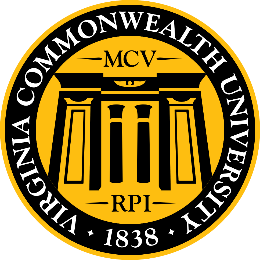
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A6a: Time Series Analysis**

**KATHRYN SHAJU**

**V01111327**

**Date of Submission: 22-07-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Objectives | **2** |
| **3.** | Business Significance | **3** |
| **4.** | Results & Interpretations | **4-12** |

**INTRODUCTION**

In this study we focus on conducting a time series analysis on Infosys Ltd data on its stock prices for the time period, 2021-04-01 to 2024-06-30. We conduct the time series in both R and Python to determine exact results.

Time series analysis is the process of examining time-ordered data points in order to derive significant statistics and pinpoint the features of the data. It is extensively utilized in many disciplines, including as engineering, environmental science, finance, and economics.

Along with this we have also done Univariate and Multivariate Forecasting using different models within these to accurately determine the results.

Analyzing a single time-dependent variable is known as univariate time series analysis. Forecasting and comprehending the underlying structure are frequently the objectives.

For this purpose we have made use of the Holt Winters model to forecast for the next year.  
A forecasting technique for time series data, the Holt-Winters model, also called the Holt-Winters Exponential Smoothing, builds on basic exponential smoothing to incorporate trend and seasonality. It works especially well with data that exhibit a distinct seasonal rhythm.

We are also fitting a ARIMA model to the daily data and check validity of the model.

The ARIMA (AutoRegressive Integrated Moving Average) model is a kind of time series forecasting models. It blends three elements together: Dependencies between an observation and several lagged observations are used in the autoregression (AR) model.  
Integration (I): This is the process of making the data steady by differencing it. Utilizing dependencies between an observation and a residual error from a moving average model applied to lagged observations, moving average (MA) is a statistical technique.  
The formula for the ARIMA model is ARIMA(p, d, q), where:  
The number of lag observations in the model (AR portion) is represented by p.  
The number of times the raw observations are differed (I portion) is denoted by 𝑑d. We have also fitted ARIMA model for monthly series as well.

Similarly we have also fitted a SARIMA model. The SARIMA (Seasonal ARIMA) model extends ARIMA by explicitly modeling seasonal effects.

The Multivariate analysis addresses a number of time-dependent variables. The correlations among these variables can enhance predicting accuracy and yield more insightful information.

Under this we have fitted the LSTM(long short term memory). Recurrent neural networks (RNNs) of the LSTM type are made especially to handle sequential data and long-term dependencies. Because of the vanishing gradient problem—where gradients get too small for the network to learn effectively—traditional RNNs have trouble handling long-term dependence. LSTMs use a unique architecture that incorporates gating mechanisms and memory cells to handle this.

We have also made use of Tree models for our study such as Decision Tree and Radom Forest.

Useful for problems involving regression and classification, decision trees resemble flowcharts. Based on feature values, it divides the data into subsets and branches the data until a predetermined condition is satisfied (e.g., all samples belong to the same class). A feature is represented by an internal node, a decision rule by a branch, and an outcome is represented by a leaf node.

Several decision trees are combined in random forests, an ensemble learning technique, to increase prediction accuracy and reduce overfitting. In a random forest, every tree utilizes a random collection of characteristics for splitting and is trained on a separate sample of the data (bootstrapping)

**OBJECTIVES**

* 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**

The aforementioned models can be used to analyze Infosys stock price data in order to estimate future prices, comprehend trends and seasonality, and find factors that influence stock price movements, among other benefits.

LSTM is an excellent tool for processing time-series and sequential data, which makes it ideal for a variety of business applications that require dynamic trends and predictions. LSTMs perform exceptionally well in forecasting future stock prices by utilizing historical data. They are essential for stock market analysis because they may identify long-term trends and dependencies in the time series data.

Decision trees provide clear insights and simple interpretation for decision-making. They are straightforward yet effective for classification and regression applications. The most influential aspects on Infosys stock prices, such as trading volume, economic indicators, and news mood, can be determined using the use of decision trees.

Because they are ensemble in nature, random forests: Produce reliable and accurate forecasts; they are particularly helpful in risk management and feature importance analysis. Because random forests combine the output from several decision trees, they can improve the accuracy of stock price forecasts. When compared to individual decision trees, they lessen over fitting.

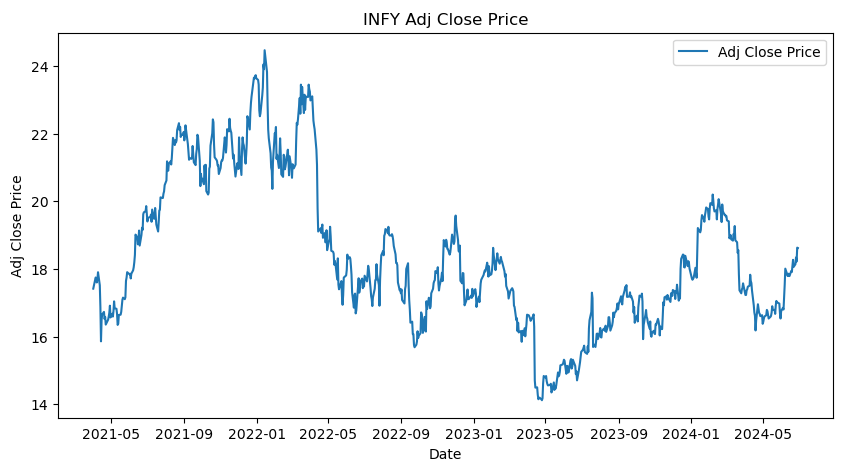
Holt-Winters Model: Helps with resource allocation, financial planning, inventory management, and seasonality-based data forecasting. The seasonal patterns in Infosys stock prices can be identified and predicted by the Holt-Winters model.  
helpful in predicting short-term stock prices when seasonality is present.

These models enable companies to use data to get strategic insights, increase customer satisfaction, streamline operations, and spur overall growth and profitability. Through the utilisation of these models, analysts can enhance their forecasting precision, obtain a more profound understanding of the Infosys stock price dynamics, and make more knowledgeable investing choices.

**RESULTS AND INTERPRETATIONS**

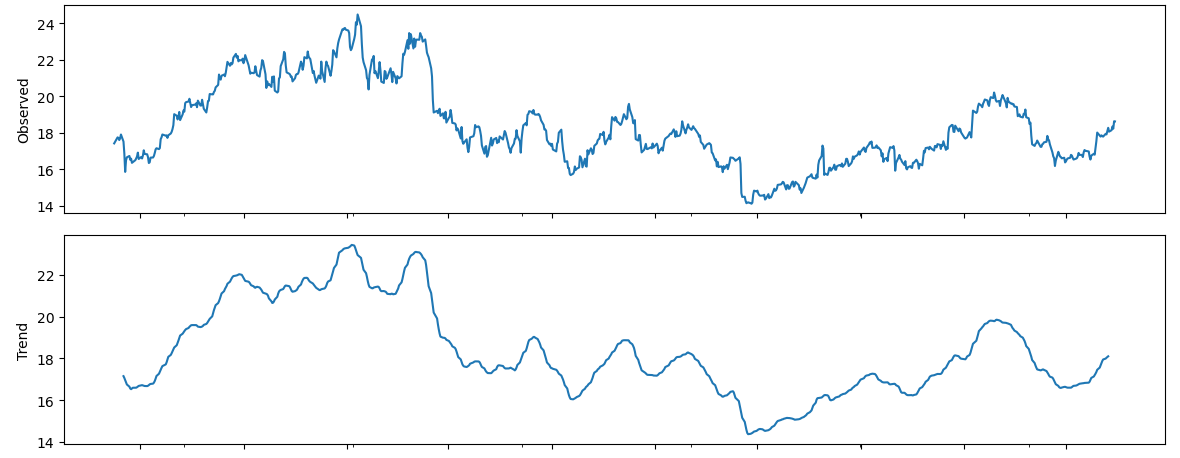
* Clean the data, check for outliers and missing values, interpolate the data if there are any   
  missing values, and plot a line graph of the data neatly named. Create a test and train data set out of this data.

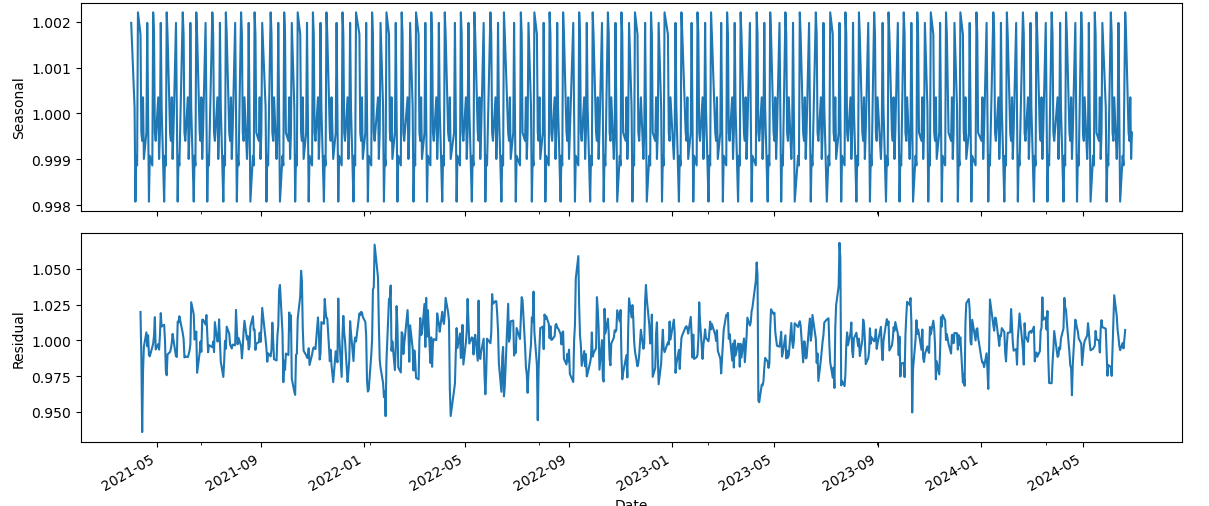
We have firstly downloaded the data into python using yfinance, wherein we have downloaded the data for Infosys using the ticker “INFY” and gotten the data from the time period 2021-04-01 to 2024-06-30. Once we have the data, we have then gone ahead and checked for missing values and found none, besides this we have also chosen our target variable Adjusted Close Price. Now we have plotted a line graph of the same data, as shown below.



* Convert the data to monthly and decompose time series into the components using additive and multiplicative models

Next we have decomposed the time series, using multiplicative model, the results show are as follows:





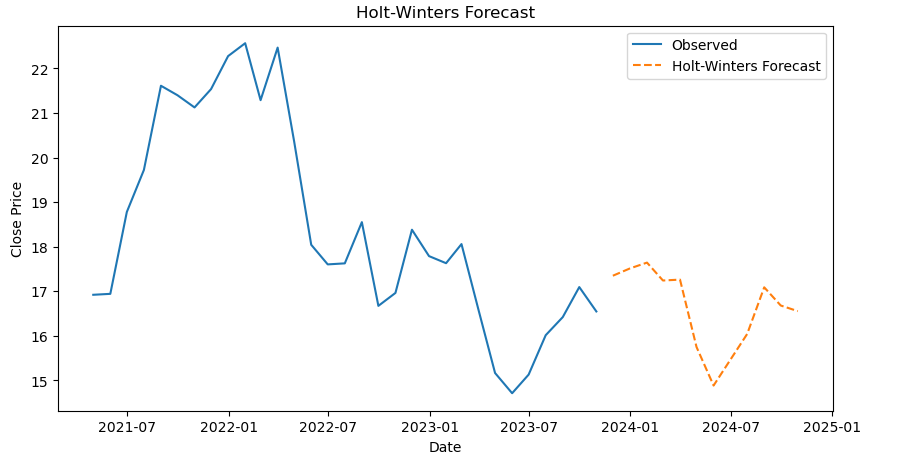
* 1. Univariate Forecasting - Conventional Models/Statistical Models  
  − Fit a Holt Winters model to the data and forecast for the next year.

Firstly we have fitted a Holt Winters model to the data and forecasted for the next year by resampling the data to monthly data. We have tested the accuracy of the model by computing the mean squared error, mean absolute error, r2 score. We get,

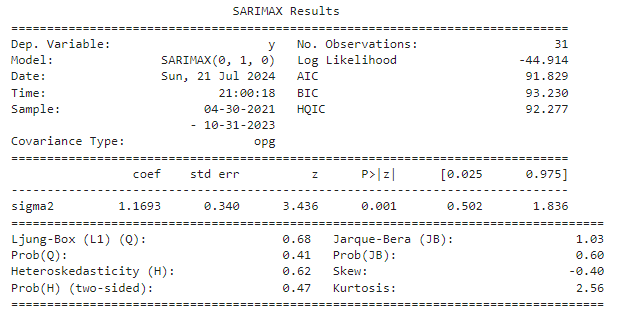
|  |  |
| --- | --- |
| RMSE | 1.6075285023086332 |
| MAE | 1.4225884040304297 |
| MAPE | nan |
| R-squared | -1.4949821267715495 |

INTERPRETATION - Poor Performance: Both RMSE and MAE are high, indicating that the model’s predictions are significantly off from the actual stock prices. Invalid MAPE: The nan value for MAPE suggests issues with the dataset, such as zero or near-zero actual values. Negative R-squared: A negative R-squared indicates that the model is not only failing to explain the variance in the stock prices but is also performing worse than a naive mean prediction.

Plotting the forecast we get, the idea of this model being not a good fit,

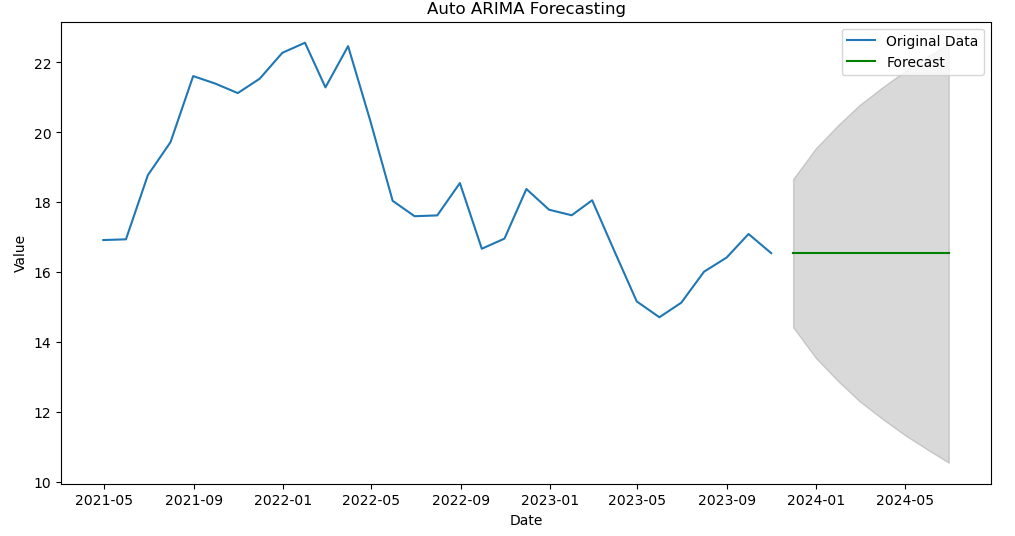


* Fit an ARIMA model to the daily data and do a diagnostic check validity of the model. See whether a Seasonal-ARIMA (SARIMA) fits the data better and comment on your results. Forecast the series for the next three months.   
  − Fit the ARIMA to the monthly series.

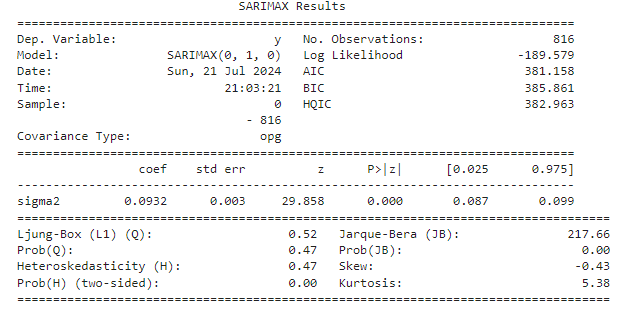


* INTERPRETATION - The SARIMAX model is a simple (0, 1, 0) model with first differencing. This indicates that the data was differenced once to achieve stationarity.
* The coefficient for sigma2 is significant, indicating that the model has captured the variance in the residuals well.
* The fit statistics (AIC, BIC, HQIC) are reasonably low, suggesting a decent model fit.
* Diagnostic tests suggest that the residuals do not exhibit autocorrelation, are normally distributed, and exhibit homoscedasticity (constant variance).

Overall, the model appears to be a good fit for the Infosys stock price data with significant coefficients and acceptable diagnostic test results.



* Fitting ARIMA and SARIMA for daily data, the results are as follows,



INTERPRETATION

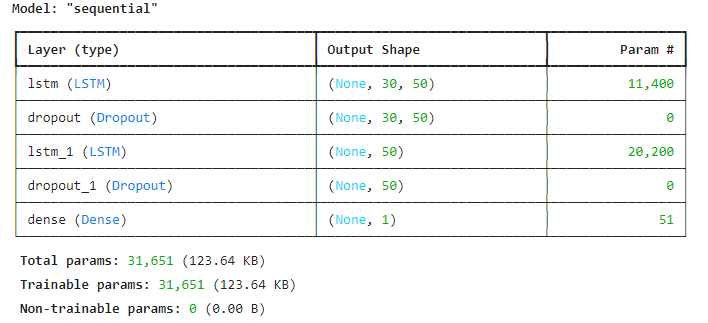
 The SARIMAX model used is a simple (0, 1, 0) model with first differencing, which means the data was differenced once to achieve stationarity.

 The coefficient for sigma2 is highly significant, indicating that the model has captured the variance in the residuals well.

 The fit statistics (AIC, BIC, HQIC) are reasonably low, suggesting a decent model fit, though comparing these values to those from other models would provide a better assessment.

2. Multivariate Forecasting - Machine Learning Models  
− NN (Neural Networks) -Long Short-term Memory (LSTM)

We have built an LSTM model, the results are as follows,

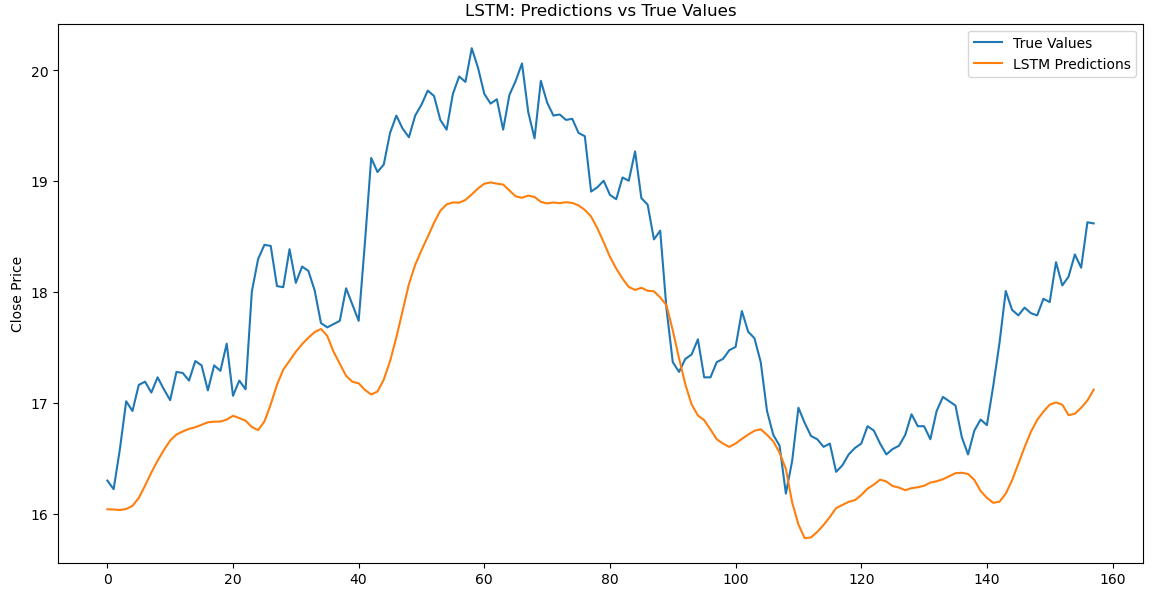


INTERPRETATION - The LSTM model consists of two LSTM layers and two dropout layers followed by a dense layer:

* The first LSTM layer has 50 units and processes sequences of length 30 with 6 features each.
* The dropout layer helps prevent overfitting by randomly setting a fraction of the input units to 0 at each update during training.
* The second LSTM layer has 50 units and processes the output of the first LSTM layer.
* The second dropout layer also helps prevent overfitting.
* The dense layer outputs a single value, which is the final prediction.

The model has a total of 31,651 trainable parameters, indicating the complexity of the model. The parameters are updated during training to minimize the loss function and improve the model's performance on the given task

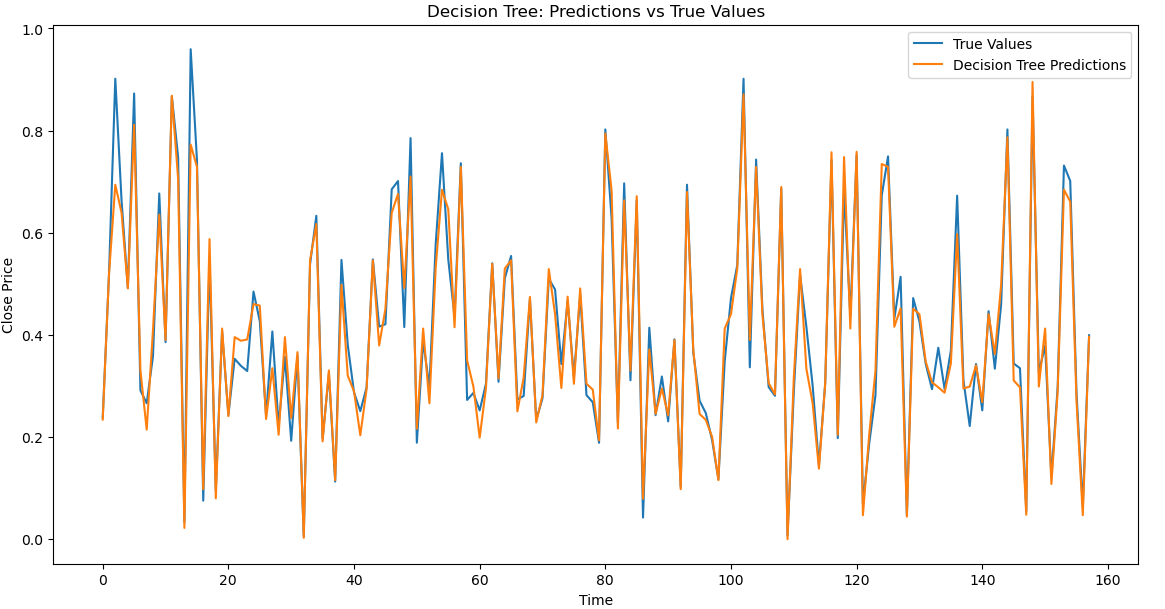
We have also plotted the prediction of LSTM model vs the true values to check the accuracy,



As seen the LSTM values are lower than the true values.

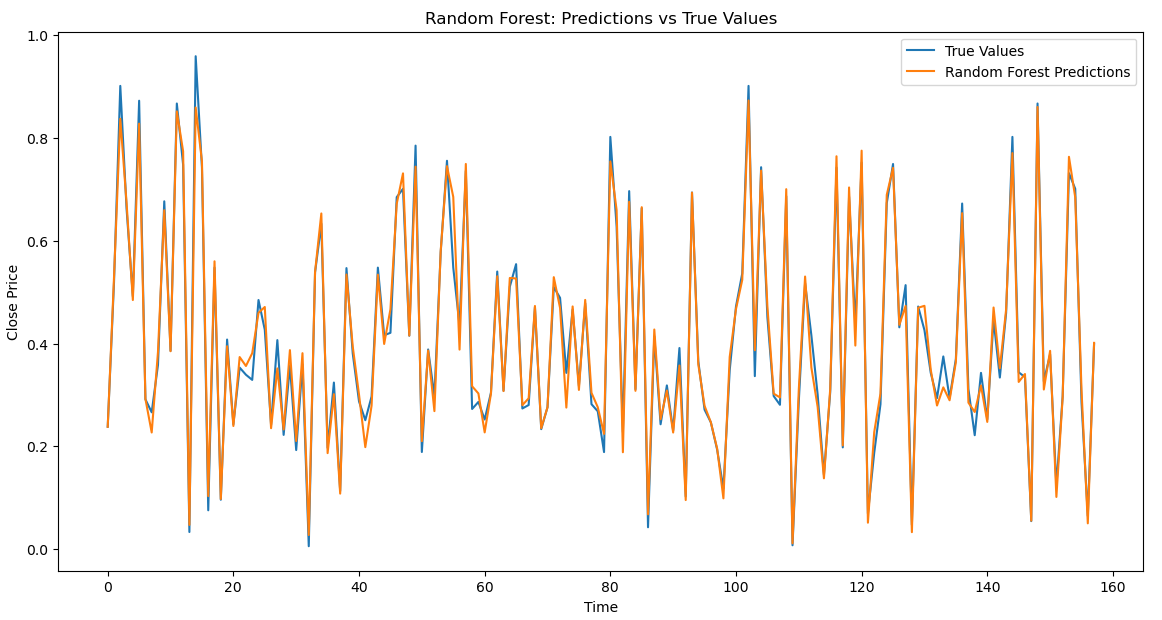
− Tree based models - Random Forest, Decision Tree

Using decision trees, for the data, when plotted against the true values we receive,



As we can see from the image above that the decision tree is able to accurately forecast the data and is more significant and good fit for forecasting.

We also have tested the Random Forest to see if it can predict as well, we get the results as follows,



As we can observe the Random forest model also showcases approximately accurate forecasting when potted against the true values.

To determine the same further, we have computed mean squared error, mean absolute error, r2 score for both the models and the results are as follows,

Decision Tree –

|  |  |
| --- | --- |
| RMSE | 0.04002366433077059 |
| MAE | 0.027433735590152306 |
| MAPE | inf |
| R-squared | 0.9642854873613813 |

INTERPRETATION - Accuracy: The model demonstrates high accuracy with a very low RMSE and MAE, suggesting small average prediction errors.

Fit: The R-squared value indicates that the model explains a large proportion of the variance in stock prices, reflecting a strong fit.

Random Forest –

|  |  |
| --- | --- |
| RMSE | 0.02728430954343284 |
| MAE | 0.01961637118148466 |
| MAPE | 7410.736756935401 |
| R-squared | 0.9834027181547437 |

INTERPRETATION - Accuracy: The model demonstrates very high accuracy with a low RMSE and MAE, suggesting small average prediction errors.

Fit: The R-squared value indicates that the model explains a large proportion of the variance in stock prices, reflecting a strong fit.