VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A6a: Time Series Analysis

Lakshmy Cherussery Jayaprakash

V01110023

Date of Submission: 26-07-2024

CONTENTS

Introduction

This assignment focuses on time series forecasting and machine learning models applied to stock price data. The primary goal is to leverage various statistical and machine learning techniques to predict future stock prices accurately. We will explore both univariate and multivariate forecasting methods, starting with traditional statistical models and advancing to sophisticated machine learning algorithms.

Objective

Data Cleaning and Preparation:

- Download stock price data of Wipro Limited (WIT) from 2021-06-21 to 2024-07-20
- Split the data into training and test sets.

Time Series Decomposition:

- Convert the data to a monthly frequency.
- Decompose the time series using additive and multiplicative models to understand its components.

Univariate Forecasting:

- Fit a Holt-Winters model to the monthly data and forecast for the next year.
- Fit an ARIMA model to the daily data and perform diagnostic checks.
- Evaluate whether a Seasonal-ARIMA (SARIMA) model provides a better fit.
- Forecast the monthly series using ARIMA.

Multivariate Forecasting:

- Apply Long Short-term Memory (LSTM) neural networks to the data.
- Use tree-based models such as Random Forest and Decision Tree for forecasting

Summary of Models and Statistical Tools

- **Holt-Winters Model**

Holt-Winters exponential smoothing is a time series forecasting method for data with seasonality. It considers three components: level, trend, and seasonality. Suitable for forecasting data with clear seasonal patterns.

- **ARIMA (AutoRegressive Integrated Moving Average)**

Combining autoregressive (AR) terms, differencing (I), and moving average (MA) terms. Useful for data without a strong seasonal component but can be extended to SARIMA for seasonal data. Effective for capturing the underlying patterns in time series data without seasonality.

- **SARIMA (Seasonal ARIMA)**

SARIMA extends ARIMA by adding seasonal components. It accounts for seasonality, trend, and noise in time series data. Ideal for time series data with seasonal patterns. Provides forecasts that account for both seasonal and non-seasonal factors.

- **LSTM (Long Short-Term Memory)**

LSTM is a type of recurrent neural network (RNN) capable of learning long-term dependencies in sequence prediction problems. Suitable for complex time series data with long-term dependencies.

- **Decision Tree**

A decision tree is a flowchart-like model used for classification and regression tasks. It splits data into subsets based on feature values. Simple and interpretable model for both classification and regression tasks. Provides a piecewise constant approximation to the target function.

- **Random Forest**

An ensemble method that constructs multiple decision trees and merges their predictions for more accurate and stable forecasts. Reduces overfitting and improves predictive performance compared to a single decision tree.

Statistical Tools

- **RMSE (Root Mean Square Error)**

  Measures the average magnitude of the error between predicted and actual values, penalizing larger errors more significantly.

- **MAE (Mean Absolute Error)**

Measures the average absolute difference between predicted and actual values, treating all errors equally.

- **MAPE (Mean Absolute Percentage Error)**

Measures the average absolute percentage error between predicted and actual values, providing a normalized measure of prediction accuracy as a percentage.

- **R-squared (Coefficient of Determination)**

Indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s). Values range from 0 to 1, with higher values indicating better model fit.

Business Significance

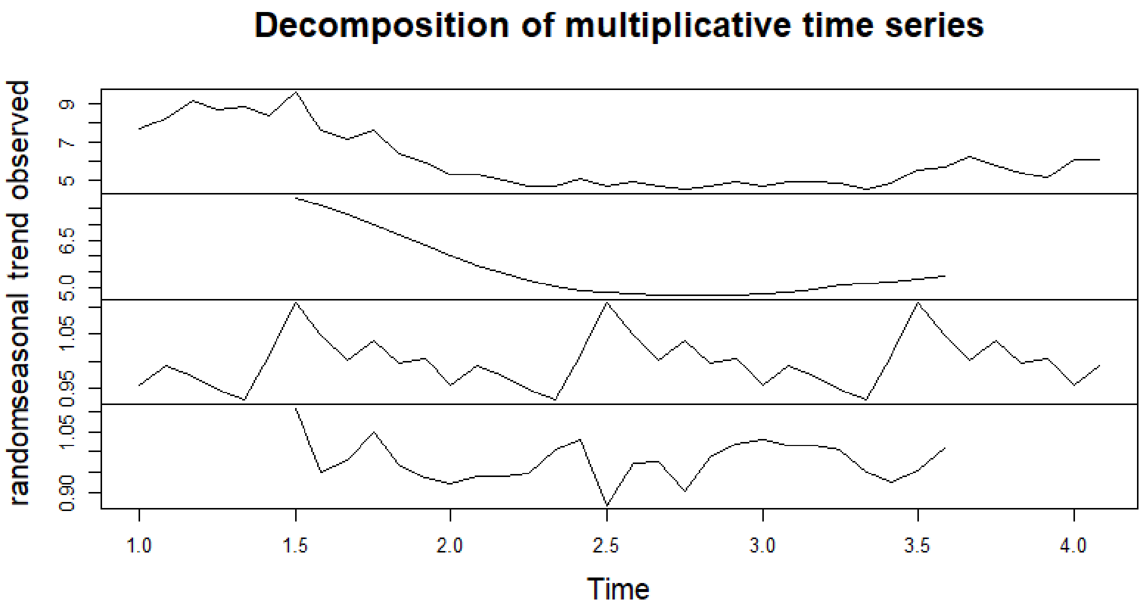
Accurate stock price forecasting is crucial for investors, financial analysts, and portfolio managers to make informed decisions. By predicting future stock prices, stakeholders can:

- Optimize investment strategies.
- Manage risk more effectively.
- Improve portfolio performance.
- Make timely buy/sell decisions to maximize returns.

Results & Interpretations- R

Part A

Univariate Forecasting - Conventional Models/Statistical Models

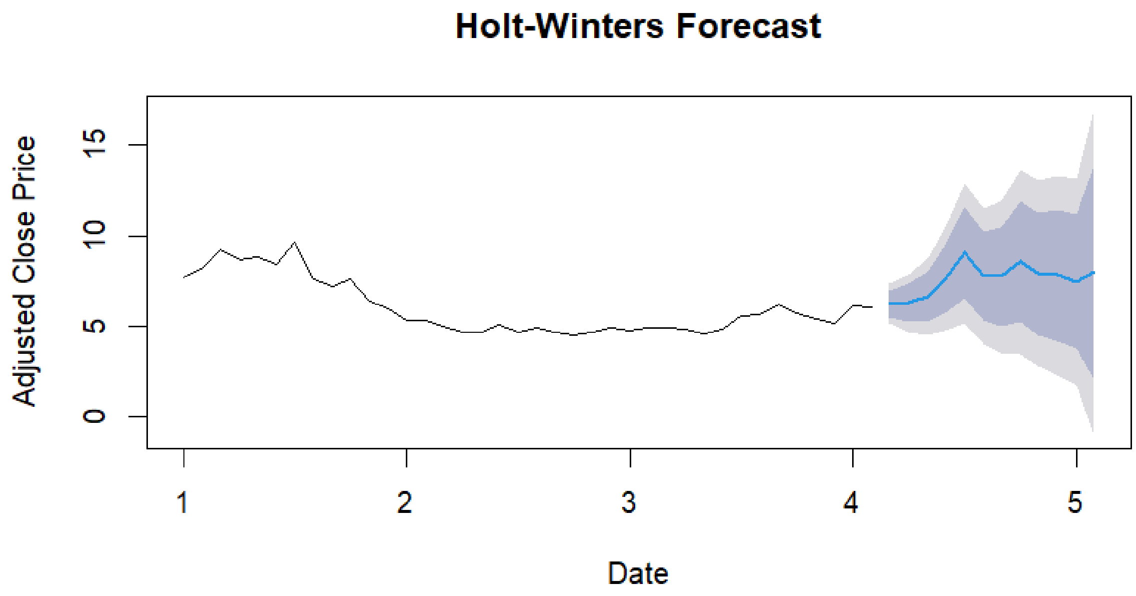


<u>Holt-Winters</u>

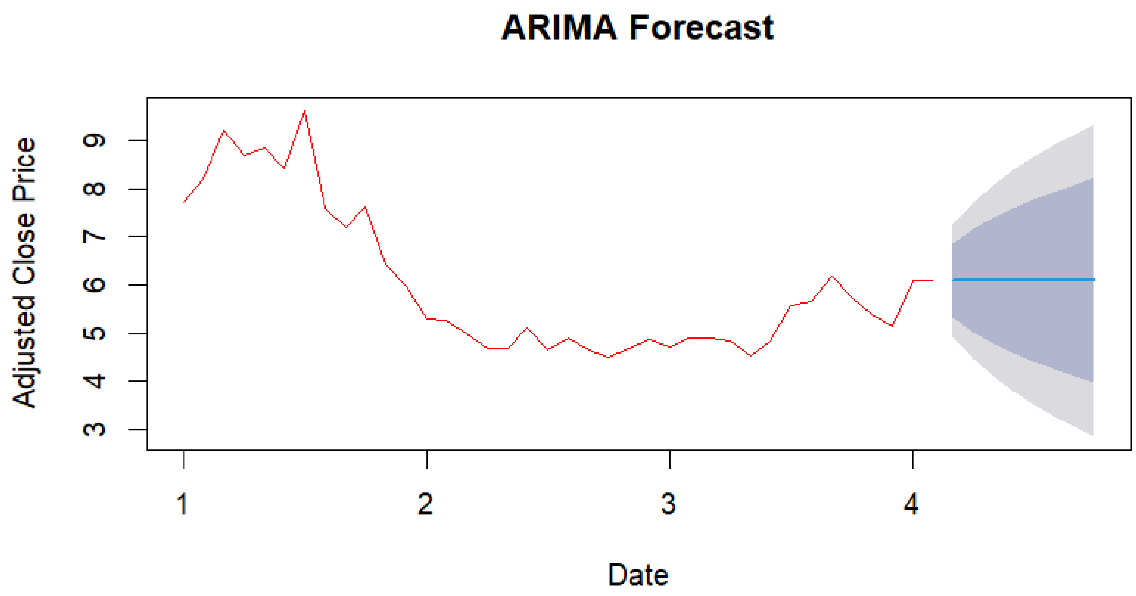
RMSE: 2.79091511252183

MAE: 2.51834166336209

MAPE: 43.4998082178039



<u>ARIMA</u>



Series: train_data

ARIMA(0,1,0)

sigma^2 = 0.368: log likelihood = -26.66

AIC=55.31   AICc=55.46   BIC=56.68

Training set error measures:

ME   RMSE   MAE   MPE

Training set -0.0949777 0.5964657 0.413488 -1.878373

MAPE   MASE   ACF1

Training set 6.378977 0.1917 -0.259705

forecast_values

Point Forecast   Lo 80   Hi 80   Lo 95

Jul 3   4.839878 4.062409 5.617347 3.6508414

Aug 3   4.839878 3.740370 5.939386 3.1583263

Sep 3   4.839878 3.493262 6.186494 2.7804062

Oct 3   4.839878 3.284940 6.394816 2.4618048

Nov 3   4.839878 3.101404 6.578352 2.1811113

Dec 3   4.839878 2.935475 6.744281 1.9273450

Jan 4   4.839878 2.782888 6.896868 1.6939828

Feb 4   4.839878 2.640863 7.038893 1.4767746

Mar 4   4.839878 2.507470 7.172286 1.2727682

Apr 4   4.839878 2.381304 7.298452 1.0798141

May 4   4.839878 2.261304 7.418452 0.8962897

Jun 4   4.839878 2.146646 7.533110 0.7209344

Hi 95

Jul 3 6.028915

Aug 3 6.521430

Sep 3 6.899350

Oct 3 7.217951

Nov 3 7.498645

Dec 3 7.752411

Jan 4 7.985773

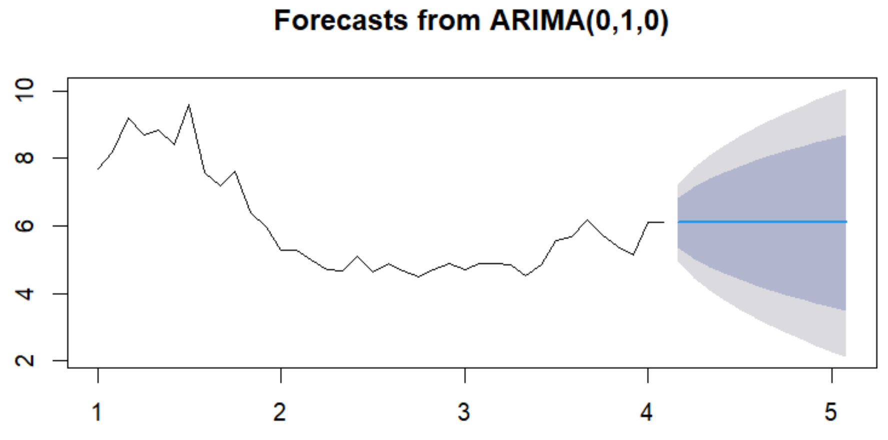
Feb 4 8.202981

Mar 4 8.406988

Apr 4 8.599942

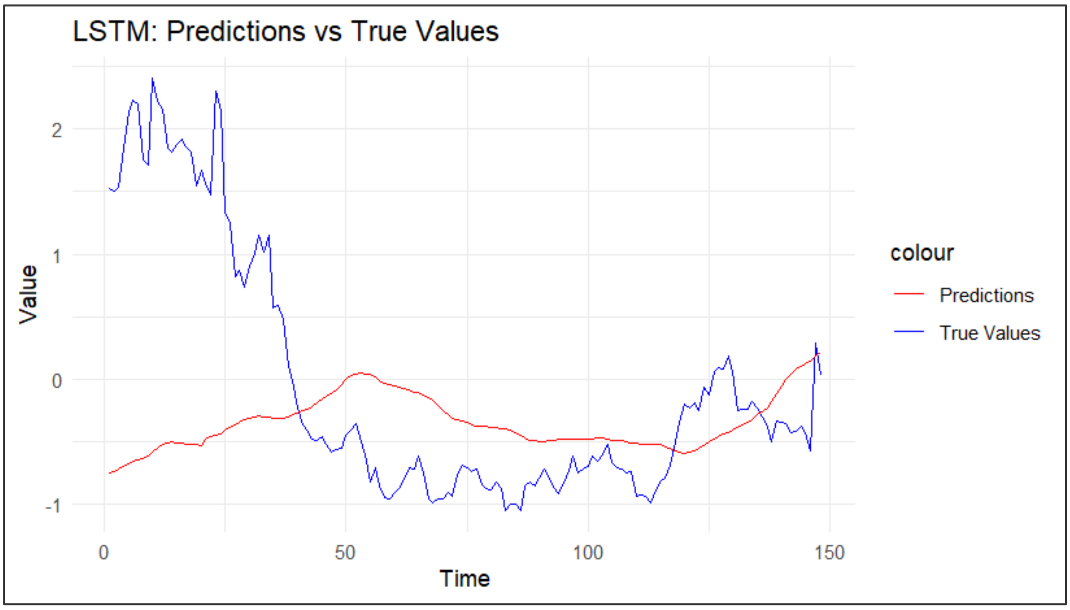
May 4 8.783466

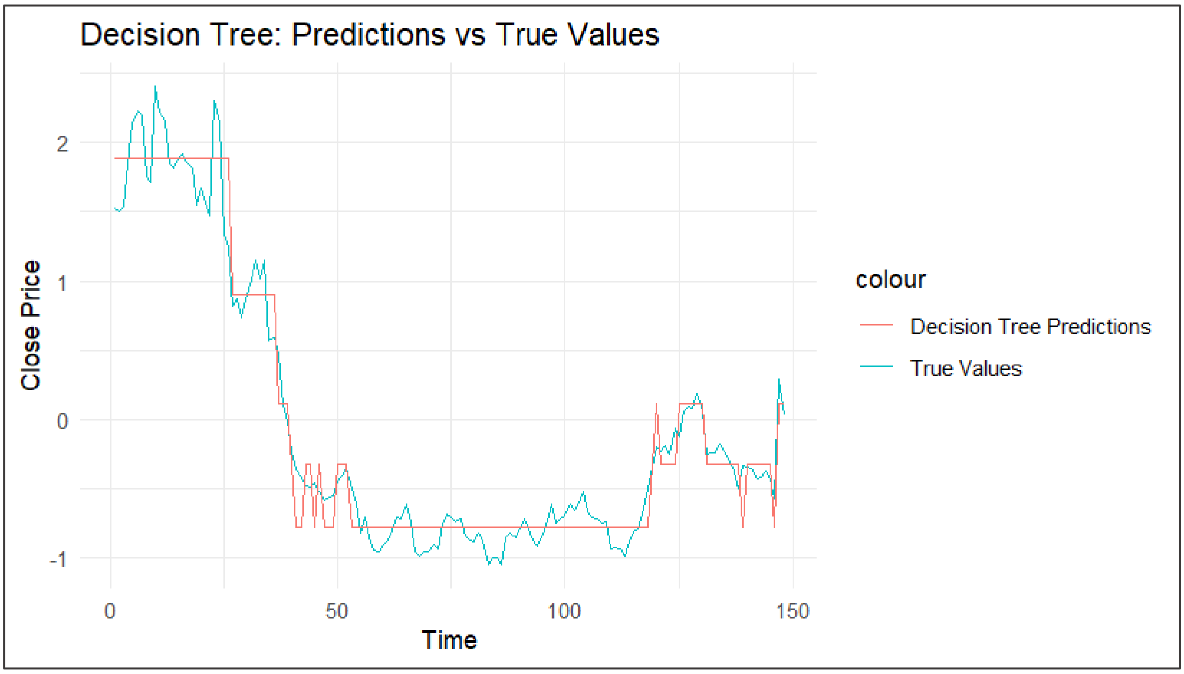
Jun 4 8.958822

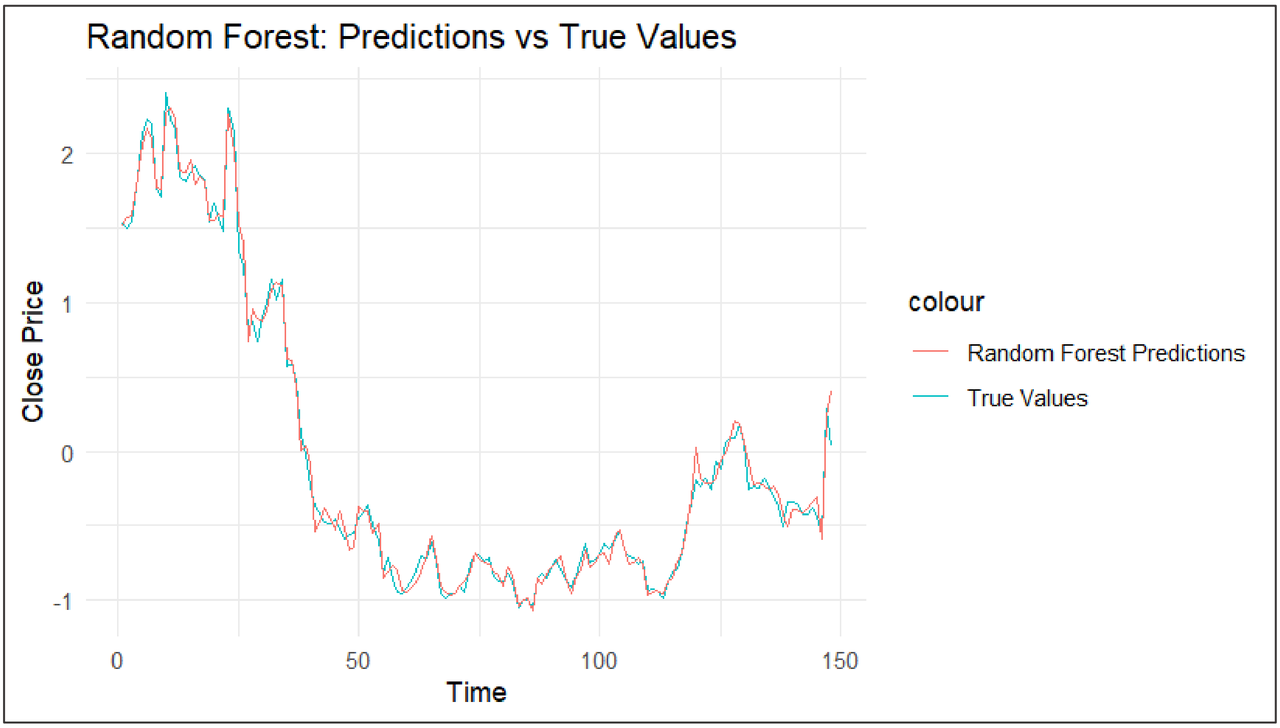


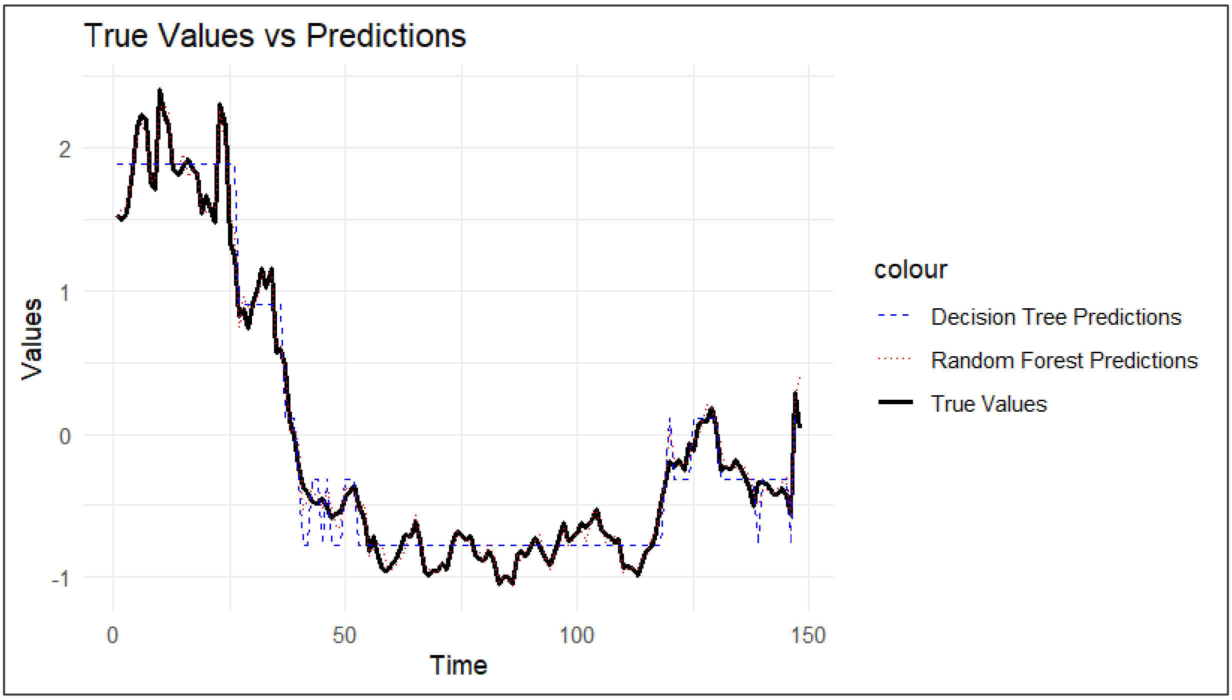
Part B

Multivariate Forecasting - Machine Learning Models




Decision Tree: Predictions vs True Values
Close Price
2
1
0
-1
0
50
100
150
Time
colour
Decision Tree Predictions
True Values

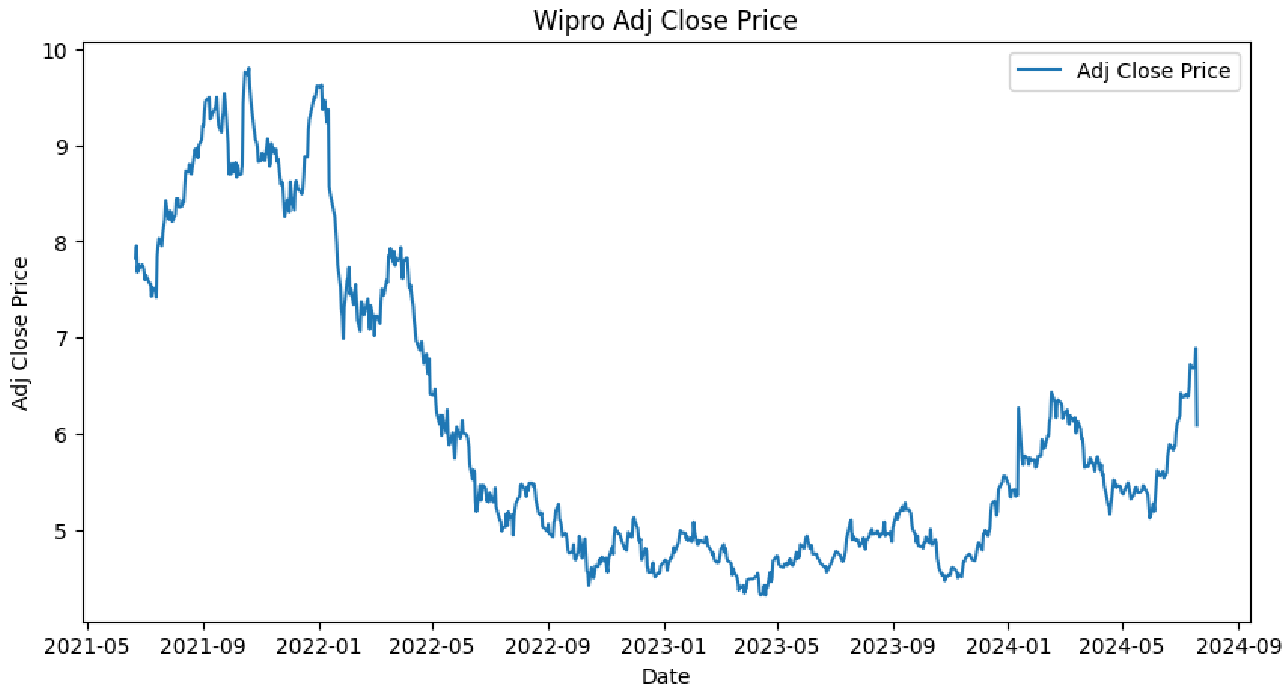

Random Forest: Predictions vs True Values
Close Price
2
1
0
-1
0
50
100
150
Time
colour
Random Forest Predictions
True Values


True Values vs Predictions
Values
Time
colour
Decision Tree Predictions
Random Forest Predictions
True Values
2
1
0
-1
0
50
100
150

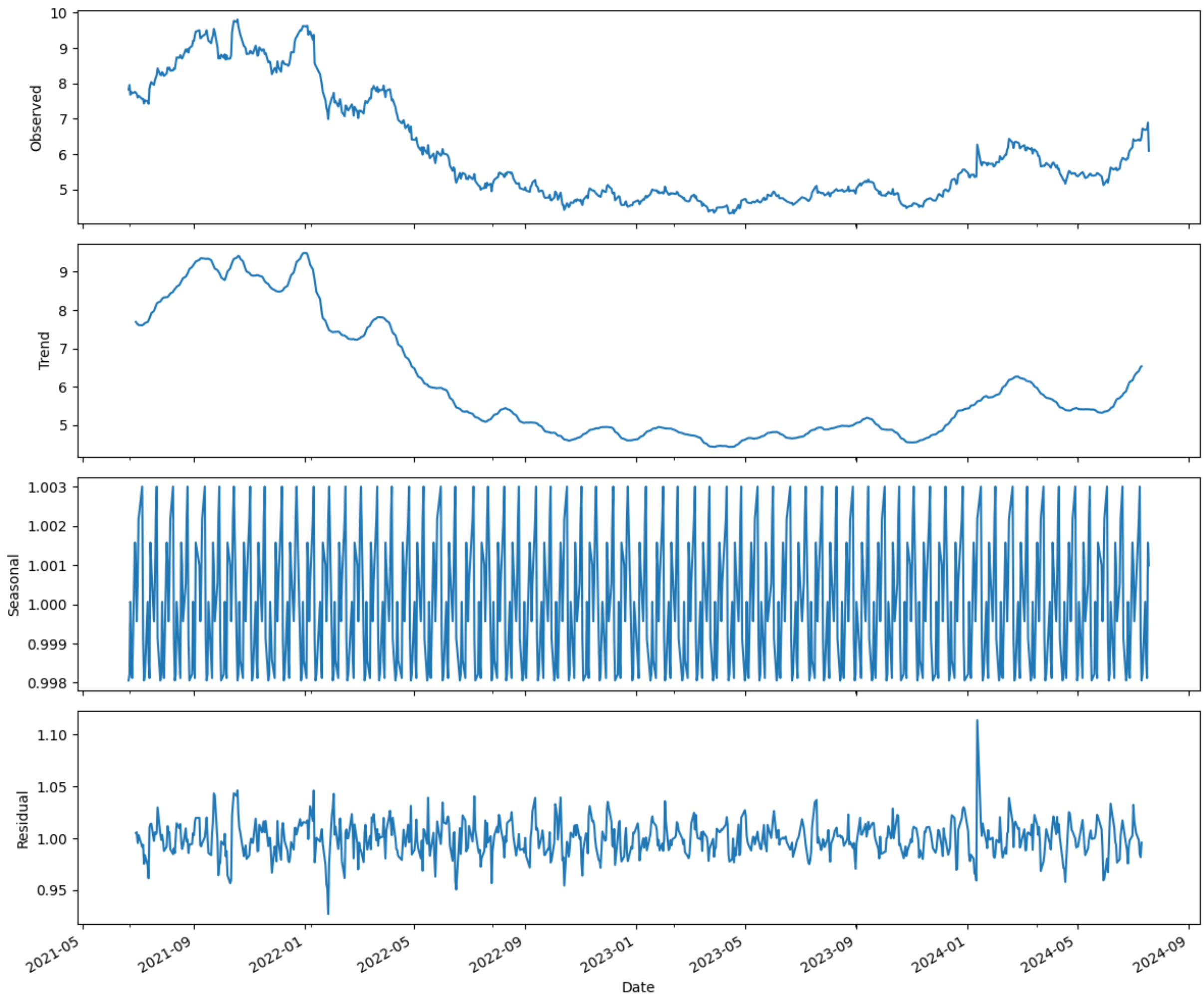
Results & Interpretations- Python

Part A

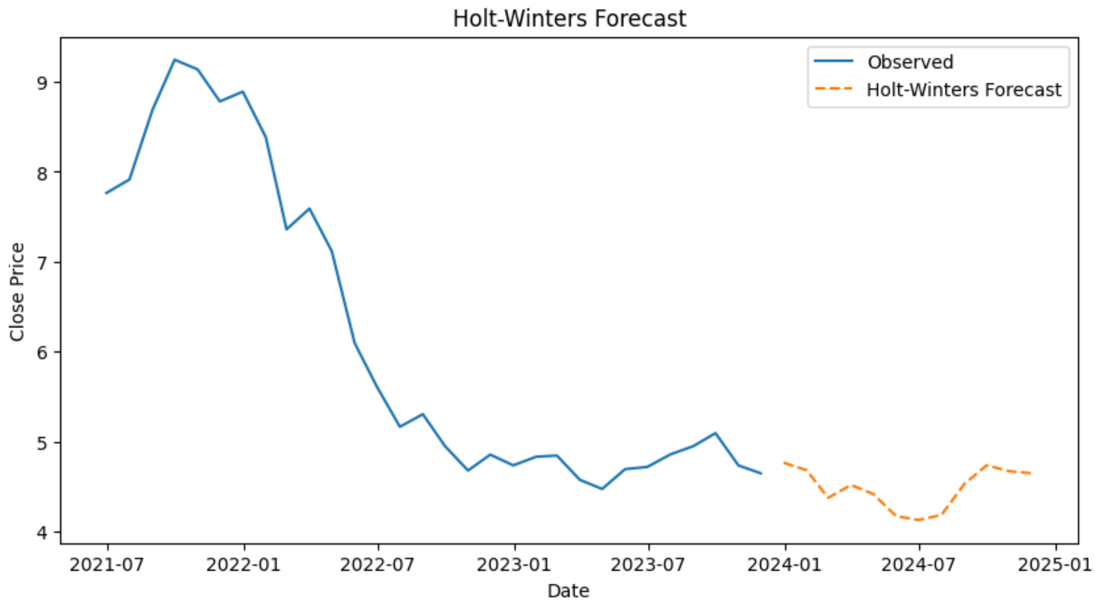
Univariate Forecasting - Conventional Models/Statistical Models



<u>Decompoed Time series</u>



Holt Winters model



RMSE: 1.4310576307821365

MAE: 1.329078487794442

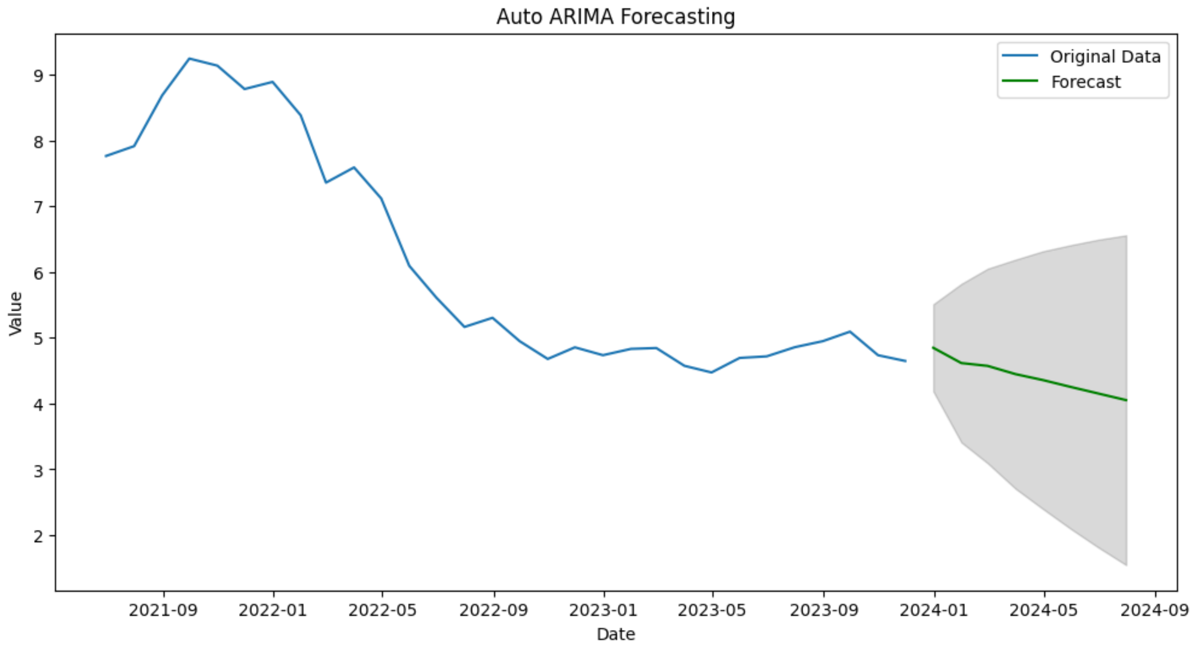
MAPE: nan

R-squared: -11.794681430910016

<u>ARIMA</u>

SARIMAX Results

```
==============================================================================
Dep. Variable:                 y   No. Observations:                 30
Model:            SARIMAX(1, 1, 1)   Log Likelihood              -10.445
Date:            Thu, 25 Jul 2024   AIC                          28.891
Time:                    18:22:04   BIC                          34.360
Sample:                06-30-2021   HQIC                         30.604
                     - 11-30-2023
Covariance Type:              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -0.1445      0.128     -1.130      0.259      -0.395      0.106
ar.L1         -0.4382      0.214     -2.050      0.040      -0.857     -0.019
ma.L1          0.9640      0.521      1.851      0.064      -0.057      1.985
sigma2         0.1136      0.057      1.994      0.046       0.002      0.225
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 0.75
Prob(Q):                              0.89   Prob(JB):                         0.69
Heteroskedasticity (H):               0.27   Skew:                            -0.21
Prob(H) (two-sided):                  0.05   Kurtosis:                         2.34
===================================================================================
```

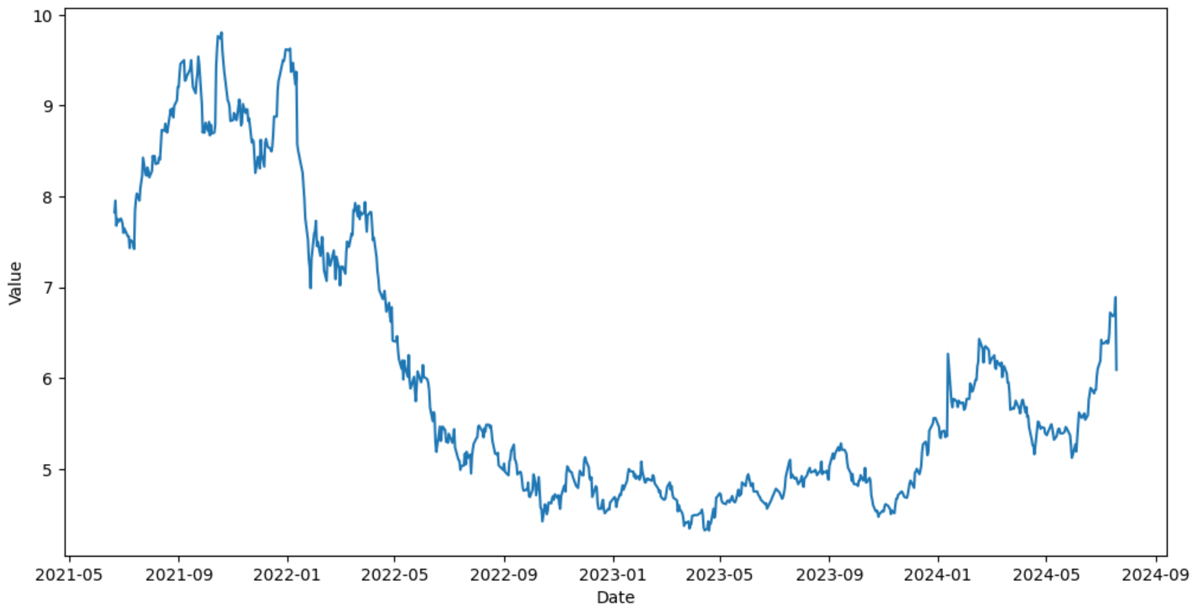


RMSE: 1.4366380194613422

MAE: 1.3214077560424142

MAPE: nan

R-squared: -11.894661339665367



SARIMAX Results

==

Dep. Variable: y No. Observations: 775

Model: SARIMAX(0, 1, 0) Log Likelihood 553.754

Date: Thu, 25 Jul 2024 AIC -1105.509

Time: 18:27:04 BIC -1100.857

Sample: 0 HQIC -1103.719

- 775

Covariance Type: opg

==

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	0.0140	0.000	49.931	0.000	0.013	0.015

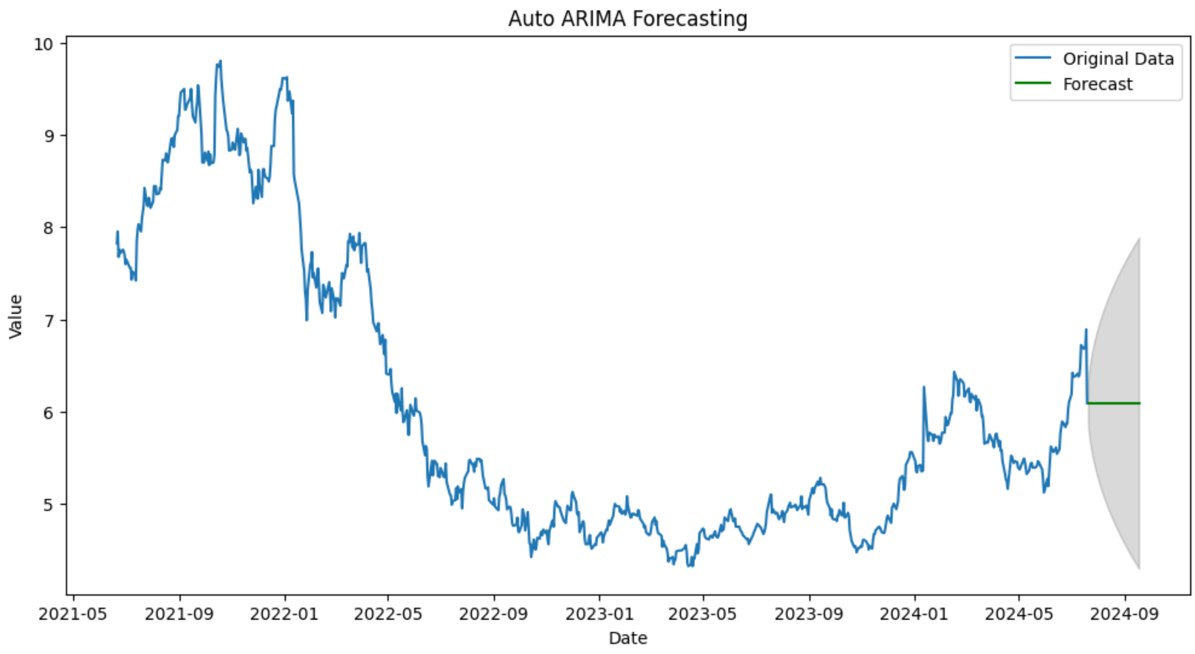
===

Ljung-Box (L1) (Q): 2.53 Jarque-Bera (JB): 3821.41

Prob(Q): 0.11 Prob(JB): 0.00

Heteroskedasticity (H): 0.54 Skew: -0.09

Prob(H) (two-sided): 0.00 Kurtosis: 13.88



Part B

Multivariate Forecasting - Machine Learning Models

Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 30, 50)	11400
dropout (Dropout)	(None, 30, 50)	0
lstm_1 (LSTM)	(None, 50)	20200
dropout_1 (Dropout)	(None, 50)	0
dense (Dense)	(None, 1)	51

Total params: 31651 (123.64 KB)

Trainable params: 31651 (123.64 KB)

Non-trainable params: 0 (0.00 Byte)

Predictions vs True Values:

Prediction: 4.786262284488004, True Value: 5.18914794921875

Prediction: 4.812307136515379, True Value: 5.268980979919434

Prediction: 4.849132705721722, True Value: 5.2989182472229

Prediction: 4.891068905197912, True Value: 5.249022960662842

Prediction: 4.929000719197909, True Value: 5.149231433868408

Prediction: 4.955670495325194, True Value: 5.18914794921875

Prediction: 4.9744693009228556, True Value: 5.418667793273926

Prediction: 5.0048299379582915, True Value: 5.4985013008117685

Prediction: 5.045570068996334, True Value: 5.558375835418701

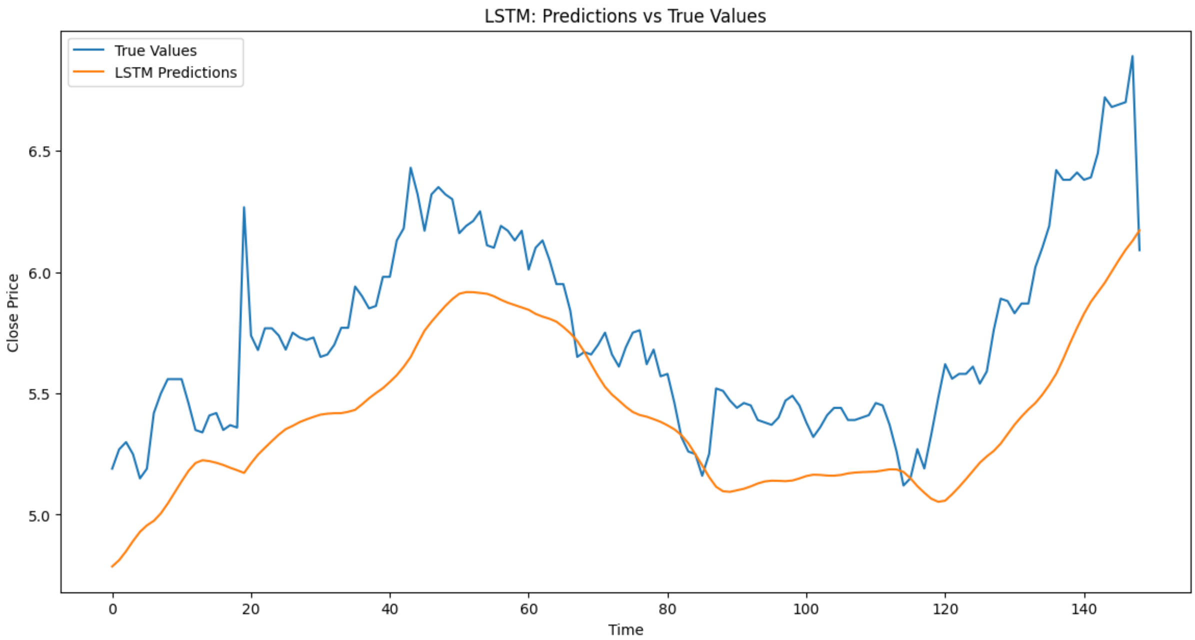
Prediction: 5.090864397754892, True Value: 5.558375835418701

RMSE: 0.3919427264693785

MAE: 0.34329631692921836

MAPE: 6.337672937465531

R-squared: 0.015139553147401341



Tree Based Models

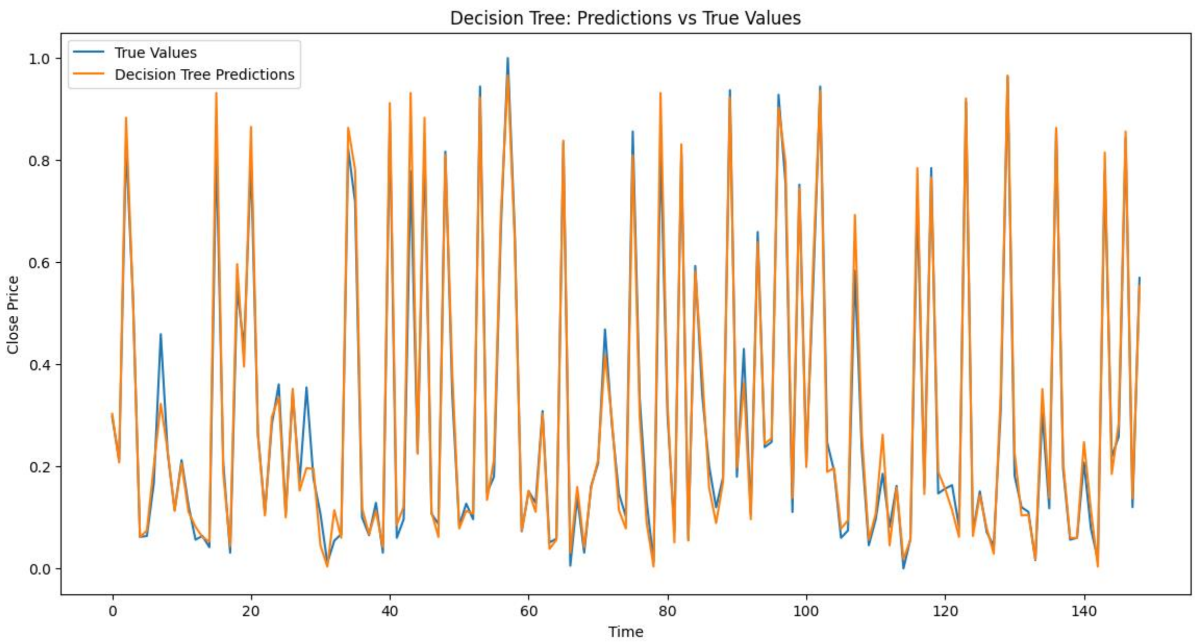
- Decision Tree

  MSE (Decision Tree): 0.001380986569392743

  RMSE: 0.03716162764724849
  MAE: 0.025027981113996005

MAPE: 7273.613917038064
R-squared: 0.9835931984596074

Decision Tree Predictions vs True Values:
Prediction: 0.3030239040496213, True Value: 0.29757887202735966
Prediction: 0.20766573016833578, True Value: 0.21227444254992645
Prediction: 0.8832103094391363, True Value: 0.8311039445337512
Prediction: 0.526130250785484, True Value: 0.529729989596427
Prediction: 0.06163023735394724, True Value: 0.06185844307710331
Prediction: 0.07459397420397273, True Value: 0.0636777425698607
Prediction: 0.20138437850540292, True Value: 0.16689966833232206
Prediction: 0.3225254523276505, True Value: 0.45911296739683516
Prediction: 0.22954374764567753, True Value: 0.22560130904018927
Prediction: 0.11280065452009458, True Value: 0.11426482243986424



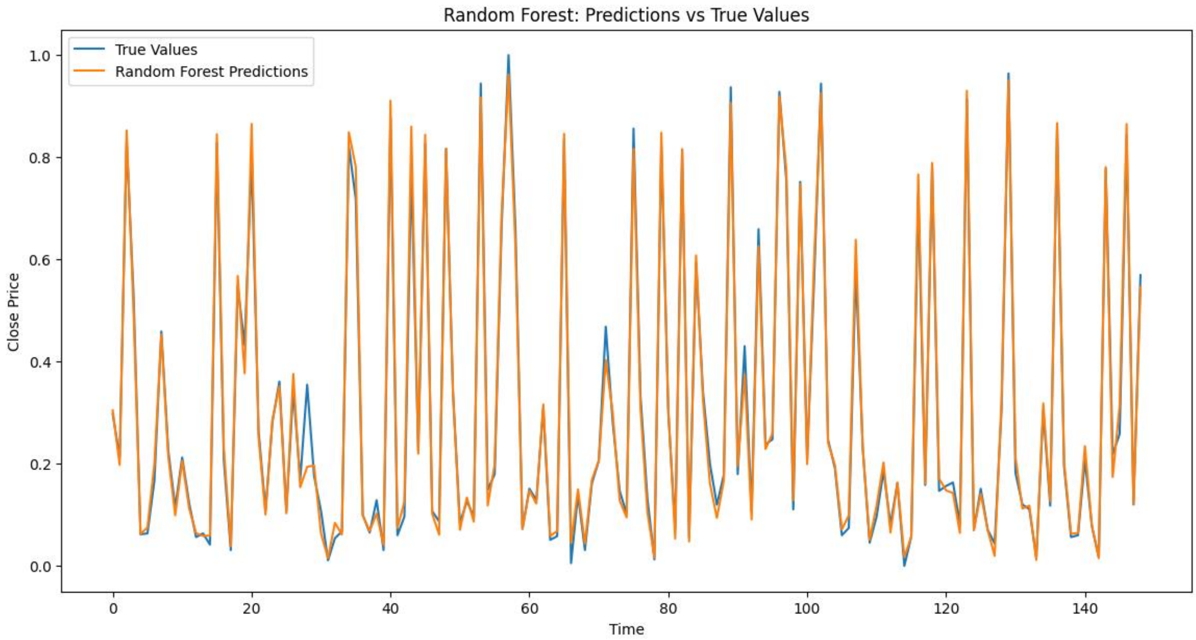
- <u>Random Forest</u>

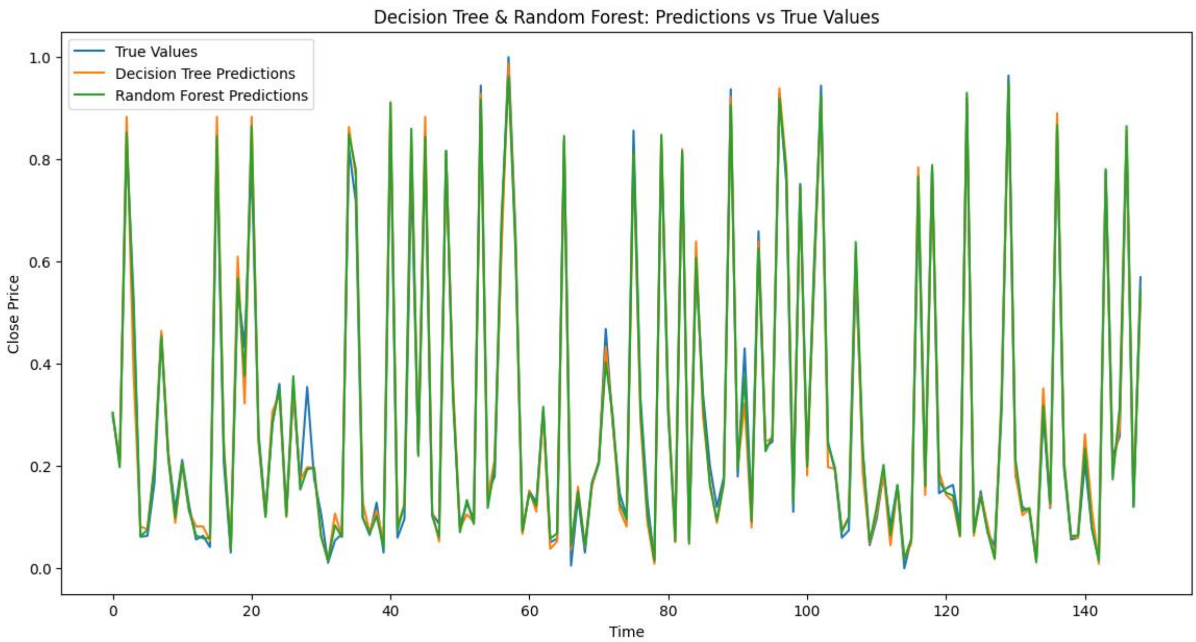
Random Forest Mean Squared Error: 0.0006889511731164984

RMSE: 0.026247879402277404
MAE: 0.018068252366894782
MAPE: 5044.562557345967
R-squared: 0.9918149202759347

Random Forest Predictions vs True Values:
Prediction: 0.30472012116026576, True Value: 0.29757882028702465
Prediction: 0.19765283543171006, True Value: 0.21227440564155753
Prediction: 0.8525572485471125, True Value: 0.831103626157883

Prediction: 0.497548386777707597, True Value: 0.5297299844272443
Prediction: 0.06273833258235566, True Value: 0.061858432321714285
Prediction: 0.07520491870618078, True Value: 0.06367773149814815
Prediction: 0.19998501753619408, True Value: 0.16689963931331042
Prediction: 0.4530938379162397, True Value: 0.4591128875704058
Prediction: 0.21455480461637932, True Value: 0.22560126981466477
Prediction: 0.09894463260667145, True Value: 0.1142648025725258





Recommendations

For future work, focus on improving data quality and incorporating additional relevant features to enhance model performance. Perform extensive hyperparameter tuning and explore advanced and hybrid models. Implement time series cross-validation and real-time forecasting capabilities while continuously monitoring and retraining models. Conduct scenario and sensitivity analyses to understand key drivers and impacts.