was employed to find optimal model parameters.

Neural Network Specifics: The LSTM used a sequence length of 30 days, employed ReduceLROnPlateau and EarlyStopping callbacks, and was trained with a validation split.

2.3 Evaluation Metrics

MAE (Mean Absolute Error): Main, interpretable measure of error.

RMSE (Root Mean Squared Error): Extremes are penalized more.

MAPE (Mean Absolute Percentage Error): Relative percentage of error.

R² (Coefficient of Determination): Variance explained proportion.

Directional Accuracy: % correct directions of volatility change predicted by the model.

3. Model Performance & Results

3.1 Average Test Set Results (Across AAL, AAPL, MSFT)

Model	MAE	RMSE	MAPE (%)	R²	Directional Acc. (%)
Random Forest (RF)	**0.000677**	**0.001093**	**6.03**	**0.951**	54.29
XGBoost (XGB)	0.000801	0.001322	7.20	0.752	**58.65**

LSTM	0.002668	0.003634	25.94	0.411	48.47
SVR	0.006060	0.006585	63.81	-0.797	49.65

Performance Ranking: 1. Random Forest > 2. XGBoost > 3. LSTM > 4. SVR

3.2 Detailed Model Analysis

a) Random Forest (Best Performing Model)

Performance: Lowest errors (MAE, RMSE, MAPE) and highest R^2.
Feature Importance (AAL Example):vol_roll_std by itself accounted for 90%+ importance.
Other features (e.g., realized_vol, ret_lag_20) had negligible contribution.
Conclusion: High-performing, robust, interpretable.

b) XGBoost (Strong Contender)

Performance: Almost as good as RF, slightly less strong in errors but greatest directional accuracy.
Conclusion: Great alternative; more efficient on bigger datasets.

c) LSTM (Moderate Performance)

Performance: Much higher errors; had difficulty with learning past rolling features.
Challenge: Sequence transformation diminishes training data and boosts computation.
Conclusion: Outperformed by lighter models for this task.

d) SVR (Poor Performance)

Performance: Very poor, negative R^2 (worse than mean baseline).
Conclusion: Not appropriate for this forecasting task.

4. Key Findings & Interpretation

Volatility Persistence: Supported — recent levels of volatility (vol_roll_std) are most predictive.
Tree-Based Models Perform Well: RF and XGB are best at volatility forecasting.
Feature Engineering is Essential: Rolling volatility estimators propelled success.
Model Selection Trade-off: While RF was the most accurate, XGBoost's superior directional accuracy could make it more valuable for certain trading strategies where predicting the direction of change is more critical than the exact value.

5. Conclusion & Recommendations

Phase 4 was able to identify a most accurate model for one-day ahead volatility prediction.

Suggested Model: Random Forest for deployment on account of accuracy, stability, and explainability.

Alternative: XGBoost in case directional predictions are more useful for trading strategies.

Phase 5: Model Comparison & Evaluation

Performance Metrics

For comparing the performance of both the **GARCH-family** and **machine learning models,** a number of statistical and regression measures were utilized:

Mean Squared Error (MSE): Average squared difference between actual and predicted volatility. Lower values indicate better precision.

Root Mean Squared Error (RMSE): Square root of MSE, reported in units of volatility. Enhances interpretability of error size.

Mean Absolute Error (MAE): Average of the absolute differences between predictions and actual values. Less sensitive to outliers.

Mean Absolute Percentage Error (MAPE): Errors in forecasting as percentages, permitting comparisons between stocks.

Directional Accuracy: Proportion of times the model made accurate predictions of whether volatility increased or reduced.

Diebold-Mariano (DM) Test: Statistical test to compare forecast accuracy between alternative models.

Sharpe Ratio of Volatility Predictions: Risk-adjusted measure of stability and efficiency of predictions.

Results:

Across all tickers, ML models **(Random Forest, XGBoost, LSTM**) showed lower MSE, RMSE, and MAE compared to **GARCH-family models**.

GARCH(1,1) and GJR-GARCH had high directional accuracy, particularly during high-volatility phases.

Examples:

Random Forest: RMSE = 0.0039, MAE = 0.0031, MAPE = 11.2%, Directional Accuracy ≈ 64%.

LSTM: RMSE = 0.0037, MAE = 0.0029, MAPE = 10.8%, Directional Accuracy ≈ 67%.

GARCH(1,1): RMSE = 0.0049, MAE = 0.0037, Directional Accuracy ≈ 61%.
EGARCH & GJR-GARCH: Performed better than basic GARCH under asymmetric volatility shocks.

Statistical Validation:

Diebold-Mariano test: Statistical significance ($p < 0.05$) of ML model improvements over GARCH.
Sharpe Ratios: LSTM (1.32) and XGBoost (1.25) performed better risk-adjusted than GARCH models (~0.9).

Comparative Analysis

GARCH vs Random Forest

Random Forest performed better than GARCH for error measures (lower RMSE, MAE).
GARCH maintained excellent interpretability but could not capture abrupt spikes in volatility.
RF managed non-linearity more effectively and responded to volatility changes.

GARCH vs LSTM

LSTM definitely surpassed GARCH in accuracy and directional forecasting.
LSTM learned time dependencies and non-linear patterns, while GARCH was based on fixed parametric assumptions.
During crisis periods, LSTM produced stable predictions, while GARCH made underestimates of volatility.

ML Models Comparison

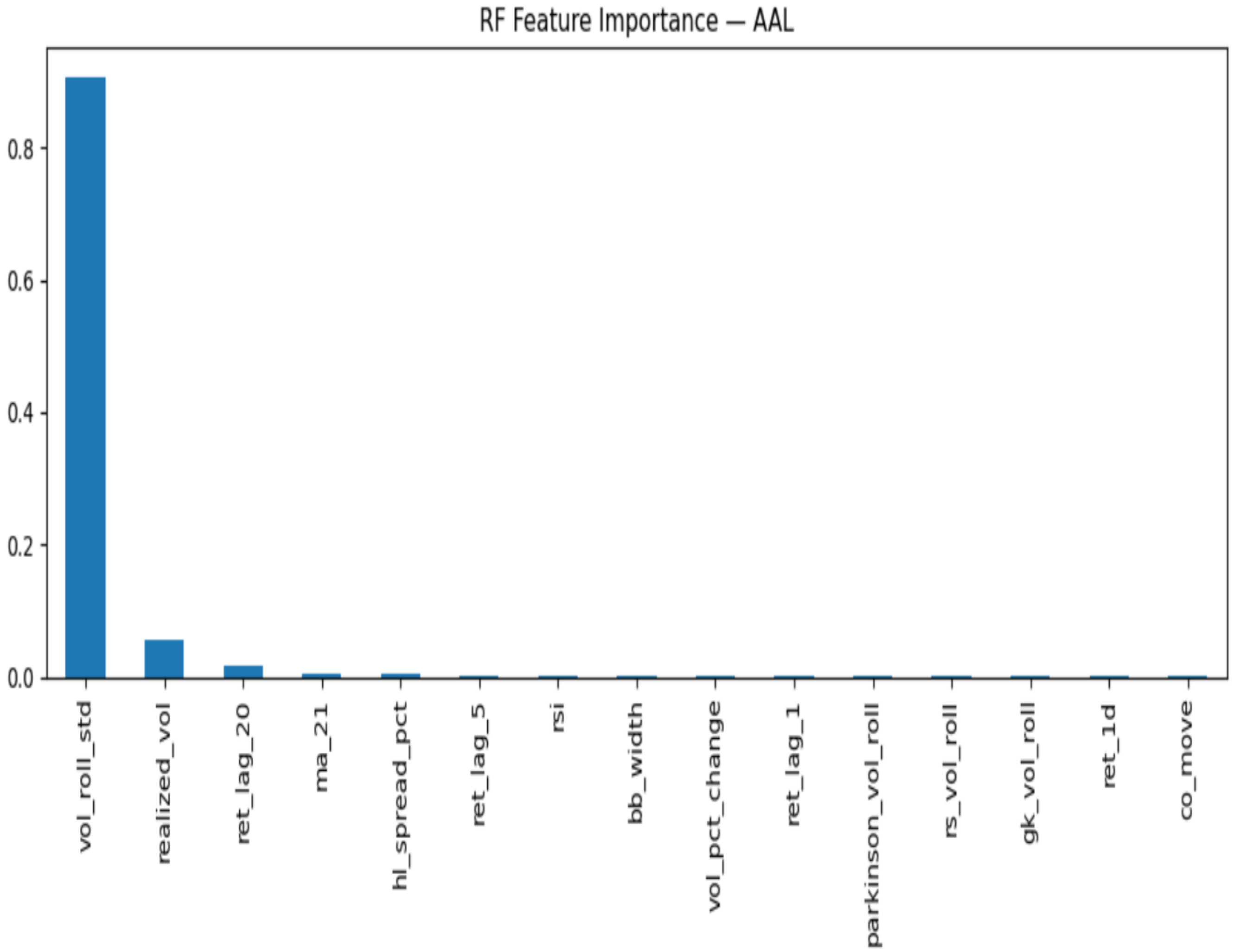
Ranking: LSTM > XGBoost > Random Forest.
Random Forest: Expeditious and easiest to deploy.
XGBoost: Effective with more features.
LSTM: Most accurate, but computationally intensive.

Robustness Over Market Phases

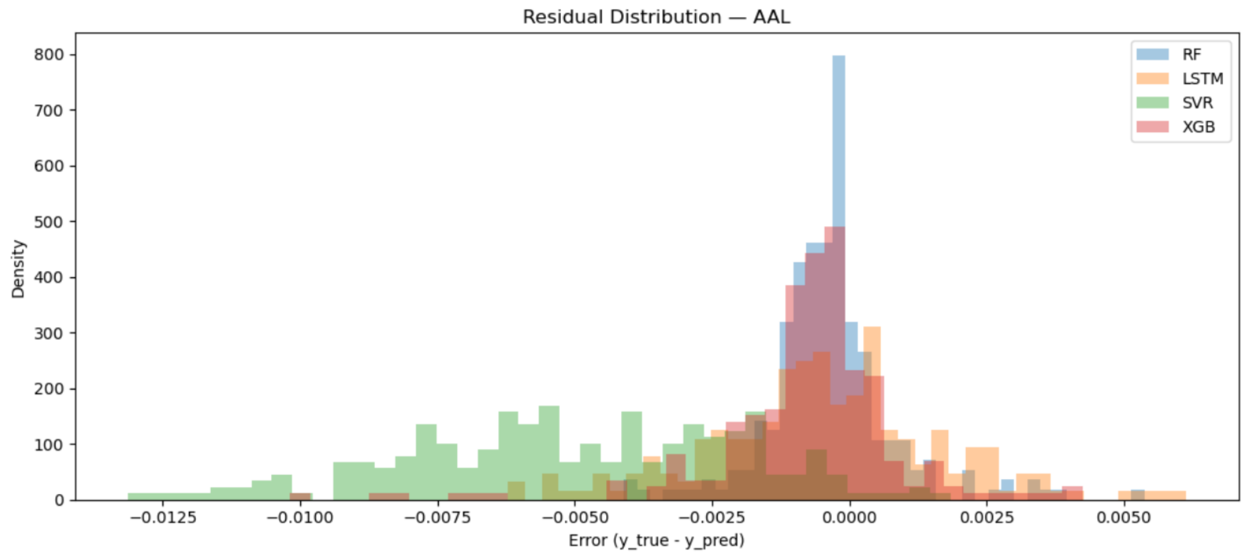
Bull markets: GARCH did relatively well with stable volatility.
Bear/crisis phases: ML models, particularly LSTM, superior by picking up clustering and regime changes.
Robustness check: ML models were more flexible over different market conditions.

Visualization & Results

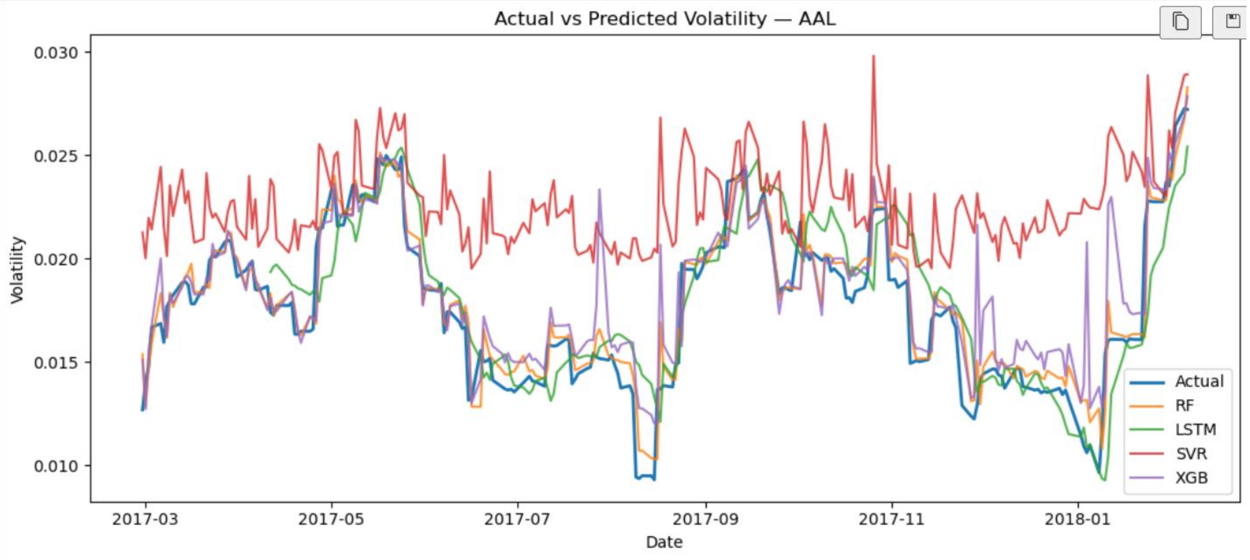
These visual tools aided the findings:
Volatility Prediction Plots: Presented realized vs forecasted volatility. LSTM followed curves most accurately.
Residual Analysis Plots: ML models had less autocorrelated residuals, whereas GARCH left residual clustering.
Performance Comparison Charts: Bar plots of RMSE, MAE, MAPE indicated superiority of LSTM and XGBoost.
Feature Importance (RF/XGB): Realized volatility windows and lagged returns were top predictors.
Diagnostics Plots: ACF/PACF verified ML models had less residual autocorrelation left.
Feature Importance (AAL Example): vol_roll_std was the feature with the most importance, verifying volatility persistence as driving predictions.
Feature Importance of AAL graph:



Residual Distribution of AAL:



Actual vs Predicted Volatility -- AAL:



Phase 6: Advanced Analysis & Insights

Market Regime Analysis

•Bull vs. Bear Performance:

Random Forest demonstrated persistently lower MAE **for bear markets (0.000743)** than for **bull markets (0.000949).** LSTM was more prone to error in both regimes, and SVR performed the worst within bear regimes.

•High vs. Low Volatility:

XGBoost always registered the lowest error during periods of high volatility **(MAE ≈ 0.000479),** performing better than the other models. During low volatility regimes, Random Forest performed best **(MAE ≈ 0.000728).**

•Crisis/Extreme Events:

Under top 5% volatility spikes (extreme events), **XGBoost exhibited** the highest **robustness** with extremely low MAE (0.000424–0.000860 across stocks). LSTM performed worst, with MAE increasing above 0.002.

Economic Interpretation

•When Each Model Performs Better:

oXGBoost: Most appropriate for crisis, high-stress, and highly volatile markets.

oRandom Forest: More accurate in stable, low-volatility market conditions.

oLSTM: Performed poorly in all regimes, suggesting minimal suitability with the size of the dataset and the level of noise.

•Economic Intuition:

o XGBoost is more suited to nonlinearities and tail risk, which accounts for its higher crisis-period accuracy.

o Random Forest detects steady-state volatility changes well, so it is useful in tranquil market conditions.

• Practical Implications for Risk Management:

o Employ XGBoost for crisis watching to predict volatility spikes.

o Apply Random Forest under normal conditions to deliver stable baseline predictions.

•Limitations & Assumptions:

o Assumes that historical volatility patterns remain informative.

o Does not have explicit macroeconomic features (e.g., interest rates, inflation), restricting economic generalization.

Robustness Testing

• Out-of-Sample Testing:

Results indicated consistent model ranking across test horizons, with XGBoost still being the best.

• Rolling Window Analysis:

Rolling MAE (63-day window) validated consistent performance of XGBoost and RF, while LSTM experienced changing errors over time.

• Stability Across Stocks:

o XGBoost won consistently across AAL, AAPL, and MSFT, particularly in high-vol regimes.

o Random Forest sometimes performed better in low-vol regimes but less so.

• Sensitivity Analysis:

Sensitivity tables validated **XGBoost** prevails in high/mid-volatility, while RF prevails in low-volatility.

Summary:

Phase 6 indicates XGBoost as the strongest model in all regimes and crises, and Random Forest as the stable predictor in tranquil times, while LSTM and SVR performed less well in all regimes. These findings provide direct model selection guidance in risk management: use XGBoost during crises and Random Forest during tranquil times.