

Statistical and Machine Learning Models for Stock Price and Return Forecasting: An Out-of-Sample Comparison

Smaila Amoanu

`smaila.amoanu@aims.ac.rw`

University of Nevada <https://orcid.org/0009-0002-1641-5422>

Research Article

Keywords: Financial Time Series; Stock Price Forecasting; Return Forecasting; ARIMA; ARMA–GARCH; Random Forest; XGBoost; Random Walk; Out-of-Sample Forecasting; S&P 500

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-7837766/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Comparing Statistical and Machine Learning Models for Stock Price and Return Forecasting: Evidence from the S&P 500

Smaila Amoanu

Department of Statistics and Data Science, University of Nevada, Reno

Abstract. This study compares classical time-series models and tree-based machine-learning methods for one-step-ahead forecasting of S&P 500 prices and returns. The analysis is deliberately separated into two forecasting tasks because a persistent price level and a stationary return series represent different statistical problems. For price forecasting, a random-walk benchmark is compared with ARIMA, Random Forest, and XGBoost. For return forecasting, a zero-return benchmark is compared with ARMA, ARMA–GARCH, Random Forest, and XGBoost. Models are evaluated using a chronological out-of-sample design, with 2025 reserved as the main holdout and 2022–2025 used for an annually re-estimated robustness analysis. In the 2025 holdout, ARIMA has slightly lower price error than the random walk, but the difference is not statistically significant. Random Forest and XGBoost perform poorly when the index moves beyond the price range observed during estimation. For returns, rankings vary by loss measure and year, and no fitted model shows stable improvement over the zero-return benchmark. The results indicate that model rankings depend materially on the forecasting target and that simple benchmarks remain important in financial forecasting comparisons.

Keywords: financial time series; S&P 500; ARIMA; GARCH; Random Forest; XGBoost; forecast evaluation

1 Introduction

Forecasting financial markets is difficult for a simple reason: the object being forecast matters. A stock-index level is highly persistent and may appear predictable over short horizons even when its daily increments are not. Returns, by contrast, fluctuate around a much more stable level and typically contain considerably less short-run signal. Treating these two targets as though they were the same forecasting problem can therefore lead to misleading model comparisons.

Classical time-series methods provide a natural starting point. ARIMA models describe dependence in the conditional mean through autoregressive and moving-average terms (Box et al., 2015), while GARCH models allow conditional variance to change over time and capture the volatility clustering commonly observed in financial returns (Engle, 1982; Bollerslev, 1986). Machine-learning methods such as Random Forest and XGBoost offer a different approach by estimating nonlinear relationships from lagged predictors without imposing a fixed parametric mean function (Breiman, 2001; Chen and Guestrin, 2016). These methods have been used widely in financial prediction, including recent work on S&P 500 return forecasting (Htun et al., 2024).

The existence of more flexible models does not remove the need for simple benchmarks. The random-walk view of asset prices has a long history in financial economics (Fama, 1970), and a model that cannot improve on a no-change price forecast may have limited practical forecasting value even if it outperforms another fitted model. A similar point applies to daily returns, where a zero-return forecast provides a transparent reference. Recent comparative work also shows that rankings across econometric and machine-learning methods can change with the forecast horizon, target, and evaluation design (Bunnag, 2026). This makes the experimental setup at least as important as the choice of algorithm.

The purpose of this paper is therefore modest. We do not propose a new forecasting method. Instead, we ask how ARIMA, ARMA–GARCH, Random Forest, and XGBoost compare when price forecasting and return forecasting are evaluated separately under the same chronological out-of-sample design. The analysis emphasizes three points. First, price models are compared against a random-walk benchmark. Second, return models are compared against a zero-return benchmark. Third, the main 2025 holdout is supplemented by annual holdouts for 2022–2025 so that conclusions do not depend on a single test period.

The results are clear. The random walk remains difficult to improve upon for one-day-ahead price forecasting. ARIMA is numerically competitive, but its advantage in 2025 is small and not statistically significant. Direct Random Forest and XGBoost price forecasts deteriorate sharply once the index moves above the price range represented in the training data. Return forecasting is more difficult still. ARMA–GARCH, Random Forest, and XGBoost each perform well in some periods or under some metrics, but none shows stable dominance over the zero-return benchmark. These findings support a restrained conclusion: forecast performance depends on the target being predicted, and model complexity alone does not guarantee better out-of-sample forecasts.

2 Data and methods

2.1 Data and forecasting targets

The empirical series is the daily closing level of the S&P 500 index from January 2015 through December 2025. S&P Dow Jones Indices LLC is the underlying source of the index, and FRED reports the series as a daily closing price index rather than a total-return index, so dividends are not included (S&P Dow Jones Indices LLC, 2026). The fixed data file used in the reproducibility package combines

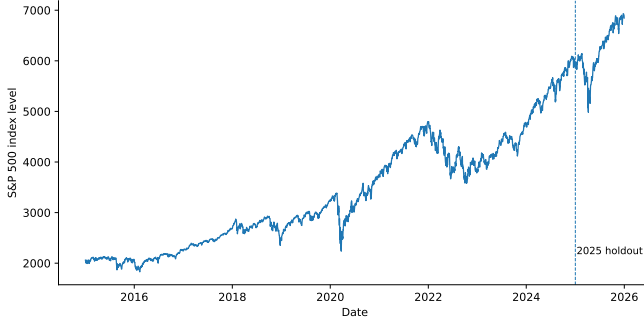


Figure 1: Daily S&P 500 closing index level, 2015–2025. The dashed vertical line marks the beginning of the 2025 holdout used in the main analysis.

the recent FRED history with archived historical closing values for the earlier part of the sample.

Let P_t denote the closing index level on trading day t . Daily log returns are defined in percentage terms as

$$r_t = 100 \log \left(\frac{P_t}{P_{t-1}} \right). \quad (1)$$

The two response variables are therefore P_{t+1} for the price experiment and r_{t+1} for the return experiment.

Figure 1 shows the full price series. The index exhibits a pronounced upward movement over the sample, together with several periods of substantial market decline and recovery. A preliminary Augmented Dickey–Fuller diagnostic gives a p -value of 0.938 for log prices and a value below 0.001 for daily returns. This distinction is consistent with treating the price level as a nonstationary process and the return series as stationary for the present forecasting exercise.

2.2 Price forecasting models

The price experiment contains four models. The first is the random-walk benchmark,

$$P_{t+1} = P_t + \varepsilon_{t+1}, \quad \hat{P}_{t+1|t} = P_t. \quad (2)$$

This benchmark is intentionally simple. It asks whether a fitted model improves materially on the most recent observed price.

The statistical competitor is an ARIMA($p, 1, q$) model,

$$\Phi(B)(1 - B)P_t = c + \Theta(B)\varepsilon_t, \quad (3)$$

where B is the backshift operator. The orders p and q are selected by Akaike’s information criterion over $p, q \in \{0, 1, 2\}$. For the 2015–2024 estimation sample, this procedure selects ARIMA(2, 1, 0).

Random Forest and XGBoost are fitted directly to the price level using the five most recent observed prices as predictors. Thus, with $L = 5$,

$$X_t = (P_{t-1}, \dots, P_{t-L}), \quad P_t = f(X_t) + \varepsilon_t. \quad (4)$$

The same lag depth is used for both machine-learning methods so that differences are not driven by different

information sets. Random Forest uses 200 trees, maximum depth 8, and minimum leaf size 3. XGBoost uses 200 boosting rounds, maximum depth 3, learning rate 0.05, and subsampling rate 0.90. These settings are deliberately modest. The purpose is a controlled comparison rather than an exhaustive hyperparameter search.

2.3 Return forecasting models

The return experiment begins with the zero-return benchmark,

$$\hat{r}_{t+1|t} = 0. \quad (5)$$

The first fitted model is an ARMA(p, q) specification for the conditional mean. As with prices, p and q are selected by AIC over $\{0, 1, 2\}$. The 2015–2024 sample selects ARMA(2, 2).

The ARMA–GARCH model retains the same ARMA conditional-mean order and allows conditional variance to vary over time. In the GARCH(1, 1) case,

$$r_t = \mu_t + \varepsilon_t, \quad (6)$$

$$\varepsilon_t = \sigma_t z_t, \quad (7)$$

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \quad (8)$$

where μ_t follows the selected ARMA structure. Parameters are estimated jointly by Gaussian quasi-maximum likelihood subject to stationarity, invertibility, positivity, and $\alpha + \beta < 1$. In the main estimation sample, $\hat{\alpha} + \hat{\beta} = 0.969$, indicating substantial volatility persistence. The GARCH component is included because time-varying variance is a central feature of financial returns, but forecast comparisons in this paper concern the conditional mean return rather than volatility itself.

Random Forest and XGBoost use the same five-lag structure as in the price experiment, now applied to returns. Their hyperparameters are held fixed across price and return tasks. This keeps the comparison transparent and avoids giving one method a larger tuning budget than another.

2.4 Out-of-sample design and evaluation

The main experiment estimates all models using 2015–2024 and reserves calendar year 2025 as an untouched holdout. Forecasts are one-step ahead. At each trading day, the model uses information available through the previous trading day. After the realized value is observed, it becomes part of the information set for the next forecast. Statistical states are updated without re-estimating coefficients during the holdout, and the fitted tree structures are also kept fixed.

A robustness analysis repeats the same procedure for 2022, 2023, 2024, and 2025. For each holdout year, all models are estimated using only observations available through the preceding December. ARIMA and ARMA orders are reselected by AIC using the relevant estimation sample. This expanding annual design follows the general principle that time-series validation should preserve

Table 1: One-step-ahead S&P 500 price forecasts, 2025 holdout.

Model	MAE	RMSE	MAPE (%)
Random walk	44.09	66.77	0.734
ARIMA	43.50	66.19	0.724
Random Forest	314.33	429.61	4.784
XGBoost	319.98	434.91	4.872

temporal ordering and avoid future information leakage (Bergmeir et al., 2018).

Forecast accuracy is summarized by mean absolute error and root mean squared error,

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |Y_t - \hat{Y}_t|, \quad (9)$$

$$\text{RMSE} = \left\{ \frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2 \right\}^{1/2}. \quad (10)$$

For positive price levels we also report mean absolute percentage error. Directional accuracy is included for fitted return models as a descriptive measure, not as the principal selection criterion.

Differences in squared forecast loss are examined using a Diebold–Mariano-style comparison (Diebold and Mariano, 1995). For a fitted model M and benchmark B , define

$$d_t = e_{M,t}^2 - e_{B,t}^2. \quad (11)$$

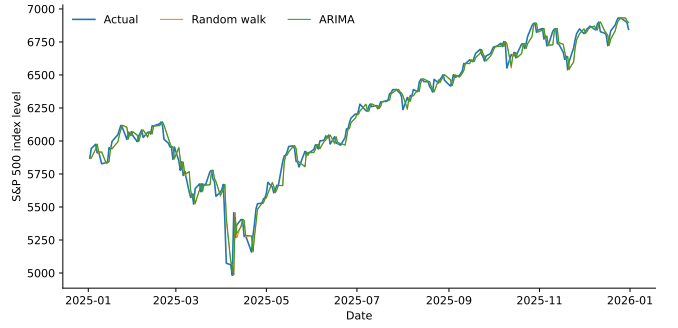
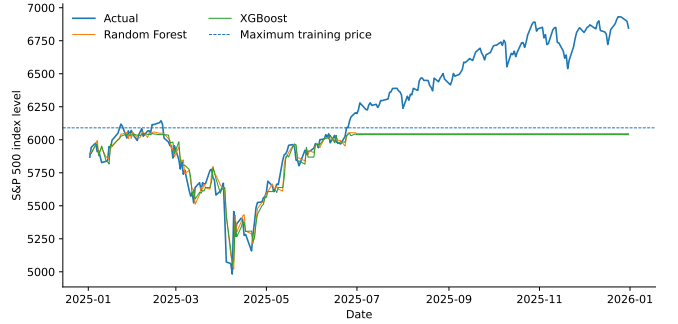
The null hypothesis is $E(d_t) = 0$. A negative mean differential favors the fitted model, while a positive value favors the benchmark. Standard errors are based on a Newey–West heteroskedasticity and autocorrelation consistent estimate with five lags (Newey and West, 1987).

3 Results

3.1 Main 2025 holdout

Table 1 reports the one-step-ahead price results. ARIMA has the smallest numerical errors, with RMSE 66.19 compared with 66.77 for the random walk. The difference is small. The loss-differential test comparing ARIMA with the random walk gives $p = 0.380$, so the numerical advantage is not statistically significant. Figure 2 shows why the two forecasts are so similar. Both track the observed index closely at the one-day horizon because yesterday’s price contains most of the short-run level information.

The tree-based price forecasts behave differently. Their errors are much larger than those of the statistical models, and the loss-differential tests strongly favor the random walk ($p < 0.001$ for both models). Figure 3 provides a direct explanation. The maximum S&P 500 level in the 2015–2024 estimation sample is 6090.27. During 2025, the index rises as high as 6932.05, while the largest Random Forest and XGBoost forecasts are only 6059.39 and 6048.81. Once the observed index moves materially beyond the training range, the tree forecasts flatten close

**Figure 2:** Actual S&P 500 levels and one-step-ahead forecasts from the random walk and ARIMA models in 2025.**Figure 3:** Actual S&P 500 levels and direct Random Forest and XGBoost price forecasts in 2025. The dashed horizontal line marks the maximum price observed in the 2015–2024 estimation sample.

to the upper part of that range. This result should be interpreted as a limitation of the direct lagged-price specification used here, not as a general statement that tree methods are unsuitable for financial forecasting.

Return forecasting produces a different ordering. Table 2 shows that XGBoost has the smallest RMSE, while ARMA–GARCH has the smallest MAE. Random Forest has the highest directional accuracy at 57.2%. These numerical differences are not sufficient to establish a stable forecasting advantage. Relative to the zero-return benchmark, the squared-loss comparison gives $p = 0.650$ for ARMA, $p = 0.507$ for ARMA–GARCH, $p = 0.842$ for Random Forest, and $p = 0.459$ for XGBoost. None of the fitted return models therefore provides a statistically significant improvement in mean squared forecast loss during 2025.

3.2 Robustness across 2022–2025

The annual holdouts confirm that the 2025 findings are not driven entirely by one test year. Table 3 reports price RMSE for each holdout. The random walk has the smallest RMSE in 2022, 2023, and 2024, while ARIMA is marginally lower in 2025. When the four holdouts are pooled, the random walk has RMSE 53.32 and ARIMA has RMSE 53.89. The pooled loss-differential test gives $p = 0.065$, so the difference is not significant at the conventional 5% level.

The large tree-model errors in 2024 and 2025 are con-

Table 2: One-step-ahead S&P 500 return forecasts, 2025 holdout. Returns are in percentage points.

Model	MAE	RMSE	Directional accuracy (%)
Zero return	0.734	1.174	—
ARMA	0.753	1.192	50.4
ARMA-GARCH	0.726	1.165	54.8
Random Forest	0.744	1.178	57.2
XGBoost	0.745	1.125	54.8

Table 3: Annual one-step-ahead price RMSE under re-estimated holdouts.

Model	2022	2023	2024	2025
Random walk	61.78	34.75	43.52	66.77
ARIMA	63.58	35.12	44.30	66.19
Random Forest	70.23	37.64	754.87	429.61
XGBoost	73.49	37.20	759.09	434.91

sistent with the same extrapolation issue observed in the main holdout. In earlier years, when the test range remains closer to the estimation range, the tree forecasts are much more competitive. This variation is useful because it shows that the weakness is conditional on the level-forecasting design and market path rather than a uniform failure across all periods.

For returns, the ranking changes from year to year. The zero-return benchmark has the smallest RMSE in 2022, XGBoost is marginally lowest in 2023, Random Forest is marginally lowest in 2024, and XGBoost is lowest in 2025. Figure 4 summarizes this variation. No single fitted return model remains best across the four holdouts.

Pooling all four years gives RMSE 1.121 for the zero-return benchmark, 1.148 for ARMA, 1.124 for ARMA-GARCH, 1.134 for Random Forest, and 1.134 for XGBoost. ARMA is significantly worse than the zero-return benchmark under the pooled squared-loss comparison ($p = 0.026$). The differences for ARMA-GARCH ($p = 0.582$), Random Forest ($p = 0.180$), and XGBoost ($p = 0.554$) are not statistically significant. The pooled directional accuracies are 49.5%, 51.5%, 53.4%, and 54.1% for ARMA, ARMA-GARCH, Random Forest, and XGBoost, respectively. These values provide some descriptive evidence of directional information in the tree models, but they do not overturn the broader result that improvements in magnitude-based forecast loss are small and unstable.

4 Discussion

The main finding is not that one model universally dominates the others. It is that the conclusion changes with the forecasting target.

For the S&P 500 price level, the random walk is a strong one-day-ahead benchmark. ARIMA tracks it closely and is occasionally numerically better, but the improvement is small enough that the loss-differential tests do not establish a clear advantage. This is consistent with the high

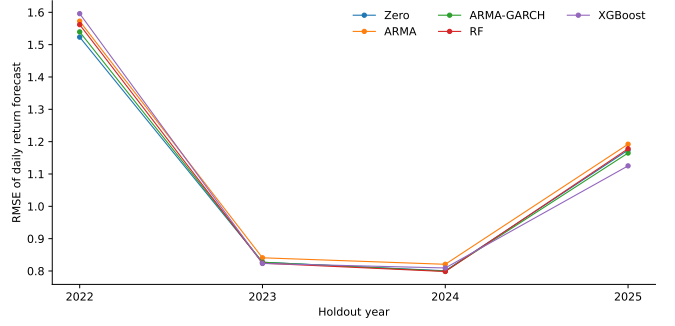


Figure 4: One-step-ahead return RMSE across annual holdouts. Model rankings vary across years.

persistence of the index level. At such a short horizon, a model must improve on an already informative forecast, namely the most recent observed price.

The direct Random Forest and XGBoost price forecasts reveal a second issue. Both methods use lagged price levels as predictors and learn from the range represented in the estimation sample. Their 2025 forecasts flatten when the index rises substantially above that range. The same effect is visible in 2024, whereas the models are far more competitive in 2022 and 2023. This is an important practical distinction. Flexible nonlinear models can fit local relationships well, but direct price-level forecasting may still be poorly suited to a market that moves into a new level regime. Transforming the prediction target to returns largely removes this particular level-extrapolation problem, although it introduces the more fundamental difficulty that returns themselves contain little predictable mean structure.

The return results make that difficulty explicit. ARMA-GARCH produces the smallest MAE in 2025 and remains close to the zero-return benchmark in the pooled analysis. Its estimated variance persistence is consistently high across annual fits, with $\hat{\alpha} + \hat{\beta}$ between approximately 0.963 and 0.980. This confirms that conditional volatility is persistent in the sample. It does not, however, imply that the conditional mean return is correspondingly easy to forecast. The distinction is important: GARCH can be useful for modeling risk even when improvements in mean return forecasts are weak.

The machine-learning return models show a similar pattern. XGBoost has the lowest 2025 RMSE and the highest pooled directional accuracy, while Random Forest performs best by RMSE in 2024. Yet neither method significantly improves on zero return in the pooled squared-loss comparison. The results therefore caution against interpreting a small error ranking as evidence of stable superiority. Formal forecast comparison and repeated holdouts provide a more reliable picture than a single test split.

This study has several deliberate limitations. It considers one equity index, one-day-ahead forecasts, and a compact lag-only predictor set. The tree models are not tuned through a large hyperparameter search, and no macroeconomic, valuation, or sentiment variables are in-

cluded. These choices are intentional because the objective is a controlled comparison of model classes rather than the construction of a production forecasting system. Future work could examine whether the conclusions change across assets, longer horizons, or richer information sets, but those extensions are not necessary for the question addressed here.

5 Conclusion

This paper compared statistical and tree-based machine-learning models for one-step-ahead S&P 500 price and return forecasting under a common out-of-sample design. Separating the two targets changes the interpretation of model performance. For prices, the random walk remains a difficult benchmark to beat, and ARIMA does not provide a statistically clear improvement. Direct Random Forest and XGBoost forecasts are vulnerable when the index moves outside the level range represented in the training sample. For returns, model rankings vary across years and loss measures, and no fitted model demonstrates stable superiority over a zero-return benchmark.

The practical conclusion is therefore restrained. Forecasting performance should be evaluated relative to the target, the benchmark, and the out-of-sample period. More flexible models can be useful, but complexity by itself is not evidence of predictive improvement. In this application, simple benchmarks remain central to a defensible comparison of financial forecasting methods.

Data and code availability

The fixed analysis data, executed Jupyter notebook, forecast outputs, and robustness tables used to produce the results are included in the accompanying reproducibility package.

Declarations

Funding. The author received no specific funding for this study.

Competing interests. The author declares no competing interests.

Use of generative AI. Generative AI tools were used to assist with code development, manuscript organization, language editing, and formatting. All analyses, numerical results, interpretations, and final manuscript content were reviewed and verified by the author, who takes full responsibility for the work.

References

- Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Bunnag, T. (2026). Can geopolitical risk improve the forecasting of Thai stock-market returns? Evidence from econometrics and machine-learning models. *Journal of Risk and Financial Management*, 8(5), 87.
- Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.
- Htun, H. H., Biehl, M., and Petkov, N. (2024). Forecasting relative returns for S&P 500 stocks using machine learning. *Financial Innovation*, 10, 118.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- S&P Dow Jones Indices LLC (2026). S&P 500 [SP500]. Federal Reserve Bank of St. Louis, FRED.