STOCK MARKET PREDICTION USING MACHINE LEARNING MODEL

ELVIN NG KIAN HOU

DIPLOMA IN COMPUTER SCIENCE
METHODIST COLLEGE KUALA LUMPUR

JANUARY 2023

ACKNOWLEDGMENT

I would like to express my profound gratitude to Ms. Anis Zahirah binti Azman, my supervisor, and Ms. Nur Shameen Aina Bt Abdul Rahim, my co-supervisor for their contributions to the completion of my project titled ‘Stock Market Prediction Using Machine Learning Model’. I want to give a particular thank you to my supervisors for their time and work over the course of the semester. I found their recommendations and assistance to be helpful as I finished the job. In this aspect, I am eternally grateful to them.

ABSTRACT

The advancement of machine learning techniques has led to an increase in the number of models that can be used for stock market prediction. However, due to the wide range of models available, data scientists often have different perspectives, understandings, and preferences on which model is the best to use in predicting the stock market. This massive number of professional opinions can lead to confusion, as it can be challenging to differentiate between the different models and choose the best one for a given situation. To address this issue, I have developed projects aimed at comparing the performance of different machine-learning models for stock market prediction. One such project includes collecting and analyzing stock market data, which would then be fitted into two popular models for stock market prediction, Linear Regression, and LSTM. The models can be used to forecast the future price of the stock after being fitted. The project would then compare the performance of these two models using the same evaluation metrics. The performance of each model can then be evaluated using metrics such as Mean Squared Error (MSE) or Root Mean Squared Error (RMSE). These metrics can be used to compare the accuracy of the two models and determine which one performs better in predicting stock prices. In the end, it was discovered that LSTM slightly outperformed Linear Regression in terms of stock price prediction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

LSTM	Long Short-term Memory
RNN	Recurrent Neural Network
RSI	Relative Strength Index
EMA	Exponential Moving Average
EMAF	Exponential Moving Average Fast
EMAM	Exponential Moving Average Medium
EMAS	Exponential Moving Average Slow
CSV	Comma-Separated Value
EDA	Exploratory Data Analysis
3D	3-Dimensional
MAE	Mean Absolute Error
RSME	Root Mean Squared Error

CHAPTER 1

INTRODUCTION

1.1 Background of The Study

A marketplace where stock in publicly traded corporations can be bought and sold is the stock market [1]. A business can offer shares of stock to the public in exchange for funding when it needs to raise money. They also reflect a piece of the company's ownership. These shares represent a portion of ownership in the company and give shareholders the right to vote on certain company decisions and receive a portion of the profits in the form of dividends [2].

Investors take part in the stock market when they buy and sell shares of stock. A few variables, such as business performance, prevailing economic conditions, market trends, and political developments, can affect a stock's price [3]. Investors may be able to profit from these price swings in stocks by buying and selling shares.

Machine learning is a subset of artificial intelligence that involves the development of algorithms and statistical models that enable computer systems to automatically improve their performance on a specific task through experience or data [4]. Machine learning is essentially a technique for teaching computers to learn from data rather than being explicitly programmed to perform tasks.

In the realm of stock market analysis and forecasting, machine learning approaches are gaining importance. One frequent use of machine learning in finance is the prediction of stock prices. Machine learning algorithms can be used to examine enormous volumes of financial data in order to spot trends and forecast future stock values [5].

The fundamental concept is to forecast future stock prices by training a machine learning model on historical stock price data. A range of input features, such previous stock prices, trading volumes, news items, or economic indices, can be used to train the model. Predicting stock prices with machine learning is a difficult problem, and precise forecasts is challenging to make. Yet, machine learning models can be useful instruments for assisting investors in making wise choices regarding when to buy or sell stocks.

1.2 Problem Statement

In the complex and dynamic environment of the stock market, various models can be utilised to analyse and forecast stock value **[5]**. The ideal model is not always evident, though, depending on the situation. So, which model outperforms the others regarding reliability and accuracy of predictions in the stock market?

1.3 Objectives

The objectives of this project are:

1. To clean the data by using python.
2. To analyse the cleaned data through data exploration and implement the different models into it.
3. To evaluate the models using the performance evaluation metrics.
4. To predict the actual stock price using the best performance model.

1.4 Scopes

The scopes of this project are:

1. Collect the historical prices dataset of Apple company.
2. Test and run the data using the LSTM (Long Short-term Memory) and Linear Regression model.
3. Compare and select the best-trained model based on the performance evaluation metrics.
4. Predict the stock price within a month.

CHAPTER 2

LITERATURE REVIEW

2.1 LSTM (Long Short-term Memory)

Long Short-Term Memory, often known as LSTM, is a form of artificial neural network frequently used for time series data analysis for data such as stock prices, weather data, and speech signals. It is a specialized type of RNN that is designed to address the "vanishing gradient" problem that can occur in traditional RNNs [6].

When the gradient, a measure of how quickly the error changes in relation to the weights of the network, gets very small, an issue known as the vanishing gradient problem arises. As a result, the network may find it challenging to learn links between input and output data that persist over extended time periods [6]. LSTM networks address this problem by using a system of "gates" that can selectively remember or forget information over time.

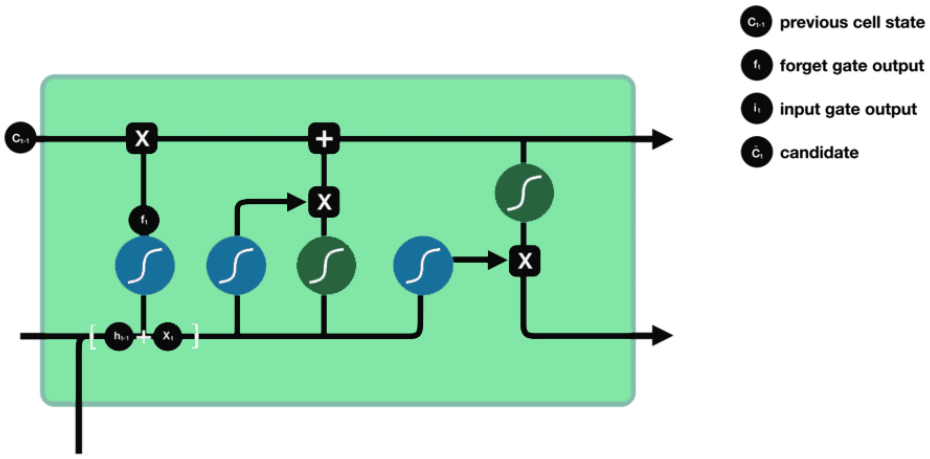


Figure 1: LSTM component.

The key components of an LSTM network are memory cells, input gates, output gates, and forget gates. The input gates control how much fresh information is allowed into the memory cells, while the memory cells gradually store information about the input data. The output gates regulate the information flow from the cells to the network's output, while the forget gates decide which information should be kept in the cells and which should be deleted [6].

Because they can learn to recognise patterns and relationships in the data across extended time periods and can modify their memory cells to keep or forget information as necessary, LSTM networks are well suited for evaluating time series data. Many applications, including as speech recognition, natural language processing, and stock price prediction, have made use of them.

2.2 Linear Regression

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables [7]. Using a straight line to depict the relationship between the input factors and the output variable is the fundamental concept behind linear regression.

The goal of linear regression is to find the best-fit line that minimizes the sum of the squared differences between the predicted values and the actual values of the dependent variable. The least squares estimate method can be used for this [7]. Making predictions about the dependent variable based on the values of the independent variable is possible once the best-fit line has been identified.

In the case of stock market price prediction, the input features might include factors such as the historical stock prices, trading volumes, or economic indicators [8]. The coefficients of the input features that best predict the stock price would then be estimated using the linear regression model. Once the coefficients have been estimated, the model can be used to make predictions about future stock prices based on the input features.

One limitation of linear regression is that it assumes a linear relationship between the input features and the output variable, which may not always be the case in real-world situations [9]. Moreover, complicated nonlinear correlations between input data and stock prices may be difficult for linear regression models to capture. Yet, when paired with other machine learning methods like time-series analysis or deep learning, such as a model of the stock market, linear regression can be a valuable tool for predicting price movements.

2.3 Indicators of Stock Market

Indicators in the stock market are statistical tools used by traders and analysts to make predictions about future price movements of stocks or other financial instruments [10]. The price, volume, or open interest of an asset over time are typically used in these indicators' mathematical derivations. Some commonly used indicators in the stock market include moving averages (such as Simple Moving Average and Exponential Moving Average) and Relative Strength Index (RSI) [10]. These indicators are introduced because the machine learning models will utilise them as input to study the data's pattern and trend.

2.3.1 EMA (Exponential Moving Average)

The EMA is a technical analysis tool used to track the average price of an asset over time [11]. Traders frequently use it to spot trends in the stock market and base their trading decisions on those trends.

In instance, EMA is more responsive to recent price movements than other moving averages since it pays greater weight to recent prices. This may be helpful in spotting transient trends and trading chances. For instance, it may be a bullish indicator if an asset's current price is above its EMA, and it may be a negative signal if it is below its EMA [11].

EMAF, EMAM, and EMAS are variants of the EMA that differ in how they weight previous prices. To spot trends and trading opportunities, they can both be used in a similar manner, but which one to use may depend on the trading method being used [11].

2.3.2 RSI (Relative Strength Index)

Relative Strength Index, or RSI for short, is a well-liked technical analysis indicator that gauges the strength of a stock's price movement over a certain time frame [10]. The average gains and losses of an asset over a predetermined period, often 14 days, are compared to determine the RSI.

Because it enables traders to spot probable trend reversals and trading opportunities, RSI is helpful to traders. Depending on the trader's particular trading technique, when the RSI indicates that a stock is overbought or oversold, it may be a signal to buy or sell the stock. To make wise trading decisions, RSI should, however, be used in conjunction with other technical and fundamental analysis techniques [10].

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Project Development

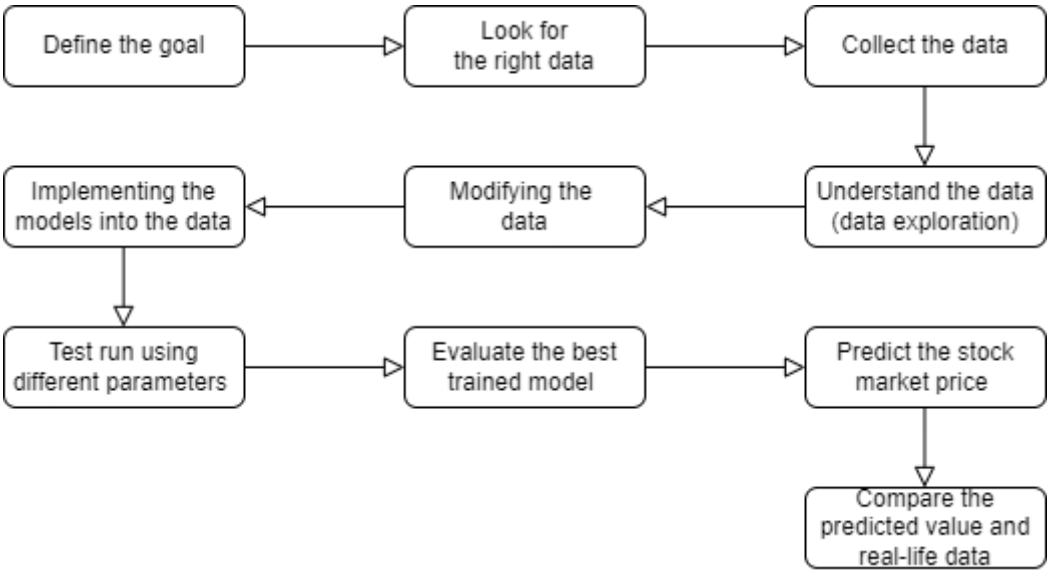


Figure 2: Flow chart of the project.

Figure 2 shows the whole picture of the project. The project is partitioned into ten small parts. The first part is to define the goal of the project. The purpose of this project is to find out the best model out of the chosen models for stock market prediction by comparing the evaluation metrics. To ensure the project is fairly implemented, the data will be the same for both models.

Next step is looking for the right data. It is easy to get the data of any companies that go public, but it is important to make sure the source of the data is verified or certified. In this case, the historical stock price data of Apple company is collected through the website, NASDAQ. It is an online electronic marketplace where investors can buy or sell the shares of public company on a computer network [11]. It can be trusted because it operates as a national securities exchange in the year of 2006.

Next, the data is collected from the website by downloading the data. In this part, it is important to download the data as a CSV file. It is much easier for the Python to handle the data in CSV form. It is also important to only collect the data that are related, in this case, the data is collected from the past 10 years of Apple company's historical data.

The fourth step is understanding the data through EDA (exploratory data analysis). According to an article, Exploratory data analysis is the crucial process of doing preliminary analyses on data to find patterns, identify anomalies, test hypotheses, and double-check assumptions with the help of summary statistics and graphical representations [12]. In a dataset, it contains different variables. Different tools, charts and graphs are used to understand the data by using the library Matplotlib in Python programming language.

The fifth step involves data modification, where the type of variables imported may not be suitable for calculations, especially if they are labelled as object variables. In this step, the variables are converted to numerical types, normalized, and additional aggregated variables such as EMAM, EMAF, EMAS, RSI, and Target are included for more information.

In the next step, the prepared data is fitted into the model while ensuring the consistency of the data shape, as this affects the shape of the predicted value.

The seventh step involves test runs using different parameters such as the number of iterations, batch size, and learning rate. Parameters must be tuned appropriately to avoid training errors.

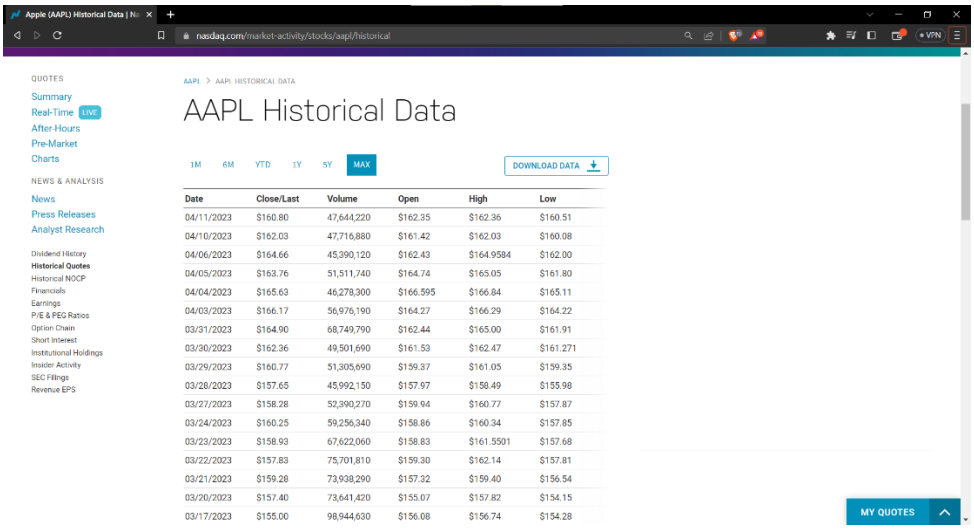
Evaluation metrics are then used to measure the model's performance, and both models are compared to determine the best trained model. The best model will be used to predict the stock market price, and its predictions will be compared to the actual stock price.

3.2 Basic Operation of the Proposed System

3.2.1 Source of Data

The source of data is important to investigate because the quality and reliability of the data can significantly impact the results and conclusions drawn from the analysis. Making projections and judgments based on faulty or inadequate data may adversely affect firms, investors, and other stakeholders. Therefore, it is essential to guarantee that the data used for analysis is reliable, pertinent, and current. The source of the data should also be looked at because it may reveal biases and constraints that could affect the analysis. For instance, the data may not be indicative of the entire market or economy if it is gathered from a particular location or industry. Additionally, the data may reflect the views or viewpoints of the organization or group that collected it. Therefore, knowing the source of the data can aid researchers and analysts in assessing the data's strengths and limitations and in providing a more accurate and insightful interpretation of the findings.

The data of the stock market price from NASDAQ is generally considered to be reliable. One of the biggest stock exchanges in the world, NASDAQ offers a trading platform for a variety of equities and other securities [13]. To maintain the quality and integrity of the trading data, the exchange uses cutting-edge technology and processes and has strict regulations and standards for the businesses and securities that are listed on it. Additionally, NASDAQ offers its own data products and services, including the NASDAQ Data Store, which provides access to a range of historical and real-time market data sets [13].



Date	Close/Last	Volume	Open	High	Low
04/11/2023	$160.80	47,644,220	$162.35	$162.36	$160.51
04/10/2023	$162.03	47,716,880	$161.42	$162.03	$160.08
04/06/2023	$164.66	45,390,120	$162.43	$164.9584	$162.00
04/05/2023	$163.76	51,511,740	$164.74	$165.05	$161.80
04/04/2023	$165.63	46,278,300	$166.595	$166.84	$165.11
04/03/2023	$166.17	56,976,190	$164.27	$166.29	$164.22
03/31/2023	$164.90	68,749,790	$162.44	$165.00	$161.91
03/30/2023	$162.36	49,501,690	$161.53	$162.47	$161.271
03/29/2023	$160.77	51,305,690	$159.37	$161.05	$159.35
03/28/2023	$157.65	45,992,150	$157.97	$158.49	$155.98
03/27/2023	$158.28	52,390,270	$159.94	$160.77	$157.87
03/24/2023	$160.25	59,256,340	$158.86	$160.34	$157.85
03/23/2023	$158.93	67,622,060	$158.83	$161.5501	$157.68
03/22/2023	$157.83	75,701,810	$159.30	$162.14	$157.81
03/21/2023	$159.28	73,938,290	$157.32	$159.40	$156.54
03/20/2023	$157.40	73,641,420	$155.07	$157.82	$154.15
03/17/2023	$155.00	98,944,630	$156.08	$156.74	$154.28

Figure 3: Interface of website of NASDAQ [14].

To collect the data, it can be access through the NASDAQ website. Searching the desired company's price stock data and add on NASDAQ as keyword on Google will lead to the website of the data. The data can then be downloaded as a CSV file. The website can also set the time of the data according to the users' needs, such as one year, five years, and full period.

3.2.2 Train & Test Data

In machine learning, the performance of a model is compared using train and test data. The dataset was split into two subsets primarily for the purpose of training the model on one subset and evaluating its performance on the other. The model's generalizability can be assessed—its ability to apply to fresh, untested data—through this procedure. The training dataset is used to fit the model's parameters and learn patterns from the input data. The testing dataset, on the other hand, is used to evaluate the model's performance by comparing its predictions against the actual output [15].

It is crucial to divide time-series data into training and testing sets while taking into account the fact that the data are temporal. Using a "rolling window" method, which uses more recent data for testing and older data for training, is one popular strategy [16]. Following are the steps for utilizing the rolling window approach to divide time-series data into training and testing sets:

First is to establish the rolling window size. The rolling window size controls the number of data points utilized for training and testing. Utilizing a window size of 20–30% of the entire data is a typical strategy [16].

Next, create training and test sets from the data by using the initial chunk of the time-series data as the training set and the rest as the testing set to get started. The first 80 data points of a data set with 100 total data points, for instance, would be utilized for training, and the final 20 data points would be used for testing.

3.2.2.1 LSTM (Long Short-term Memory)

The time-series data is divided into a training set and a test set in order to apply train and test data to an LSTM model. A greater chunk of the data is used for training and a smaller amount for testing. Typically, 70–80% of the data is used for training, and the remaining 20–30% is used for testing.

The data is pre-processed for the LSTM model once it has been separated. In order to let LSTM understand the data, the data must be scaled, time windows must be made, and the data must be transformed into a 3D format.

After pre-processing, an appropriate optimizer and loss function are being used to train the LSTM model on the training dataset. To enhance the model's performance, its hyperparameters are adjusted, such as the number of hidden layers, the number of neurons in each layer, and the number of time steps in each window.

After the model has been trained, its effectiveness is measured using the test dataset. To compare the predicted values with the actual values and assess how well the model is functioning on unobserved data, I used a variety of evaluation measures, such as mean squared error (MSE) or root mean squared error (RMSE).

3.2.2.2 Linear Regresion

To apply train and test data on a linear regression model, we also need to split the data into a training set and a test set.

The pre-processed data for the linear regression model once it has been divided. Because the data must be converted into a format that the linear regression model can comprehend. After pre-processing, the linear regression model is trained on the training dataset. To enhance the model's performance.

After the model has been trained, its effectiveness is tested by using the test dataset. To compare the predicted values with the actual values and assess how well the model is functioning on unobserved data, we can use a variety of evaluation measures, such as mean squared error (MSE) or root mean squared error (RMSE).

CHAPTER 4

RESULT AND DISCUSSION

4.1 Result

Upon the completion of the project, there will be two outcomes to report. Firstly, the EDA results will be presented in the form of graphs and charts to aid in better comprehension of the data. Secondly, the machine learning result will be the evaluation metrics and comparative charts to show the performance of the model in predicting stock prices in comparison to actual stock prices.

Exploratory data analysis, or EDA, is the process of looking at and analyzing data to find patterns, connections, and trends [12]. EDA aims to improve data comprehension and utilize that understanding to guide the choice of suitable statistical techniques for additional analysis.

Data transformations, descriptive statistics, and visualizations are only a few of the methodologies and techniques used in EDA. Making graphs and plots to graphically display the data, such as histograms, scatterplots, and boxplots, is a frequent EDA approach. Data patterns and trends that may not be obvious from numerical summaries alone can be found using these visualizations. Another important aspect of EDA is data cleaning and pre-processing [12]. The quality and validity of statistical models may be impacted by missing values, outliers, and other anomalies in the data, which must be identified and handled. EDA aids in ensuring that the data used for analysis are accurate and devoid of biases or errors.

For machine learning, the result is typically a predictive model that can be used to make predictions or decisions based on new data. The algorithm chosen, the features used to train the model, the quantity and caliber of the training data, and the hyperparameters of the model are all variables that affect the model's quality. Depending on the problem and the type of data, the model's output can be assessed using a variety of performance metrics.

4.2 EDA (Exploratory Data Analysis)

```
DatetimeIndex: 2516 entries, 2013-03-04 to 2023-02-28
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Close/Last  2516 non-null   object
 1   Volume      2516 non-null   int64
 2   Open        2516 non-null   object
 3   High        2516 non-null   object
 4   Low         2516 non-null   object
dtypes: int64(1), object(4)
memory usage: 117.9+ KB
```

Figure 4: Metadata of the raw data.

Figure 4 reveals that some columns, such as “Close/Last”, “Open”, “High” and ‘Low”, are in object format and thus cannot be used for calculations. To address this issue, the data is converted into float format, which enables Python to treat the data as an integer with decimal points.

```
Out[32]: Close/Last    float64
         Open          float64
         High          float64
         Low           float64
         RSI           float64
         EMAF          float64
         EMAM          float64
         EMAS          float64
         Target        float64
         dtype: object
```

Figure 5: Metadata of the data after transformation.

After the necessary conversion from object to float, the dataset is now in a suitable format for further analysis. In addition to the conversion new calculated variables are added into the dataset to provide additional information for the model to learn from. These variables are all calculated using the raw data and can help to capture important patterns or trends in the stock market data. For instance, the RSI variable is calculated using the “Close/Last” data, which provides a measure of the stock’s strength and momentum.

Table 1: Data dictionary.

Variables	**Type**	**Description**
Close/Last	Float	The last price at which a stock trades during a regular trading session.
Open	Float	The first price at which a stock trades during a regular trading session.
High	Float	The highest price at which a stock trades during a regular trading session.
Low	Float	The lowest price at which a stock trades during a regular trading session.
RSI	Float	Relative strength index
EMAF	Float	Exponential moving average fast.
EMAM	Float	Exponential moving average medium.
EMAS	Float	Exponential moving average slow.
Target	Float	The close price of the next day. The value we are trying to predict.

[17]

Out[18]:

	Close/Last	Open	High	Low	RSI	EMAF	EMAM	EMAS	Target
count	2366.000000	2366.000000	2366.000000	2366.000000	2366.000000	2366.000000	2366.000000	2366.000000	2366.000000
mean	67.422620	67.373748	68.144076	66.636706	55.446731	66.894882	64.730922	63.313071	67.477626
std	49.338553	49.297715	49.959213	48.660287	12.497875	49.044053	47.957957	47.104044	49.355155
min	17.176400	17.280700	17.307100	17.081400	22.304008	17.195509	16.526686	15.987272	17.176400
25%	28.533425	28.505000	28.849375	28.315625	46.348006	28.340299	27.907821	27.621345	28.544025
50%	43.750000	43.736250	43.965000	43.472500	55.662196	43.325848	44.035493	43.662638	43.751250
75%	117.232500	117.627500	119.120000	115.995000	64.284670	116.663273	105.810540	98.887157	117.335000
max	182.010000	182.630000	182.940000	179.120000	89.779292	174.825981	165.738864	162.090712	182.010000

Figure 6: Statical information of the data.

According to Figure 6, The data can be conclude that it is suitable for analysis. Firstly, all columns have the same number of data points, indicating that there are no missing values. Typically, the stock prices of opening, closing, highest, and lowest prices do not vary much due to the stability of large companies such as Apple. Therefore, it is expected to see similar values for the minimum, maximum, and mean prices. The mean, minimum, and maximum values appear to be normal, indicating that there are likely no outliers in the data.

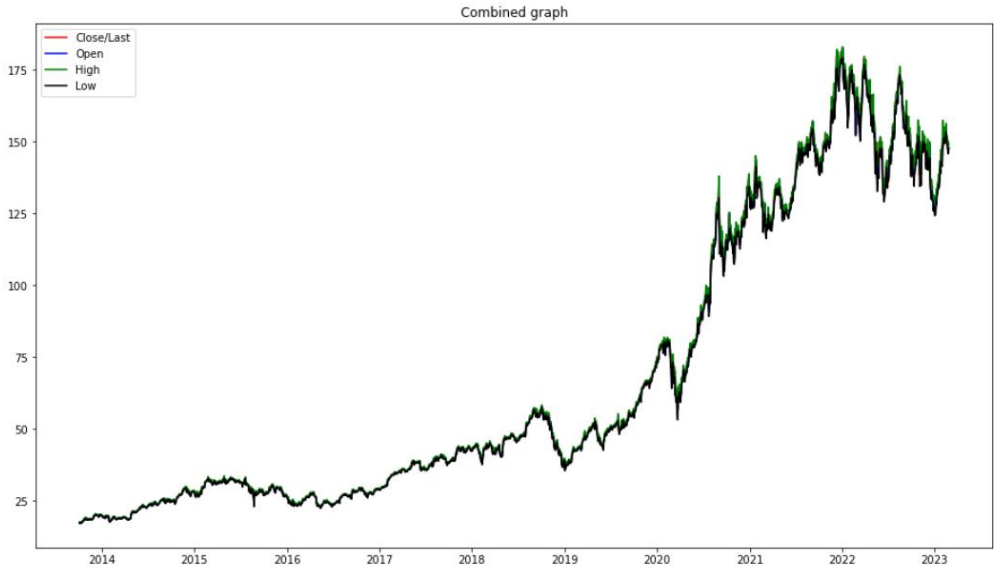


Figure 7: Line graph of prices of Apple company for the past 10 years

`The graph displays the stock price on the x-axis and the date on the y-axis, with different types of prices represented by various colours. However, due to the large amount of data and the small differences between prices, it is difficult to distinguish between the colours on the graph. The overlapping of the lines also makes it difficult to interpret the information accurately.

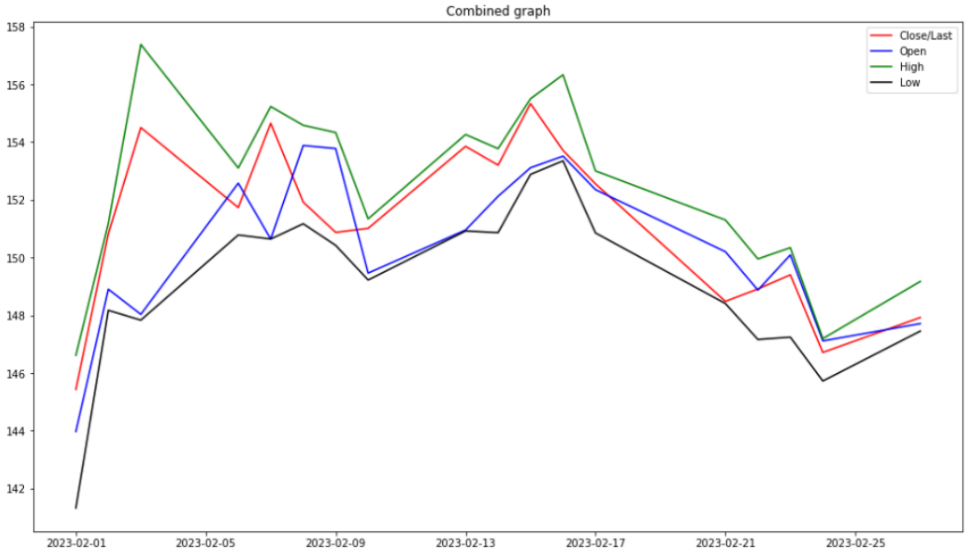


Figure 8: Line graph of prices of Apple company for the month of February.

Figure 8 is a line graph similar to Figure 7, but it only shows data for a month. The purpose of Figure 8 is to verify the normality of the prices. The graph shows that the prices vary from day to day. This supports our previous assumptions based on Figure 7. The overlapping of line and 10 years of time period causes the unclear line graph in Figure 8.

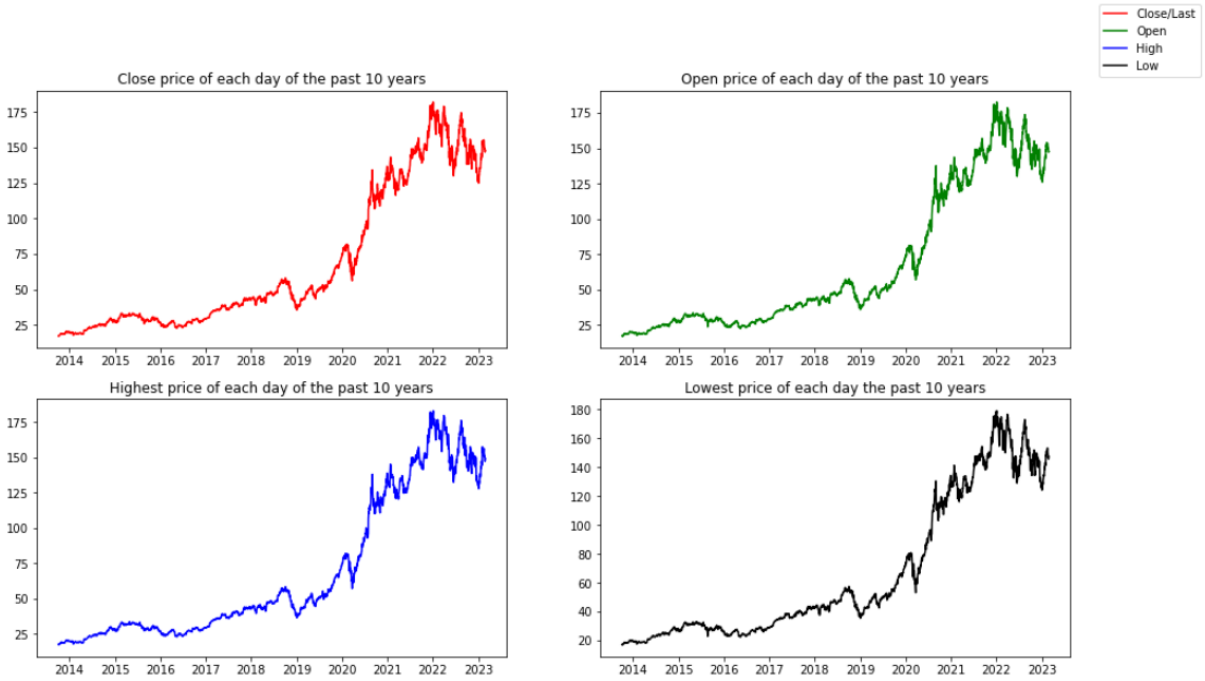


Figure 9: Line graphs the separated based on type of prices for 10 year.

Figure 9 provides a comprehensive view of the Apple company's stock prices. Prior to 2020, the stock prices of Apple fluctuated between 25 and 75. However, between 2020 and 2021, there was a significant surge in Apple's stock prices, which have remained steady in the 100 to 175 range since then. This growth can be attributed to the increased demand for technology due to the pandemic, which acted as a catalyst for people to embrace the digital environment. Apple has benefited from this situation because of the products and services they provide.

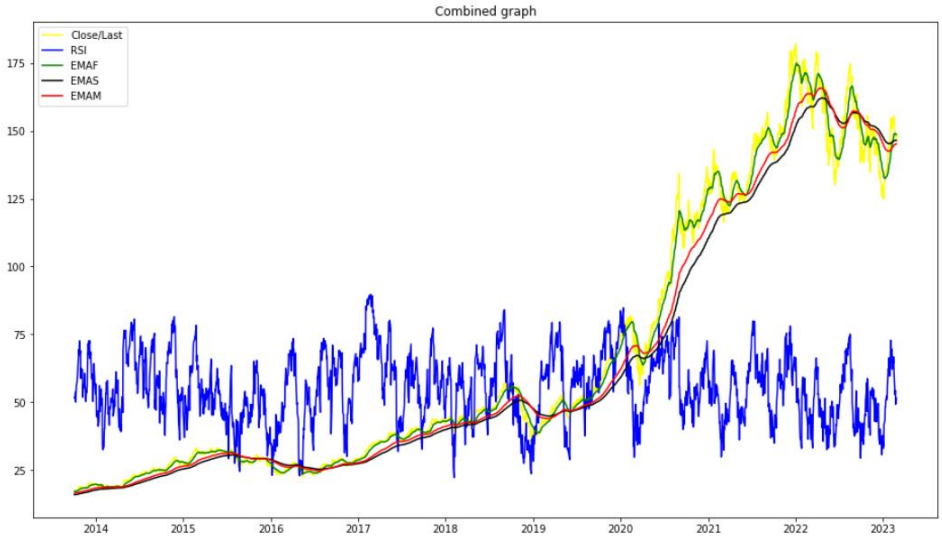


Figure 10: Line graph of the indicators and the close price for the past 10 years.

In Figure 10, the x-axis represents the time period, and the y-axis represents the value. The yellow line representing the closing price is closely aligned with the lines for EMAF, EMAS, and EMAM. This is expected because the EMA represents the average of the stock price over a period. Therefore, if the stock price is increasing, the EMA will show a similar trend. There are no issues with the line graphs for the EMAs. As for the RSI indicator, it stays within the range of 0 to 100, which is normal since the RSI is normalized to this range.

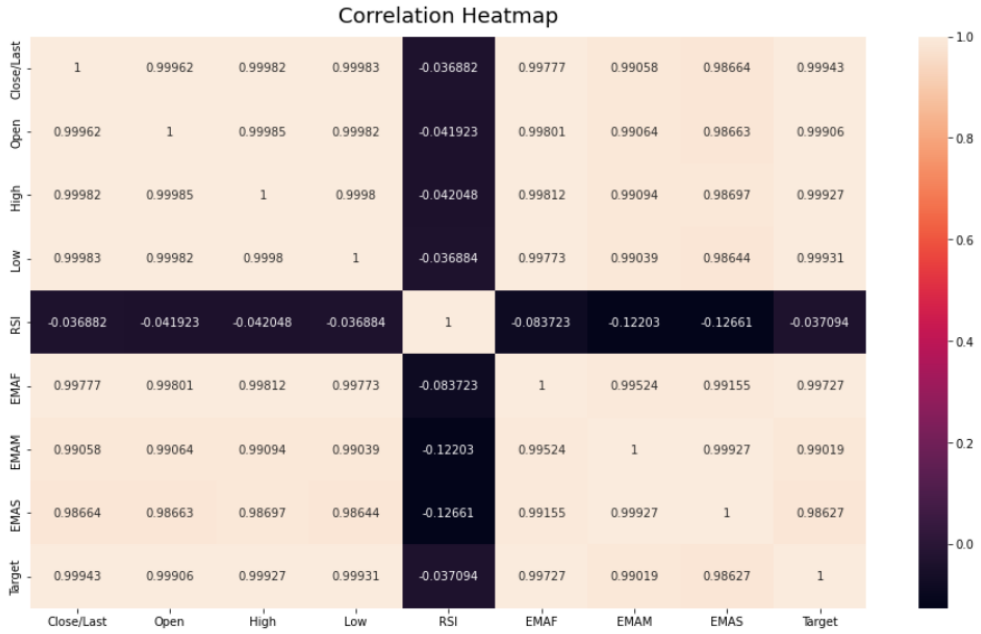


Figure 11: Correlation heatmap between all variables.

To interpret Figure 11, the colour of each square is important and needs to be focused on. The darker the colour, the stronger the correlation between the variables. The heatmap reveals that most variables in the dataset are highly correlated with each other, except for the RSI variable. This is expected because the stock prices have small differences, and RSI is normalized to a range of 1 to 100, limiting its correlation with other variables.

4.3 Evaluation Metrics

At the end of this project, graphs of the predicted and actual value of the stock price are produced. It can better visualize the differences between the value.

4.3.1 LSTM (Long Short-term Memory)

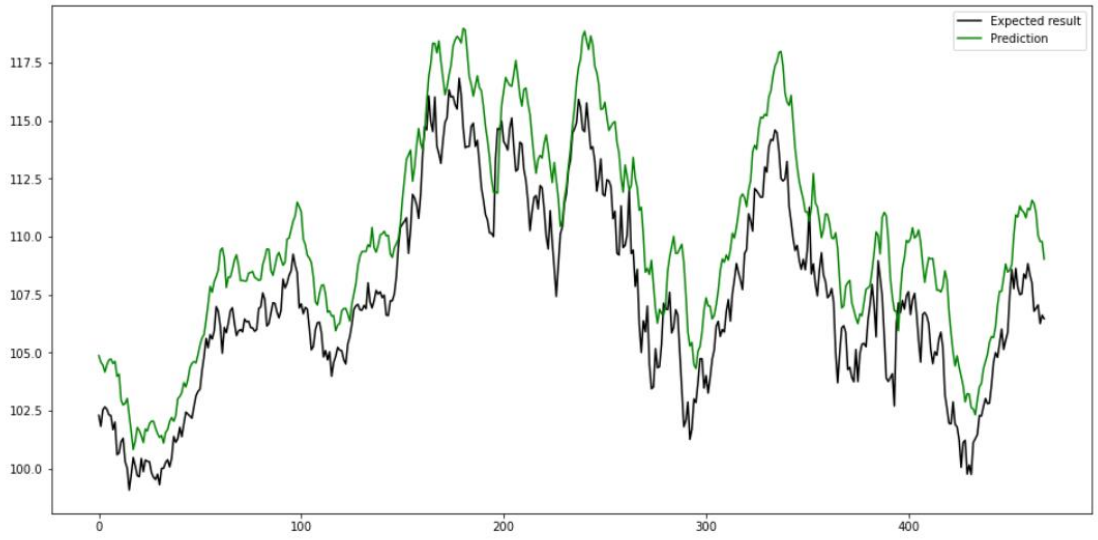


Figure 12: Combined line graph of predicted and actual value of stock price (LSTM)

Figure 12 shows the prediction value from LSTM model and the actual value of stock price. Figure 12 displays the red line that shows the predicted values of the stock price, while the black line indicates the actual value. The predicted values deviate from the actual value by a considerable amount. However, this issue can be resolved by subtracting a fixed value from the predicted stock price. Nonetheless, the pattern of the predicted line closely follows the actual line, capturing most of the ups and downs, which is a positive sign. Nonetheless, this trained model might be subject to overfitting.

When a statistical model is overly complicated and fits the noise in the data rather than the underlying pattern, the phenomenon of overfitting takes place [18]. In other words, the model doesn't generalize well to new, unforeseen data since it is too precisely tuned to the unique data that was used to train it. The model may perform poorly in terms of prediction because of overfitting, and conclusions may be incorrect or misleading.

4.3.2 Linear Regression

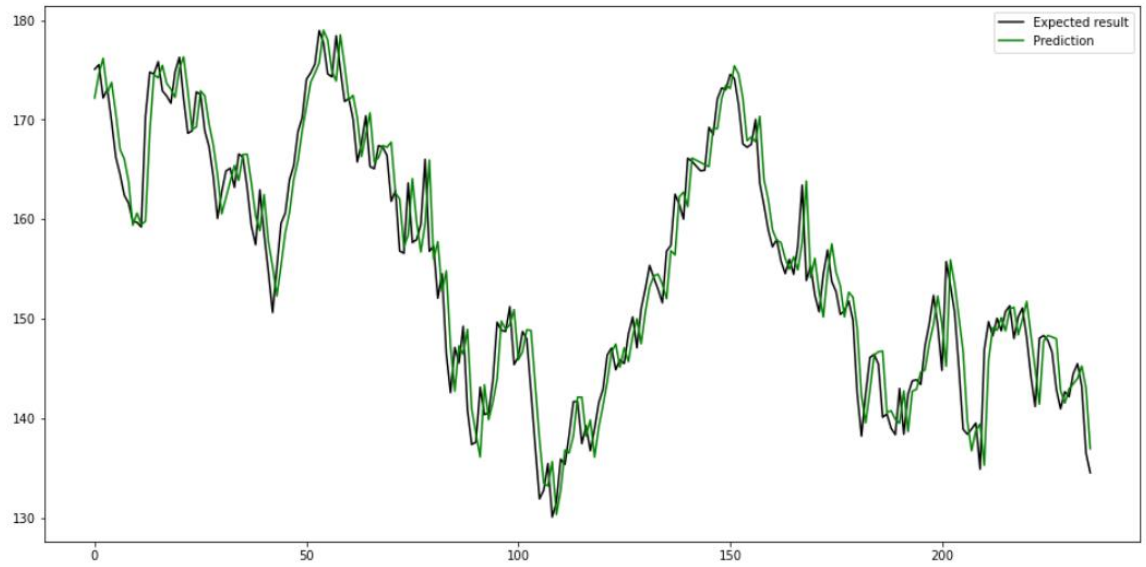


Figure 13: Combined line graph of predicted and actual value of stock price (Linear Regression)

Figure 13 displays the actual stock price values and the predicted values from the Linear Regression model. The red line and black line are almost perfectly aligned, which indicates that the predicted values closely match the actual values. The model is able to capture the pattern of the ups and downs of the stock price. However, we still need to be cautious of the possibility of overfitting.

Comparing these two models, it is difficult to determine which one is better without using evaluation metrics. We can use formulas to calculate evaluation metrics based on the predicted and actual values of the stock price to determine the best model. In this project, MAE and RSME will be used as the main metrics to evaluate.

Table 2: Evaluation table for the two models.

Evaluation metrics/Model	LSTM	Linear Regression
MAE	2.51	3.49
RSME	2.81	2.76

The LSTM model performs better than the linear regression model according to the MAE and RSME evaluation criteria. The lower the MAE and RSME, the better the model performance. In this case, the LSTM model has an MAE of 2.51 and RSME of 2.81, while the linear regression model has an MAE of 3.49 and RSME of 2.76.

The MAE calculation averages the absolute differences between the actual and anticipated values. Without taking into account their direction, it calculates the average magnitude of the errors in a series of forecasts. A lower MAE value denotes greater model performance, and a value of 0 denotes no error between the anticipated and actual data [19].

The average of the squared discrepancies between the expected and actual values is known as the root mean square error, or RMSE. It is comparable to MAE but places more emphasis on larger errors. RMSE is more sensitive to outliers than MAE since it penalises greater mistakes more severely. An improved performance of the model is indicated by a decreased RMSE number, similar to MAE. There is no difference between the expected and actual values, as shown by a value of 0.

The MAE for the LSTM model is significantly lower than that of the linear regression model, indicating that the LSTM model has a smaller average error in predicting the stock price. The RSME for the LSTM model is slightly higher than that of the linear regression model, but the difference is not substantial.

Overall, based on these evaluation measures, the conclusion can be stated that the LSTM model outperforms the linear regression model at forecasting Apple stock price. It is crucial to remember that there might be additional considerations to make when selecting a model, such as the model's complexity and interpretability.

CHAPTER 5

CONCLUSION

In conclusion, after implementing and comparing both LSTM and linear regression models on the dataset, it was found that LSTM outperforms linear regression in terms of evaluation metrics. However, it was also observed that both models perform well and are capable of predicting the stock market price accurately. In addition, the familiarity of the user with the models plays a crucial role in the selection of the best model, as users can input extra parameters to hyper tune the models for better results. It highly improves the performance of the models based on the parameters. Furthermore, both models have a potential overfitting issue, which should be considered and addressed in future iterations. Overall, the study highlights the importance of careful consideration of the model selection process and the need for regular evaluation and tuning to ensure accurate and reliable predictions of stock market prices.

The end of the report, LSTM outperforms Linear Regression in this Apple stock price prediction comparison.

REFERENCES

[1] J. Chen, "What Is the Stock Market, What Does It Do, and How Does It Work?," Investopedia, 7 7 2022. [Online]. Available: https://www.investopedia.com/terms/s/stockmarket.asp. [Accessed 13 2 2023].

[2] A. Hayes, "Stocks: What They Are, Main Types, How They Differ From Bonds," Investopedia, 6 7 2022. [Online]. Available: https://www.investopedia.com/terms/s/stock.asp. [Accessed 13 2 2023].

[3] D. R.Harper, "Forces that move stock prices," Investopedia, 22 7 2022. [Online]. Available: https://www.investopedia.com/articles/basics/04/100804.asp. [Accessed 7 4 2023].

[4] G. L. Team, "What is Machine Learning? Defination, Types, Applications, and more," Great Learning, 7 2 2023. [Online]. Available: https://www.mygreatlearning.com/blog/what-is-machine-learning/. [Accessed 7 4 2023].

[5] A. Biswal, "Stock Price Prediction Using Machine Learning: An Easy Guide!," simplilearn, 4 4 2023. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learning. [Accessed 7 4 2023].

[6] Colah, "Understanding LSTM Networks," colah's blog, 27 8 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 10 4 2023].

[7] V. Kanade, "What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022," spicework, 3 4 2023. [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/. [Accessed 10 4 2023].

[8] A. Wilson, "Stock Prediction Using Linear Regression," Medium, 6 12 2020. [Online]. Available: https://medium.com/analytics-vidhya/stock-prediction-using-linear-regression-cd1d8351f536. [Accessed 10 4 2023].

[9] D. Madhugiri, "Linear Regression in Machine Learning: A Comprehensive Guide," knowledgehut, 6 1 2023. [Online]. Available: https://www.knowledgehut.com/blog/data-science/linear-regression-for-machine-learning. [Accessed 10 4 2023].

[10] T. i. team, "7 Technical Indicators to Build a Trading Toolkit," Investopedia, 31 3 2023. [Online]. Available: https://www.investopedia.com/top-7-technical-analysis-tools-4773275. [Accessed 10 4 2023].

[11] A. Hayes, "What Nasdaq Is, History, and Financial Performance," Investopedia, 5 5 2022. [Online]. Available: https://www.investopedia.com/terms/n/nasdaq.asp. [Accessed 11 4 2023].

[12] P. Patil, "What is Exploratory Data Analysis?," Medium, 24 3 2018. [Online]. Available: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15. [Accessed 12 4 2023].

[13] A. Hayes, "What Nasdaq Is, History, and Financial Performance," Investopedia, 5 5 2022. [Online]. Available: https://www.investopedia.com/terms/n/nasdaq.asp. [Accessed 26 4 2023].

[14] "AAPL Historical Data," NASDAQ, 1 5 2023. [Online]. Available: https://www.nasdaq.com/market-activity/stocks/aapl/historical. [Accessed 2 5 2023].

[15] Minewiskan and TimeShererWithAquent, "Training and Testing Data Sets," Microsoft, 12 10 2022. [Online]. Available: https://learn.microsoft.com/en-us/analysis-services/data-mining/training-and-testing-data-sets?view=asallproducts-allversions. [Accessed 26 4 2023].

[16] S. Yildirim, "Time Series Analysis: Resampling, Shifting and Rolling," Medium, 15 4 2020. [Online]. Available: https://towardsdatascience.com/time-series-analysis-resampling-shifting-and-rolling-f5664ddef77e. [Accessed 26 4 2023].

[17] L. Smigel, "What Is Open High Low Close in Stocks?," Analyzing Alpha, 24 6 2022. [Online]. Available: https://analyzingalpha.com/open-high-low-close-stocks. [Accessed 26 4 2023].

[18] C. Team, "Overfitting," CFI, 7 12 2022. [Online]. Available: https://corporatefinanceinstitute.com/resources/data-science/overfitting/. [Accessed 26 4 2023].

[19] A. Chugh, "MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?," Analytics Vidhya, 8 12 2020. [Online]. Available: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e. [Accessed 30 4 2023].

[20] "Stock Price Prediction using Machine Learning with Source Code," Project pro, 2 2 2023. [Online]. Available: https://www.projectpro.io/article/stock-price-prediction-using-machine-learning-project/571#:~:text=Machine%20learning%20models%20such%20as,%2C%20of%20course%2C%20stock%20prices.. [Accessed 13 2 2023].

[21] J. Maverick, "Most Commonly-Used Periods in Creating Moving Average (MA) Lines," Investopedia, 30 6 2021. [Online]. Available: https://www.investopedia.com/ask/answers/122414/what-are-most-common-periods-used-creating-moving-average-ma-lines.asp. [Accessed 10 4 2023].