Forecasting S&P 500 Realized Volatility

Rocco Pio Lorenzo Ventruto
EPFL
MSc in Financial Engineering
rocco.ventruto@epfl.ch

***Abstract*—This study performs a strict walk-forward comparison of econometric (GARCH, EGARCH) and machine-learning (Elastic Net, XGBoost, LSTM) models for one-day-ahead forecasting of S&P 500 realized volatility, using a comprehensive feature set comprising lagged realized-volatility measures, implied volatility (VIX), microstructure proxies, and cross-asset regime indicators. We establish a clear performance hierarchy in which XGBoost delivers the most accurate forecasts, with gains concentrated in volatility spikes where linear specifications tend to underreact. Feature-relevance analysis shows strong agreement across model classes: both Elastic Net and XGBoost consistently identify implied volatility and short-horizon realized-volatility lags as the primary drivers. SHAP analyses reveal that XGBoost's advantage arises less from identifying new key predictors than from learning nonlinear, state-dependent and interaction effects among the same core features, especially during high-stress regimes. Finally, augmenting the feature space with EGARCH-based forecasts and fitted parameters yields no material performance improvements, suggesting that these inputs are largely redundant given the already rich predictor set.**

I. Introduction

Volatility forecasting plays a central role in financial econometrics, risk management, option pricing, and portfolio allocation. Despite decades of research, accurately predicting market volatility remains challenging due to its latent nature, abrupt regime changes, asymmetric responses to shocks, and highly persistent dynamics. Traditional econometric models such as GARCH variants capture key stylized facts of volatility, yet their parametric structure can limit their ability to adapt to complex nonlinear patterns. In contrast, machine-learning methods can accommodate large engineered feature sets and naturally allow for nonlinear responses, threshold effects, and interactions. The empirical question, however, is not whether machine learning is flexible in principle, but whether it delivers economically meaningful gains in a strict walk-forward forecasting setting in which models are repeatedly re-estimated and evaluated out of sample.

This study conducts a rigorous comparative analysis of one-day-ahead S&P 500 realized volatility forecasting, computed from high frequency data, across a progressive set of model classes. We escalate complexity systematically from parametric baselines (GARCH and EGARCH) to regularized linear learning (Elastic Net), nonlinear ensemble methods (XGBoost), and deep recurrent architectures (LSTM). Predictors are constructed from a comprehensive information set combining high-frequency realized measures, option-implied volatility (VIX), market microstructure signals (volume and true range), and cross-asset regime indicators (gold, oil, and interest rates).

Overall, the paper addresses three questions: (i) which model class delivers the most accurate out-of-sample forecasts of S&P 500 realized volatility under a consistent rolling evaluation protocol; (ii) which predictors and market conditions primarily drive these forecasts, and to what extent nonlinearities and interaction effects explain performance differences between models; and (iii) whether augmenting the feature space with EGARCH-based volatility forecasts and fitted parameters improves predictive accuracy relative to the baseline specification.

II. Literature Review

A large empirical literature documents that financial-market volatility is time-varying and exhibits strong persistence, with periods of calm followed by periods of turbulence. The ARCH framework [1] and its generalization to GARCH [2] provide a parametric representation of this clustering through a conditional variance recursion. While GARCH models are effective baselines, volatility dynamics are also known to be asymmetric with negative return shocks that tend to raise future volatility more than positive shocks of the same magnitude. This leverage effect motivates asymmetric specifications such as EGARCH [3].

For machine-learning models, forecast accuracy depends on the information set available. When many predictors are potentially informative and correlated, regularized linear models provide a disciplined way to exploit a richer feature set while controlling overfitting. The Elastic Net [4] stabilize estimation under multicollinearity and perform variable selection, making it an attractive bridge between classical econometric interpretability and high-dimensional predictors.

Empirically, recent work applying machine learning to realized-volatility forecasting finds that nonlinear methods can improve predictive accuracy, especially when rich predictors or cross-sectional information are available [5] [6]. Related evidence suggests that boosting can be competitive for short-horizon realized-volatility forecasts [7]. This motivates our progressive modeling strategy, moving from linear regularization to nonlinear boosting and then to deep recurrent architectures within a unified out-of-sample protocol.

An hypothesis is that combining econometric structure with flexible machine learning can yield further gains. Hybrid approaches augment machine learning inputs with GARCH-type forecasts or parameters, aiming to provide summaries of persistence and asymmetry that may be difficult to learn robustly in small samples. For example, [8] propose hybrid LSTM–GARCH-type constructions and report improvements in certain settings. Nevertheless, whether hybridization improves forecasts is ultimately an empirical question and may depend on how informative the baseline feature set already is, and whether the learning model can internalize the same mechanisms from the raw predictors.

A key difficulty with nonlinear forecasting models is economic interpretation: split-based importances do not provide direction or regime dependence. SHAP values provide an additive decomposition of individual predictions into feature contributions grounded in Shapley values [9]. This enables analysis of nonlinear marginal effects and interaction effects, which is central to understanding why nonlinear models (such as boostedtrees) might outperform linear regularized benchmarks in volatility forecasting.

Relative to the existing literature, the main interest of our approach is offering a coherent, comparative and interpretable, evaluation under realistic forecasting constraints: all models are trained and assessed in a strictly out-of-sample walk-forward environment using the same rolling window design, the same engineered feature set, and the same evaluation metrics. This provides a clean attribution of performance differences to modeling capacity (linearity vs nonlinearity, interaction learning, and regime adaptation), rather than to differences in data handling or information sets. Moreover, the analysis relies on a deliberately rich feature set ensuring that any performance gains reflect differences in modeling capacity rather than limitations of the information set.

III. Data Description and Sourcing

To construct a robust forecasting framework for S&P 500 volatility, we aggregate information from multiple sources, combining high-frequency realized volatility measures with daily macro-financial and cross-asset indicators. The dataset spans the period from 3 january 2000 to 16 february 2021 ensuring that the analysis captures several major market regimes.

The core of our target variable, Realized Volatility, is sourced from the Oxford-Man Institute's Realized Library. This database is a standard in academic volatility research, providing realized measures constructed from high-frequency (intraday) tick data of S&P 500. In addition, we utilize the Yahoo Finance API as the primary source for all daily-frequency predictor variables used in our model. From this interface, we extracted historical time-series data including the Open, High, Low, and Close (OHLC) prices, as well as total trading volume. We use data for S&P 500 Index (Thicker:ˆGSPC), Gold Futures contracts (GC=F), Crude Oil West Texas Intermediate Futures contracts (CL=F) and Implied Volatility from CBOE volatility Index (ˆVIX). Moreover, the 10 year treasury yield (DGS10) is sourced from the Federal Reserve Economic Data API.

Because the S&P 500, VIX, gold futures, crude oil futures, and the 10-year Treasury yield follow different trading calendars, merging them into a single dataset with outer joins naturally produces missing observations. Equity markets close on U.S. holidays and weekends, whereas futures markets often operate on extended or overnight schedules, resulting in dates on which gold or oil prices are available but equity-based variables are not, and vice versa. Similarly, Treasury yields are not published on certain market holidays. These missing values were forward-filled to maintain a fully aligned daily panel. This procedure affects less than 0.5% of the sample and prevents unnecessary loss of observations during feature construction while avoiding look-ahead biais.

IV. The Target Variable

The primary objective of this study is to forecast the one day ahead volatility of the S&P 500 index. Unlike asset prices, volatility is a latent variable that cannot be directly observed but must be estimated. In traditional econometrics, squared daily returns (r_t^2) are often used as unbiased but noisy proxy for daily variance [10]. To overcome this limitation, we utilize Realized Volatility (RV), a non-parametric estimator constructed from high-frequency intraday data. By summing squared returns over 5 minuts interval. The formula for Realized Variance on day t is given by:

$$RV_t = \sum_{i=1}^{N} r_{t,i}^2$$

where $r_{t,i}$ is the logarithmic return of the i-th 5-minute interval on day t.

To obtain our definitive target variable, we compute the square root of the daily Realized Variance to obtain the Realized Volatility ($\sqrt{RV_t}$), and subsequently apply a logarithmic transformation $y_t = \ln\left(\sqrt{RV_t}\right)$, wich ensures stability in the optimization process. Forecast are then exponentiated to obtain volatility in levels.

V. The Features

We construct a comprehensive set of predictors, grouped into five main categories, to capture multiple dimensions of financial markets that may influence future S&P 500 volatility. Whenever appropriate, the features are expressed in logarithmic form to stabilize variance, reduce skewness, and improve stationarity.

A. S&P 500 Volatility Features

To exploit the well-established empirical fact that volatility clusters, we construct logarithmic lagged features of our intraday realized volatility measure at the 1-day, 5-day (weekly), and 22-day (monthly) horizons. To complement historical volatility information, we also incorporate the VIX index to obtain a forward-looking measure of market-implied volatility. Specifically, we use the logarithm of the VIX level and the daily logarithmic change in the VIX, capturing both the market's expectation of future volatility and short-term shifts in risk sentiment.

B. S&P 500 Microstructure Features

To capture intraday trading dynamics and market microstructure conditions, we start by computing the Relative Volume, which is defined as daily trading volume scaled by its 60-day moving average, providing a scale-free indicator of abnormal trading activity often associated with liquidity stress or order-flow imbalance. Second, the True Range captures total daily price variation by incorporating both intraday movements and overnight gaps, allowing the model to account for jump risk and news-driven price discontinuities. Together, these microstructure features complement traditional volatility measures by capturing information about intraday trading intensity and overnight volatility pressures.

C. S&P 500 Momentum and Distribution Features

To capture nonlinear return dynamics that may precede volatility shifts, we incorporate several distributional and momentum-based features. First, the 60-day rolling kurtosis measures the heaviness of return tails and provides a forward signal of potential volatility surges. We also include short-, medium-, and long-horizon momentum indicators computed over 1-, 5-, and 22-day windows, which reflect directional pressure in price movements and may signal trend continuation or reversal mechanisms relevant for volatility forecasting. Finally, we compute a 14-day Relative Strength Index (RSI), a bounded oscillator designed to quantify overbought or oversold conditions that might accompany shifts in market sentiment and subsequent changes in volatility.

D. Macro-Financial Regime Features

To capture broader macro-financial conditions that influence equity volatility, we incorporate a set of cross-asset indicators designed to detect transitions between risk-on and risk-off environments. We first compute daily log returns for the S&P 500, gold and crude oil futures, as well as basis-point changes for the 10-year U.S. Treasury yield. These assets are highly sensitive to macroeconomic forces such as inflation, growth expectations, and monetary policy, and their co-movements often reveal shifts in market sentiment that precede changes in equity volatility. In addition, we construct 60-day rolling correlations between the S&P 500 and gold, crude oil, and the 10-year yield, respectively. These correlations act as regime detectors. We further include True Range measures for gold and oil to capture intraday stress, liquidity conditions, and jump risk in macro-sensitive asset classes. Finally, we compute Garman–Klass volatility estimators for gold and oil, which provide efficient daily volatility measures by exploiting the full Open–High–Low–Close structure of prices. The Garman–Klass estimator combines intraday range information with opening and closing levels to reduce estimation noise relative to purely close-to-close measures [11].

In the end we drop the first 60 rows of our dataset to avoid Nan values due to rolling correlations and moving average from all of our features construction.

E. EGARCH Features

To assess the effectiveness of hybrid econometric–machine learning models within our feature framework, we incorporate parameters derived from an estimated E–GARCH(1,1) specification. Specifically, we retain the four estimated model coefficients:ω, α, β, and γ, which respectively capture the long-run volatility level, the sensitivity of volatility to the magnitude of return shocks, the persistence of conditional variance, and the asymmetric leverage effect. In addition to these structural parameters, we include the model-implied conditional log-variance series. Details on the model will follow in the Section VI.

VI. General Methodology

A. Rolling Window Framework

Financial time series are inherently non-stationary, exhibiting structural breaks, regime shifts, and evolving volatility dynamics. To account for this, we employ a rolling-window (walk-forward) forecasting framework, which ensures that each model is trained only on the most recent data and evaluated on strictly out-of-sample observations.

At each iteration, a fixed-length window of past observations is used as the training set. Let the window contain $W = 2000$ days of historical data. For day t, the model is estimated using data from $(t - 2000 + 1)$ to t, and a one-day-ahead forecast $\hat{y}_{t+1}$ is generated. Although forecasts are produced every day, the model parameters are re-estimated only every 20 days (the refit frequency). After each refit, the training window is shifted forward and the model is retrained. This rolling procedure is repeated across the entire sample, which spans, after feature construction and cleaning, the period from 2 April 2001 to 16 February 2021. For XGBoost and LSTM models, each training window is further divided into a training segment (80%) and a validation segment (20%), used solely for hyperparameter tuning and early stopping, while testing remains strictly out-of-sample. This setup replicates a realistic real-time forecasting environment. The rolling-window procedure is applied uniformly across all models, ensuring a fair and consistent performance comparison under identical forecasting constraints. Note that models are retrained, with proper hyperparameter optimisation, every 20 days to reflect real computational constraints, especially for deep learning architectures, which require significant computational resources, and ensure a fair comparison between models.

B. Preprocessing Pipeline

To limit the influence of extreme outliers, we apply a mild winsorization procedure that preserves the integrity of the underlying distribution. Values falling below the 0.5th percentile or above the 99.5th percentile are clipped to these respective bounds. These percentile thresholds are computed exclusively on the training set of each rolling window and then applied on the train set and to the corresponding test set, as well as to the validation set for the XGBoost and LSTM models. In addition, input features are standardized using Z-score normalization to ensure comparability across variables with different scales.

The mean and standard deviation are estimated on the training set only, and the resulting transformation is applied to the train, test and validation sets. Importantly, winsorization or standardization are not applyed to the target variable, nor do we preprocess the return series used as inputs for the GARCH and EGARCH models. The entire preprocessing pipeline described above is repeated dynamically for each rolling window, ensuring that no look-ahead bias is introduced.

C. Overview of Model Classes

We select our set of models to follow a coherent methodological progression, moving from traditional econometric approaches to machine-learning and deep-learning techniques. This structured escalation allows us to isolate the incremental contribution of each modeling paradigm to the volatility forecasting problem. We begin with the **GARCH(1,1)** model, a classical benchmark that captures volatility clustering. We then extend this framework to **EGARCH(1,1)**, which additionally accounts for the leverage effect: the asymmetric response of volatility to positive and negative shocks.

Next, we evaluate the predictive power of a linear machine-learning model, the **Elastic Net**, which combines L1 and L2 regularization and leverages our full set of engineered features. This model enables us to assess how much forecasting performance can be gained from a richer feature space under linear constraints. To explore nonlinear relationships, we introduce **XGBoost**, a gradient-boosted tree algorithm capable of capturing complex interactions and regime-dependent effects that linear models cannot detect. We then move to a deep-learning architecture, the **LSTM**, designed to model long-range temporal dependencies and nonlinear dynamics in sequential data, thereby allowing us to evaluate whether deep learning provides a tangible improvement over both Econometric and classical ML models.

Finally, as an additional layer of complexity, we construct **Hybrid Versions** of the Machine Learning and Deep Learning models by augmenting their feature sets with the one-day-ahead EGARCH volatility forecast and the associated fitted parameters. This allows us to assess whether combining Econometric structure with Machine Learning methods can enhance predictive performance within our feature ecosystem.

D. Building EGARCH features

Our realized volatility target data begins on 3 January 2000, and this early availability allows us to estimate the EGARCH model well before the start of the machine-learning training period. We therefore initialize the procedure with an expanding estimation window starting from this date, ensuring that the model benefits from a sufficiently long burn-in period and produces stable parameter estimates by the time our main training sample begins. Once the expanding window reaches a maximum length of 2000 observations, the procedure transitions to a fixed-size rolling window, as described in Subsection VI-A. Model fitting is performed by maximum likelihood under Gaussian innovation assumptions. The corresponding fitted parameters and the associated one-step-ahead volatility forecasts are then stored as regular features for use in the machine-learning models.

E. Evaluation Metrics and Comparison Criteria

To assess predictive performance, we rely on evaluation metrics commonly used in the volatility forecasting literature. Although the models are trained on the logarithm of realized volatility to stabilize variance and improve numerical behavior, all evaluation metrics are computed in volatility space: both forecasts and true values are exponentiated back to their original scale before calculating RMSE and out-of-sample R^2. Our primary accuracy measure is the Root Mean Squared Error (RMSE), which penalizes large forecasting errors more heavily than small ones and is therefore particularly appropriate in a volatility context where missing a volatility spike is far more consequential than misestimating a tranquil period. By contrast, the Mean Absolute Error (MAE), which treats small and large errors linearly, is less suited for applications in which tail events carry significantly greater financial risk.

We also report the out-of-sample coefficient of determination, R^2_{OOS}, which measures each model's performance relative to a benchmark forecast. Because volatility is highly persistent, the appropriate baseline is not the historical mean but a random-walk predictor in which tomorrow's volatility level is set equal to today's. A positive R^2_{RW} indicates that the model captures predictive information beyond this persistence benchmark, while a negative value signals inferior performance. This relative metric follows the Campbell-Thompson formulation [12] and provides a transparent, model-agnostic comparison across approaches with fundamentally different structures. For completeness, we also report the conventional out-of-sample coefficient of determination, R^2_{mean}, computed relative to the historical mean. This metric is included for descriptive purposes, as the mean provides a poor benchmark in highly persistent volatility series .

Finally, to guarantee a fair comparison, all models rely on the same feature set, forecasting horizon, rolling-window procedure, and preprocessing pipeline. Any observed differences in predictive accuracy therefore reflect intrinsic modeling capabilities rather than inconsistencies in experimental design.

VII. Models

A. GARCH(1,1)

The Generalized Autoregressive Conditional Heteroskedasticity model is the classical benchmark for modeling and forecasting financial market volatility.

The model assumes that returns follow

$$r_t = \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

with a conditional variance evolving according to

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

Here, ω governs the long-run variance level, α measures the impact of new information through the previous squared shock r_{t-1}^2, and β captures the persistence of volatility via the lagged conditional variance σ_{t-1}^2.

In our implementation, the model is estimated by maximum likelihood under a Gaussian innovation assumption. Although financial returns are not normally distributed, the Gaussian likelihood is interpreted as a quasi-likelihood and yields consistent parameter estimates under standard regularity conditions. For numerical stability in the optimization routine, returns are scaled before estimation and re-scaled after forecasting to recover interpretable volatility levels.

B. EGARCH(1,1)

To incorporate the asymetric response of volatility to posite and negative return, we introduce the EGARCH(1,1) specification, that models the logarithm of the conditional variance:

$$\log(\sigma_t^2) = \omega + \beta \log(\sigma_{t-1}^2) + \alpha \left(\frac{|\varepsilon_{t-1}|}{\mathbb{E}|\varepsilon_{t-1}|} - 1 \right) + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}}$$

Here, β captures volatility persistence, α measures the symmetric impact of shock magnitude, γ governs the asymmetric response and ω contributes to the long-run mean of log conditional variance. Since the model is specified in terms of $\log(\sigma_t^2)$, no positivity constraints are required on the parameters. Implementation is done in the same way as GARCH(1,1).

C. Elastic Net

To complement the econometric models, we employ the Elastic Net regression, a linear machine-learning method that combines the strengths of the Lasso and Ridge penalties.

The estimator solves the penalized least-squares problem

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|^2 + \lambda \left(\alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2 \right) \right\}$$

where λ controls the overall strength of regularization and $\alpha \in [0,1]$ determines the relative contribution of the Lasso (ℓ_1) and Ridge (ℓ_2) penalties. The ℓ_1 component performs variable selection by shrinking irrelevant coefficients to zero, while the ℓ_2 term stabilizes the estimates in the presence of multicollinearity.

Elastic Net offers a flexible and interpretable benchmark that can exploit the rich feature set used in this study while remaining more robust than ordinary least squares or pure Lasso regression.

For the general setup, we refer to the general methodology in Section VI. Hyperparameters are selected dynamically using a time-series–aware cross-validation procedure. We perform a grid search over the ElasticNet mixing parameter $\alpha \in \{0.1, 0.5, 0.7, 0.9, 0.99\}$ and a sequence of 100 regularization strengths λ, implemented via `ElasticNetCV` with a `TimeSeriesSplit` scheme (5 folds). Hyperparameter tuning select (α,λ) by minimizing MSE in log-space. This procedure respects the temporal ordering of the data and avoids look-ahead bias in hyperparameter selection.

D. XGBoost

As a flexible non-linear benchmark, we employ Extreme Gradient Boosting (XGBoost) to forecast log realized volatility. XGBoost constructs an ensemble of shallow regression trees

$$\hat{y}_t = \sum_{m=1}^{M} f_m(x_t), \qquad f_m \in \mathcal{F},$$

where each tree f_m is fitted sequentially to the negative gradients of the loss function from the previous ensemble under a regularised squared-error objective. The overall loss function can be written as

$$\mathcal{L} = \sum_t (y_t - \hat{y}_t)^2 + \sum_{m=1}^{M} \Omega(f_m),$$

where $\Omega(\cdot)$ penalises tree complexity via ℓ_1 and ℓ_2 regularisation. This framework naturally captures non-linearities and high-order interactions between predictors while controlling overfitting.

We perform a grid search over the key hyperparameters: maximum tree depth `max_depth` $\in \{2,3\}$, learning rate `learning_rate` $\in \{0.05, 0.10\}$, row and column subsampling (`subsample`, `colsample_bytree`) $\in \{0.7, 0.9\}^2$, and regularisation strengths $\lambda \in \{5, 10\}$ and $\alpha \in \{0, 1\}$. This restricted set of hyperparameters is deliberately chosen to adapt the model bias–variance trade-off while keeping computational and methodological complexity under control.

For each combination, we estimate an XGBoost model with 300 trees and select the specification that minimises the validation MSE on log volatility. The final model for that window is then re-fitted on the full training sample. Unlike the Elastic Net, which relies on `ElasticNetCV` combined with a `TimeSeriesSplit` scheme, XGBoost does not provide a built-in time-series–aware cross-validation routine within the scikit-learn API. We therefore implement a rolling-window hyperparameter optimisation procedure manually.

E. LSTM

Volatility is known to exhibit persistent, long-range dependence and complex temporal patterns that may not be fully summarised by a small set of engineered lags. Recurrent neural networks, and LSTMs in particular, are specifically designed to learn such sequential dependencies through a gated memory mechanism. For this reason, we complement our analysis with an Long Short-Term Memory (LSTM) forecaster that processes sequences of past observations rather than single-day feature vectors.

We rely on the standard LSTM architecture, Each LSTM cell contains three gates (input, forget and output) that use a sigmoid activation to decide how much past information should be kept or discarded at each step. The internal memory of the cell is updated through a `tanh` nonlinearity, which allows the network to store a smooth representation of recent volatility dynamics. Once the sequence has been processed, the last hidden state is passed through a simple linear layer to produce the forecast.

Model estimation proceeds in two stages. **Phase 1** performs a global hyperparameter search over the LSTM architecture, including the sequence length `lookback`, treated here as a structural parameter. We evaluate combinations of `lookback` $\in \{5, 10, 20, 30\}$, `num_layers` $\in \{1, 2\}$, `units` $\in \{16, 32, 64\}$, together with dropout rates `dropout` $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and learning rates `learning_rate` $\in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. Each specification is trained on the first 2000 observations and evaluated chronologically. Training minimises the mean squared error (MSE) in standardised log-volatility space, using the Adam optimiser and early stopping. Early stopping is implemented by monitoring the validation MSE and terminating training when no improvement is observed for 10 consecutive epochs, retaining the parameter values that minimise the validation loss. The combination yielding the lowest validation MSE is retained as the global model structure. The global tuning phase selects an LSTM architecture characterised by a 20-day lookback window, a single recurrent layer with 64 hidden units, a dropout rate of 0.4, and a learning rate of 0.01. **Phase 2** implements a rolling walk-forward forecasting procedure as described in Section VI. For each new window, only the hyperparameter dropout rate is tuned, while the global architecture defined in phase 1 is held fixed. The best specification is again selected by minimising validation MSE.

This two-stage framework balances flexibility and methodological rigour: architectural complexity is determined globally, while local retraining allows the LSTM to adapt to evolving market regimes.

F. Hybrids Econometrics - Machine Learning Models

To investigate whether combining parametric econometric structure with machine-learning methods yields additional predictive gains, we construct hybrid forecasting models. Specifically, we enrich the input space of the Elastic Net, XGBoost, and LSTM models with additional features as described in Subsection VI-D.

All hybrid models are estimated using the same preprocessing pipeline, rolling-window design, hyperparameter tuning strategy, and evaluation metrics as their non-augmented counterparts.

VIII. RESULTS

A. Model Performance

TABLE I: Out-of-Sample Forecast Performance Across Models

Model	RMSE	MAE	R^2_{MEAN}	R^2_{RW}
GARCH(1,1)	0.00429	0.00323	0.43694	−0.36719
E-GARCH(1,1)	0.00391	0.00297	0.53306	−0.13562
Elastic Net	0.00336	0.00187	0.61027	0.14338
XGBoost	0.00299	0.00182	0.69260	0.32401
LSTM	0.00407	0.00222	0.42925	−0.25479

Table I compares the out-of-sample forecasting accuracy of all models across four evaluation criteria described in Subsection VI-E . The traditional GARCH(1,1) benchmark exhibits the weakest performance, with the highest RMSE and MAE and a negative R^2_{RW}, indicating that it performs worse than a naïve random-walk benchmark. Introducing asymmetry through the E-GARCH specification improves accuracy reducing both RMSE, by 8.9%, and MAE, by 8.1% and yielding a substantially higher R^2_{mean}. As shown in Figure 1, EGARCH forecasts tend to be slightly closer to the realised series with respect to GARCH forecasts. This illustrates the relevance of the leverage effect. However, its R^2_{RW} remains negative, signalling persistent difficulty in capturing short-term volatility reversals.

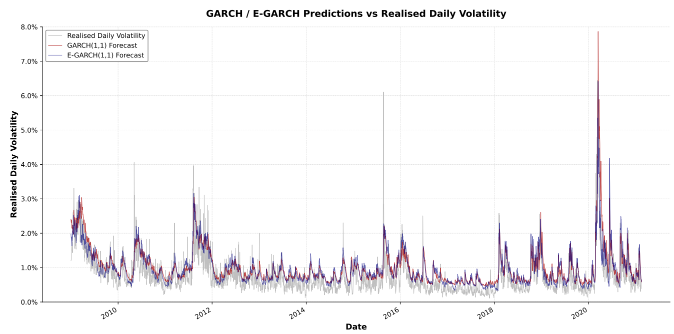

Fig. 1: GARCH / EGARCH volatility forecast VS Realised Volatility

Machine-learning models clearly outperform both econometric benchmarks. The Elastic Net achieves a notable improvement in predictive accuracy, with a 13.97% reduction in RMSE and 37.03% reduction in MAE relative to E-GARCH, but most importantly a positive R^2_{RW}, suggesting that even a sparse linear model benefits greatly from the richer feature set. XGBoost delivers the best overall performance amongs all models, attaining the lowest RMSE and MAE and a markedly positive $R^2_{\text{RW}} = 0.324$. XGBoost reduces by 11.2% RMSE and 2.73% MAE with respect to Elastic Net. These improvements highlight the presence of predictive nonlinearities and interaction effects that cannot be captured by linear regularised models. As seen in Figure 2, both XGBoost and the Elastic Net track realised volatility much more closely than the econometric benchmarks.

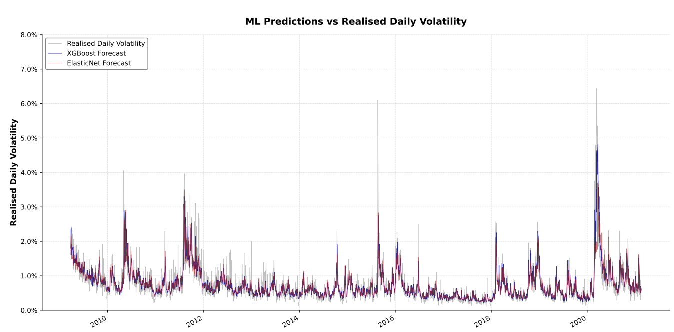

Fig. 2: Elastic Net / XGBoosst volatility forecast VS Realised Volatility

Interestingly, the gains from Elastic Net to XGBoost are asymmetric between MAE and RMSE. The much larger improvement in RMSE indicates that the main benefit of XGBoost lies in its ability to better capture large volatility spikes, precisely the episodes that dominate squared–error metrics. In other words, nonlinear effects and interactions matter primarily during extreme market conditions, where linear models tend to underreact. This behaviour is also visible in Figure 2, where XGBoost responds more sharply to pronounced volatility swings, both during typical turbulence episodes and in the Covid Crisis extreme spike observed in 2020–2021.

In contrast, the LSTM performs less favourably. RMSE and MAE improve only with respect to GARCH, MAE only with respect to EGARCH, and most importantly the model yields a negative R^2_{RW}, indicating that it fails to outperform the simple persistence benchmark. This suggests that the LSTM has difficulty adapting to rapid shifts in volatility regimes, a limitation that is likely exacerbated by the relatively small effective training sample available for deep recurrent architectures. Despite an extensive hyperparameter search designed to mitigate overfitting, we deliberately restrict the rolling window to 2000 observations to ensure comparability across models, which further constrains the LSTM's capacity to learn stable long–term temporal patterns.

As seen in Figure 3, the LSTM also displays weaker visual performance. While it captures broad low-frequency movements, its forecasts appear oversmoothed and react too slowly to abrupt volatility surges, consistent with its negative R^2_{RW}. This behaviour aligns with the model's structural difficulty in handling fast regime changes under a limited and constantly shifting training window.

A further explanation for the LSTM's weaker performance lies in the well-known behaviour of recurrent networks in noisy environments. LSTMs have a natural tendency to produce conservative, smoothed forecasts because their gating mechanism favours the retention of past information and penalises abrupt changes in the hidden state. In settings with limited data and high noise-to-signal ratios, such as daily volatility, the model often defaults to a "slow adaptation" regime, reacting only mildly to sudden shifts. This structural inertia contributes to the oversmoothing observed in Figure 3 and helps explain the negative R^2_{RW}, as the model struggles to adjust quickly enough during volatile episodes.

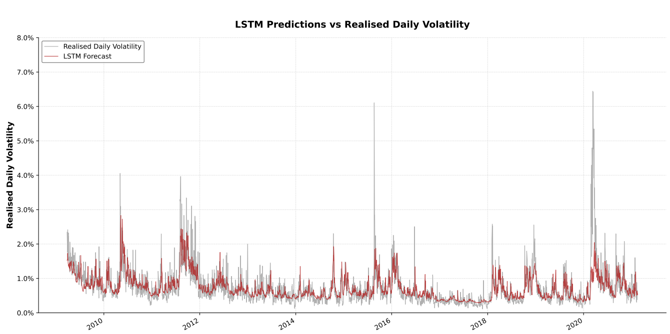

Fig. 3: LSTM Forecast VS Realised Volatility

B. Elastic Net Feature Relevance

A key advantage of the Elastic Net, compared with nonlinear machine-learning models such as XGBoost or LSTMs, is its high degree of interpretability. Because the estimator is linear, the fitted model takes the form

$$\widehat{y}_t = \beta_0 + \sum_{j=1}^{p} \beta_j x_{t,j},$$

so that each coefficient β_j measures the marginal effect of predictor x_j on the predicted log-volatility. A positive coefficient indicates that increases in the predictor are associated with higher future volatility, whereas a negative coefficient has the opposite effect. The magnitude of β_j reflects the strength of this relationship and provides a natural measure of feature importance.

By retaining only a subset of informative predictors while shrinking the remaining coefficients toward zero, the Elastic Net yields a transparent and economically interpretable feature-importance structure. Elastic Net coefficients admit a direct economic interpretation, allowing us to identify which variables contribute most to the forecast of future volatility and in what direction.

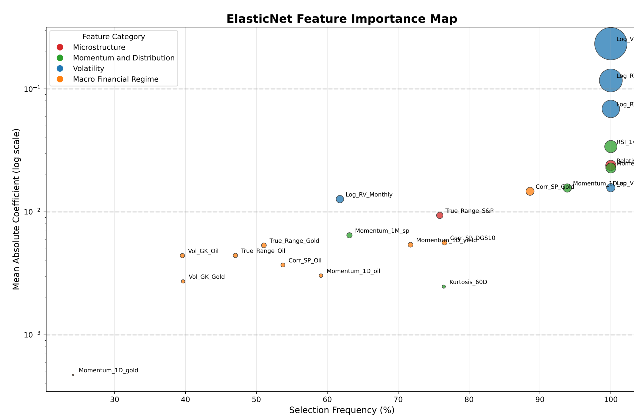

Fig. 4: Elastic Net Feature Importance Map by Predictor Category

Figure 4 summarises the information contained in the rolling Elastic Net coefficients. For each predictor, the horizontal axis reports its *selection frequency*, defined as the percentage of rolling windows in which the coefficient is non-zero. This enables us to distinguish variables that consistently matter from those that the model considers relevant only in specific market environments.The vertical axis shows the mean absolute value of the corresponding coefficient, which captures the average strength of the effect on predicted log-volatility when the variable is selected. The y axis is plotted in log scale to ensure readability due to important disparity in features contributions and we compute it with the aboslute value to have a meaningfull value even when the coefficient sign might be unstable. The bubble size reflects the magnitude of the mean coefficient so that we are able to highlight with variable matter the most in the predictions when its selected. Points are colour-coded by feature category (microstructure, momentum/distribution, volatility and macro–financial regime). Hence Variables located in the upper-right corner are both *frequently* selected and associated with *large* coefficients; these are the key drivers of the Elastic Net forecasts.

A first observation from the feature-importance map is the clear emergence of a distinct subset of predictors, namely the volatility-related variables. In particular, `Log_VIX`, `Log_RV_Weekly`, and `Log_RV_Daily` stand out both by their *100% selection frequency* and by their relatively large coefficient magnitudes. Among them, `Log_VIX` appears as the single most influential predictor, with a magnitude that exceeds that of all other features. This dominance highlights the strong autoregressive behaviour of realised volatility at the daily and weekly horizons, while the much smaller importance of the monthly lag reflects the mean-reverting nature of volatility. The prominent role of `Log_VIX` further underscores the informational advantage of implied volatility over purely historical measures: as a forward-looking indicator, VIX encapsulates market expectations and therefore provides a more powerful signal for forecasting short-term volatility than past realised volatility alone.

Beyond the dominant volatility predictors, a second group of variables also exhibits a 100% selection frequency but with substantially smaller coefficient magnitudes. This set includes `Relative_Volume`, `Momentum_1W_sp`, `Log_VIX_Change`, and `RSI_14`. Although their individual impact is considerably weaker than that of volatility lags or implied volatility, their systematic selection suggests that they provide complementary information that consistently improves forecast accuracy. The Elastic Net therefore treats these variables as stable but low-intensity signals, contributing marginal adjustments rather than driving the predictions. Economically, these predictors relate to trading activity, short-term market momentum, and pressure imbalances, all of which influence market microstructure and short-horizon volatility dynamics.

Within the macro-financial category, a clear hierarchy also emerges. `Corr_SP_Gold` is by far the most frequently selected predictor, appearing in almost all estimation windows with a non-negligible coefficient. By contrast, the other macro or cross-asset variables display much lower selection frequencies and extremely small magnitudes.

Focusing on the coefficient sign, Table III reveals that the 3 biggest mean signed coefficient variables are `Log_VIX`, `Log_RV_Weekly`, and `Log_RV_Daily`. These variables also display perfect sign stability (Pos Share = 1), highlighting their structural and consistently positive contribution to future volatility.

In contrast, the largest negative effects correspond to equity momentum and technical indicators such as `Momentum_1W_sp`, `RSI_14`, and `Momentum_1M_sp`. Their sign is also extremely stable (Neg Share close to 1), indicating that positive price trends tend to be associated with lower subsequent volatility, observation consistent with the classical leverage effect.

Finally, a subset of variables such as cross–asset correlations and commodity volatility measures exhibit mixed sign behaviour. Their Pos/Neg shares are balanced, suggesting that these predictors are regime–dependent: they may have a positive effect in some market phases and a negative effect in others. Most of them exhibit absolute mean coefficients disproportionately small with respect to most important predictor.

However, within this group, an interesting case is `Corr_SP_Gold`. Although it is the only macro–financial indicator that consistently emerges as informative, its coefficient is far from stable: it is positive in roughly 30% of rolling windows and negative in the remaining 70%. This asymmetric sign pattern indicates a strongly regime–dependent effect.

C. XGBoost Feature Relevance

Whereas the Elastic Net identifies predictors that exert a stable and linear influence on future volatility, XGBoost allows for a much richer functional form, capturing non-linearities, threshold effects, and interactions between variables. Studying feature relevance in this setting therefore serves a complementary purpose: it enables us to assess whether the predictors highlighted by the Elastic Net remain important once linearity constraints are relaxed and to determine whether additional variables become influential through mechanisms that a linear model cannot capture.

While split-based feature importance measures provide a first indication of which variables are frequently used by the XGBoost model, they do not convey the direction, magnitude, or state-dependence of a predictor's effect on volatility forecasts. To obtain a more granular and economically interpretable view of the model's decision process, we rely on SHapley Additive exPlanations (SHAP).

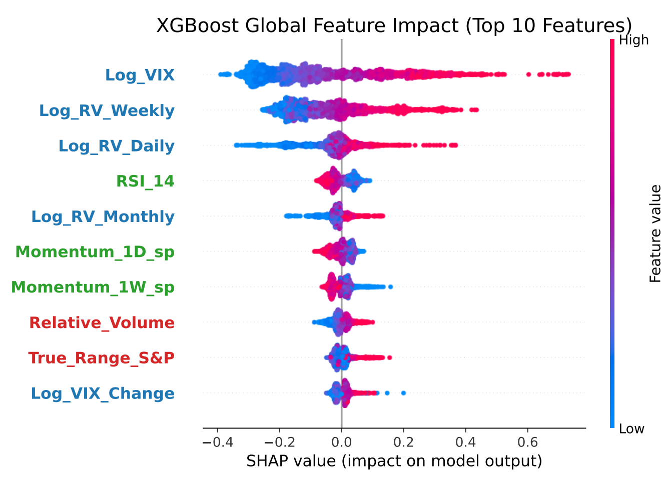

Fig. 5: Predictors ranked by decreasing mean absolute contribution. Blue : Volatility , Green : Momentum and Distribution, Red: Microstructure, Orange: Macro Financial Regime

SHAP values decompose each individual prediction into additive contributions from each feature, allowing us to quantify how and under which market conditions a given predictor increases or decreases the predicted level of future volatility. Importantly, SHAP values are computed on the out-of-sample test observations across all rolling windows, ensuring that the analysis reflects the drivers of actual forecasts rather than in-sample fitting behaviour.

Compared with the Elastic Net, as highilghted by Figure 5, XGBoost largely preserves the hierarchy of the most influential predictors, while altering their relative importance and the way their effects enter the forecast. In both models, volatility-related variables, most notably `Log_VIX`, `Log_RV_Weekly` and `Log_RV_Daily`, remain the dominant drivers of short-term volatility, confirming the central role of volatility persistence. The structure of the least influential predictors also remains very similar across models; in particular, macro–financial indicators consistently exhibit low explanatory power, as illustrated in Figure 8.

Beyond this broad similarity in predictor ranking, Figure 5 reveals substantial nonlinearities in how individual variables affect the volatility forecast under XGBoost. In contrast to the linear and symmetric contributions implied by the Elastic Net coefficients, the SHAP distributions exhibit pronounced asymmetries and state-dependent responses. For the main volatility predictors, the dispersion of SHAP values increases sharply at higher feature levels: large feature values are associated with disproportionately large positive contributions to predicted volatility, whereas low values generate comparatively modest negative effects. This pattern is especially pronounced with of `Log_VIX` and `Log_RV_Weekly` and indicates a convex response of the forecast to volatility measures. Nonlinear effects are also evident for the other class of predictors, even if their impact is way smaller.

Overall, the SHAP analysis highlights that XGBoost improves forecasting performance not by replacing the dominant predictors identified by the Elastic Net, but by modelling nonlinear transformations of the same core information set. While the present discussion focuses on marginal nonlinear effects within individual variables, the pronounced asymmetries observed in Figure 5 suggest that the impact of key predictors may also depend on the state of other variables. This interpretation is consistent with the empirical finding that XGBoost yields substantial gains in RMSE, indicating that its advantages arise primarily from better modelling of extreme forecast errors. This suggest that the nonlinear mechanisms captured by XGBoost become particularly relevant during periods of elevated market stress, thereby motivating a closer examination of interaction effects across predictors.

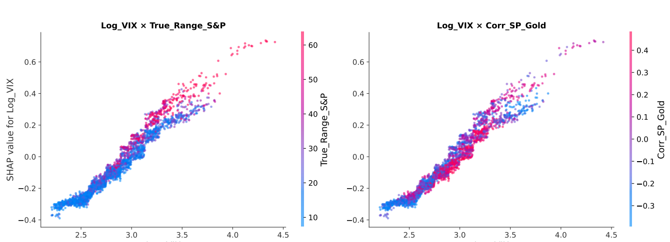

Fig. 6: SHAP interaction effects for `Log_VIX` with `True_Range_SP` and `Corr_SP_Gold`.

Figure 6 highlights pronounced interaction effects involving `Log_VIX`. On the left-hand panel, for a given level of `Log_VIX`, the contribution assigned by the model varies systematically with intraday price dispersion, with higher true range values associated with larger SHAP values. This pattern indicates that the effect of implied volatility on realized volatility forecasts is conditioned by contemporaneous market microstructure stress and becomes stronger during periods of elevated intraday price variation. In addition, the right-hand panel indicates that higher equity/gold correlation is associated with a steeper conditional SHAP profile for `Log_VIX`, implying a stronger contribution of implied volatility to realized volatility predictions in high-comovement regimes.

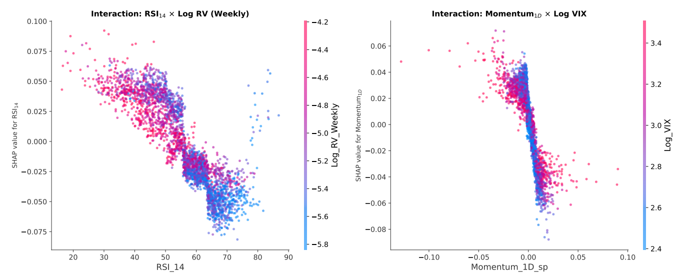

Fig. 7: SHAP interaction effects for `RSI_14` with `Log_RV_Weekly` and `Momentum_1D_sp` with `Log_VIX`.

Figure 7 illustrates SHAP interaction effects involving short-term technical indicators. In the left-hand panel, for a given RSI level, higher weekly realized volatility is associated with a rotation in SHAP values functional form, indicating that the informational content of RSI is modified during volatile periods. In the right-hand panel, the conditional SHAP profile of `Momentum_1D_sp` varies with the level of implied volatility. During low-volatility periods, the relationship between momentum and its SHAP contribution is steeper, indicating that small changes in short-term price pressure have a relatively large impact on volatility forecasts. In contrast, when implied volatility is high, the SHAP profile flattens, suggesting that the informational content of momentum is attenuated as aggregate volatility risk dominates directional signals.

Taken together, the interaction analyses highlight the presence of pronounced nonlinear and state-dependent relationships in volatility dynamics. The contribution of key predictors such as implied volatility, technical indicators, and cross-asset variables depends critically on market conditions, including intraday price dispersion and volatility regimes. These patterns are inherently difficult to capture with linear specifications, even when augmented with rich feature sets. In contrast, the tree-based structure of XGBoost allows the model to adapt flexibly to such conditional relationships, selectively emphasizing different predictors across regimes. This ability to accommodate nonlinearities and interaction effects provides a natural explanation for the substantial reduction in out-of-sample RMSE observed for XGBoost relative to Elastic Net, as the model more accurately captures extreme volatility realizations and regime-dependent forecast errors.

D. Hybrid Models

TABLE II: Out-of-sample forecast performance for EGARCH-augmented models

Model	RMSE	MAE	R^2_{MEAN}	R^2_{RW}
Elastic Net	0.00338	0.00186	0.60579	0.133212
XGBoost	0.00299	0.00182	0.69260	0.32401
LSTM	0.00406	0.00224	0.431486	−0.24871

Table II reports the out-of-sample forecasting performance of the different model classes when the feature set is augmented with EGARCH-based volatility forecasts and parameters. A comparison with Table I, which reports results for the baseline feature specification, shows that the augmented feature set does not lead to material improvements in forecast accuracy across models.

The lack of improvement from augmenting the feature set with EGARCH-based volatility forecasts and parameters indicates substantial informational overlap with the baseline predictors. The existing feature set already includes multiple measures of volatility persistence, intraday price variation, and forward-looking risk expectations through implied volatility. In addition, the nonlinear structure of models such as XGBoost enables them to internalize asymmetric and regime-dependent volatility dynamics directly from the data. As a result, the EGARCH model does not provide sufficiently orthogonal information to yield further gains in out-of-sample forecast accuracy.

IX. Conclusion

This study investigated the performance, drivers, and interpretability of econometric, machine-learning, and hybrid models for one-day-ahead forecasting of S&P 500 realized volatility in a strictly out-of-sample, walk-forward setting. By evaluating all models under a common rolling-window protocol, using the same engineered feature set and identical evaluation metrics, we ensured that observed differences in forecasting accuracy reflect intrinsic modeling capabilities rather than differences in data handling or information availability.

Addressing the first research question, we find clear and robust evidence that machine-learning models outperform classical econometric benchmarks. While the GARCH(1,1) specifications capture volatility persistence and EGARCH(1,1) the leverage effects, both fail to consistently beat the random-walk benchmark. In contrast, the Elastic Net achieves substantial gains by exploiting a rich feature set under linear regularization, yielding positive out-of-sample performance relative to persistence. Among all models considered, XGBoost delivers the most accurate forecasts, achieving the lowest RMSE, MAE and the highest out-of-sample R^2_{RW}. The LSTM, by contrast, underperforms relative to simpler machine-learning approaches, suggesting that deep recurrent architectures are ill-suited to daily volatility forecasting under short rolling windows.

The second research question concerns the predictors and mechanisms driving forecast performance. Feature relevance analyses reveal a striking degree of consistency across model classes: volatility-related variables, particularly implied volatility (VIX) and short-horizon realized volatility measures, dominate the information set. Elastic Net results highlight the stability and economic interpretability of these predictors, while SHAP analyses show that XGBoost largely relies on the same core variables but exploits them in fundamentally different ways. In particular, XGBoost captures pronounced nonlinearities, asymmetries, and interaction effects that become especially relevant during periods of elevated market stress. Interaction analyses demonstrate that the contribution of key predictors, such as implied volatility and technical indicators, is strongly state-dependent, varying with intraday price dispersion and cross-asset comovement regimes. These nonlinear mechanisms provide a natural explanation for XGBoost's superior RMSE performance, which is driven primarily by improved forecasting of extreme volatility realizations.

Finally, addressing the third research question, we find no evidence that augmenting the feature space with EGARCH-based volatility forecasts and fitted parameters improves predictive accuracy. Across all model classes, hybrid specifications perform nearly identically to their non-augmented counterparts. This result suggests substantial informational redundancy: the baseline feature set already embeds multiple proxies for volatility persistence, asymmetry, and forward-looking risk expectations.

Taken together, the results indicate that gains in volatility forecasting accuracy stem less from access to new information than from the ability to model nonlinear, regime-dependent mappings between a well-designed feature set and future volatility. In this context, gradient-boosted tree models offer a compelling trade-off between predictive accuracy and interpretability, outperforming both classical econometric models and deep recurrent architectures under realistic real-time forecasting constraints. Future research may explore whether deep learning models regain an advantage at higher sampling frequencies or with substantially longer training histories. Additional extensions include incorporating alternative high-frequency market signals, assessing model performance across a broader cross-section of assets, examining loss-function dependence, and evaluating the economic value of improved volatility forecasts in risk-management applications.

REFERENCES

[1] R. F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.

[2] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.

[3] D. B. Nelson, "Conditional heteroskedasticity in asset returns: A new approach," *Econometrica*, vol. 59, no. 2, pp. 347–370, 1991.

[4] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[5] C. Zhang, Y. Zhang, M. Cucuringu, and Z. Qian, "Volatility forecasting with machine learning and intraday commonality," *Journal of Financial Econometrics*, vol. 22, no. 2, pp. 492–530, 2024.

[6] V. Pyrlik, P. Elizarov, and A. Leonova, "Forecasting realized volatility using machine learning and mixed-frequency data (the case of the russian stock market)," CERGE-EI Working Paper Series, Tech. Rep. wp713, 2021.

[7] A. Teller, U. Pigorsch, and C. Pigorsch, "Short- to long-term realized volatility forecasting using extreme gradient boosting," SSRN, Tech. Rep., 2022, working paper.

[8] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models," *Expert Systems with Applications*, vol. 103, pp. 25–37, 2018.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.

[10] T. G. Andersen and T. Bollerslev, "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts," *International Economic Review*, vol. 39, no. 4, pp. 885–905, Nov. 1998.

[11] M. B. Garman and M. J. Klass, "On the estimation of security price volatilities from historical data," *The Journal of Business*, vol. 53, no. 1, pp. 67–78, 1980.

[12] J. Y. Campbell and S. B. Thompson, "Predicting excess stock returns out of sample: Can anything beat the historical average?" *The Review of Financial Studies*, vol. 21, no. 4, pp. 1509–1531, 2008.

APPENDIX

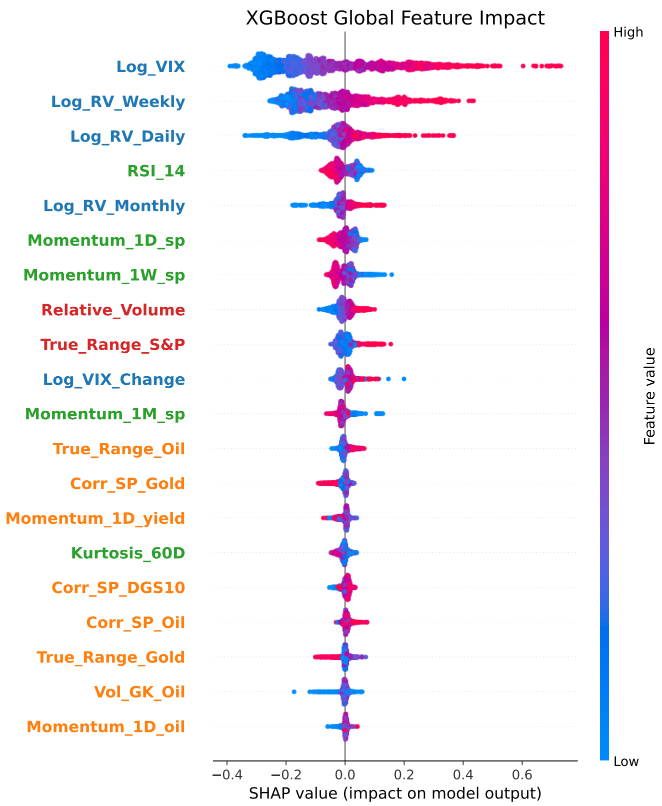

Fig. 8: Predictors ranked by decreasing mean absolute contribution. Blue : Volatility , Green : Momentum and Distribution, Red: Microstructure, Orange: Macro Financial Regime

TABLE III: Elastic Net Coefficient Diagnostics

Feature	Sel. Freq (%)	Mean Abs Coeff	Mean Signed Coeff	Pos Share	Neg Share
Relative_Volume	100.00	0.0239	0.0239	1.0000	0.0000
Momentum_1W_sp	100.00	0.0228	-0.0228	0.0000	1.0000
Log_RV_Daily	100.00	0.0688	0.0688	1.0000	0.0000
Log_RV_Weekly	100.00	0.1172	0.1172	1.0000	0.0000
Log_VIX_Change	100.00	0.0157	0.0157	1.0000	0.0000
Log_VIX	100.00	0.2335	0.2335	1.0000	0.0000
RSI_14	100.00	0.0340	-0.0340	0.0000	1.0000
Momentum_1D_sp	93.87	0.0167	-0.0166	0.0071	0.9929
Corr_SP_Gold	88.60	0.0166	-0.0108	0.2963	0.7037
Corr_SP_DGS10	76.53	0.0074	0.0063	0.8336	0.1664
Kurtosis_60D	76.43	0.0032	-0.0002	0.5614	0.4386
True_Range_S&P	75.86	0.0124	0.0124	1.0000	0.0000
Momentum_1D_yield	71.74	0.0075	-0.0070	0.1776	0.8224
Momentum_1M_sp	63.12	0.0102	-0.0102	0.0000	1.0000
Log_RV_Monthly	61.78	0.0206	0.0206	1.0000	0.0000
Momentum_1D_oil	59.10	0.0051	0.0044	0.8849	0.1151
Corr_SP_Oil	53.74	0.0069	0.0061	0.7754	0.2246
True_Range_Gold	51.06	0.0105	-0.0098	0.1051	0.8949
True_Range_Oil	47.03	0.0094	0.0091	0.9287	0.0713
Vol_GK_Gold	39.66	0.0069	0.0018	0.5435	0.4565
Vol_GK_Oil	39.56	0.0112	0.0064	0.8305	0.1695
Momentum_1D_gold	24.14	0.0020	-0.0019	0.1111	0.8889

HELPER TOOLS

The following helper tools were used :

- ChatGPT (OpenAI): assistance with Python code debugging and implementation, language polishing, including grammar corrections and translation from French to English.
- Gemini (Google): assistance with Python code debugging and implementation, language polishing, including grammar corrections and translation from French to English.