

**Stock Market Prediction using Time Series Analysis and Neural Networks**

Misati Nyambane - 670145

Bethlehem Kebede

Hoziyana Rachael

MTH1060: Analytical Computational Foundation

Dr. Francis Oketch

July 29, 2024

# Abstract

Table of Contents

[Abstract 2](#_Toc173158486)

[Introduction 4](#_Toc173158487)

[Methodology 5](#_Toc173158488)

[Data Acquisition 5](#_Toc173158489)

[Data Preprocessing 7](#_Toc173158490)

[Machine Learning Models 8](#_Toc173158491)

[Times Series Analysis 8](#_Toc173158492)

[Neural Networks 10](#_Toc173158493)

[Results and Discussion 13](#_Toc173158494)

[Time Series Analysis 13](#_Toc173158495)

[Results 13](#_Toc173158496)

[Discussion 14](#_Toc173158497)

[Neural Networks 18](#_Toc173158498)

[Training 18](#_Toc173158499)

[Testing 18](#_Toc173158500)

[Validation 19](#_Toc173158501)

[Conclusion and Recommendations 20](#_Toc173158502)

[Time Series Analysis 20](#_Toc173158503)

[Neural Networks 20](#_Toc173158504)

[References 20](#_Toc173158505)

[Appendices 21](#_Toc173158506)

# Introduction

Stock market prediction has long been a topic of interest for investors, financial analysts, and researchers. With the advent of machine learning and advanced statistical techniques, new approaches to forecasting stock prices have emerged. This study explores the application of Autoregressive Integrated Moving Average (ARIMA) and Feedforward Neural Network models for predicting stock prices of 20 different companies.

ARIMA models are a class of statistical models used for analyzing and forecasting time series data. They combine three components: Autoregression (AR), which uses past values to predict future ones; Integration (I), which makes the time series stationary by differencing; and Moving Average (MA), which incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations (Box et al., 2015).

Feedforward Neural Networks (FNNs) are a type of artificial neural network, which is a machine learning model inspired by the structure and function of the human brain. FNNs are called "feedforward" because information flows in one direction, from the input layer, through any hidden layers, to the output layer, without any feedback loops or connections (Goodfellow et al., 2016)

Key terms:

* Time Series: A sequence of data points indexed in time order.
* ARIMA: Autoregressive Integrated Moving Average, a statistical analysis model that uses time series data to predict future trends.
* MSE: Mean Squared Error, a measure of the average squared difference between the estimated values and the actual value.
* MAE: Mean Absolute Error, a measure of the average magnitude of the errors in a set of predictions.
* MAPE: Mean Absolute Percentage Error, a measure of prediction accuracy in statistics.
* **AR (AutoRegressive) Value:**Represents the number of lagged observations included in the model. It indicates how the current value of the time series is influenced by its past values. The AR parameter is usually denoted by p in the ARIMA(p,d,q) model, where p is the order of the autoregressive part. For example, an AR(1) model uses only one lagged term.
* **MA (Moving Average) Value:**Represents the number of lagged forecast errors included in the model. It captures the influence of past forecast errors on the current value of the time series. The MA parameter is usually denoted by q in the ARIMA(p,d,q) model, where q is the order of the moving average part. For example, an MA(1) model uses only one lagged forecast error term.
* **Input Layer:** The layer that receives the input data
* **Hidden Layer:** The intermediate layers between the input and output layers, where the data is transformed and weighted
* **Output Layer:** The final layer that produces the predicted output
* **Neuron:** An interconnected node in the neural network that performs a simple computation
* **Weight:** A value that determines the strength of the connection between neurons
* **Bias:** An additional parameter that shifts the activation function of a neuron
* **Activation Function:** A function that determines the output of a neuron based on its input
* **Loss Function:** A function that measures the difference between the predicted output and the true output, which is minimized during training
* **Gradient Descent:** An optimization algorithm that adjusts the weights and biases of the network to minimize the loss function
* **Backpropagation:** An algorithm used to compute the gradient of the loss function with respect to the weights and biases of the network, which is used in gradient descent

# Methodology

## Data Acquisition

Data was obtained from Yahoo Finance. The data spans from 1st July 2014 to 28th June 2024. The data is in the general form:

*Date,Open,High,Low,Close,Adj Close,Volume*

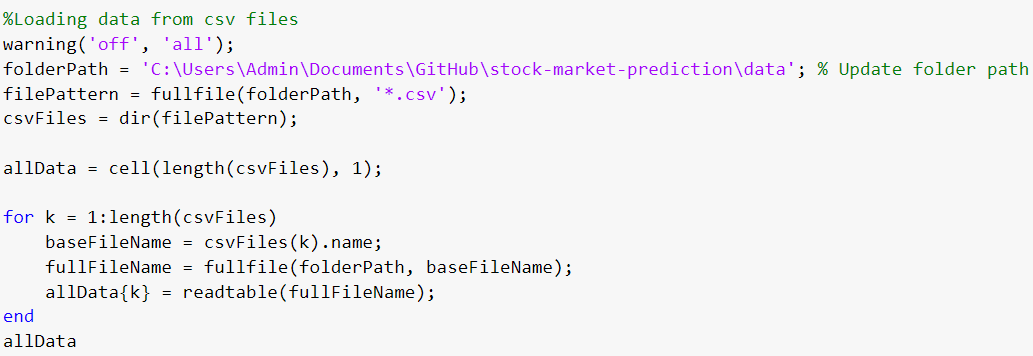
*2014-07-01,23.3799991607666,23.517499923706055,23.282499313354492,23.3799991607666,20.68042755126953,152892000*

*2014-07-02,23.467500686645508,23.514999389648438,23.272499084472656,23.3700008392334,20.671592712402344,113860000*

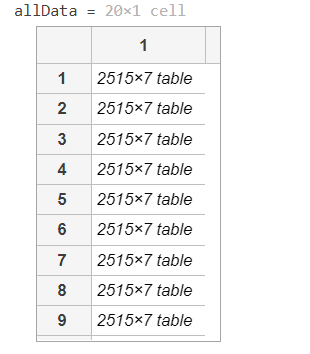
*2014-07-03,23.417499542236328,23.524999618530273,23.299999237060547,23.50749969482422,20.793209075927734,91567200*

The data was obtained using the Yahoo finance package in Python. By use of a dictionary the 20 companies selected were iterated through and their data written to CSV files. Yahoo Finance was preferred since it offers data for each company separately.

To import data the following code was used:



This code imports all CSV files from the specified folder into a cell array allData. Each element of allData contains a table with the stock data for one company.



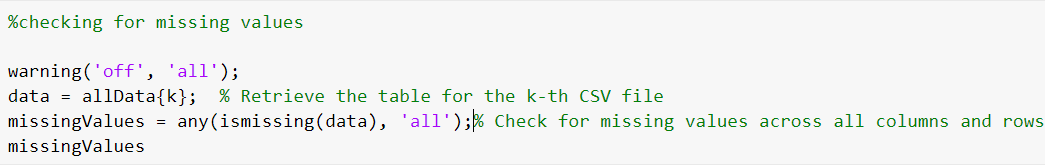
## Data Preprocessing

Data was imported into MATLAB in bulk and stored in an array. This enables the use of iteration and control structures to access the data. Warnings are suppressed for cleaner output and if there are any missing values they are pointed out.

This allows for missing values to be dropped to ensure that operations are not skewed or curtailed by this data.

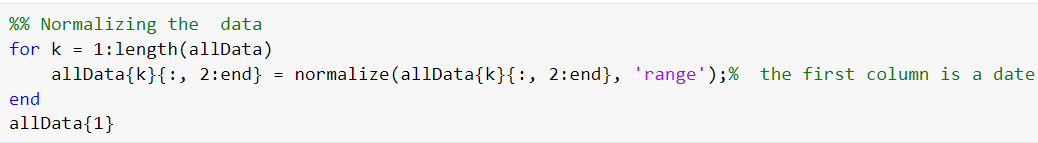
Furthermore, data is normalized between the range of 0 and 1 for our Time Series Analysis and Neural Network models to be able to work with our data. Only numerical columns are normalized. The column *date* is not affected in any way.

Data integrity was checked using:



This code checks for missing values across all columns and rows in each dataset. No missing values were found in the data sets.

Data normalization was performed using:

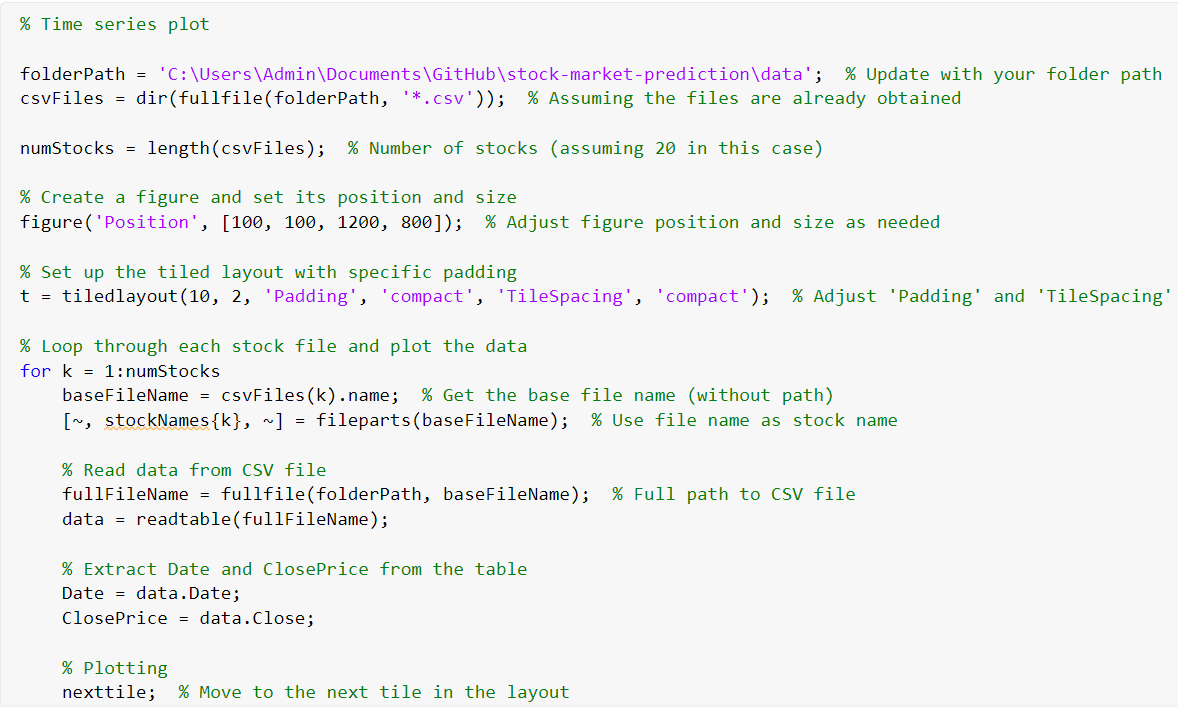


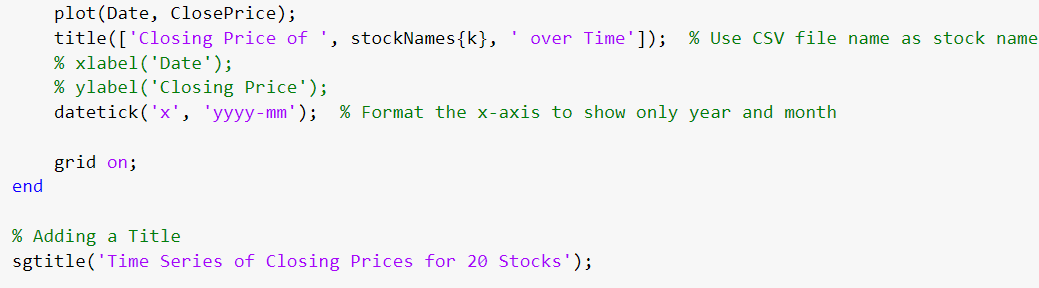
This normalizes all numerical columns to the range [0, 1], facilitating comparison between stocks with different price ranges.

## Machine Learning Models

### Times Series Analysis

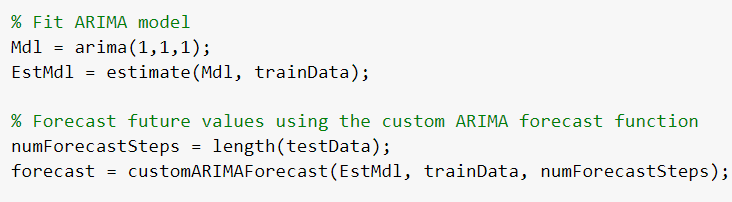
The following code was used to create time series plots for all 20 stocks:





This code creates a 10x2 grid of subplots, each showing the closing price over time for one stock.

The ARIMA(1,1,1) model was implemented for each stock using:



Here, arima(1,1,1) specifies the ARIMA model order. The estimate function fits the model to the training data, and customARIMAForecast (a custom function) generates predictions for the test period.

### Neural Networks

The neural network uses different functions to compute the values. In simplified terms a neural network is a polynomial function whose parameters we do not know. That is, we do not know the degree, variables or coefficients. Nonetheless, what we do know are the weights and biases. These weights and biases are updated using polynomial functions that differentiated and integrated as needed. Furthermore, an activation function and delta variable are employed to control the rate at which the weights and biases are updated. This modification of weights and biases is how the model ‘learns’.

#### Forward Propagation

For forward propagation the functions employed are:

The function above computes a linear combination of the weights and biases then this linear combination is transformed using a non-linear function. This enables complex relationships to be learnt.

The function receives the output from the entry node and then computes the linear combination of weights and biases which are then transformed by the Sigmoid function (defined below).

This function is the Rectified Linear Unit (ReLU) activation function which outputs a value directly if it is positive. Otherwise, it outputs zero.

This function is the Sigmoid function which maps any real-valued number into the range (0, 1).

#### Back Propagation

The function below is the loss function which calculates the Mean Squared Error (MSE) between the actual values and the predicted values. The aim is to minimize the value obtained from the loss function to get a more accurate neural network.

The function below is the derivative of the loss which indicates how much the loss changes with a small change in the predicted output. If  is greater than *y*, the derivative will be positive, suggesting that decreasing   will reduce the loss. Conversely, if   is less than *y*, the derivative will be negative, indicating that increasing  ​ will help reduce the loss. This will inform the network on the way weights and biases should be updated.

This function is the derivative with respect to weights. This term assesses the sensitivity of the output with respect to the input of the activation function. The product with *h* reflects how changes in the weights affect the output based on the current output from the first layer. This derivative is then employed in backpropagation to adjust the weights in the direction that reduces the loss.

The last function is the derivative with respect to bias. Similar to the derivative with respect to weights function this one is used to adjust the bias in the direction that minimizes the loss.

#### Gradients

The functions are the derivatives with respect to the hidden layer output, weights and bias. These derivatives are then used to update their respective variables. This backpropagation aims to reduce the error between the predicted and the actual values.

#### Weight Update Rules

The weights are then updated using the following functions to minimize the value obtained from the loss function.

#### Bias Update Rules

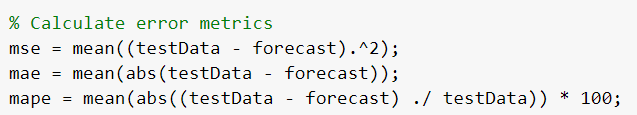
Lastly, the biases are updated.

# Results and Discussion

## Time Series Analysis

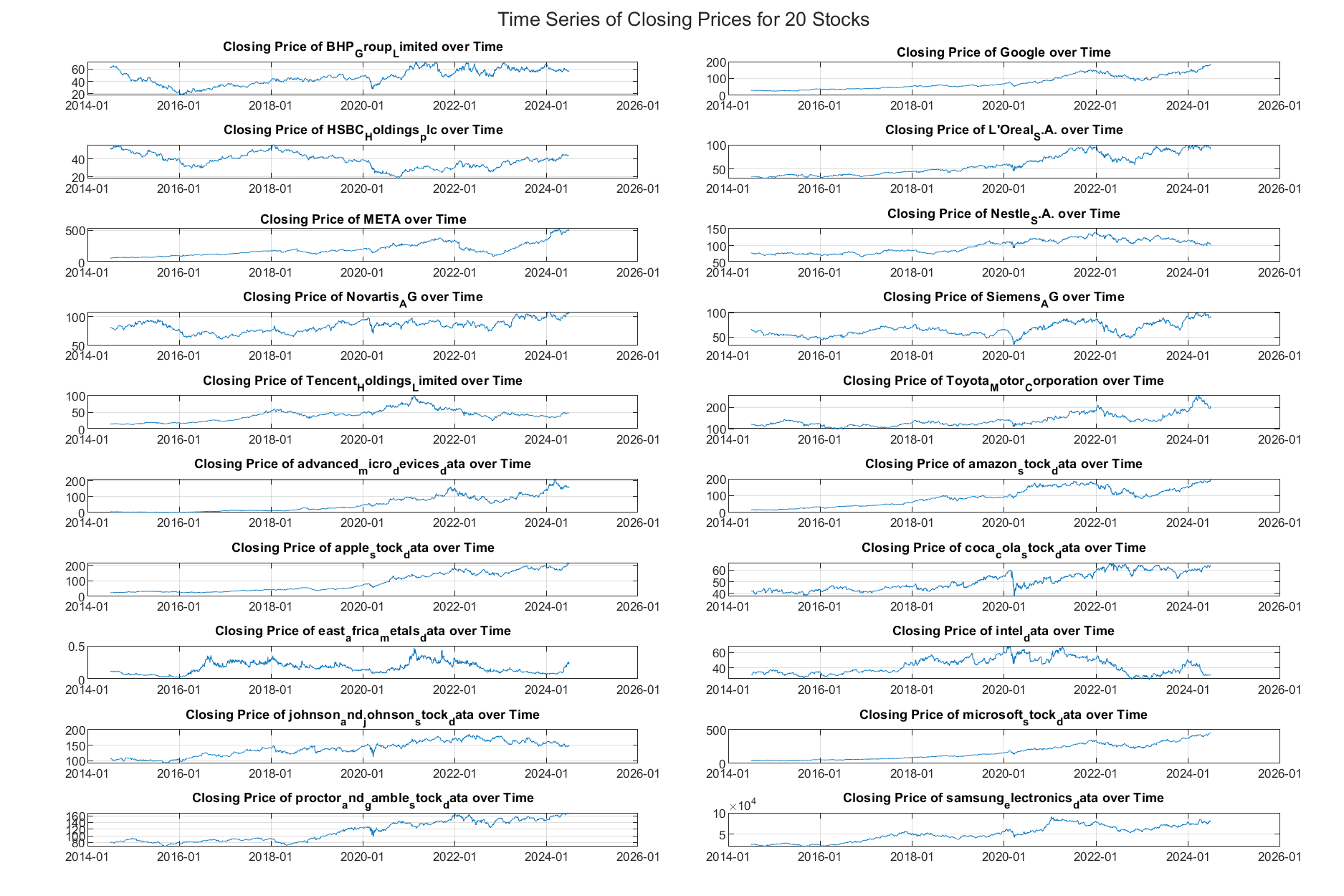
### Results

Performance metrics were calculated using:



These metrics provide different perspectives on the model's prediction accuracy.

The time series plots (Image 1) provide valuable insights into the price movements of the 20 stocks over the period from 2014 to 2024:



### Discussion

#### Key Observations

1. Overall Market Trend: Most stocks show a general upward trend from 2014 to 2024, indicating broad market growth over this decade. This aligns with the overall bullish trend in global markets during this period.
2. Tech Stock Performance: Technology stocks like Google, Amazon, Microsoft, and Apple demonstrate particularly strong and consistent growth. This reflects the increasing dominance of the tech sector in global markets.
3. Volatility Differences: Some stocks, such as Intel and Johnson & Johnson, show more pronounced fluctuations, indicating higher volatility. In contrast, stocks like Nestlé and Procter & Gamble display more stable price movements, typical of consumer staples companies.
4. COVID-19 Impact: A sharp decline is visible across most stocks around early 2020, corresponding to the onset of the COVID-19 pandemic. However, the recovery patterns vary:

* Tech stocks like Amazon and Microsoft show rapid recovery and accelerated growth post-dip.
* Traditional industries like Toyota and Johnson & Johnson show a slower, more gradual recovery.

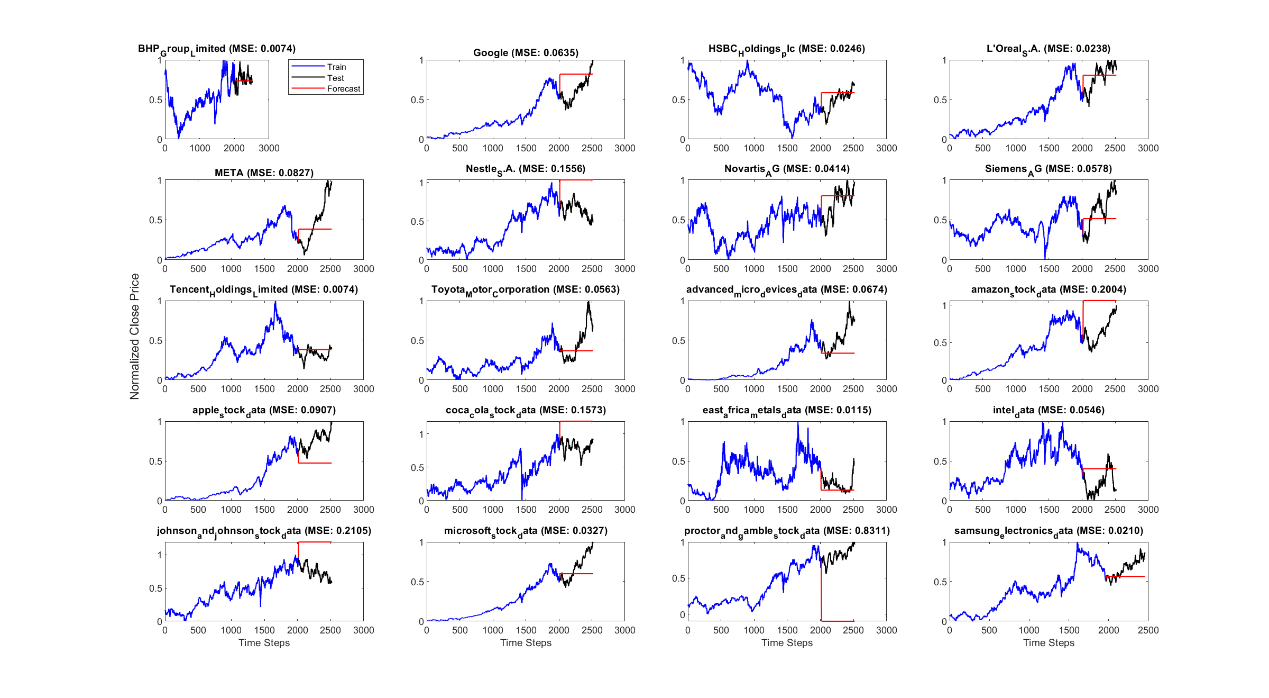
1. Sector-Specific Trends:

* Financial stocks (e.g., HSBC) show more cyclical patterns.
* Consumer goods companies (e.g., Coca-Cola, Procter & Gamble) display relatively stable growth with less pronounced peaks and troughs.
* Pharmaceutical companies (e.g., Novartis) show steady growth, likely influenced by healthcare sector dynamics.

1. Regional Variations: Stocks from different regions (e.g., Tencent from China, Samsung from South Korea) show some unique patterns, possibly reflecting local economic conditions and regulations.

#### Model Performance

The ARIMA(1,1,1) model performance varied significantly across the 20 stocks analyzed:



1. Best Performing Stocks: a. BHP Group Limited (MSE: 0.0074) b. Tencent Holdings Limited (MSE: 0.0074) c. L'Oreal S.A. (MSE: 0.0238) For these stocks, the ARIMA model's predictions closely align with actual price movements. The low MSE values indicate that the model captured the underlying trends and patterns effectively. These stocks likely have more predictable price movements or stronger adherence to historical patterns.
2. Worst Performing Stocks: a. Proctor & Gamble (MSE: 0.8311) b. Coca-Cola (MSE: 0.1573) c. Johnson & Johnson (MSE: 0.2105) The high MSE values for these stocks suggest that the ARIMA model struggled to accurately predict their price movements. This could be due to higher volatility, more frequent external shocks, or complex underlying patterns that a simple ARIMA(1,1,1) model cannot capture adequately.

The performance distribution exhibited the following characteristics:

* The wide range of MSE values (from 0.0074 to 0.8311) indicates that the ARIMA model's performance is highly stock-specific.
* Stocks with MSE values below 0.05 (e.g., BHP, Tencent, L'Oreal, HSBC) show good model fit.
* Stocks with MSE values above 0.1 (e.g., Proctor & Gamble, Coca-Cola, Johnson & Johnson) indicate poor model fit.

In terms of sector based performance it is evident that:

* The model seems to perform better for some technology and financial stocks.
* Consumer goods companies (Proctor & Gamble, Coca-Cola) show poorer performance, possibly due to their more stable price movements which might not benefit from the ARIMA model's strength in capturing trends and seasonality.

By visual analysis:

* For well-performing stocks, the forecast line (red) closely follows the actual price line (black) in the test period.
* Poorly performing stocks show significant divergence between forecast and actual prices, often with the model failing to capture rapid price changes or reversals.

#### Model Parameters and Statistical Significance

Examining the ARIMA model outputs for each stock reveals interesting patterns. The following is a sample of the outputs of the ARIMA model.

ARIMA(1,1,1) Model (Gaussian Distribution):

**Value** **StandardError** **TStatistic** **PValue**

**\_\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_\_\_**

**Constant** -3.6643e-05 0.00092815 -0.039479 0.96851

**AR{1}**  2e-12 22.291 8.9722e-14 1

**MA{1}**  0.00075248 22.291 3.3757e-05 0.99997

**Variance** 0.00028467 5.4987e-06 51.771 0

The key observations are:

1. AR and MA Coefficients:

* AR{1} coefficients range from highly negative (e.g., -0.65488 for L'Oreal) to positive (e.g., 0.63026 for Samsung Electronics).
* MA{1} coefficients show similar variability, with values ranging from -0.71849 (Samsung Electronics) to 0.61574 (L'Oreal).
* The variability in these coefficients suggests different underlying price dynamics across stocks.

1. Statistical Significance:

* Many stocks show statistically significant AR{1} and MA{1} coefficients (p-value < 0.05), indicating that these components contribute meaningfully to the model.
* The constant term is often not statistically significant (p-value > 0.05), suggesting that the differencing component adequately captures the trend in most cases.

1. Variance:

* The variance term is consistently small and statistically significant across all stocks, indicating that the model captures a significant portion of the price variability.

1. Model Interpretation:

* Negative AR{1} coefficients suggest a tendency for the series to oscillate, while positive coefficients indicate momentum in price movements.
* The MA{1} term captures the impact of past forecast errors, with its sign and magnitude indicating how quickly the model adjusts to these errors.

## Neural Networks

### Training

We decided to use a Feedforward Neural Network that:

* Has 40 hidden nodes
* Trains for a maximum of 1000 epochs
* Uses the adam optimizer
* Utilizes 75% of the data to train
* Utilizes 25% of the data to test
* Uses the Mean Squared Error to test the error of the results
* Utilizes validation data from 1st July 2024 to 26th July 2024

### Testing

In testing the MSE’s obtained were as detailed in the table below.

|  |  |
| --- | --- |
| **Company** | **MSE** |
| HSBC Holdings | 0.000289854 |
| Tencent Holdings Limited | 0.000328486 |
| Samsung Electronics | 0.000384859 |
| Amazon | 0.000831416 |
| Alphabet | 0.000862567 |
| Nestle | 0.000945337 |
| Intel | 0.001157845 |
| Proctor and Gamble | 0.001207186 |
| BHP Group Limited | 0.001295159 |
| East Africa Metals | 0.003050332 |
| AMD | 0.005859265 |
| Coca Cola | 0.006657724 |
| Apple | 0.007479319 |
| Novartis AG | 0.007599726 |
| Siemens AG | 0.021622344 |
| L'Oreal | 0.026000365 |
| Microsoft | 0.041602214 |
| Johnson and Johnson | 0.077928339 |
| Meta | 0.123486883 |
| Toyota Motor Corporation | 0.156136595 |

The smallest MSE is the one from HBSC Holdings at 0.000289854 and the largest is by Toyota Motor Corporation at 0.156136595. An interesting trend that emerges that for companies that:

1. Have a wide range of products for sale.
2. Were not affected by the AI boom.
3. Were not really affected by COVID 19.

The model can use data for the first seven and a half years to predict the remaining two and a half years. This explains why Samsung Electronics or Alphabet despite being technological companies have such low MSEs.

Nonetheless,

### Validation

|  |  |
| --- | --- |
| **Company** | **MSE** |
| AMD | 0.011286 |
| Alphabet | 0.036998 |
| Amazon | 0.041988 |
| Apple | 0.044326 |
| BHP Group Limited | 0.015347 |
| Coca Cola | 0.051388 |
| East Africa Metals | 0.483313 |
| HSBC Holdings | 0.102384 |
| Intel | 0.072831 |
| Johnson and Johnson | 0.049414 |
| L'Oreal | 0.174741 |
| Meta | 0.028277 |
| Microsoft | 0.024886 |
| Nestle | 0.049855 |
| Novartis AG | 0.028158 |
| Proctor and Gamble | 0.126288 |
| Samsung Electronics | 0.020738 |
| Siemens AG | 0.154387 |
| Tencent Holdings Limited | 0.01493 |
| Toyota Motor Corporation | 0.017593 |

# Conclusion and Recommendations

## Time Series Analysis

## Neural Networks

# References

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

# Appendices

|  |  |  |
| --- | --- | --- |
| Company | Testing MSE | Validation MSE |
| AMD | 0.005859265 | 0.011285854 |
| Alphabet | 0.000862567 | 0.03699829 |
| Amazon | 0.000831416 | 0.041988428 |
| Apple | 0.007479319 | 0.044326173 |
| BHP Group Limited | 0.001295159 | 0.015347291 |
| Coca Cola | 0.006657724 | 0.051387616 |
| East Africa Metals | 0.003050332 | 0.483313323 |
| HSBC Holdings | 0.000289854 | 0.102384032 |
| Intel | 0.001157845 | 0.072831488 |
| Johnson and Johnson | 0.077928339 | 0.049413927 |
| L'Oreal | 0.026000365 | 0.174740765 |
| Meta | 0.123486883 | 0.028276792 |
| Microsoft | 0.041602214 | 0.024885528 |
| Nestle | 0.000945337 | 0.049855178 |
| Novartis AG | 0.007599726 | 0.028157622 |
| Proctor and Gamble | 0.001207186 | 0.126287847 |
| Samsung Electronics | 0.000384859 | 0.020737847 |
| Siemens AG | 0.021622344 | 0.154387357 |
| Tencent Holdings Limited | 0.000328486 | 0.014930324 |
| Toyota Motor Corporation | 0.156136595 | 0.017593062 |

*Fig 1.1: Testing and Validation MSE’s*