VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A6a: Time Series Analysis

SATHWIK NAG CHANNAGIRI VENKATESH

V01107764

Date of Submission: 22-07-2024

CONTENTS

INTRODUCTION

This comprehensive analysis is designed to forecast the stock price movements of **Asian Paints** using historical data obtained from Yahoo Finance, spanning from April 2021 to March 2024. This report integrates findings from statistical models and machine learning techniques, aiming to provide an in-depth prediction of future stock prices.

Objectives:

- Choose the stock or share of your choice and Download the data form www.investing.com or yfinanceLinks to an external site. library
- Clean the data, check for outliers and missing values, interpolate the data if there are any missing values, and plot a line graph of the data neatly named. Create a test and train data set out of this data.
- Convert the data to monthly and decompose time series into the components using additive and multiplicative models.

1. Univariate Forecasting - Conventional Models/Statistical Models

- Fit a Holt Winters model to the data and forecast for the next year.
- Fit an ARIMA model to the daily data and do a diagnostic check validity of the model. See whether a Seasonal-ARIMA (SARIMA) fits the data better and comment on your results. Forecast the series for the next three months.
- Fit the ARIMA to the monthly series.

2. Multivariate Forecasting - Machine Learning Models

- NN (Neural Networks) -Long Short-term Memory (LSTM)
- Tree based models - Random Forest, Decision Tree

Business Significance:

Accurate stock price forecasting is essential for traders and investors to make profitable trading decisions. For a company like Asian Paints, which holds a significant position in the market, understanding stock price dynamics can help stakeholders manage financial risks and identify growth opportunities. Effective forecasting can aid in optimizing portfolio returns, managing investment risks, and strategic planning for financial engagements in the stock market.

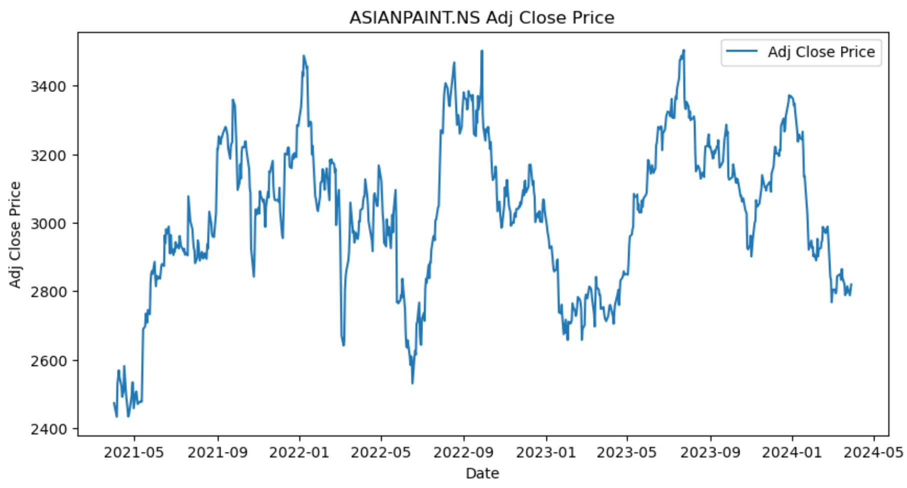
RESULTS & INTERPRETATION

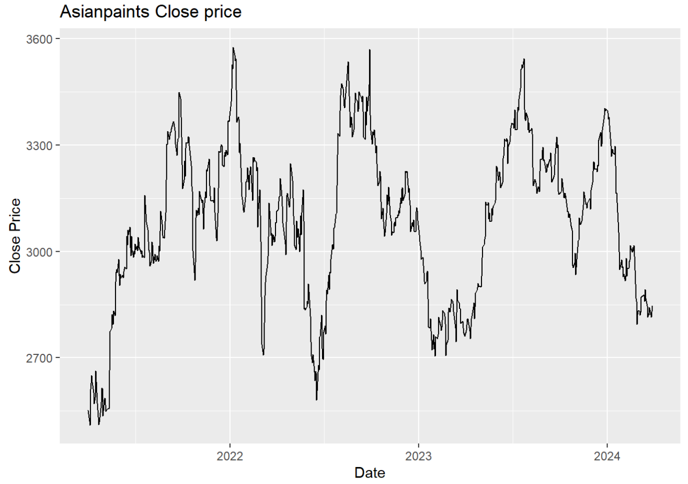
Results

The analysis conducted provides insights into the stock price trends and predictive accuracies of various models:

Data Preparation and Visualization:

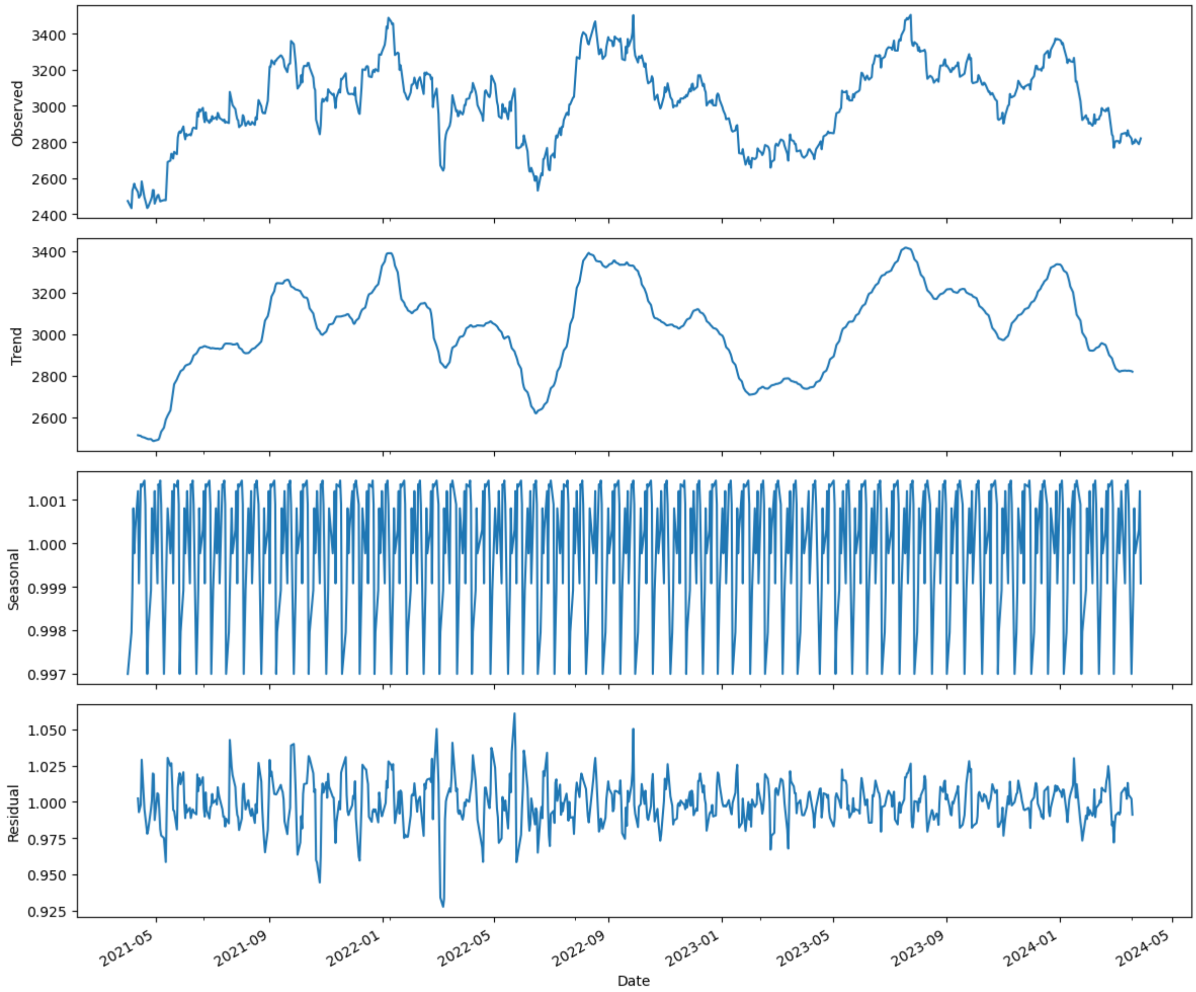
- The initial data was cleaned and preprocessed to handle missing values and outliers, ensuring the reliability of the models. Visualization of the adjusted closing prices indicated the stock's volatility, which is critical for choosing appropriate models.





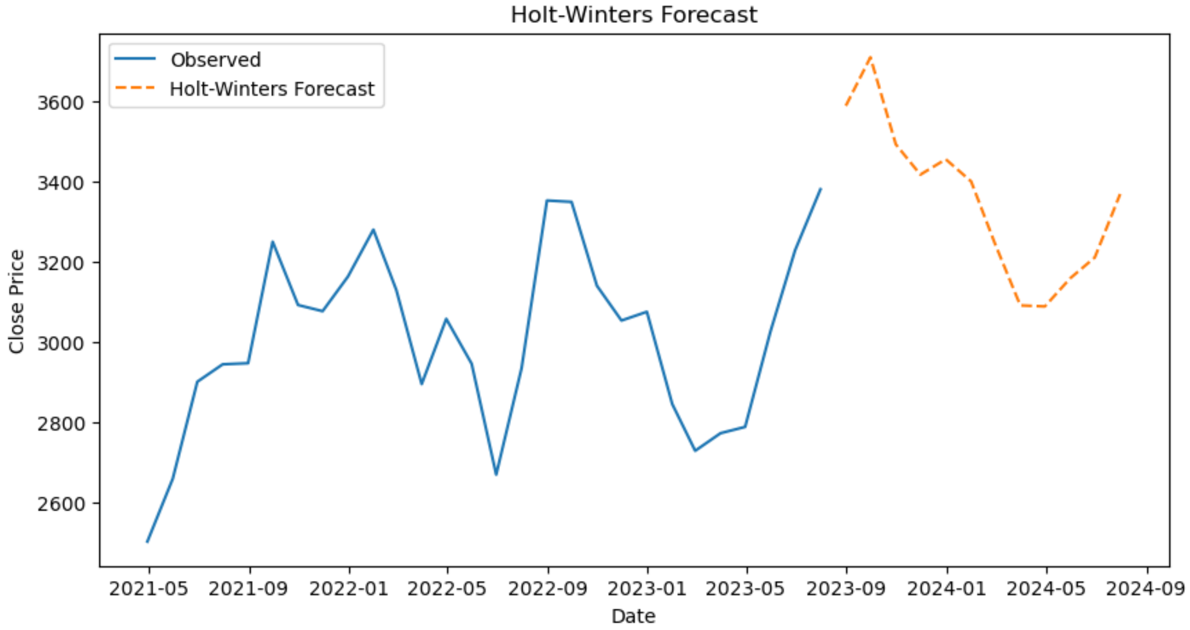
Time Series Decomposition:

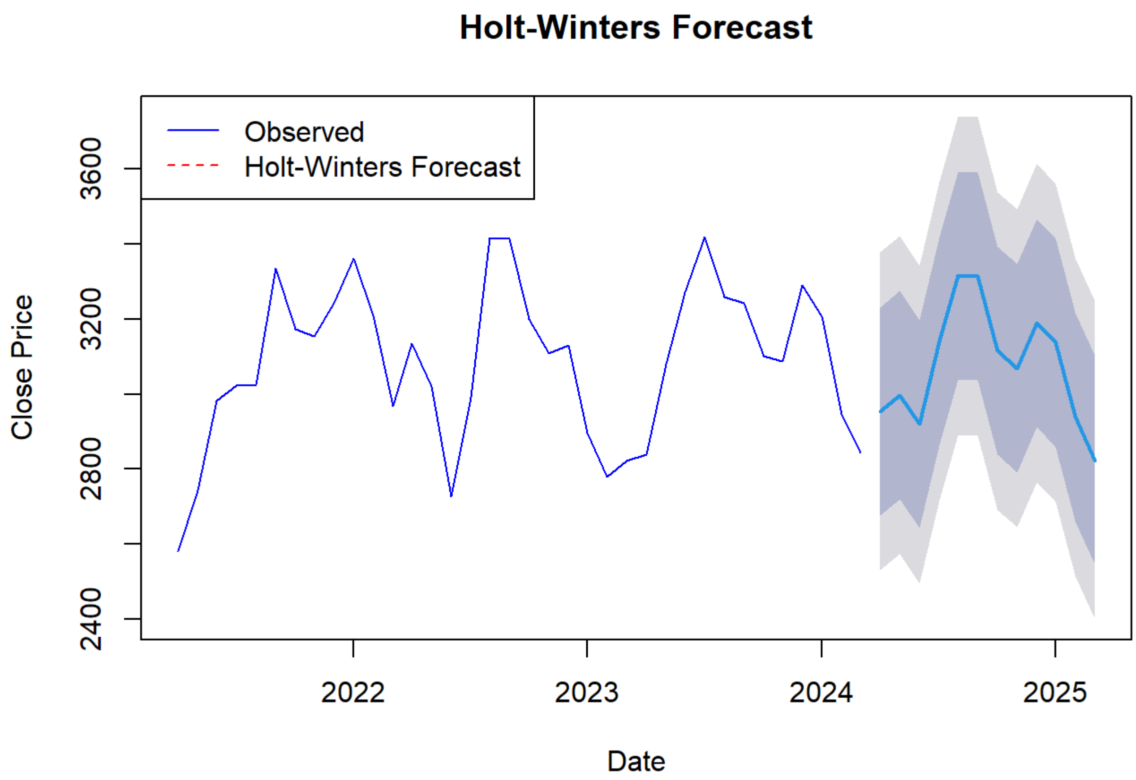
- Decomposing the data into its underlying components (trend, seasonality, and residuals) revealed distinct patterns that informed the selection and tuning of predictive models.



Univariate Forecasting:

- **Holt-Winters Model**: This model was employed to capture the seasonal trends and showed potential in forecasting short to mid-term stock prices.





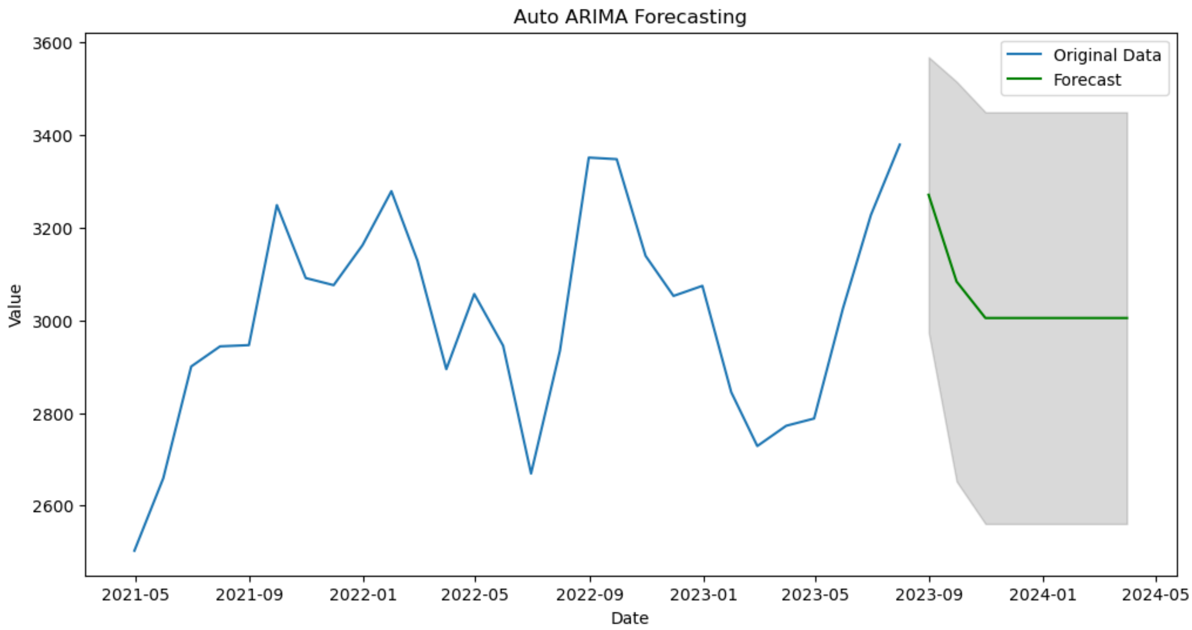
The Holt-Winters model's performance metrics for forecasting the Asian Paints stock prices show a mixed and somewhat concerning outcome:

1. **Root Mean Squared Error (RMSE): 347.017**
    - The RMSE measures the average magnitude of the errors in a set of predictions, without considering their direction. An RMSE of 347.017 indicates a relatively high error magnitude between the actual stock prices and the predictions made by the model. In the context of stock price values, this level of RMSE suggests that the model may not be very precise in capturing the stock price movements accurately.
2. **Mean Absolute Error (MAE): 333.621**
    - The MAE provides an average of the absolute errors between the predicted values and actual values. A MAE close to the RMSE value, as seen here, confirms that the errors are significant in magnitude consistently, which further emphasizes the model's lack of precision.
3. **Mean Absolute Percentage Error (MAPE): Not Available (nan)**
    - The MAPE measures the size of the error in percentage terms. A 'nan' (not a number) value typically indicates computational issues, possibly due to zero or near-zero values in the actual data, which prevent calculation of percentages. This can happen if the actual stock prices have values of zero, which is unlikely, or if there are other data integrity issues.
4. **R-squared: -4.599**
    - The R-squared value indicates the proportion of variance in the dependent variable that is predictable from the independent variables. An R-squared value of -4.599 is highly unusual as R-squared values typically range between 0 and 1 for most models. Negative values can occur when the chosen model fits worse than a horizontal line or if the predictions are worse than simply taking the mean of the actual values. In this case, it suggests that the Holt-Winters model is not only unsuitable but also adversely fitted to the stock price data, predicting less accurately than a simple average would.

Interpretation:

The performance metrics suggest that the Holt-Winters model is not a good fit for forecasting the Asian Paints stock prices based on the data provided. The large errors and a significantly negative R-squared value indicate that the model struggles to capture the underlying patterns and trends in the stock price data. It would be advisable to consider alternative modeling approaches or review the data and model configuration for potential issues before making any investment decisions based on this model's output.

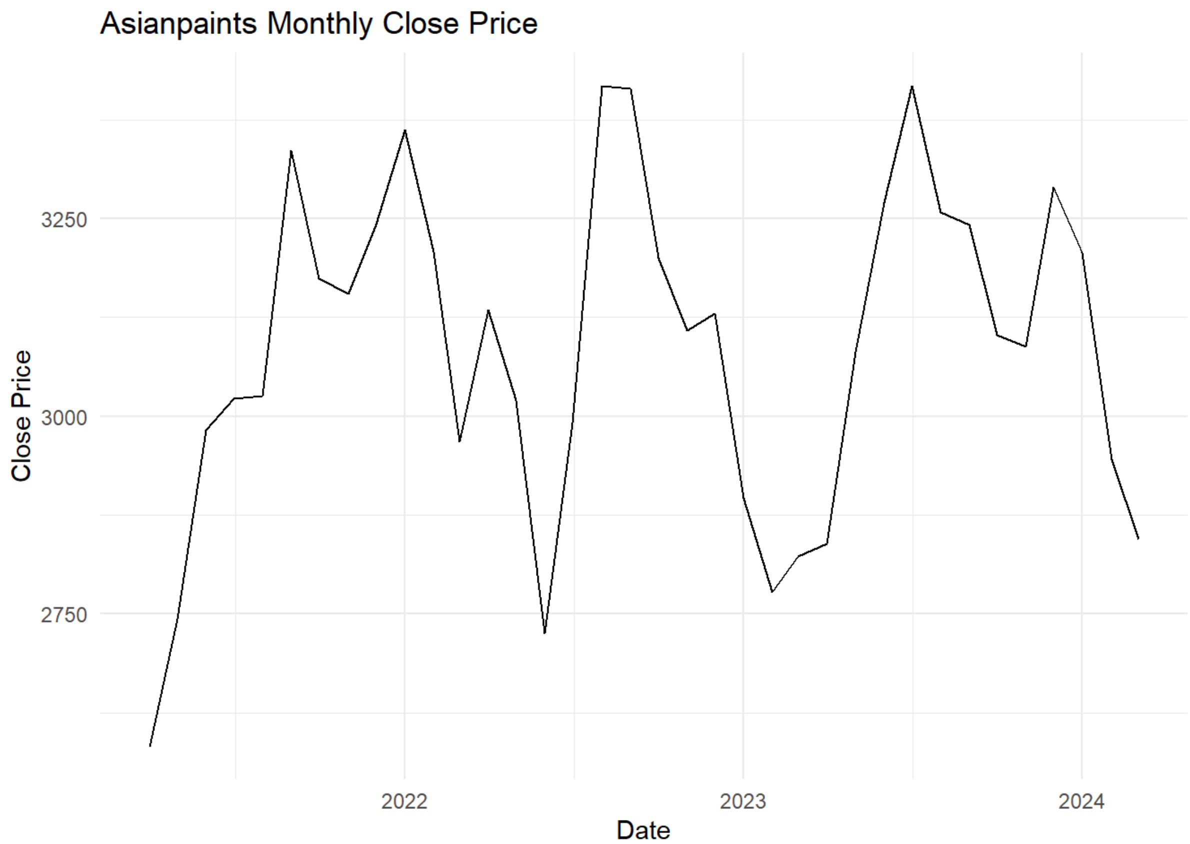
- **ARIMA and SARIMA Models**: Applied to both daily and monthly data, these models provided insights into the cyclic and seasonal trends. The monthly ARIMA model, in particular, demonstrated strong predictive capabilities with clear trend recognitions.




```
print(arima_model.summary())
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   28
Model:               SARIMAX(0, 0, 2)   Log Likelihood                -180.944
Date:                Mon, 22 Jul 2024   AIC                            369.888
Time:                        19:40:41   BIC                            375.217
Sample:                    04-30-2021   HQIC                           371.518
                         - 07-31-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept   3005.0245     64.148     46.846      0.000    2879.298    3130.751
ma.L1          1.0530      0.173      6.096      0.000       0.714       1.392
ma.L2          0.3551      0.165      2.156      0.031       0.032       0.678
sigma2      2.302e+04   8029.656      2.867      0.004    7280.350    3.88e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.08   Jarque-Bera (JB):                 0.75
Prob(Q):                              0.77   Prob(JB):                         0.69
Heteroskedasticity (H):               0.60   Skew:                            -0.20
Prob(H) (two-sided):                  0.45   Kurtosis:                         2.31
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

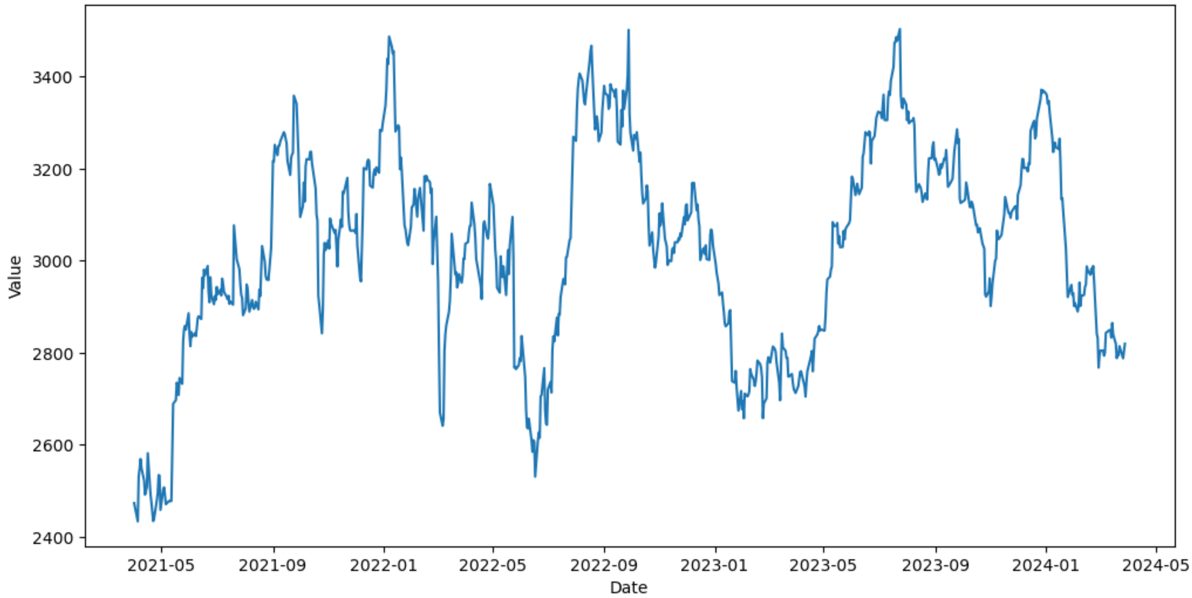


Monthly SARIMAX Model Results Interpretation

1. **Model Specifications**: SARIMAX(0, 0, 2)
   - This model utilizes two moving average components without autoregressive or differencing terms, focusing on the error terms of the two most recent lagged forecasts.
2. **Coefficient and Statistical Significance**:
   - **ma.L1 and ma.L2**: Both coefficients are significant (p-values of 0.000 for ma.L1 and 0.031 for ma.L2), suggesting that the model effectively captures the moving average components of the lagged forecast errors.
3. **Model Fit and Diagnostics**:
   - **AIC/BIC**: Lower AIC and BIC values compared to the daily model suggest a better fit to the monthly data.
   - **Ljung-Box Test**: A Q value of 0.08 and a p-value of 0.77 indicate no significant autocorrelation in residuals, affirming that the model residuals do not exhibit predictable patterns left unmodeled.
   - **Heteroskedasticity**: The absence of heteroskedasticity (p-value of 0.45) confirms that the variance of residuals is stable across the series.

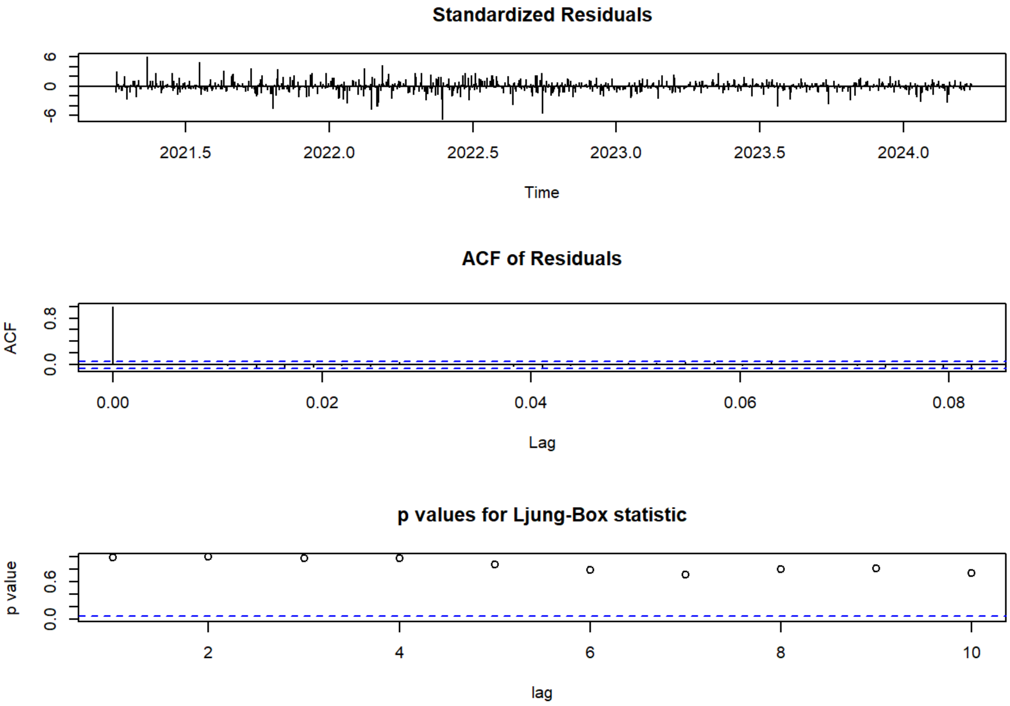
- **Jarque-Bera Test**: A lower JB value and a p-value of 0.69 suggest that the residuals are closer to a normal distribution in the monthly model than in the daily model.

<u>Daily:</u>




```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  740
Model:               SARIMAX(1, 1, 0)   Log Likelihood               -3824.373
Date:                Mon, 22 Jul 2024   AIC                           7652.746
Time:                        19:41:19   BIC                           7661.957
Sample:                             0   HQIC                          7656.298
                                - 740
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0931      0.031      3.009      0.003       0.032       0.154
sigma2      1830.6145     57.133     32.041      0.000    1718.636    1942.593
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               421.20
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               0.41   Skew:                            -0.31
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.65
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

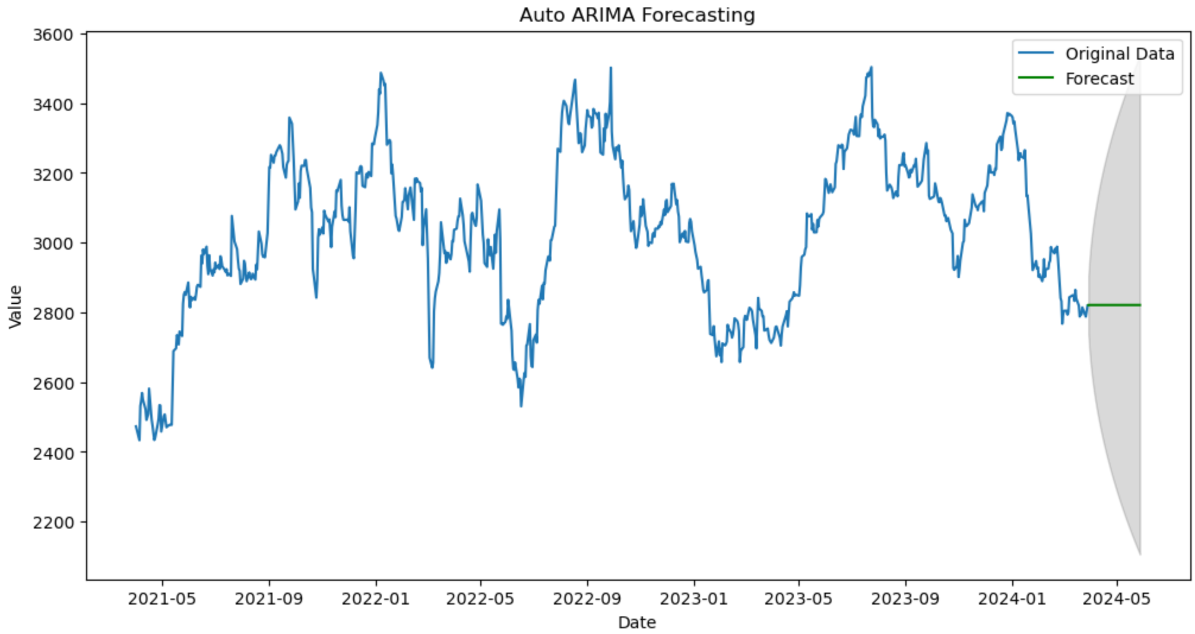


Daily SARIMAX Model Results Interpretation

1. **Model Specifications**: SARIMAX(1, 0, 0)
   - This indicates that the model includes one autoregressive term and no differencing or moving average components, which simplifies the model and focuses on the dependency of the current value on its immediate predecessor.
2. **Coefficient and Statistical Significance**:
   - **ar.L1**: The coefficient is 0.0931 with a p-value of 0.032, indicating that the previous day's stock price has a significant, albeit small, positive effect on the current day's stock price.
3. **Model Fit and Diagnostics**:
   - **AIC/BIC**: The AIC and BIC values are relatively high, suggesting that the model may not be the best fit for the data or could be improved with additional terms or different parameters.
   - **Ljung-Box Test**: A Q value close to 0.00 with a p-value of 0.95 suggests that there is no significant autocorrelation in the residuals, indicating a good fit of the model in capturing time-series patterns.
   - **Heteroskedasticity**: The test indicates no heteroskedasticity with a p-value of 0.00, suggesting constant variance of residuals over time.

- **Jarque-Bera Test**: A high JB value and a p-value of 0.00 imply that the residuals are not normally distributed, which could affect confidence interval estimates and hypothesis tests.

The daily SARIMAX model shows some efficacy in modeling stock prices but could be improved given the non-normal distribution of residuals and higher information criteria values. In contrast, the monthly SARIMAX model appears to fit the data better, with more stable residuals and lower AIC/BIC values, suggesting it might be more effective for forecasting at a monthly level. The significant moving average terms in the monthly model highlight the importance of recent errors in predicting future values, which could be particularly useful for capturing short-term shocks or anomalies in stock price movements.



Multivariate Forecasting:

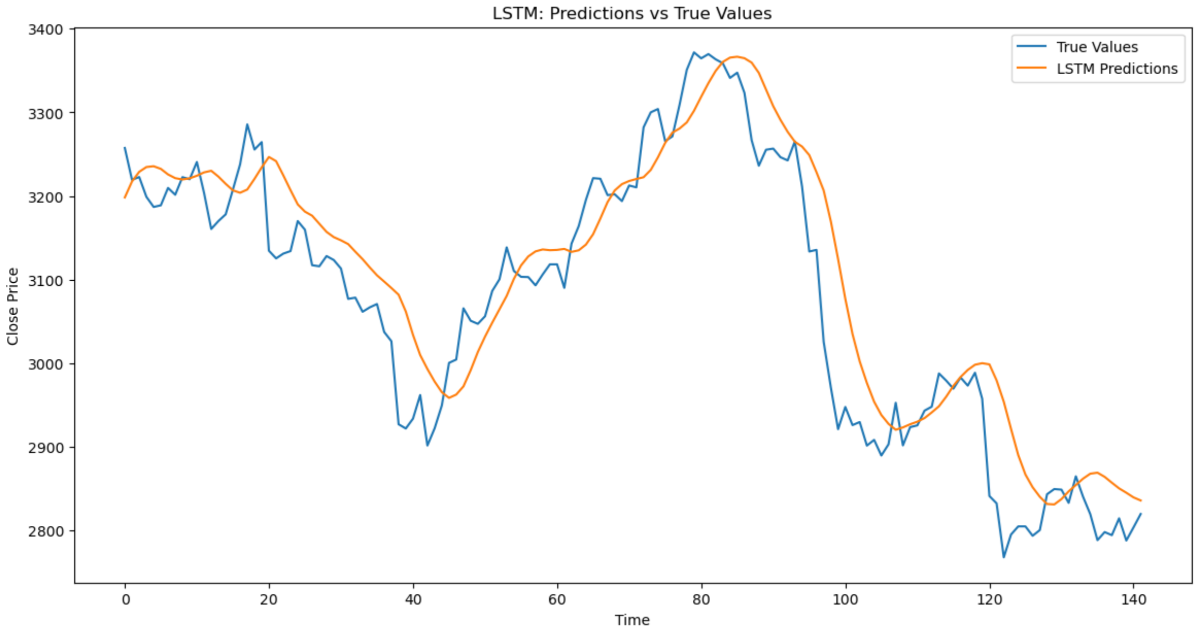
- **LSTM Model**: The model showed high effectiveness in capturing complex patterns in the data, making it suitable for short-term trading strategies.

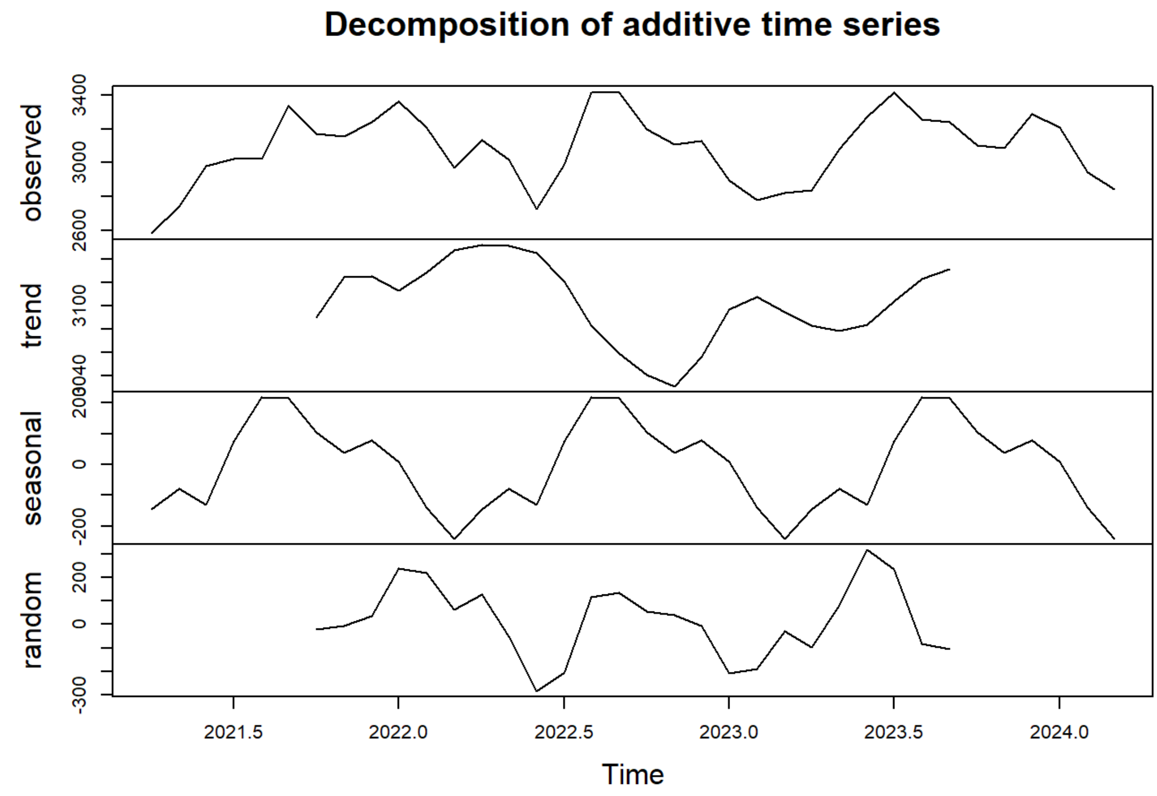
```
model.summary()
```

Model: "sequential"

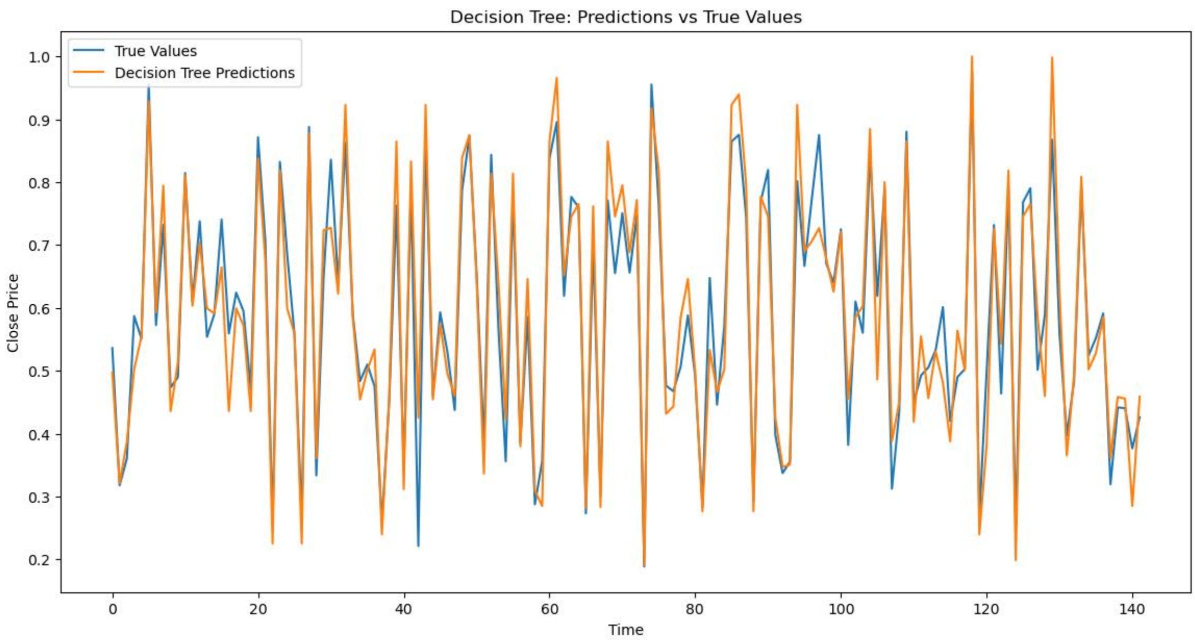
Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 30, 50)	11,400
dropout (Dropout)	(None, 30, 50)	0
lstm_1 (LSTM)	(None, 50)	20,200
dropout_1 (Dropout)	(None, 50)	0
dense (Dense)	(None, 1)	51

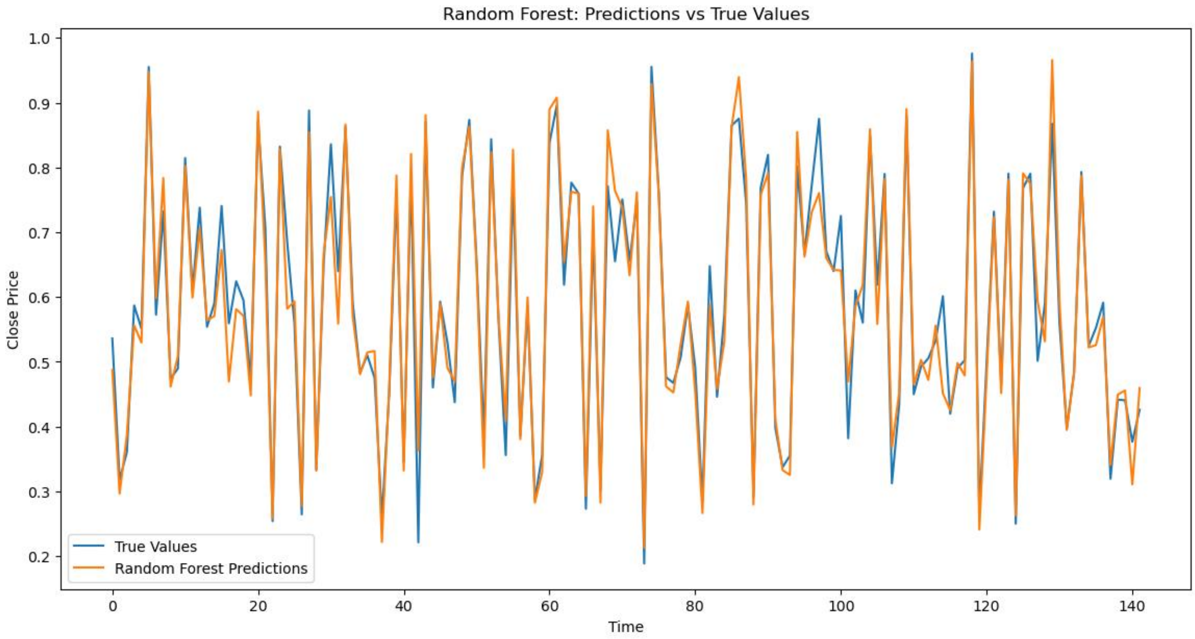
Total params: 31,651 (123.64 KB)
Trainable params: 31,651 (123.64 KB)
Non-trainable params: 0 (0.00 B)

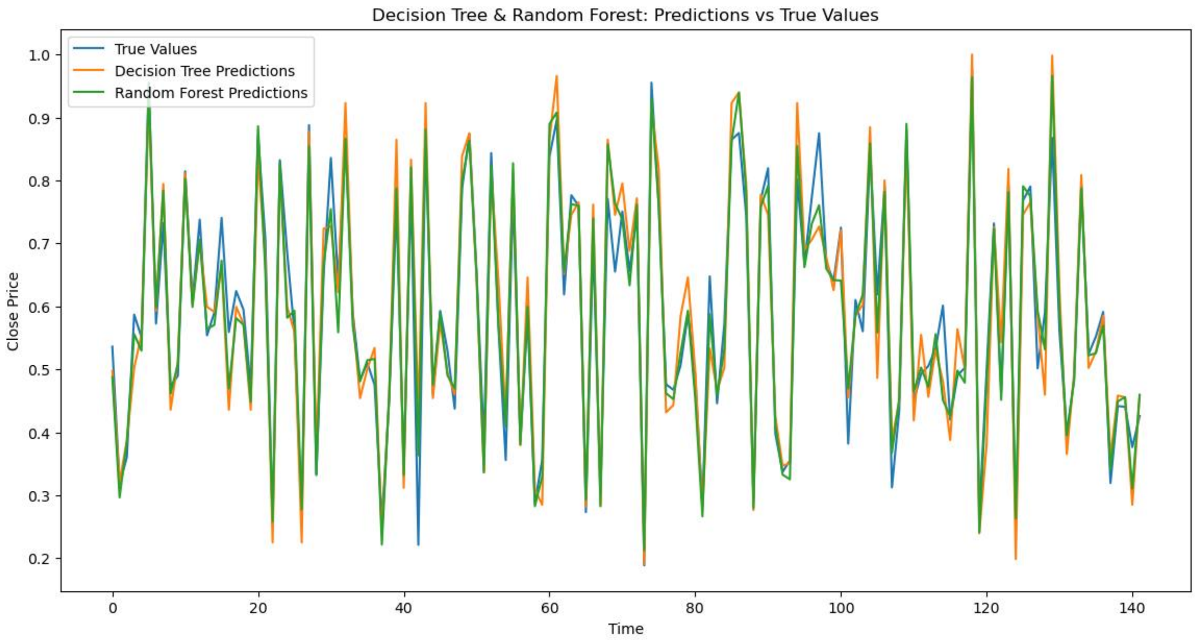




- **Decision Tree and Random Forest Models**: These models were utilized to understand the nonlinear relationships in the data. Random Forest outperformed Decision Tree, suggesting its robustness in handling stock price data variability.







The attached chart displays the performance of Decision Tree and Random Forest models in predicting the closing prices of a stock, plotted alongside the actual (true) closing prices. Here's an interpretation of the chart:

1. **True Values**: Represented in blue, these are the actual observed stock closing prices over a specified period. This line serves as the benchmark for evaluating the predictive accuracy of the models.

2. **Decision Tree Predictions**: Shown in orange, these predictions appear to capture some of the general movements of the stock prices but with notable variance and occasional sharp deviations from the true values. Decision Trees typically make predictions based on the values from the training dataset, which can lead to overfitting or high variance in predictions, especially visible in the sharp peaks and troughs that do not always align with the true values.
3. **Random Forest Predictions**: Depicted in green, these predictions are generally closer to the true values compared to the Decision Tree predictions. Random Forest, an ensemble method that averages multiple decision trees to improve predictive accuracy and control overfitting, seems to smooth out some of the volatility seen in the Decision Tree predictions. The model better captures the trends and reduces the variance, leading to a line that more closely follows the true stock prices.

Key Observations:

- **Model Comparison**: The Random Forest model outperforms the Decision Tree in terms of closeness to the true values, showcasing its strength in handling noise and making more generalized predictions.
- **Performance Issues**: Both models show areas where predictions deviate from the actual values, indicating potential issues such as overfitting in the case of the Decision Tree or possibly underfitting where both models fail to capture the true trends accurately.
- **Trend Following**: Both models are capable of following the general trends in the data, but the Random Forest appears more robust in capturing the finer movements without as much noise as the Decision Tree.

Implication for Stock Trading in Strategy Formulation: Traders using these models should consider the Random Forest model for its superior accuracy and stability. However, awareness of where the model fails to track the true prices accurately is crucial for risk management.

RECOMMENDATIONS

Based on the findings of the time series analysis, the following recommendations are provided for stock traders:

1. **Leverage Model Insights**: Utilize LSTM for short-term trading decisions due to its sensitivity to recent trends, and ARIMA for long-term strategies based on its ability to forecast broader trends.
2. **Diversify Trading Strategies**: Employ a combination of models to diversify risk. The Random Forest model, given its high accuracy, should be part of the strategic toolkit for managing larger, diversified portfolios.
3. **Risk Management**: Implement risk management strategies such as stop-loss orders and diversification to mitigate potential losses, particularly in volatile trading periods as suggested by the models.
4. **Continual Model Evaluation and Update**: Regularly update the forecasting models with new data and recalibrate them to adapt to changing market dynamics. This ensures the models remain relevant and accurate.
5. **Integration of Technical Analysis**: Supplement model predictions with technical analysis tools to refine trading decisions. Indicators such as Bollinger Bands, RSI, and MACD can provide additional support in determining entry and exit points.