Forecasting S&P 500 Volatility with an ARIMA-EGARCH and LightGBM Pipeline

Integrating a conditional log-sigma signal into a multi-asset machine-learning model

English version prepared from the original French report for international / YC review

November 21, 2025

Contents

Forecasting S&P 500 Volatility with an ARIMA-EGARCH and LightGBM Pipeline

Abstract

This report studies whether an econometric volatility signal can improve a tabular machine-learning model for next-day volatility forecasting on S&P 500 equities.

The core idea is simple: a LightGBM model already captures a large set of lagged returns, technical indicators and volatility features, but it may still benefit from a conditional volatility estimate produced by an ARIMA-EGARCH model. The study therefore introduces a conditional `log_sigma_garch` feature and tests whether it provides measurable predictive value in several controlled feature sets.

The pipeline is designed to reduce the main methodological risks in financial forecasting: temporal leakage, non-causal feature construction, unrealistic validation, and over-interpreting isolated performance gains. The evaluation uses chronological splits, walk-forward logic, refit discipline, ablation datasets, Diebold-Mariano tests, bootstrap confidence intervals for R^2 differences, SHAP analysis and permutation importance.

The main result is positive but nuanced. In the complete feature set, adding the conditional log-sigma signal improves RMSE from `0.0113` to `0.0109` and R^2 from `0.749` to `0.765`. The difference is statistically significant under Diebold-Mariano tests with $p <$ `0.01`. In a reduced autoregressive setup, the same signal improves R^2 from `0.497` to `0.538`. However, the insight alone performs poorly and does not beat a persistence baseline. The evidence therefore supports the feature as a complementary risk signal, not as a standalone forecasting model.

This report is a forecasting and methodology study. It is not an investment strategy or a financial recommendation.

1. Research Question

Financial volatility is both persistent and unstable. It depends on recent realized movements, market regimes, leverage effects, macro shocks, and changing risk appetite. Classical econometric models such as GARCH and EGARCH are built to model conditional heteroskedasticity, while gradient boosting models such as LightGBM can learn nonlinear interactions from large tabular feature sets.

The question studied here is:

> Does a conditional volatility signal estimated by ARIMA-EGARCH improve next-day volatility forecasts when added to a LightGBM model?

The hypothesis is not that EGARCH replaces machine learning. The hypothesis is that EGARCH captures a structured volatility component that can be useful as an additional feature inside a stronger model.

Three hypotheses guide the study:

- **H1**: Adding `log_sigma_garch` to the complete feature set improves predictive performance.
- **H2**: The EGARCH insight is powerful enough to work well alone.
- **H3**: The contribution of the insight depends on the surrounding feature set.

The empirical results validate H1, reject H2, and do not reject H3.

2. Data and Prediction Target

The study uses daily S&P 500 equity data over the 2013-2024 period. The target is next-day volatility at the security level. The final test set contains `213,021` observations.

The modeling problem is formulated as a supervised regression task. Each row corresponds to a stock-date observation with features available at the decision time and a next-day volatility target. The feature engineering process explicitly avoids future information.

The project focuses on the methodological value of combining econometric volatility structure with machine-learning features. It does not attempt to produce a deployable trading strategy, transaction-cost-aware portfolio, or live risk engine.

3. Pipeline Overview

The pipeline contains five major stages:

1. **Data preparation**: load market data, align dates, construct security-level panels, and enforce temporal consistency.
2. **Econometric modeling**: fit ARIMA-EGARCH specifications and extract conditional volatility estimates.
3. **Feature construction**: create lagged volatility, returns, technical indicators and the conditional `log_sigma_garch` insight.
4. **Machine-learning evaluation**: train LightGBM models on controlled feature sets and evaluate out of sample.
5. **Statistical and interpretability analysis**: compare models using error metrics, Diebold-Mariano tests, bootstrap R^2 intervals, SHAP and permutation importance.

The important design constraint is causality. Every feature must be available at the date where the prediction would have been made. This is especially important for financial time series where small leakage can create large artificial gains.

4. ARIMA-EGARCH Signal

The econometric component estimates conditional volatility. EGARCH is useful because it can model asymmetric volatility responses, including leverage effects where negative shocks can have a different impact from positive shocks.

The resulting signal is transformed into a log-scale feature:

`log_sigma_garch`

This feature is then added to different LightGBM feature sets. The signal is not evaluated only as a standalone model; it is evaluated as an input that may help a nonlinear model better represent risk regimes.

The original French report documents the EGARCH diagnostics and the estimated parameters. The fitted specification uses a skewed Student distribution. Reported parameters include:

Parameter	Value
omega	−0.31
alpha	0.23
gamma	−0.18
beta	0.97
nu	7.30
lambda	−0.30

The high beta value is consistent with volatility persistence. The gamma term captures asymmetric effects. The skewed distribution is more appropriate than a normal assumption for financial returns.

5. Feature Sets and Ablation Design

The experiment compares several datasets in order to isolate the contribution of the conditional volatility insight.

The most important comparisons are:

- Complete dataset with `log_sigma_garch` versus complete dataset without the insight.
- Technical indicators with and without insights.
- Autoregressive volatility lags with and without insights.
- Insights alone versus a persistence baseline.

This ablation structure matters because a feature can appear useful in one context and redundant in another. The study therefore avoids relying on a single global metric.

6. Main Results

The complete dataset with the EGARCH insight achieves the best reported performance.

Dataset	RMSE	MAE	MSE	R^2	Features
Complete dataset	0.0109	0.0062	0.000119	0.765	55
Without insight	0.0113	0.0065	0.000127	0.749	51
Indicators only	0.0117	0.0070	0.000136	0.730	41

Dataset	RMSE	MAE	MSE	R^2	Features
Indicators + insights	0.0117	0.0068	0.000136	0.731	45
Insights alone	0.0221	0.0142	0.000487	0.037	5
Log-volatility lags only	0.0160	0.0095	0.000255	0.497	4
Log-volatility + insights	0.0153	0.0091	0.000234	0.538	8
Persistence baseline	0.0185	0.0107	0.000344	0.320	n/a

The headline result is the improvement from 0.0113 to 0.0109 RMSE and from 0.749 to 0.765 R^2 in the complete dataset. This is a modest absolute improvement, but it is meaningful given the size of the test set and the difficulty of the task.

The reduced autoregressive comparison is also important: adding insights to volatility lags improves R^2 from 0.497 to 0.538. This shows that the signal carries information in a constrained setting.

The insights-alone result is weak. With R^2 0.037, it does not beat the persistence baseline. This rejects the idea that the EGARCH signal alone is sufficient.

7. Statistical Tests

The study uses Diebold-Mariano tests to compare predictive losses between model variants.

Key comparisons:

Comparison	DM MSE	DM MAE	p-value	ΔR^2
Complete vs without insight	−29.8	−63.5	< 0.01	+0.0154
Indicators + insights vs indicators only	n.s. for MSE/R^2	significant for MAE	mixed	+0.001 approx.

Comparison	DM MSE	DM MAE	p-value	ΔR^2
Log-vol + insights vs log-vol only	−20.8	−41.5	< 0.01	+0.0413

For the complete comparison, the bootstrap confidence interval for the R^2 gain is approximately [0.0144, 0.0165], supporting a stable positive effect.

For the autoregressive comparison, the R^2 gain is approximately +0.0413, with confidence interval [0.0372, 0.0455].

The indicators-only comparison is more nuanced. The MAE improvement is significant, but the MSE and R^2 gains are not decisive. This supports the interpretation that the value of the insight depends on the surrounding feature set.

8. Interpretability

SHAP and permutation importance provide complementary views of feature relevance.

The log_sigma_garch feature appears inside the top features but not at the very top in the complete dataset:

- Rank 13/55 in the complete dataset.
- Rank 9/45 in the technical indicators + insights dataset.
- Rank 3/8 in the autoregressive + insights dataset.

This pattern is consistent with the ablation results. The signal is useful, especially when the feature set is constrained, but it is not the sole driver of predictions.

Permutation importance shows relatively modest degradation when log_sigma_garch is permuted. This does not necessarily invalidate the feature because boosted trees can partially compensate through correlated volatility features. The interpretation is therefore architectural compensation rather than pure redundancy.

9. Interpretation

The conditional log-sigma feature behaves like a forward-looking risk proxy. It summarizes conditional volatility dynamics in a way that can complement raw lags and technical indicators.

The results support a pragmatic conclusion:

- EGARCH alone is not enough.
- LightGBM already captures much of the structure.
- The econometric insight still adds measurable information in the complete setting.

- The value is strongest when the model has fewer competing volatility features.

This is a useful result because it frames econometrics and machine learning as complementary rather than mutually exclusive. A structured econometric feature can act as a compressed risk signal inside a broader nonlinear model.

10. Limitations

The study has several important limitations:

- The prediction horizon is limited to next-day volatility.
- The market universe is S&P 500 equities over a fixed 2013-2024 sample.
- The ARIMA component is intentionally minimal in the current setup.
- Only one main random seed is reported.
- The project is a forecasting study, not a trading strategy.
- It does not include transaction costs, portfolio construction or live execution constraints.
- Data quality and survivorship effects must be considered when extending the study.

These limitations are not minor details. They define the boundary of what can be claimed from the work.

11. Extensions

Natural extensions include:

- Testing GJR-GARCH, FIGARCH or HAR-RV variants.
- Comparing LightGBM with XGBoost, CatBoost, LSTM and GRU models.
- Extending horizons to 5, 10 and 21 trading days.
- Evaluating other markets and asset classes.
- Adding operational VaR backtests.
- Testing robustness across regimes and crisis periods.

The broader research direction is to evaluate whether better volatility forecasts improve downstream economic decisions, not only predictive error metrics.

12. Conclusion

The study validates the main hypothesis: adding a conditional `log_sigma_garch` signal from ARIMA-EGARCH improves LightGBM volatility forecasts in the complete feature set.

The gain is statistically significant and economically interpretable as an additional risk-regime signal. However, the insight is not sufficient by itself. Its value comes from integration into a broader feature set and model architecture.

The most defensible conclusion is therefore:

> Econometric volatility structure can improve tabular machine-learning forecasts when used as a carefully constructed, leakage-free feature.

For an international or YC reader, the most relevant aspect of this project is not the specific S&P 500 result alone. It is the engineering and scientific discipline behind the work: causal feature construction, controlled ablations, statistical testing, interpretability, and honest reporting of limitations.

Detailed Translation by Original Structure

The following section expands the English version using the structure of the original French report. It is designed for readers who want a fuller technical view than the website case study, while keeping the exact numerical results and methodological claims aligned with the source document.

1. Introduction

1.1 Context and Motivation

Conditional volatility forecasting is central to risk management. It affects regulatory capital, Value-at-Risk limits, exposure sizing and asset-allocation decisions. Econometric volatility models remain a historical reference because they are parsimonious and interpretable. However, constant-parameter models can struggle with regime breaks across twelve years of daily S&P 500 data.

Machine-learning models are more flexible. They handle high-dimensional feature spaces and nonlinear interactions naturally. But they remain dependent on the quality of the features they receive. This creates a useful complementarity: an ARIMA-EGARCH block can provide a structured signal about conditional volatility, while LightGBM can combine this signal with a broader set of market features.

The project therefore tests the value of joining econometrics and machine learning in a strictly out-of-sample forecasting setup.

1.2 Research Question

The central question is whether the conditional log-sigma forecast produced by an ARIMA-EGARCH pipeline trained on S&P 500 index log-returns improves title-level LightGBM forecasts of next-day volatility.

This is evaluated by comparing multiple datasets. Some include the conditional volatility insight; others deliberately remove it. This makes the contribution of the insight observable instead of relying on a single global model score.

1.3 Notation and Conventions

The study uses daily returns, log-volatility transformations and next-day targets. The main econometric insight is denoted `log_sigma_garch`. It represents the logarithm of the conditional volatility estimate produced by the EGARCH model.

All feature construction follows the same principle: a feature is allowed only if it would have been available at the forecast date. Future observations are excluded from the feature set. The evaluation is chronological and does not rely on random shuffling.

1.4 Methodological Protocol

The protocol is built around reproducibility and leakage prevention. The data are split chronologically. Refits are performed in a controlled walk-forward logic. Model comparisons are made through ablation datasets rather than informal inspection.

The evaluation combines standard regression metrics, statistical tests and interpretability tools. RMSE, MAE, MSE and R^2 summarize predictive performance. Diebold-Mariano tests compare forecast losses. Bootstrap confidence intervals quantify the stability of R^2 differences. SHAP and permutation importance help interpret the role of the conditional volatility signal.

2. Data and Variable Construction

2.1 Universe and Source

The empirical universe is based on S&P 500 equities over the 2013-2024 period. The study is daily. The final test set contains `213,021` stock-date observations.

The analysis is intentionally focused on volatility forecasting, not on portfolio construction. The dataset is therefore used to test whether a conditional volatility signal improves next-day predictions, not whether it directly generates a tradable strategy.

2.2 Cleaning and Validation

Financial data require strong validation. Missing values, inconsistent dates, splits, outliers and survivorship effects can all distort results. The original project includes cleaning and validation steps before model estimation.

The most important validation rule is temporal consistency. The model must never use information that would not have been available at the forecast time. This constraint applies to lagged variables, technical indicators, ARIMA-EGARCH estimates and all target transformations.

2.3 Market Index

The ARIMA-EGARCH component is estimated on market-index log-returns. This creates a market-level conditional volatility signal. The LightGBM model then uses this signal at the stock level together with title-specific features.

The methodological choice is deliberate: the market index provides a common risk-regime proxy, while the machine-learning layer handles the richer cross-sectional structure.

3. ARIMA Model

3.1 Theoretical Approach

ARIMA models describe the conditional mean of a time series through autoregressive and moving-average components. In this project, the ARIMA component is intentionally kept minimal. The purpose is not to claim that ARIMA alone explains the full return process. The purpose is to provide a disciplined mean specification before modeling conditional heteroskedasticity through EGARCH.

3.2 Assumption Checks

The original report checks the usual time-series requirements: stationarity, autocorrelation and partial autocorrelation. These diagnostics help determine whether the return series can be modeled in a stable way.

Stationarity matters because non-stationary behavior can invalidate model interpretation. ACF and PACF diagnostics help evaluate persistence and the need for autoregressive or moving-average terms.

3.3 Forecasting Role

The ARIMA part is not the final forecasting model. It supports the conditional volatility pipeline. Its output is used inside the broader ARIMA-EGARCH procedure, which then produces the volatility signal integrated into LightGBM.

4. EGARCH Model

4.1 Theoretical Motivation

EGARCH models conditional volatility in logarithmic form. This is useful because volatility is positive by construction and because the log specification can capture asymmetric responses to shocks. In financial markets, negative returns often have a different impact on future volatility than positive returns of similar magnitude.

4.2 Model Hypotheses

The project tests whether a conditional volatility estimate is informative for the downstream machine-learning task. The hypotheses distinguish between three ideas: the insight improves the complete model, the insight can work alone, and the insight's value depends on the feature context.

The results support the first idea, reject the second, and suggest that feature context matters.

4.3 Optimization

The EGARCH model is optimized using a log-QLIKE objective and Bayesian optimization. The goal is to obtain a robust conditional volatility signal, not simply to fit in-sample noise. Walk-forward logic and regular refits reduce overfitting to a single static period.

4.4 Estimation

The reported EGARCH specification uses a skewed Student distribution. This choice is more realistic for financial returns than a Gaussian assumption because returns are heavy-tailed and often skewed.

The reported parameters show strong persistence through beta, an asymmetric term through gamma, and heavy tails through the degrees-of-freedom parameter. These are economically plausible for an equity-index volatility process.

4.5 Diagnostics

The original report documents several diagnostic checks: residual distribution, autocorrelation functions, stability checks and distributional validation. These diagnostics are necessary because a poorly calibrated econometric block could inject noise into the LightGBM model rather than useful structure.

The diagnostic conclusion is that the EGARCH signal is usable as a structured volatility proxy. It is not perfect, and the report does not treat it as a standalone solution.

4.6 EGARCH Evaluation

The EGARCH component is evaluated through forecasting metrics and VaR-oriented diagnostics. The operational interpretation is risk calibration: the conditional volatility signal can help dimension buffers, estimate stress periods and improve downstream forecasts.

This section is important because it links statistical modeling to a real risk-management use case. Even when the final model is LightGBM, the econometric signal keeps a direct economic interpretation.

5. Machine-Learning Model: LightGBM

5.1 Gradient Boosting Foundations

LightGBM is a gradient-boosting framework based on decision trees. It can model nonlinear interactions, handle heterogeneous feature scales and perform well on structured tabular data. This makes it a strong candidate for stock-level volatility forecasting.

The model is not used as a black-box shortcut. It is embedded in a controlled experimental protocol with ablations and statistical tests.

5.2 Problem Framing

The supervised task predicts next-day stock-level volatility. Each observation contains only information available at the forecast date. The conditional `log_sigma_garch` feature is treated as one feature among others.

The target and features are constructed to respect temporal availability. This prevents a common failure mode in financial ML: using future realized information accidentally.

5.3 Seven Dataset Variants

The report compares seven main dataset variants:

1. Complete dataset.
2. Complete dataset without the EGARCH insight.
3. Technical indicators only.
4. Technical indicators plus insights.
5. Insights alone.
6. Log-volatility lags only.
7. Log-volatility lags plus insights.

This design isolates whether the insight adds information in a rich setting, a technical-indicator setting, a minimal autoregressive setting and a standalone setting.

5.4 Training and Evaluation

The LightGBM models are trained on temporal splits and evaluated out of sample. The objective is not just to minimize a metric but to understand where the signal adds value.

The main test set is large enough for small metric differences to be meaningful when statistically supported. This is why the study uses Diebold-Mariano tests rather than relying only on visual score comparisons.

6. Results

6.1 Complete Dataset

The complete dataset with the conditional volatility signal achieves RMSE `0.0109`, MAE `0.0062`, MSE `0.000119` and R^2 `0.765`.

The same architecture without the insight obtains RMSE `0.0113`, MAE `0.0065`, MSE `0.000127` and R^2 `0.749`.

The gain is modest in absolute terms but robust. The Diebold-Mariano comparison reports p < 0.01, and the bootstrap R^2 difference is positive with a confidence interval around [0.0144, 0.0165].

6.2 Technical Indicators

The technical-indicator comparison is less decisive. The version with insights improves MAE, but the MSE and R^2 differences are not as strong. This suggests that part of the EGARCH information may overlap with technical volatility indicators.

This is not a failure of the insight. It clarifies where it is most useful: in settings where the feature set does not already contain many correlated volatility proxies.

6.3 Autoregressive Feature Set

The reduced autoregressive comparison is one of the clearest pieces of evidence. Log-volatility lags alone obtain R^2 0.497. Adding the insights improves R^2 to 0.538.

The Diebold-Mariano tests are significant with p < 0.01, and the R^2 improvement is approximately +0.0413. This supports the idea that the conditional volatility estimate adds information beyond simple lag persistence.

6.4 Insights Alone

The insights-alone dataset performs poorly: RMSE 0.0221, MAE 0.0142, MSE 0.000487 and R^2 0.037. It does not beat the persistence baseline, which has R^2 0.320.

This rejects the second hypothesis. The conditional volatility signal should not be interpreted as a standalone forecasting engine. It is useful as a complementary feature.

6.5 Interpretability Results

SHAP ranks `log_sigma_garch` in the top 20 features of the complete dataset, around rank 13/55. It ranks higher in reduced settings: 9/45 in the technical indicators + insights dataset and 3/8 in the autoregressive + insights dataset.

This ranking pattern is coherent with the ablation results. The more constrained the feature set, the more visible the EGARCH signal becomes.

Permutation importance shows only modest degradation in the complete model. This likely reflects compensation by correlated volatility features rather than a total absence of usefulness.

7. Discussion

7.1 Economic Interpretation

The conditional log-sigma feature can be read as a market-level risk-regime indicator. It compresses information about conditional volatility persistence, asymmetry and heavy tails into a single feature that the machine-learning model can combine with stock-level signals.

This is especially relevant during stress periods. A model that recognizes conditional volatility regimes can be more useful for VaR calibration, capital buffers and risk-control decisions.

7.2 Prediction Versus Decision

The report is framed around predictive performance, but the deeper question is decision quality. Better volatility forecasts matter if they improve downstream risk decisions. This is why the work includes statistical tests and honest limitations instead of treating a small RMSE gain as a complete economic result.

The natural next step would be to evaluate the effect of improved volatility forecasts on operational risk decisions: VaR exceedances, position sizing, drawdown control or capital allocation.

7.3 Methodological Lessons

The main methodological lesson is that econometric features can still matter inside modern ML systems when they are constructed causally and tested through ablations.

The second lesson is that a feature's value is contextual. A signal can be statistically useful in one feature set, redundant in another, and weak as a standalone model.

The third lesson is that reporting negative or mixed results is part of the scientific value of the project. H2 is rejected, and the technical-indicator comparison is nuanced.

8. Limitations and Future Work

The current work is limited to a D+1 horizon and a fixed S&P 500 universe. It does not model transaction costs, portfolio construction, live execution or multi-horizon decision-making.

The ARIMA component remains minimal and could be strengthened. The EGARCH family could be expanded to GJR-GARCH, FIGARCH or HAR-RV specifications. The machine-learning comparison could include XGBoost, CatBoost and neural architectures such as LSTM or GRU.

Future work should also test other horizons, especially 5, 10 and 21 trading days. It should evaluate other asset classes and stress periods. Most importantly, it should link forecasting improvements to downstream economic decisions.

9. Final Conclusion

The empirical evidence supports a clear but bounded conclusion: ARIMA-EGARCH conditional volatility can improve LightGBM forecasts when used as a leakage-free feature.

The signal is not enough by itself. It is valuable because it contributes structured econometric information to a broader nonlinear model. This makes the project a good example of disciplined applied ML: start from an economic question, build causal features, run ablations, test statistical significance, inspect interpretability and report limitations.

For a YC or international reader, the key takeaway is not that this project is a finished financial product. The key takeaway is that the builder can connect econometrics, machine learning, software engineering and honest evaluation in a reproducible research pipeline.