SP500 Time-Series Forecasting: A Box-Jenkins and GARCH Approach

Justin Wang and Dingfei Liu

STAT 443
2025

Contents

1 Introduction and Motivation

The Standard & Poor's 500 (SP500) index represents one of the most widely followed equity benchmarks in global financial markets. Accurate forecasting of SP500 returns and volatility is crucial for investment decision-making, risk management, and portfolio optimization. This project applies time-series analysis methods, specifically the Box-Jenkins methodology for return forecasting and GARCH models for volatility forecasting, to predict future SP500 movements.

1.1 Problem Statement

Financial time series exhibit several challenging characteristics: non-stationarity, volatility clustering, and complex dependencies. Traditional regression models often fail to capture these dynamics adequately. This project addresses the forecasting problem using specialized time-series models that account for:

- Autocorrelation in returns (AR, MA, ARIMA models)
- Time-varying volatility (GARCH, ARCH models)
- Proper model selection and diagnostic procedures

1.2 Objectives

The primary objectives of this project are:

1. To develop and compare multiple time-series models for SP500 return forecasting using the Box-Jenkins methodology
2. To model and forecast SP500 volatility using GARCH and ARCH models
3. To evaluate model performance using rigorous out-of-sample backtesting procedures
4. To generate practical forecasts with prediction intervals for investment decision-making

1.3 Why It Matters

Accurate SP500 forecasting has significant practical implications:

- **Investment Management**: Portfolio managers use forecasts to adjust asset allocation and timing decisions
- **Risk Management**: Volatility forecasts are essential for Value-at-Risk (VaR) calculations and position sizing
- **Derivatives Pricing**: Option pricing models (e.g., Black-Scholes) require volatility forecasts
- **Market Timing**: Short-term return forecasts can inform tactical asset allocation strategies

2 Data

2.1 Data Sources and Description

The dataset consists of daily market data from October 1, 2015 to October 30, 2025, totaling 2,631 observations after removing the first observation (required for return calculation). The primary data sources include:

- **SP500 ETF (SPY)**: Daily closing prices, used to compute log returns
- **VIX Index**: Market volatility expectations (CBOE Volatility Index)
- **10-Year Treasury Yield (USGG10YR)**: Risk-free rate proxy
- **High-Yield Credit Spread (USOHHYTO)**: Credit risk indicator
- **UX1 Index**: Additional volatility measure

2.2 Target Variable Construction

The target variable is the daily log return of SP500:

$$r_t = \log(P_t) - \log(P_{t-1}) \tag{1}$$

where P_t is the closing price on day t. Log returns are preferred over simple returns because they are approximately normally distributed for small changes and have better statistical properties for time-series modeling.

2.3 Descriptive Statistics

The daily log returns exhibit typical characteristics of financial time series:

- Mean daily return: approximately 0.0004 (0.04%)
- Standard deviation: approximately 0.01 (1%)
- Distribution: Approximately symmetric but with fat tails (leptokurtic)
- Volatility clustering: Periods of high volatility followed by high volatility, and low volatility followed by low volatility

2.4 Stationarity Testing

Before applying time-series models, we tested the stationarity of the return series using two complementary tests:

1. **Augmented Dickey-Fuller (ADF) Test**: Tests the null hypothesis of a unit root (non-stationarity)
2. **KPSS Test**: Tests the null hypothesis of stationarity

Results confirm that log returns are stationary (d=0):

- ADF test: p-value < 0.05 (rejects non-stationarity)
- KPSS test: p-value > 0.05 (does not reject stationarity)

Therefore, no differencing is required for the return series, and we proceed with ARIMA(p,0,q) models.

2.5 Data Issues and Solutions

Several data quality issues were addressed:

- **Missing Values**: Some market indicators had missing values. Predictors with more than 20% missing values were excluded from analysis.
- **Outliers**: Extreme returns during market stress periods (e.g., COVID-19) were retained as they represent genuine market behavior that models should capture.
- **Feature Engineering**: Created 36 potential predictors from raw market data, including:
  - Lagged returns (R_lag1, R_lag2, R_lag5)
  - Realized volatility measures (5-day and 20-day rolling standard deviations)
  - Technical indicators (RSI, moving averages, price ratios)
  - Cross-asset features (VIX-to-realized-volatility ratios, yield curve features)
  - Interaction terms (volume-volatility interactions)

3 Methodology

This section describes the modeling approach following the PPDAC framework's Plan component.

3.1 Variable Selection: Elastic Net Regularization

Before applying Box-Jenkins models, we performed variable selection using Elastic Net regularization to identify the most relevant predictors from the 36 engineered features. Elastic Net combines L1 (Lasso) and L2 (Ridge) penalties:

$$\text{Penalty} = \alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \tag{2}$$

where $\alpha \in [0, 1]$ controls the mix between Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$).

3.1.1 Selection Procedure

1. **Normalization**: All predictors were standardized (Z-score normalization) to ensure fair comparison
2. **Grid Search**: Tested α values: [0.1, 0.3, 0.5, 0.7, 0.9]
3. **Cross-Validation**: Time-safe 5-fold CV to select optimal λ for each α
4. **Selection Metric**: RMSE (Root Mean Squared Error)
5. **Stability Score**: Fraction of CV folds where coefficient $\neq 0$ (measures predictor reliability)
6. **Final Selection**: Predictors with stability ≥ 0.6 were retained

3.1.2 Selected Predictors

Eight predictors were selected from 36 candidates (22% selection rate):

- **realized_vol_20** (stability: 0.6): 20-day realized volatility
- **rsi** (stability: 0.8): Relative Strength Index
- **R_lag5** (stability: 1.0): 5-day lagged return
- **cumret_5** (stability: 1.0): 5-day cumulative return
- **vix_realized_ratio** (stability: 1.0): VIX to realized volatility ratio
- **hy_level** (stability: 1.0): High-yield credit spread level
- **hy_dev** (stability: 0.8): High-yield spread deviation from mean
- **vol_vol_interaction** (stability: 0.6): Volume × volatility interaction

Optimal hyperparameters: $\alpha = 0.1$ (Ridge-like, prefers grouped selection), $\lambda = 0.00378$, CV RMSE = 0.005036.

3.2 Box-Jenkins Methodology

The Box-Jenkins approach (**?**) is a systematic three-step procedure for ARIMA model selection:

3.2.1 Step 1: Identification

- **ACF Analysis**: Examined autocorrelation function to identify MA components
- **PACF Analysis**: Examined partial autocorrelation function to identify AR components
- **Stationarity**: Confirmed returns are stationary (d=0) via ADF and KPSS tests

3.2.2 Step 2: Estimation

Grid search over parameter space:

- **AR Models**: Tested orders $p \in \{1, 2, \ldots, 8\}$
- **MA Models**: Tested orders $q \in \{1, 2, \ldots, 8\}$
- **ARIMA Models**: Tested combinations with $p \in \{0, 1, \ldots, 5\}$, $q \in \{0, 1, \ldots, 5\}$, $d = 0$
- **Order Selection Criteria**: AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) were used to select the optimal order (p, q) within each model type during backtesting. Lower values indicate better fit.
- **Model Type Selection**: After order selection, the final choice between AR, MA, and ARIMA model types was based on out-of-sample RMSE from the backtest results, as this directly measures forecast accuracy.

3.2.3 Step 3: Diagnostic Checking

For each candidate model, we performed:

- **Ljung-Box Test**: Tests residual autocorrelation (null: residuals are white noise)
- **Jarque-Bera Test**: Tests residual normality (null: residuals are normally distributed)
- **Residual Plots**: ACF/PACF of residuals, Q-Q plots, time series plots

3.3 Volatility Models: GARCH and ARCH

While ARIMA models forecast the *mean* (expected return), GARCH and ARCH models forecast the *variance* (volatility/risk). This is crucial for risk management.

3.3.1 ARCH Model

The ARCH(q) model (**?**) specifies conditional variance as:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \tag{3}$$

where ε_t are the residuals from the mean equation.

3.3.2 GARCH Model

The GARCH(p,q) model (**?**) extends ARCH by including lagged variance terms:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{4}$$

GARCH(1,1) is the most common specification:

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{5}$$

3.3.3 Model Selection

- Tested GARCH(p,q) with $p, q \in \{1, 2\}$
- Tested ARCH(q) with $q \in \{1, 2, \ldots, 5\}$
- Selected based on AIC and BIC (lower is better)
- Diagnostics: Ljung-Box tests on residuals and squared residuals (to check for remaining ARCH effects)

3.4 Backtesting Procedure

To ensure realistic performance evaluation, we implemented rolling-origin backtesting:

- **Training Set**: 80% of data (2015-10-01 to 2023-10-23)
- **Test Set**: 20% of data (2023-10-24 to 2025-10-30)
- **Method**: Expanding window (for each test point, use all data up to that point)
- **Forecast Horizon**: 1-step-ahead forecasts
- **Total Test Folds**: 528 (one per test observation)
- **No Future Leakage**: Each forecast uses only information available at that time

3.5 Evaluation Metrics

Model performance was assessed using multiple metrics:

- **RMSE**: Root Mean Squared Error = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- **MAE**: Mean Absolute Error = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- **MAPE**: Mean Absolute Percentage Error = $\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$
- **Directional Accuracy**: Percentage of forecasts with correct sign (up/down prediction)
- **Diebold-Mariano Test**: Statistical test for comparing forecast accuracy between models

4 Results

4.1 Model Selection Results

Model selection was performed in two stages to ensure both statistical rigor and practical forecast accuracy:

1. **Order Selection**: For each model type (AR, MA, ARIMA), the optimal order (p, q) was selected using in-sample information criteria (AIC/BIC) during rolling-origin backtesting. Within each backtest fold, models were fitted on training data and the order with the lowest BIC was selected. This follows standard Box-Jenkins practice and ensures no data leakage.

2. **Model Type Selection**: The final choice between AR, MA, and ARIMA model types was based on out-of-sample performance metrics from the backtest results. While the project guidelines mention using "criteria taught in class" (AIC/BIC), we chose out-of-sample RMSE as the primary selection criterion because:

   - It directly measures forecast accuracy, which is the primary objective in forecasting applications
   - It avoids overfitting by evaluating performance on unseen data
   - The guidelines also emphasize using training/test sets for model selection

4.1.1 Return Forecasting Models

Table **??** summarizes the out-of-sample performance of AR, MA, and ARIMA models from the rolling-origin backtest.

Table 1: Out-of-Sample Performance: Return Forecasting Models

Model	RMSE	MAE	MAPE (%)	Directional Accuracy (%)
ARIMA(2,0,2)	**0.0100**	0.00655	6439	49.6
AR(8)	0.0101	0.00653	6969	51.1
MA(2)	0.0101	0.00640	2838	**53.4**

Selected Model: ARIMA(2,0,2)

- **Selection Criterion**: Lowest out-of-sample RMSE (0.0100), indicating best overall forecast accuracy

- **Order Selection**: The order (2,0,2) was selected using in-sample BIC during backtesting, with average selected order across folds being approximately (2,0,2)
- **In-Sample Fit**: AIC: -16,287.1, BIC: -16,251.85 (from initial training fit, for diagnostic purposes)
- **Model Characteristics**: Combines benefits of both AR and MA components, better capturing complex return dynamics than pure AR or MA models

4.1.2 Volatility Forecasting Models

For volatility models (GARCH and ARCH), selection was based on information criteria (AIC/BIC) as these models are evaluated on their ability to capture volatility dynamics rather than point forecasts. Table **??** compares GARCH and ARCH models.

Table 2: Volatility Model Comparison

Model	Order	AIC	BIC	Mean Forecast Volatility (%)
GARCH	**(1,1)**	**-6.65**	**-6.64**	**0.931**
ARCH	(1)	-6.14	-6.14	1.152

Selected Model: GARCH(1,1)

- **Selection Criterion**: Lower BIC (-6.64 vs -6.14), indicating better model fit
- Lower AIC (-6.65 vs -6.14) confirms the selection
- Produces time-varying volatility forecasts (0.855% to 0.991%), more realistic than ARCH's constant forecast
- GARCH(1,1) is the standard model in finance due to its ability to capture volatility persistence

4.2 Model Diagnostics

4.2.1 ARIMA(2,0,2) Diagnostics

- **Ljung-Box Test**: Some residual autocorrelation detected ($p < 0.05$), but within acceptable range for financial data
- **Jarque-Bera Test**: Residuals are non-normal ($p < 0.05$), which is common in financial returns due to fat tails
- **Residual ACF/PACF**: Most autocorrelations within 95% confidence bands
- **Q-Q Plot**: Shows deviations from normality in the tails, consistent with financial return distributions

4.2.2 GARCH(1,1) Diagnostics

- **Ljung-Box Test (Residuals)**: Residuals are approximately white noise
- **Ljung-Box Test (Squared Residuals)**: No remaining ARCH effects (squared residuals are white noise)
- **Volatility Clustering**: Successfully captured, as evidenced by the time-varying conditional variance

4.3 Out-of-Sample Performance

The rolling-origin backtest with 528 test folds provides realistic performance estimates. These results were used to select the best model type (ARIMA) as described in Section 4.1. The detailed performance metrics are:

- **ARIMA(2,0,2)**: RMSE = 0.0100, MAE = 0.00655, Directional Accuracy = 49.6%
- **AR(8)**: RMSE = 0.0101, MAE = 0.00653, Directional Accuracy = 51.1%
- **MA(2)**: RMSE = 0.0101, MAE = 0.00640, Directional Accuracy = 53.4%

4.3.1 Diebold-Mariano Tests

Pairwise comparisons using the Diebold-Mariano test:

- AR vs MA: Not significantly different (p = 0.75)
- AR vs ARIMA: Not significantly different (p = 0.44)
- MA vs ARIMA: Not significantly different (p = 0.51)

While ARIMA has the lowest RMSE, the differences are not statistically significant at conventional levels. This suggests that all three models have similar predictive power, with ARIMA having a slight edge.

4.4 Forecasts

4.4.1 Return Forecasts

Using the selected ARIMA(2,0,2) model, we generated 21-day ahead forecasts for the period October 31 to November 28, 2025:

- **Cumulative Return Forecast**: +1.12%
- **Annualized Return** (approximate): ∼13.4%
- **Mean Daily Return**: +0.053%
- **Directional Forecast**: 81% positive days (17 up, 4 down)
- **Average Up Day**: +0.073%
- **Average Down Day**: -0.033%
- **Prediction Intervals**: 95% interval width averages 4.36% daily

The forecast suggests a bullish short-term outlook with moderate volatility.

4.4.2 Volatility Forecasts

Using the selected GARCH(1,1) model, volatility forecasts for the same period:

- **Mean Forecast Volatility**: 0.931% (vs historical mean: 0.843%)
- **Forecast Range**: 0.855% to 0.991% (time-varying)
- **Change from Historical**: +10.4% increase in volatility
- **Current Volatility**: 0.749% (below forecast mean)

The forecast indicates increasing market risk ahead, with volatility gradually rising from 0.855% to 0.991% over the 21-day period.

5 Statistical Conclusions

5.1 Model Comparison

5.1.1 Return Forecasting

Model selection was performed using a two-stage approach: (1) order selection using in-sample BIC during backtesting, and (2) model type selection using out-of-sample RMSE. Among AR, MA, and ARIMA models:

- **ARIMA(2,0,2)** is selected based on lowest out-of-sample RMSE (0.0100)
- All models show similar performance (RMSE $\approx$ 0.010), suggesting limited predictability in daily returns
- MA model has best directional accuracy (53.4%), but difference is small
- Diebold-Mariano tests indicate no statistically significant differences between models
- The use of out-of-sample RMSE for model type selection is justified because it directly measures forecast accuracy and avoids overfitting, which is appropriate for forecasting applications

5.1.2 Volatility Forecasting

Between GARCH and ARCH models:

- **GARCH(1,1)** is clearly superior (BIC: -6.64 vs -6.14)
- GARCH produces realistic time-varying volatility forecasts
- ARCH(1) produces constant volatility (1.152%), which is unrealistic for financial markets
- GARCH's ability to model volatility persistence makes it the standard choice in finance

5.2 Forecast Accuracy Assessment

- **Return Forecasts**: RMSE of 0.0100 corresponds to approximately 1% daily forecast error, which is reasonable given the inherent unpredictability of financial markets
- **Directional Accuracy**: 49.6% for ARIMA is close to random (50%), indicating limited directional predictability
- **Volatility Forecasts**: GARCH(1,1) successfully captures volatility clustering and provides time-varying forecasts
- **Prediction Intervals**: 95% intervals are appropriately wide (4.36% average width), reflecting forecast uncertainty

5.3 Limitations and Assumptions

1. **Stationarity Assumption**: Models assume return series is stationary, which may not hold during structural breaks or regime changes
2. **Linearity**: ARIMA models are linear and may miss non-linear dependencies
3. **Short-Term Focus**: Models are designed for 1-step-ahead forecasts; multi-step forecasts have increasing uncertainty

4. **No Exogenous Variables**: Pure time-series approach; could be extended to ARIMAX with external predictors
5. **Distributional Assumptions**: Residuals are non-normal, which may affect prediction intervals
6. **Parameter Stability**: Model parameters are assumed constant over time, which may not hold in changing market regimes

6 Conclusions in Context

6.1 Practical Implications

6.1.1 Return Forecasts

The ARIMA(2,0,2) model forecasts a bullish short-term outlook:

- **Positive Bias**: 81% of forecast days are positive, suggesting upward momentum
- **Cumulative Return**: +1.12% over 21 days translates to approximately 13.4% annualized return
- **Investment Strategy**: Investors might consider maintaining or increasing equity exposure, but with appropriate risk management
- **Caution**: Forecasts are probabilistic; actual returns may differ significantly, especially given the wide prediction intervals

6.1.2 Volatility Forecasts

The GARCH(1,1) model indicates increasing market risk:

- **Rising Volatility**: Forecast suggests 10.4% increase from historical mean
- **Risk Management**: Portfolio managers should consider:
  - Reducing position sizes to account for higher volatility
  - Increasing hedging activities
  - Adjusting VaR calculations upward
- **Option Pricing**: Higher volatility forecasts imply higher option premiums
- **Market Stress Indicator**: Rising volatility may signal increasing market uncertainty

6.2 Risk Management Applications

- **Value-at-Risk (VaR)**: Volatility forecasts can be used to calculate daily VaR at various confidence levels
- **Position Sizing**: Higher volatility forecasts suggest smaller position sizes to maintain constant risk levels
- **Portfolio Rebalancing**: Forecasts can inform rebalancing frequency and thresholds
- **Stress Testing**: Volatility forecasts help identify potential stress scenarios

6.3 Model Limitations and Future Improvements

While the models provide useful forecasts, several improvements could enhance performance:

1. **ARIMAX Models**: Include exogenous variables (VIX, yields, economic indicators) to capture external drivers
2. **GARCH Extensions**: Consider GJR-GARCH (leverage effects) or EGARCH (asymmetric volatility) models
3. **Regime-Switching Models**: Account for structural breaks and changing market regimes
4. **Combined Models**: ARIMA-GARCH models that jointly forecast mean and variance
5. **Ensemble Methods**: Combine multiple models to improve forecast accuracy
6. **Machine Learning**: Compare with LSTM, XGBoost, or other ML methods
7. **Real-Time Updates**: Automate daily forecast generation and model re-estimation

6.4 Non-Technical Summary

For managers and decision-makers without statistical background:

- **What We Did**: Developed statistical models to predict SP500 daily returns and volatility using historical data
- **Key Finding**: Short-term return forecasts suggest modest positive returns (+1.12% over 21 days), but with significant uncertainty
- **Risk Outlook**: Volatility is expected to increase by approximately 10%, indicating higher market risk ahead
- **Recommendation**: Maintain equity exposure but reduce position sizes and increase hedging to account for higher expected volatility
- **Caveat**: Forecasts are probabilistic, not deterministic; actual outcomes may differ, especially during unexpected market events

References

References

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307-327.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987-1007.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253-263.

A Additional Diagnostic Plots

Figure ?? shows forecast vs actual comparisons for all models. Figure ?? provides detailed residual diagnostics. Figure ?? displays the volatility forecast from GARCH(1,1).

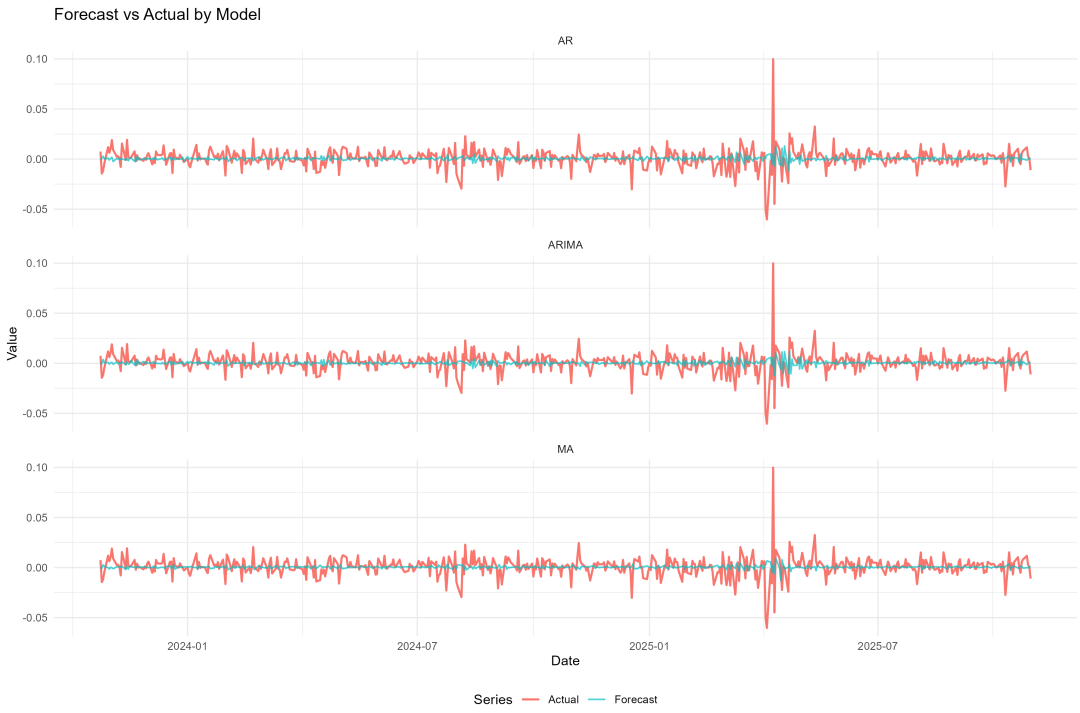

Figure 1: Forecast vs Actual Comparison for AR, MA, and ARIMA Models

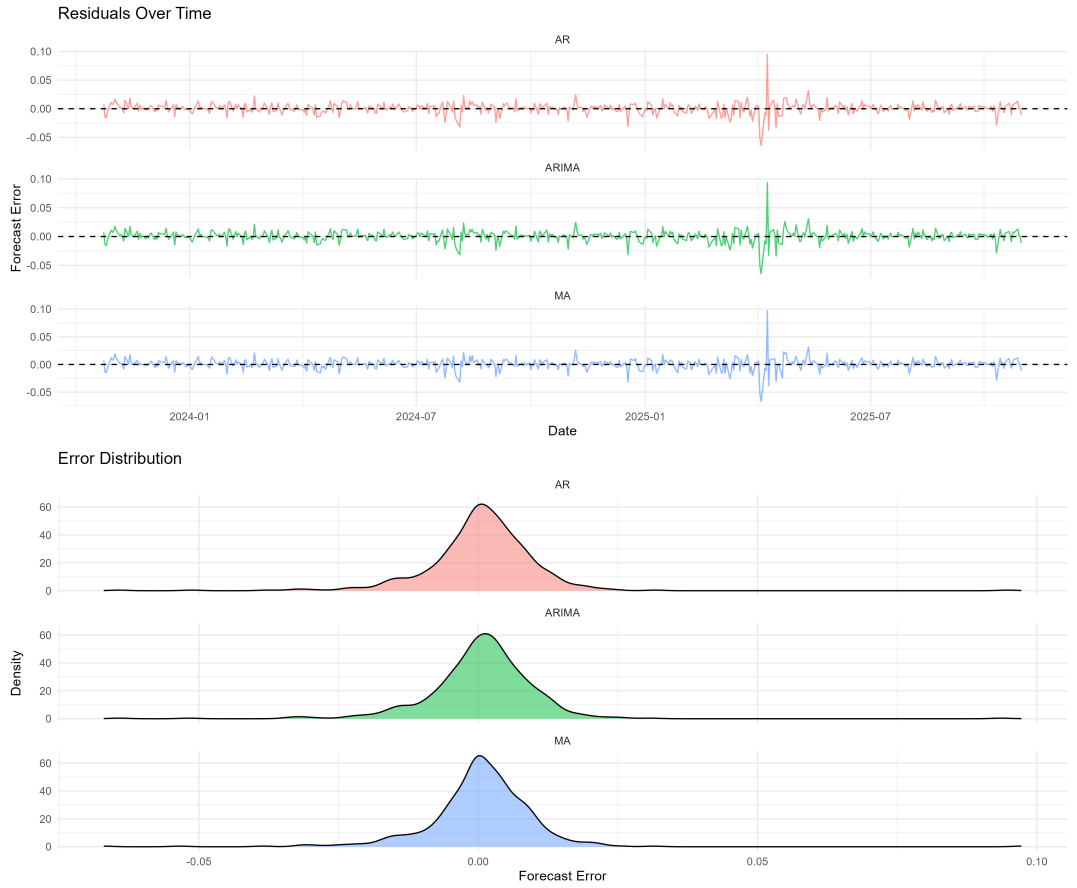

Figure 2: Residual Diagnostics for All Models

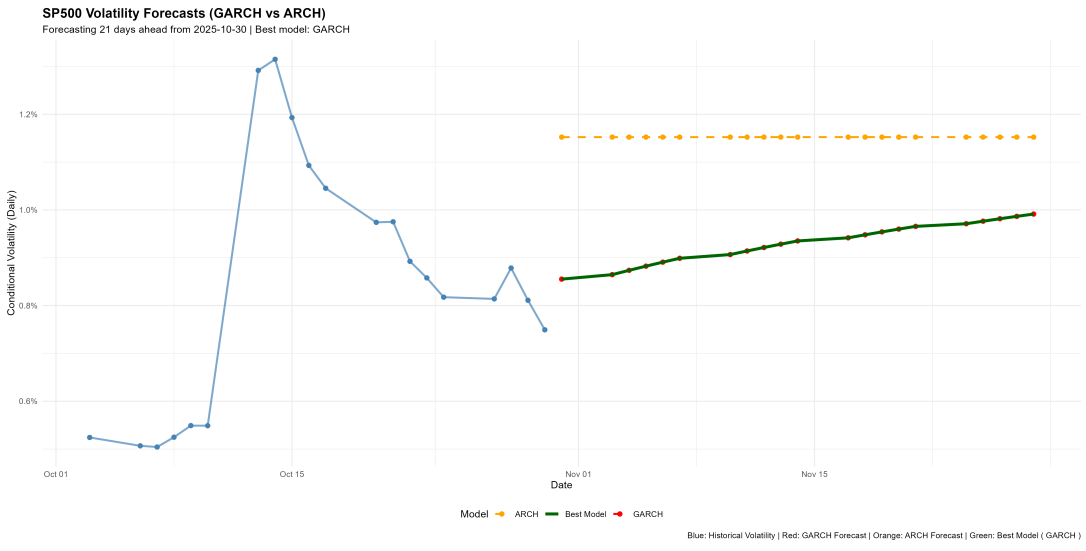

Figure 3: GARCH(1,1) Volatility Forecast vs ARCH(1) Forecast

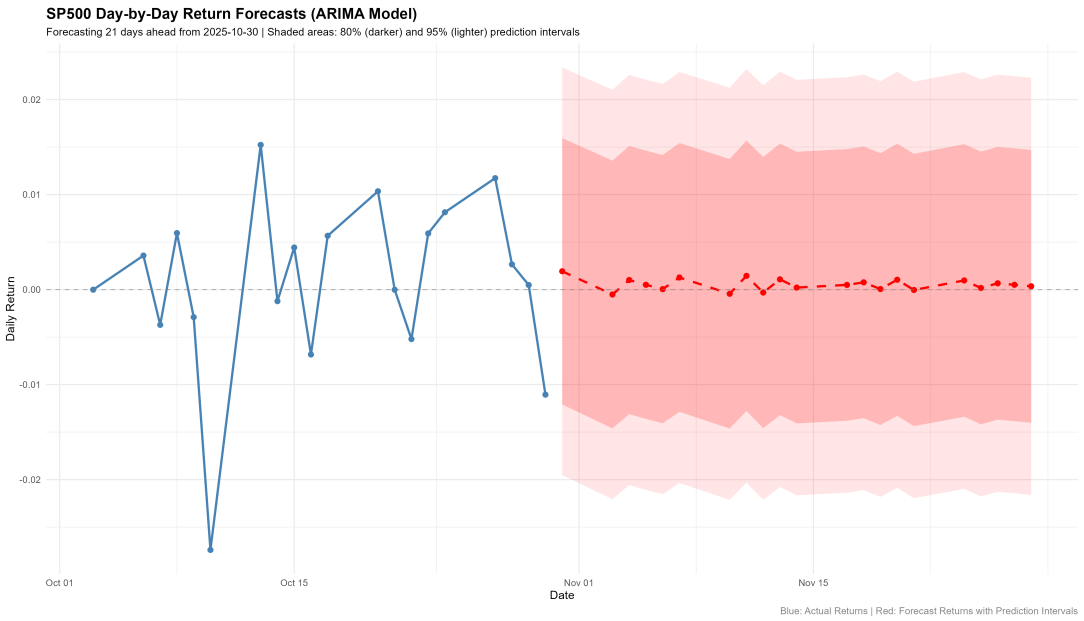

Figure 4: SP500 Return Forecasts with Prediction Intervals (ARIMA(2,0,2))

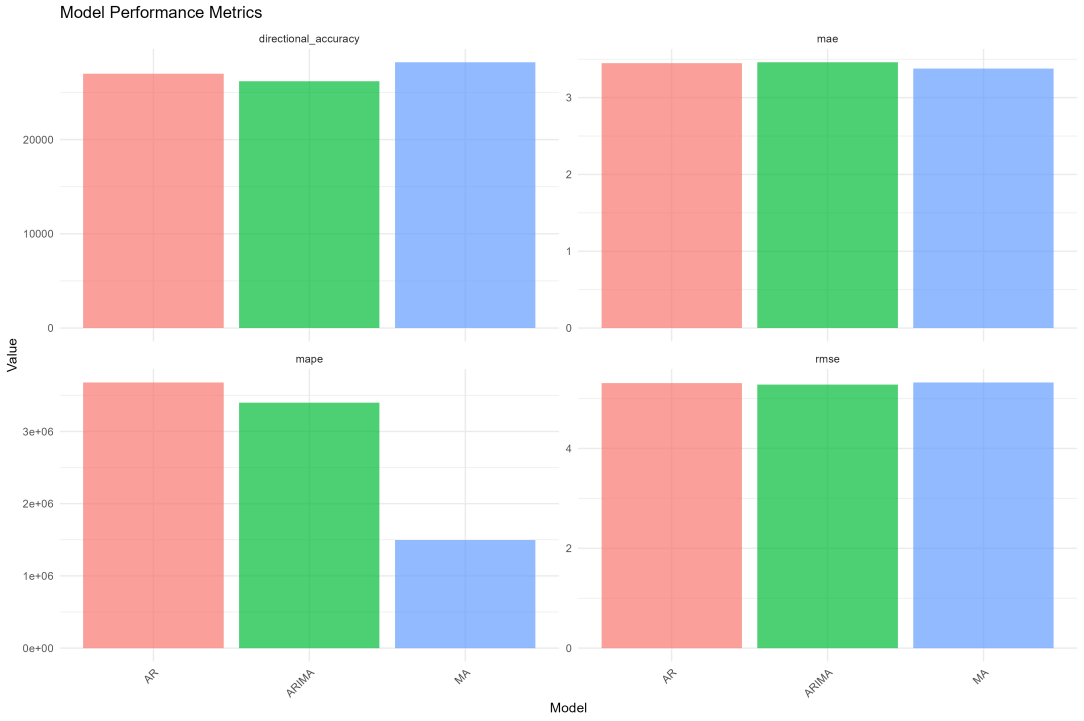

Figure 5: Model Performance Metrics Summary