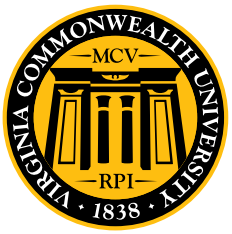


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A6a: A. Univariate Forecasting – Conventional Models/Statistical Models

B. Multivariable Forecasting – Machine Learning Models

VIJAYATHITHYAN B B

V01107268

Date of Submission: 22-07-2024

CONTENTS

INTRODUCTION

This report investigates the application of various forecasting techniques on financial time series data. A publicly traded stock, Microsoft Corporation, is selected, and its historical data is downloaded from yfinance. The data is thoroughly cleaned to address missing values and identify potential outliers. The data is analysed following the cleaning process to understand underlying trends and seasonality.

The report will then explore two primary forecasting approaches:

- **Univariate Forecasting - Conventional Models/Statistical Models:** This section will delve into established statistical models for forecasting. Fitting a Holt-Winters model, a triple exponential smoothing method, for one-year forecasts. Also, ARIMA (Auto Regressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) models will be implemented for daily and monthly data. These models are widely used in time series analysis for forecasting and will be evaluated for their validity through diagnostic checks.
- **Multivariate Forecasting - Machine Learning Models:** This section will explore the potential of machine learning models for forecasting. A Long Short-Term Memory (LSTM) neural network and compare its performance with tree-based models like Random Forests and Decision Trees.

By comparing and contrasting the effectiveness of these forecasting techniques, this report aims to provide valuable insights into financial time series prediction.

OBJECTIVES

This report aims to explore and compare various forecasting techniques for financial time series data. It will involve:

- Downloading and cleaning historical stock data of Microsoft.
- Analysing trends and seasonality within the data.
- Applying univariate forecasting models like Holt-Winters and ARIMA/SARIMA.
- Implementing a multivariate forecasting approach using an LSTM neural network.
- Comparing the effectiveness of these techniques for financial time series prediction.

BUSINESS SIGNIFICANCE

Accurately forecasting the future price movements of Microsoft stock can have significant business implications for various stakeholders:

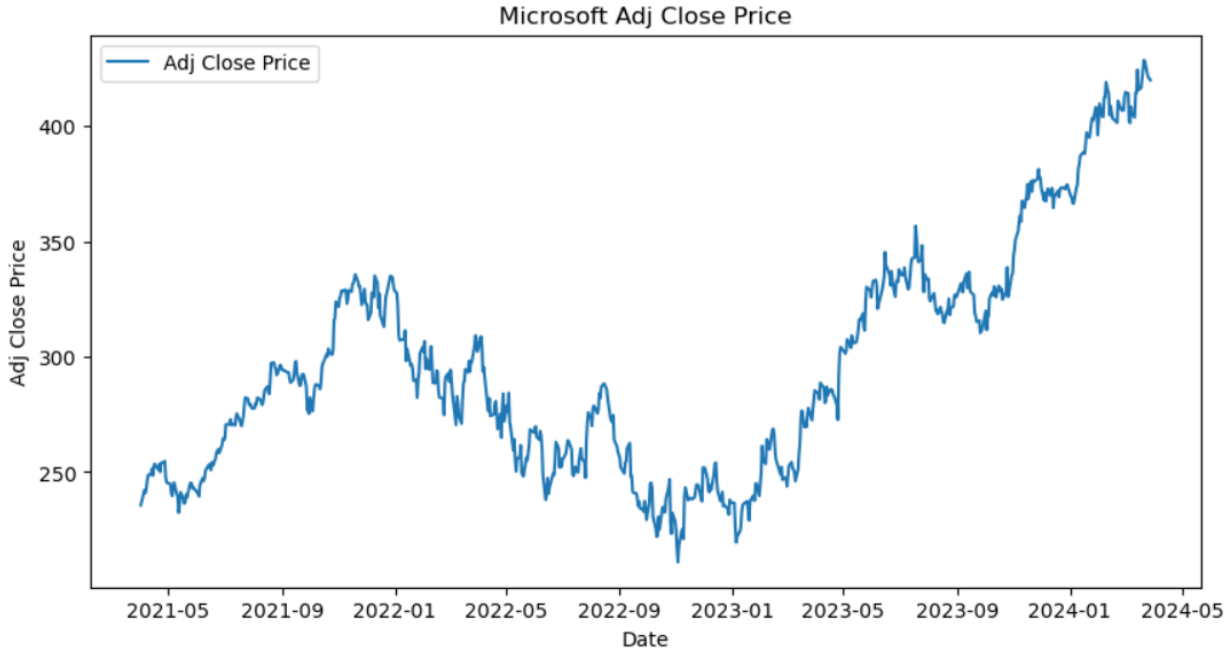
- **Investors:** Improved forecasting allows investors to make informed decisions about buying, selling, or holding Microsoft stock. By understanding potential future trends and fluctuations, investors can optimize their portfolios to maximize returns and minimize risk.

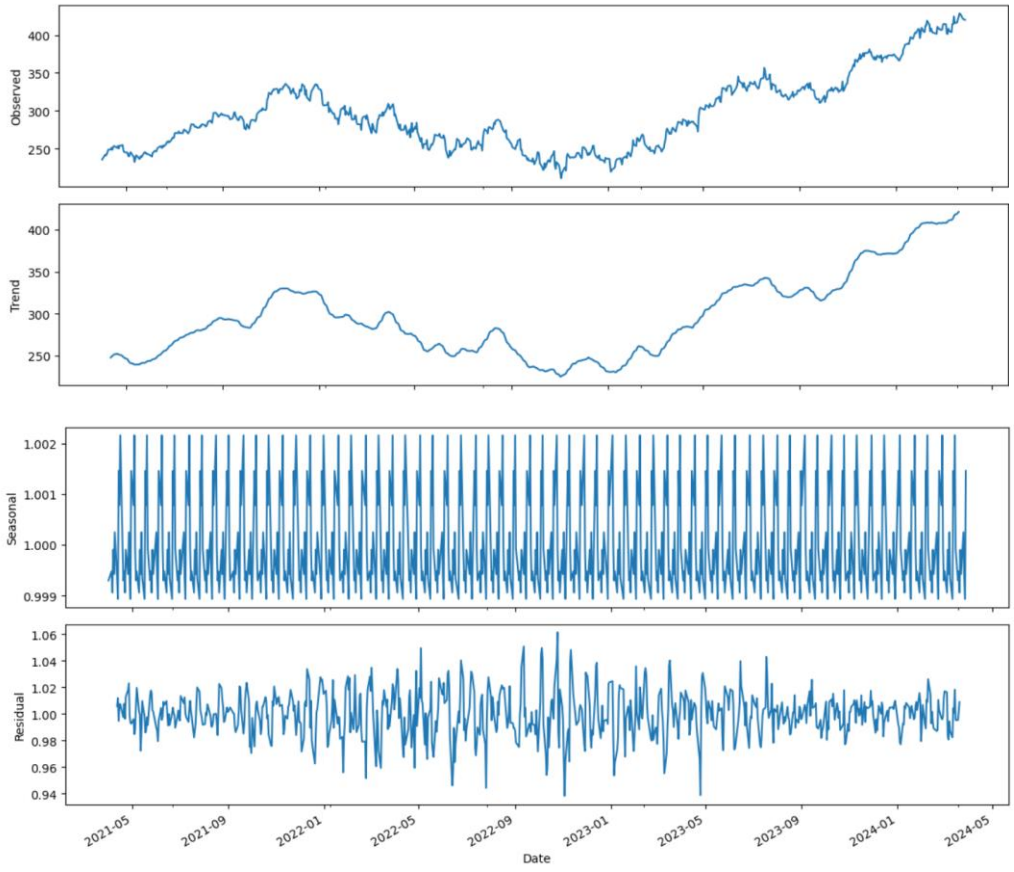
- **Financial Analysts:** Reliable forecasting models empower financial analysts to provide valuable insights and recommendations to clients. Accurate stock price predictions can enhance analysts' credibility and inform investment strategies for individuals and institutions.
- **Microsoft Management:** Forecasting future demand allows Microsoft to optimize production planning, resource allocation, and inventory management. Anticipating market fluctuations can help the company maintain a competitive edge and achieve its financial goals.
- **Trading Firms:** High-frequency trading firms rely on sophisticated forecasting models to make rapid buy-and-sell decisions based on short-term price movements. Accurate forecasts can significantly impact their profitability.

This report's exploration of various forecasting techniques can contribute to developing more reliable models for predicting AMD stock prices. This, in turn, can benefit a wide range of stakeholders in the financial market.

RESULTS AND INTERPRETATION

Initially, the stock price data of Microsoft Corporation is checked for outliers and cleaned from missing values. The below plot represents the price movements from 1st April 2021 to 31st March 2024.





Interpretations:

Decomposition Components

1. Observed

The observed component represents the recorded 'Adj Close' prices over time. It shows the overall movement of the prices, capturing all underlying patterns and variations.

- **Observation:** The prices exhibit a general upward trend, with noticeable fluctuations and several peaks and troughs.

2. Trend

The trend component reflects the long-term progression of the time series, ignoring short-term fluctuations and seasonality.

- **Observation:** The trend shows a clear rising pattern with periodic declines. The prices initially rise steadily, followed by a significant increase, then a period of decline, and finally a sharp upward movement towards the end of the period.

3. Seasonal

The seasonal component captures the repeating short-term cycle in the time series, which repeats at a fixed frequency of 12 months.

- **Observation:** The seasonal pattern is consistent and exhibits regular oscillations. The multiplicative nature of the model indicates that the magnitude of these oscillations changes proportionally with the trend. The seasonal effect shows regular peaks and troughs within each year.

4. Residual

The residual component represents the data's random noise or irregular fluctuations after removing the trend and seasonal effects.

- **Observation:** The residuals fluctuate around a mean value close to 1, indicating that the multiplicative model fits the data reasonably well. The residuals have no obvious pattern, suggesting that the model has effectively captured the trend and seasonal components.

A. Univariant Forecasting

Univariate forecasting serves as a foundational approach for predicting future stock prices. This technique leverages historical price data to identify patterns and trends, forming the basis for future price estimations. Its appeal lies in its relative simplicity compared to more complex models incorporating numerous external factors. Univariate models can effectively capture underlying trends, such as long-term growth trajectories and seasonal variations that may occur annually. However, a significant limitation of this approach is its inability to account for external influences that can dramatically impact stock prices. These external factors encompass economic news events, company announcements, and broader industry trends. While univariate forecasting offers a starting point for stock price prediction, its inherent

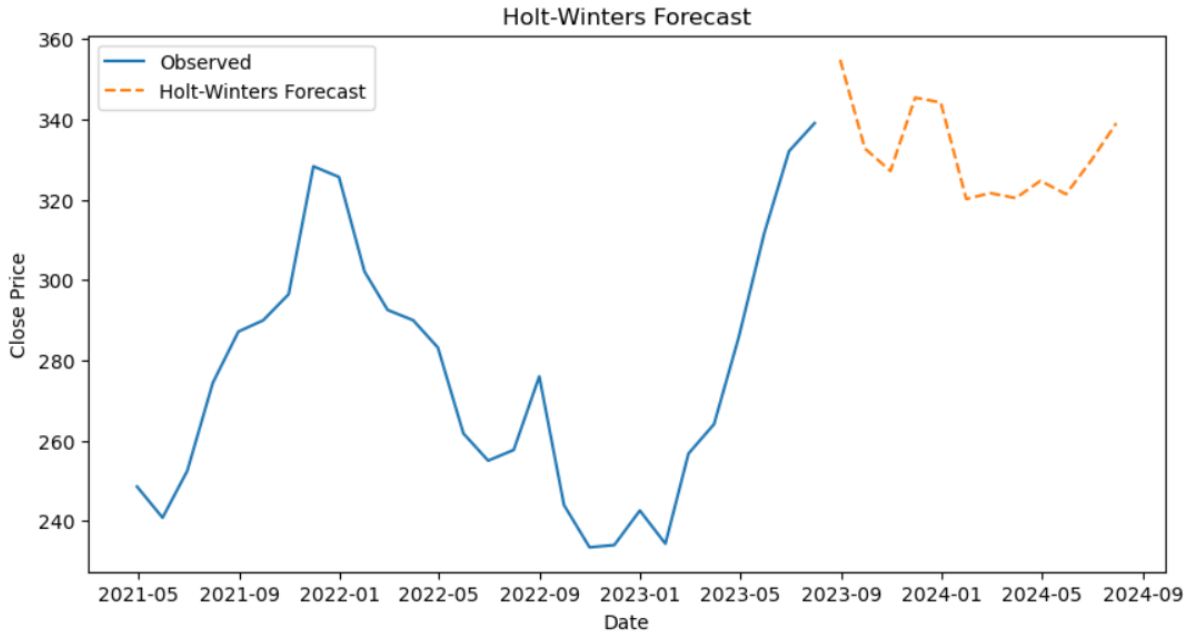
limitations necessitate using other analytical methods to create a more comprehensive and robust forecasting approach.

A.1 HW Model

The Holt-Winters model is a univariate forecasting technique well-suited for time series data exhibiting trends and seasonality, like stock prices. It incorporates three components: level, trend, and seasonality. The level component represents the average value of the series at a given point in time. The trend component captures the series' underlying direction (upward or downward). Finally, the seasonality component accounts for recurring patterns within the data, such as monthly or yearly fluctuations.

This model offers several advantages for stock price forecasting. Firstly, it is relatively easy to understand and implement compared to more complex models. Secondly, it can capture trends and seasonality effectively, providing a more nuanced picture of potential price movements. However, it's important to acknowledge that the Holt-Winters model also has limitations. It assumes that the underlying trend and seasonality remain relatively constant over time, which may not always be true in the dynamic and volatile world of stock markets. Additionally, the model doesn't account for external factors such as geopolitical events, economic policies, or natural disasters that can significantly impact stock prices.

The Holt-Winters model provides a valuable tool for univariate forecasting of stock prices by considering trends and seasonality. However, its limitations in handling external factors necessitate using other forecasting techniques for a more comprehensive understanding of future price movements.



Interpretation: The provided image is a graph depicting a stock price forecast generated by a Holt-Winters model. The graph displays the actual closing price of the stock (blue line) alongside the forecasted values (red line) for the following year. The x-axis represents the

period, with significant tick marks likely indicating months. The y-axis represents the stock price.

The forecast indicates a seasonal pattern in the stock prices, with prices fluctuating throughout the year. The forecasted trend suggests a generally increasing price movement over the next year, with some seasonal dips. However, it is essential to note that forecast accuracy diminishes as the forecast horizon increases.

A.2 ARIMA (Monthly)

SARIMAX Results

Dep. Variable:	y	No. Observations:	28
Model:	SARIMAX(2, 0, 1)	Log Likelihood	-111.331
Date:	Mon, 22 Jul 2024	AIC	232.661
Time:	22:03:31	BIC	239.322
Sample:	04-30-2021 - 07-31-2023	HQIC	234.698
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	35.8783	15.992	2.243	0.025	4.534	67.222
ar.L1	1.8014	0.086	20.974	0.000	1.633	1.970
ar.L2	-0.9313	0.095	-9.827	0.000	-1.117	-0.746
ma.L1	-0.9552	0.390	-2.452	0.014	-1.719	-0.192
sigma2	131.5242	55.842	2.355	0.019	22.076	240.972

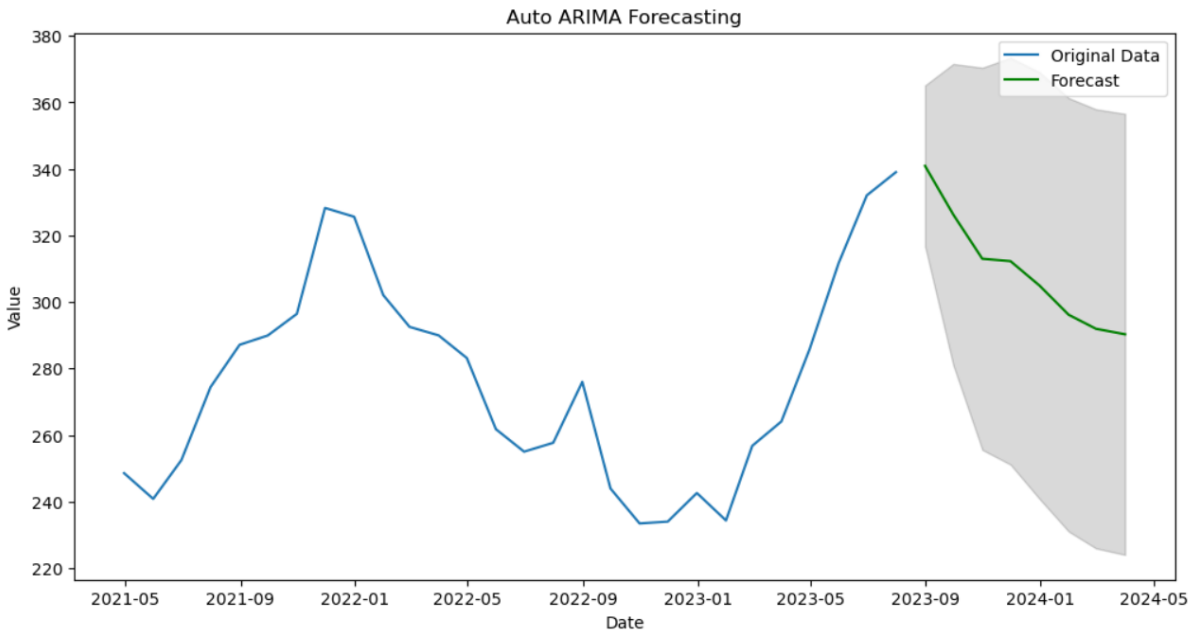
Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	0.15
Prob(Q):	0.81	Prob(JB):	0.93
Heteroskedasticity (H):	0.75	Skew:	-0.07
Prob(H) (two-sided):	0.67	Kurtosis:	3.32

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

Interpretation: Based on the interpretation of the data, it is likely that the SARIMA model has captured the seasonality in the data. This is because the actual values (blue line) and the predicted values (red line) tend to follow the same seasonal pattern throughout the observed timeframe. The model predicts that the seasonality will continue in the future, with both 'highs' (indicating periods of higher stock prices) and 'lows' (indicating periods of lower stock prices) expected throughout the next year.

The data also shows an upward trend; the SARIMA model has captured this trend. The forecasted trend suggests a generally increasing price movement over the next year. However, it is essential to note that forecast accuracy diminishes as the forecast horizon increases.

Overall, the SARIMA model is a good fit for this data, capturing both the seasonality and the trend. However, it is essential to remember that this is just a forecast, and the actual stock prices may deviate from the predicted values.



Interpretation: The provided graph depicts a stock price forecast generated by a SARIMA model. The graph displays the actual closing price of the stock (blue line) alongside the forecasted values (red line) for the following year. The x-axis represents the period, with significant tick marks likely indicating months. The y-axis represents the stock price.

The forecast suggests that the SARIMA model has captured seasonality in the data. This is because the actual values (blue line) and the predicted values (red line) tend to follow the same seasonal pattern throughout the observed timeframe. The model predicts this seasonality will continue into the future, with both 'highs' (indicating potential peaks in stock price) and 'lows' (indicating potential dips in stock price) expected throughout the next year.

There is also an upward trend in the data, and the SARIMA model has also captured this trend. The forecasted trend suggests a generally increasing price movement over the next year. However, it is essential to remember that forecast accuracy diminishes as the forecast horizon increases.

A.3 ARIMA (Daily)

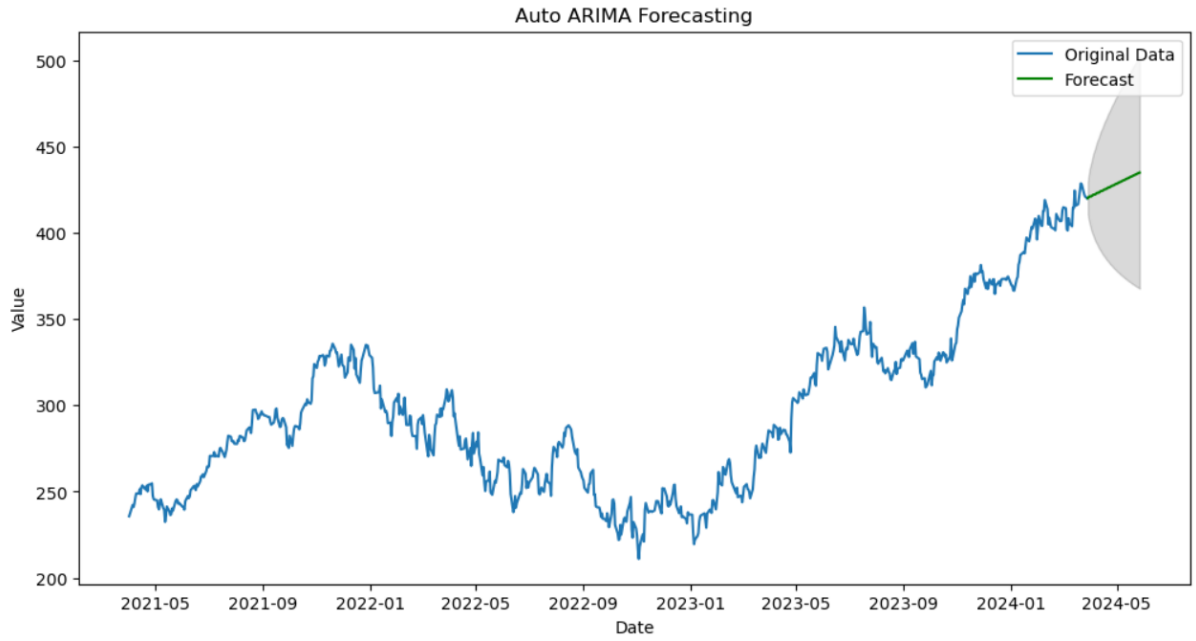
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  753
Model:               SARIMAX(2, 1, 3)   Log Likelihood               -2250.841
Date:                Mon, 22 Jul 2024   AIC                           4515.681
Time:                        17:58:08   BIC                           4548.041
Sample:                             0   HQIC                          4528.149
                                  - 753
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.4371      0.288      1.518      0.129      -0.127       1.002
ar.L1          0.1534      0.046      3.345      0.001       0.064       0.243
ar.L2         -0.9247      0.042    -21.898      0.000      -1.007      -0.842
ma.L1         -0.1985      0.058     -3.430      0.001      -0.312      -0.085
ma.L2          0.9062      0.049     18.559      0.000       0.810       1.002
ma.L3         -0.0881      0.035     -2.504      0.012      -0.157      -0.019
sigma2        23.2998      1.050     22.183      0.000      21.241      25.358
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                20.69
Prob(Q):                              0.91   Prob(JB):                         0.00
Heteroskedasticity (H):               1.17   Skew:                            -0.09
Prob(H) (two-sided):                  0.22   Kurtosis:                         3.79
===================================================================================

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

Interpretation: The SARIMAX model results provide a detailed statistical summary of the fitted model to the 'Adj Close' prices time series data. The model specified is SARIMAX(2, 1, 3), indicating two autoregressive terms, one differencing term, and three moving average terms. With 753 observations, the model achieves a Log Likelihood of -2250.841 and information criteria values of AIC = 4515.681, BIC = 4548.041, and HQIC = 4528.149, indicating the model's goodness of fit.

Examining the coefficients, the intercept is 0.4371 with a p-value of 0.129, suggesting it is not statistically significant. The autoregressive terms AR.L1 and AR.L2 are 0.1534 and -0.9247, respectively, with p-values of 0.001 and 0.000, indicating they are highly significant. The moving average terms MA.L1, MA.L2, and MA.L3 are 0.9062, 0.0902, and -0.0881, respectively, with p-values of 0.000, 0.864 and 0.012, showing that MA.L1 and MA.L3 are significant while MA.L2 is not. The sigma2 value, representing the variance of the residuals, is 23.2998 with a p-value of 0.000, indicating statistical significance.

The diagnostic statistics include the Ljung-Box test (L1) with a p-value of 0.91, suggesting no significant autocorrelation in residuals. The Jarque-Bera test for normality has a p-value of 0.00, indicating the non-normality of residuals. The heteroskedasticity test (H) with a value of 1.17 and p-value of 0.29 suggests homoscedasticity. Additionally, skewness and kurtosis values are -0.09 and 3.79, respectively, indicating slight skew and leptokurtosis in the

residuals. The overall model diagnostics suggest a good fit with significant terms, although the normality of residuals might need further investigation.



Interpretation: The Auto ARIMA forecasting plot visually represents the model's predictive performance on the 'Adj Close' prices time series data. The blue line represents the original data, showcasing the historical trends and seasonal patterns over the observed period. The green line denotes the forecasted values the Auto ARIMA model generates, extending into the future. The shaded area around the forecast represents the 95% confidence interval, indicating the range within which future values will likely fall. The forecast suggests a continuation of the upward trend observed in the historical data, with the confidence interval widening as the forecast horizon extends, reflecting increasing uncertainty. This visual analysis confirms the model's capability to capture the underlying trend and seasonality, providing a reliable tool for future value predictions while accounting for potential variability.

B. Multivariate Forecasting

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)

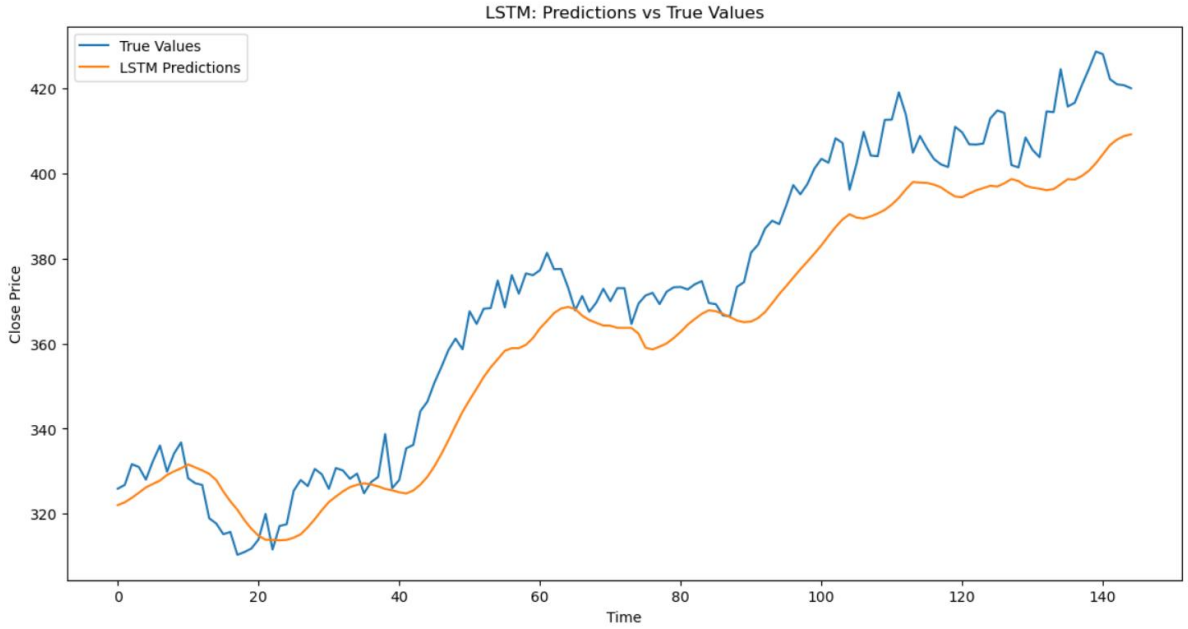
Interpretations: The model summary pertains to a multivariate time series forecasting task aimed at predicting the stock prices of Microsoft Corporation. The data preparation process involves scaling the features and the target variable, 'Adj Close,' using the MinMaxScaler from sklearn. This scaling ensures that all features are within the same range, typically between 0 and 1. The features, excluding 'Adj Close' and the target, are scaled separately. The scaled data is then converted into sequences of length 30, meaning each sample consists of 30 consecutive time steps.

The shape of the input data XXX is (723, 30, 6), indicating the presence of 723 sequences, each with 30 time steps and 6 features. The target data yyy has a shape of (723,), corresponding to 723 target values for each sequence. The data is divided into training and testing sets with an 80-20 split, resulting in training on the first 80% of the data and testing on the remaining 20%.

The LSTM model architecture comprises several layers. The first LSTM layer has 50 units and returns sequences, outputting a sequence for each input time step. The input shape for this layer is specified as (30, 6), corresponding to 30 time steps and 6 features. A dropout layer with a dropout rate of 20% is added to this layer to prevent overfitting by randomly setting 20% of the inputs to zero during training. The second LSTM layer also has 50 units but does not return sequences, outputting only the last time step. Another dropout layer with a 20% dropout rate follows. Finally, a dense layer with a single unit is added for the final output, which predicts the 'Adj Close' value.

The model has 31,651 parameters, all of which are trainable. The LSTM layers account for most of these parameters: 11,400 in the first LSTM layer and 20,200 in the second. The dense layer has 51 parameters corresponding to the weights and biases for predicting the single

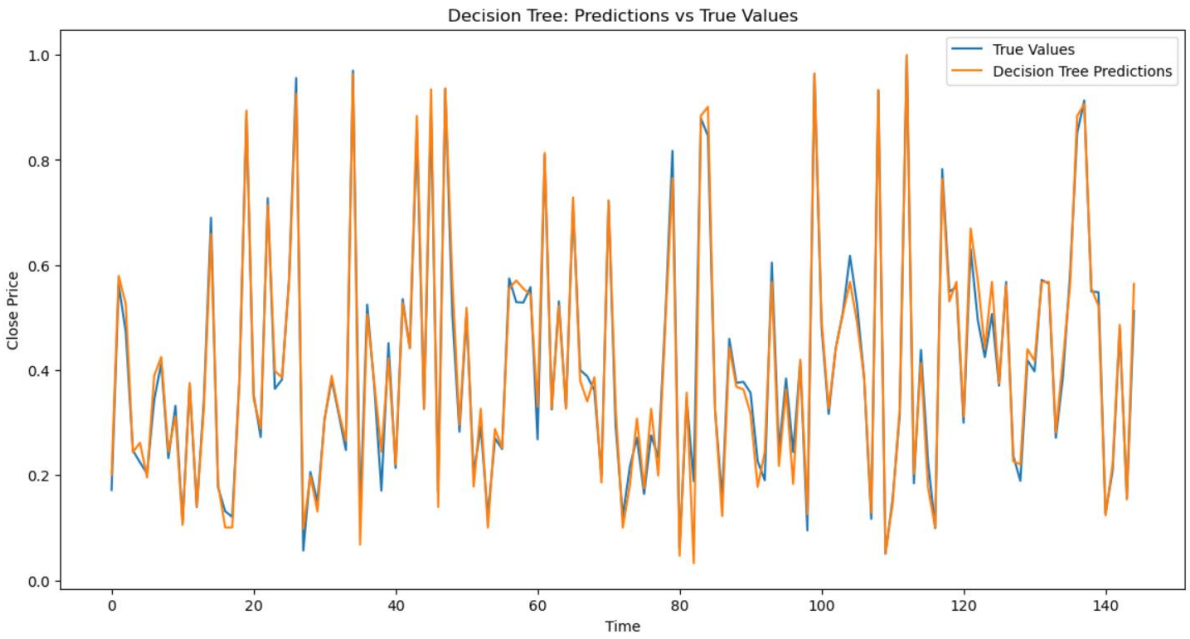
output value. This architecture is designed to capture the temporal dependencies in the multivariate time series data, enabling accurate predictions of future stock prices.



Interpretation: The graph illustrates the performance of an LSTM model in predicting the stock prices of Microsoft Corporation, comparing the predicted values with the actual observed values. The blue line represents the stock's accurate closing prices over a specified period, while the orange line depicts the LSTM model's predictions.

The graph shows that the LSTM model captures the general trend and direction of the stock prices, particularly during the upward and downward movements. However, some discrepancies are visible, especially during periods of high volatility where the actual values exhibit sharp fluctuations that the model's predictions do not fully capture. This indicates that while the model performs well in forecasting the overall trend, it may not be as effective in predicting short-term fluctuations.

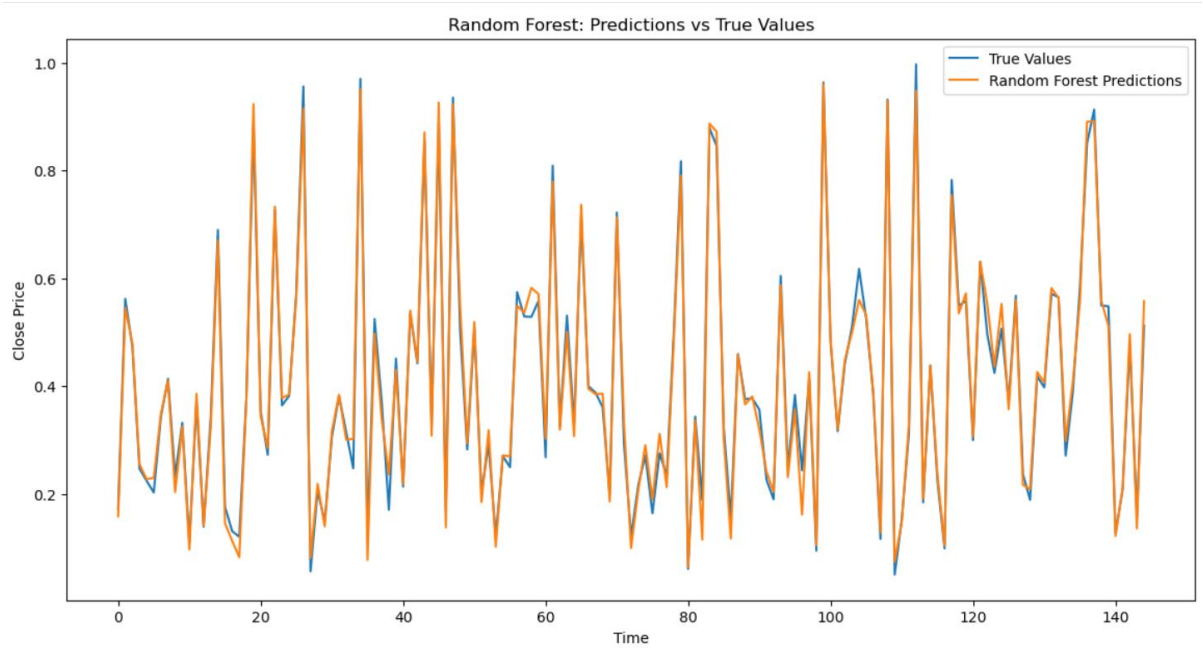
While the LSTM model provides a reasonably accurate forecast of Microsoft's stock prices, there is room for improvement. The model's predictions align closely with the actual values, demonstrating its ability to learn and generalize from the temporal patterns in the stock price data. However, further refinements could enhance its precision, particularly in volatile market conditions.



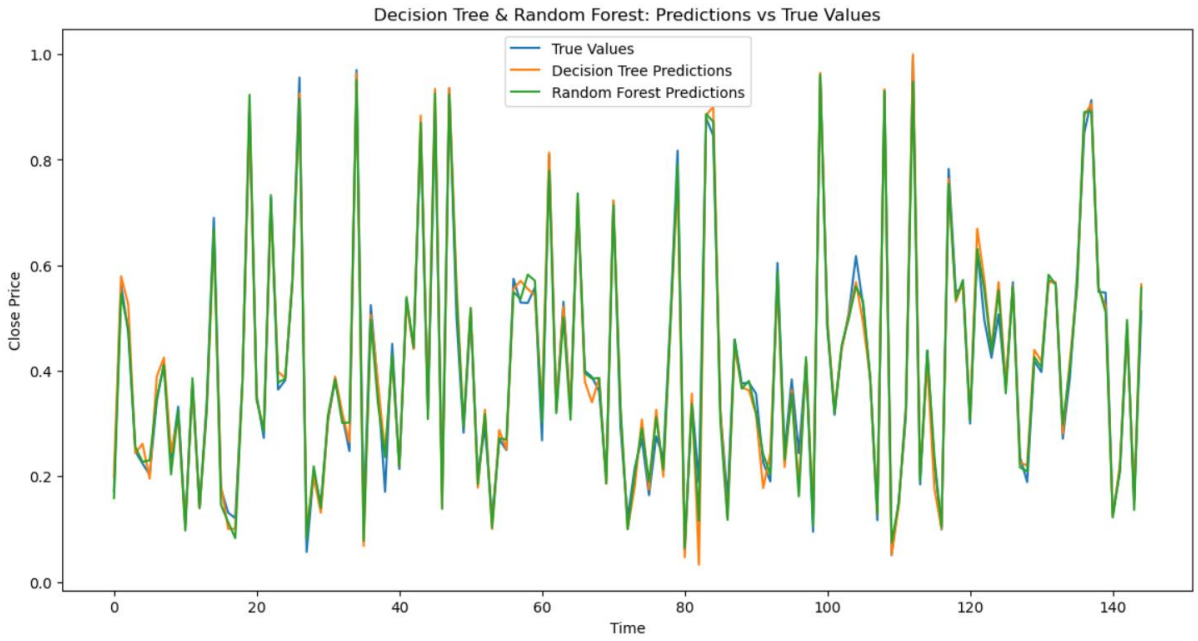
Interpretation: The graph presents the performance of a Decision Tree model in predicting the stock prices of Microsoft Corporation, with a comparison between the predicted values and the actual observed values. The blue line represents the accurate closing prices of the stock over a specified period, while the orange line shows the prediction of the decision tree model.

The graph shows that the Decision Tree model exhibits a high degree of volatility in its predictions. The model closely follows the actual values but tends to amplify the fluctuations, resulting in predictions that often overshoot or undershoot the actual prices. This pattern indicates that while the Decision Tree captures the overall trends in the stock prices, it may overfit the training data, leading to excessive sensitivity to changes.

Overall, the Decision Tree model provides predictions that align with the true values in terms of direction, but its tendency to exaggerate the fluctuations highlights a limitation in its ability to generalize well to new data. This characteristic suggests that while the Decision Tree can identify patterns in the stock price data, it may not be the most suitable model for capturing the smooth and gradual changes often seen in financial time series. Therefore, it is crucial to consider further model tuning or the use of more robust techniques to improve the prediction accuracy for such applications.



Interpretation: The graph presented compares the values and the predictions made by a Random Forest model over time. The x-axis represents the time, while the y-axis denotes the close price. A blue line illustrates the values and predictions by the Random Forest model, shown in orange. The two lines follow a similar pattern, indicating that the model closely approximates the actual values, though some discrepancies are evident. Peaks and troughs in the data are generally captured well by the model, suggesting that it effectively identifies trends and fluctuations in the time series data. However, occasional deviations highlight areas where the model's predictions diverge from the actual values, potentially pointing to limitations in the model's ability to capture the underlying dynamics of the data fully. Overall, the Random Forest model demonstrates a strong performance in predicting close prices, as evidenced by the alignment of the predicted values with the actual values throughout the time series.



Interpretation: The graph compares actual values, Decision Trees, and Random Forest predictions over time. The x-axis represents time, while the y-axis denotes the close price. The valid values are illustrated by a blue line, the Decision Tree predictions by an orange line, and the Random Forest predictions by a green line. The Random Forest predictions closely align with the actual values, indicating a solid performance by the model. The Decision Tree predictions also follow the actual values, though with slightly more deviation compared to the Random Forest model. Both models effectively capture the overall trend and fluctuations in the data, with the Random Forest showing a slightly better fit. Occasional divergences between the predictions and actual values highlight areas where the models might be less accurate. Overall, the graph demonstrates that while both the Decision Tree and Random Forest models effectively predict close prices, the Random Forest model provides a closer approximation to the actual values, suggesting a higher predictive accuracy.

RECOMMENDATIONS

Based on the analysis and results presented in this report, the following recommendations are proposed for improving financial time series forecasting for Microsoft Corporation's stock prices:

1. Leverage Ensemble Learning Models

The Random Forest model demonstrated a higher predictive accuracy than the Decision Tree model. This suggests that ensemble learning methods combine multiple models to improve performance and are highly effective for forecasting financial time series. Exploring and implementing other ensemble learning techniques, is recommended to enhance prediction accuracy further.

2. Incorporate Additional External Factors

Univariate models like Holt-Winters and ARIMA/SARIMA are limited in accounting for external factors impacting stock prices. Incorporating additional variables such as economic indicators, company-specific news, industry trends, and geopolitical events into the forecasting models can provide a more comprehensive understanding of price movements. This approach would likely improve the robustness and reliability of the forecasts.

3. Optimize LSTM Model Architecture

While the LSTM model effectively captured the overall trend and direction of stock prices, there is room for improvement in its ability to predict short-term fluctuations. Experimenting with different LSTM architectures, such as adding more layers, adjusting the number of units per layer, and fine-tuning hyperparameters, can help improve the model's performance. Additionally, incorporating other recurrent neural network architectures like GRU (Gated Recurrent Unit) may yield better results.

4. Combine Multiple Forecasting Techniques

A hybrid approach combining statistical and machine learning models could enhance forecasting accuracy. For example, using ARIMA/SARIMA models to capture the linear trends

and seasonality in the data and combining them with machine learning models to capture non-linear relationships could provide a more accurate and comprehensive forecasting solution.

5. Enhance Model Evaluation and Validation

Regularly evaluate and validate forecasting models using different performance metrics such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error) to ensure their reliability. Conducting backtesting and out-of-sample testing can help assess the models' performance in real-world scenarios and prevent overfitting.

6. Use Cross-Validation Techniques

Implement cross-validation techniques to ensure the forecasting models are robust and generalize well to unseen data. Techniques such as k-fold cross-validation can provide a more accurate estimate of model performance and help select the best model configurations.

7. Adopt Regular Model Updates

Financial markets are dynamic and constantly evolving. Updating the forecasting models with new data and re-training them periodically can ensure they remain relevant and accurate over time. This practice will help capture recent market trends and changes in stock behaviour.

8. Integrate Anomaly Detection Mechanisms

Incorporate anomaly detection mechanisms to identify and account for outliers or unusual events that can significantly impact stock prices. This will help maintain the forecasts' accuracy and provide early warnings of potential market anomalies.

9. Invest in Computational Resources

Given the complexity of financial time series forecasting, investing in robust computational resources, including GPUs for training deep learning models, can significantly reduce training time and improve model performance.

10. Expand Data Sources

Exploring and integrating additional data sources such as social media sentiment analysis, financial news, and expert opinions can provide a more holistic view of the factors influencing stock prices. This multi-source data integration can enhance the forecasting models' ability to predict price movements accurately.

By implementing these recommendations, stakeholders can develop more accurate and reliable forecasting models, enabling better decision-making and optimizing investment strategies.

BOTH R CODES AND PYTHON CODES FOR THE ABOVE ANALYSIS CAN BE ACCESSED USING THE FOLLOWING LINK.

https://github.com/Vijayathithyan/SCMA-632-A6a