STOCK MARKET PREDICTION

USING MACHINE LEARNING TECHNIQUES

By

MARIYA ELIZABETH THEROKKATTIL

Submitted to

The University of Roehampton

In partial fulfilment of the requirements

MASTER OF SCIENCE IN DATA SCIENCE

ABSTRACT

This research investigates the use of machine learning models, particularly Long Short-Term Memory (LSTM) and Linear Regression, in the context of stock market forecasting. The study compares the effectiveness of these models in predicting stock prices for three important sportswear companies: Adidas, Nike, and Puma, using historical stock price data provided from Yahoo Finance. Data preparation, model development and testing, and result interpretation are all part of the study. Metrics like Mean Squared Error(MSE), Root Mean Squared Error(RMSE), and R-squared(R2) are used to assess the prediction efficacy of both regression models. As a significant resource for investors and financial analysts looking for data-driven tools for decision-making in the turbulent world of stock trading, the project intends to give insights about the efficacy of these models for stock market prediction.

DECLARATION

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

MARIYA ELIZABETH THEROKKATTIL Signed
Date : 06/09/2023

ACKNOWLEDGEMENTS

I want to extend my sincere appreciation to everyone who helped this project be completed successfully. Your assistance, direction, and cooperation were crucial in making this project a success.

I want to start by expressing my gratitude to **Dr. Mohammad Farhan Khan** for their steadfast support and guidance during this project. Their knowledge and insightful opinions were crucial in determining the project's course and guaranteeing its success.

I appreciate the **University of Roehampton** for providing the tools, space, and funds needed to do this study. It is admirable that you continue to invest in innovation and the spread of information.

I also want to express my gratitude to my friends and family for their unfailing help and patience throughout the challenging parts of this undertaking. Your support and tolerance were quite helpful.

Finally, I want to thank everyone who participated in the study, filled out the survey, or helped with the data gathering and research. Your openness to take part and provide your ideas was essential to the project's success.

I really appreciate your contributions to our adventure. Your contributions have improved this project and provided motivation for more work in the future.

TABLE OF CONTENTS

1. Introduction

The stock market, usually referred to as the equity market, is a market where people and institutions can purchase and sell shares in publicly traded corporations. These shares give investors the chance to own a piece of the business and partake in its expansion and success. The stock market is essential to the economy because it promotes capital formation and makes it possible for companies to raise money for growth and operation. The stock market operates through stock exchanges, which are platforms where buyers and sellers come together to trade stocks. The NASDAQ, London Stock Exchange, New York Stock Exchange (NYSE), and Tokyo Stock Exchange (TSE) are a few of the well-known stock exchanges. Brokerage accounts, which serve as a middleman between investors and stock exchanges, allow investors to purchase and sell stocks. When you purchase a stock, you essentially become a stakeholder in the business, and you have the chance to profit from your investment through dividends and capital appreciation (the rise in stock price). To buy, sell, or hold stocks with the greatest potential for profit or the least amount of loss, one must make educated selections.

1.1 Research Question or Problem statement

Problem statement: In order to help investors make wise investment decisions, the issue being investigated in this study is the precise forecasting of stock market price changes.

Individual and institutional stock market investors are also impacted by the issue. In order to maximise the performance of their portfolios, financial experts, traders, and investment managers are also impacted by the difficulty of making stock price predictions. Due to the financial markets' dynamic nature, where stock values change based on different factors such as economic statistics, company performance, geopolitical events, and investor emotions, the problem persists in the stock market constantly. Accurate forecasts give investors the ability to spot stocks with potential for growth and better chances to maximise

their investment profits. By foreseeing price changes, investors can efficiently manage risk by avoiding equities that could see substantial falls or by discovering potential hedging options.

Making informed and data-driven judgements is made easier for investors when predictions are accurate since they give them significant market insights. Informed investors can make reasonable judgements, decreasing market volatility and the possibility of irrational exuberance or panic selling. As a result, better stock market predictions help to stabilise the market.

1.2 Aims

Ø Improving the institution's ability to make investment decisions is one of the main goals of applying machine learning for stock market prediction. In order to enable better informed and data-driven investment decisions, the institution uses machine learning algorithms to uncover patterns, trends, and hidden linkages in historical market data.

Ø Institutions hope to more effectively manage the risks involved with their investments by employing machine learning for stock market prediction. Machine learning models can aid in identifying potential negative risks, allowing the institution to take the necessary precautions to safeguard its portfolio during market downturns.

Ø Machine learning algorithms have the capacity to handle massive amounts of data in real-time, giving institutions quick access to market dynamics and trends. As a result, decisions may be made more swiftly and can be adjusted to rapidly changing market conditions.

1.3 Objectives

The main objective of this project is to help the investors to take appropriate decisions for the following scenarios:

Ø Investors try to make better decisions about purchasing or selling stocks by forecasting changes in stock prices. They can use predictions to find prospective investment opportunities that could prove successful or to decide when to abandon a position to limit losses.

- Investors are better able to evaluate and control the risks connected with their investments when they are aware of prospective price swings. They can modify their portfolios to match their risk appetite and investing objectives with the help of predictions.

- Some investors timing their entry and exit into the market or particular equities by using stock market predictions. The objective is to maximise returns by purchasing stocks at a discount and selling them for a higher price.

- Investors can utilise stock market forecasts to execute hedging techniques that safeguard their portfolios against probable market declines or unfavourable situations.

- Predictions can help with investment portfolio optimization by pointing out assets with anticipated good performance or low correlations to lower overall portfolio risk.

- The stock market has a large impact on the whole economy, and precise forecasts can result in more effective capital allocation, which promotes economic stability and growth.

1.4.Legal, ethical, social and professional ethics

Recently, there has been a lot of promise in the application of machine learning techniques to the field of stock market forecasting. It is crucial to understand that this project has significant legal, social, ethical, and professional ramifications that call for serious thought in addition to its technical aspects.

1.4.1.Legal considerations

Data Compliance: Strict adherence to data protection laws and financial rules is required when using financial data for predictive models. Failure to comply with this might have negative legal effects and harm the project's image.

Intellectual property rights: Respecting and acknowledging intellectual property rights helps to ensure that proprietary tools, data, and algorithms are utilised in a way that complies with the law.

1.4.2.Social consideration

Algorithmic Fairness: To make sure that the technology does not increase societal inequities or discriminate against particular groups, concerns of algorithmic bias and fairness must be taken into consideration while developing predictive models.

Technology that is accessible: To avoid putting obstacles in the way of access for people or organisations with low resources, accessibility and inclusion should be at the centre of this project.

1.4.3.Ethical consideration

Transparency and Accountability: It's crucial to maintain moral transparency in the model-development process and accountability for results. It is ethically required to explain forecasts in detail and to accept responsibility for the results.
Responsible Prediction Use: To avoid taking any acts that would endanger people or the stability of financial markets, ethical concerns should drive decision-making. The welfare of stakeholders must come first..

1.4.4. Professional consideration

Adherence to Professional standards: Maintaining the integrity of the project and promoting ethical behaviour require adhering to professional standards of conduct, such as those set by pertinent industry groups or associations.
Continuous Learning and Peer Review: Participating in peer review with the academic and professional community and remaining current on the most recent advancements in both machine learning and finance shows a dedication to professionalism and the responsible growth of knowledge.

While using machine learning to anticipate the stock market is an exciting endeavour, it is crucial to approach it with a clear understanding of the professional, social, ethical, and legal implications. Ignoring these factors can have negative effects on one's reputation and legal standing, as well as perpetuate injustices and hurt both people and society as a whole. To guarantee that its contributions to the area are not only technically good but also socially and morally sound, this initiative is devoted to navigating these issues responsibly and ethically.

1.5 Background

A major and developing topic, machine learning-based stock market prediction has deep ramifications for a number of industries, including banking, investments, and even more general parts of the global economy.
Financial Markets and Their Impact: By enabling resource allocation and asset pricing, financial markets play a crucial role in the contemporary economy. For individuals and institutions, accurate forecasts in these markets can result in large financial profits or losses. Consequently, increasing our capacity to predict stock prices and market patterns has long been a priority in the world of finance.
Accurate stock market forecasting has a number of consequences and uses, including the following:

- Investment Strategies: To make wise judgements, allocate resources, and optimise portfolios, investors and asset managers might employ prediction models.

- Risk management: By being aware of market trends and potential dangers, institutions may create risk-reduction plans and safeguard their assets.
- Economic Indicators: The performance of the stock market is frequently used as an economic indicator. Predicting market patterns can give information about more general economic developments.
- Machine learning models are essential to algorithmic trading, in which automated computers place trades based on up-to-date information and prognostic indications.

2. Literature - Technology Review

2.1 Literature Review

Accurate forecasting has always been difficult due to the stock market's complexity and dynamic nature. However, researchers have tried to increase prediction accuracy with the development of machine learning algorithms, notably LSTM and linear regression. This study of the literature intends to investigate and contrast several research articles that use linear regression and LSTM to forecast the stock market.

Article	Technologies Used	Details	reference

A LSTM-Method for Bitcoin Price Prediction: A Case Study Yahoo Finance Stock Market	RNN and LSTM	In the study, it is discussed how to forecast the behaviour of the Bitcoin stock market using LSTM , a kind of (RNN). The research considers if an automated prediction tool is necessary given how volatile cryptocurrencies are. Recurrent modules are used by LSTM for analysis, just like RNN. The process entails making predictions about Bitcoin values that will surpass $12,600 USD in the near future using tools and procedures, maybe using data from Yahoo Finance.	[1]
Stock market forecasting using machine learning algorithms	SVM	This study provides a revolutionary stock market prediction algorithm that takes use of temporal connections between various financial products and stock markets throughout the world. In order to predict the stock patterns for the following day, support vector machines (SVM) and a number of regression methods are used. With forecast accuracies of 74.4% for the NASDAQ, 76% for the S&P 500, and 77.6% for the DJIA, the figures are outstanding. The study also involves the creation of a straightforward trading model to evaluate the	[2]

		algorithm's performance in comparison to other approaches.	
Stock Market Trend Prediction with Sentiment Analysis Based on LSTM, Neural Network	LSTM	This article proposes a unique technique to stock market prediction by fusing LSTM and neural networks with sentiment analysis. The authors' goal is to increase forecast accuracy by taking into account both the long-term dependencies recorded by LSTM and exogenous inputs, such as market sentiment gathered from news articles and social media. The main contribution of this study is the creation of a hybrid model that incorporates sentiment analysis, LSTM, and neural networks. This strategy underlines the significance of taking into account both historical patterns and current market sentiment in order to make more accurate predictions while also opening up new opportunities for developing stock market prediction models. To solve the issues and constraints of this strategy, nevertheless, more investigation and testing are required.	[3]

Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation	NN, SVM,MDA, CBR	The paper surveys machine learning techniques used for stock market prediction, emphasising the challenges in predicting financial time series. It discusses recent advancements in stock market prediction models, highlighting their pros and cons. The study also explores the impact of global events on stock market prediction and underscores the significance of incorporating event information for accuracy. It suggests the need for precise event weighting methods and reliable automated event extraction systems to enhance financial time series prediction.	[4]

A machine learning model for stock market prediction	LS-SVM, PSO, LM	The Least Square Support Vector Machine (LS-SVM) method and Particle Swarm Optimisation (PSO) algorithm are combined to present a machine learning strategy for stock market price prediction in this study. In order to improve prediction accuracy by preventing overfitting and local minima problems, LS-SVM parameters are optimised using PSO. The model uses technical indicators and historical stock data to beat an artificial neural network using the Levenberg-Marquardt (LM) algorithm when evaluated on thirteen benchmark financial datasets, proving the viability of the suggested strategy for stock market forecasting.	

Deep learning for stock market prediction	LSTM,RNN,ANN, XGBoost, Adaboost	The primary goal of this essay is to forecast the future values of stock market groups, in this case, four groups from the Tehran Stock Exchange. These predictions are based on ten years of historical data and are made using a variety of machine learning algorithms, such as long short-term memory (LSTM), recurrent neural networks (RNN), artificial neural networks (ANN), bagging, random forest, adaptive boosting (Adaboost), gradient boosting, and eXtreme gradient boosting (XGBoost). Ten technical indicators are used as input data for the models, and four metrics are used to assess the performance. While tree-based models such as Adaboost, Gradient Boosting, and XGBoost all perform well, LSTM stands out as the technique with the best accuracy and model fitting capabilities.	[6]

Stock market prediction with high accuracy using machine learning techniques	Linear Regression and LSTM	to predict stock market prices with high accuracy by combining LSTM and linear regression. The model may take use of the advantages of both strategies by combining these techniques. The suggested model offers a potential strategy to take on the difficult issue of stock market prediction with high accuracy by utilising the advantages of these machine learning approaches. To investigate potential improvements and modify the model for scenarios involving real-time prediction, more study is necessary.	[7]
Stock Closing Price Prediction using Machine Learning Techniques	ANN, RMSE, MAPE	This study employs Artificial Neural Network and Random Forest, two artificial intelligence tools, to forecast stock market returns. To build input variables for the models, it uses financial data including Open, High, Low, and Close prices. The efficiency of these AI-based approaches in correctly predicting the closing stock prices of the next day for five businesses across several sectors is demonstrated by the study's evaluation of model performance using common measures like RMSE and MAPE.	[8]

Stock Price Forecasting Using Data from Yahoo Finance and Analysing Seasonal and Nonseasonal Trend	Time Series Analysis, ARIMA, Holt Winter	The goal of the study is to reduce the inherent risks involved with stock trading by investigating stock price prediction using a combination of time series analytic models, including ARIMA and Holt Winter. It aims to offer a safe investment range and improve forecast accuracy for both short and long time periods in the stock market by establishing a link between stock prices and these algorithms.	[9]
Stock Market Price Prediction Using Machine Learning Techniques	LSTM,RNN	The study and forecasting of real-time stock market data using machine learning techniques is the main topic of this research. Due to the inherent difficulties in anticipating stock market changes, the researchers developed an automated computational technique to handle the issue of predicting stock prices. They used a dataset taken from Yahoo Finance for a comparison study utilising ISTM and RNN. Their machine learning methodology showed better accuracy in stock market prediction compared to conventional approaches, and this dataset was refreshed every day.	[10]

Future study might concentrate on combining more sophisticated machine learning approaches, investigating ensemble methods, and examining the effect of additional external data sources on forecast accuracy because the stock market is very dynamic and affected by a variety of factors.

2.2 Technology Review

Uncertainties and constantly shifting trends characterise the complicated and dynamic stock market. The complex correlations and patterns seen in financial data are difficult to capture using traditional techniques to stock market forecasting. However, machine learning has come to light as a potential approach to deal with these issues. Machine learning techniques, such as supervised learning, time series analysis, recurrent neural networks (RNNs), convolutional neural networks (CNNs), natural language processing (NLP), reinforcement learning, ensembling, and transfer learning, are being used more frequently to analyse historical stock price data, technical indicators, and sentiment data from financial news and social media. Algorithms like Linear Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, Gradient Boosting, and Neural Networks are all included in supervised learning. Methods used in time series analysis include exponential smoothing, prophet, seasonal autoregressive integrated moving average (SARIMA), and autoregressive integrated moving average (ARIMA). Sequential data processing is a specialty of recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). CNNs have been used to interpret stock market data that is provided in formats that resemble images. Textual data including financial news and sentiment are analysed using NLP approaches include Sentiment Analysis utilising Bag of Words, Word Embeddings, and Transformer-based models (e.g., BERT, GPT). For algorithmic trading techniques, Proximal Policy Optimisation (PPO) and Deep Q-Networks (DQN) are used, together with reinforcement learning. Each technology provides distinct skills to glean insightful information and generate forecasts based on previous market behaviour.

This technological evaluation focuses on two crucial methodologies: LSTM from RNN and linear regression from supervised learning. Algorithms like Linear Regression, which use labelled historical data to model the connection between predictors (such as technical indicators) and stock prices, are examples of supervised learning. A RNN variation called LSTM is particularly suited for stock price forecasting since it can handle sequential data and has demonstrated promise in capturing long-term relationships in time series.

It is simple to see how different variables affect stock prices with linear regression because of its interpretability and simplicity. It offers an easy-to-understand method for simulating the

linear connection between the goal variable and the input characteristics. For datasets with a manageable number of variables, linear regression is computationally effective and well suited. But its fundamental drawback is that it can't account for the intricate non-linear linkages and temporal dependencies that are inherent in financial data. When dealing with highly erratic and turbulent stock market patterns, the linear assumption may result in less-than-ideal predictions.

For time series prediction, such as predicting stock prices, LSTM models are especially well suited since they are excellent at capturing long-term relationships and temporal patterns in sequential data. They are capable of addressing the irregularity of financial data by handling erratic time intervals and variable sequence lengths. LSTM may identify complex patterns that linear regression would overlook because it can learn from prior data and recognise non-linear correlations. However, LSTM models demand more processing power than linear regression and need more training data. Furthermore, the interpretability of LSTM may be constrained by its black-box structure, making it difficult to comprehend the precise mechanisms that underlie its predictions.

Regression analysis and LSTM in particular show potential in stock market prediction, providing useful insights into price movements and connections with pertinent variables. Before deciding to invest based entirely on forecasts from machine learning, investors must exercise prudence and take into account the inherent volatility of the financial markets.

In the future, machine learning models will continue to be developed and researched in order to produce more reliable and accurate stock market forecasts.

3. Methodology

Because so many elements have yet to be taken into account and because stock market prediction first doesn't seem statistical, it appears to be a complicated task. However, with the right machine learning techniques, it is possible to link old data to new data, teach the computer to learn from it, and train it to make accurate assumptions.

From Yahoo Finance, historical information for the three businesses has been gathered. The dataset contains Nike, Adidas, and Puma statistics for the three years between January 1, 2020, and July 10, 2023. Information regarding the stock, including its High, Low, Open, Close, Adjacent Close, and Volume, is included in the data. The stock has only been extracted for its day-wise closing price.

Although there are several models for machine learning as a whole, this study concentrates on the two that are most crucial and uses them to make predictions.

3.1. Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) is a recurrent neural network created with the goal of identifying long-term relationships and patterns in sequential input. LSTMs are often employed in a number of disciplines, including speech recognition, time series analysis, natural language processing, and stock market forecasting. In the realm of finance, predicting the stock market has always been difficult since so many different factors may affect the daily changes in stock prices. Financial analysts, traders, and investors look for precise models to predict these changes so they may reduce risks and make well-informed decisions. Regression modelling in particular has emerged as a potent method for solving this challenging issue. Long Short-Term Memory (LSTM) and Linear Regression have drawn the most interest among these models. It can be used for both classification and regression problems. Here I choose a regression problem for the prediction.[8]

The input gate, output gate, and forget gate make up a typical LSTM cell. The weights of these gates are learnt, and they choose how much of the most recent data sample should be retained in memory and how much should be forgotten from earlier lessons. Comparing this straightforward structure to the earlier, comparable RNN model is an improvement.

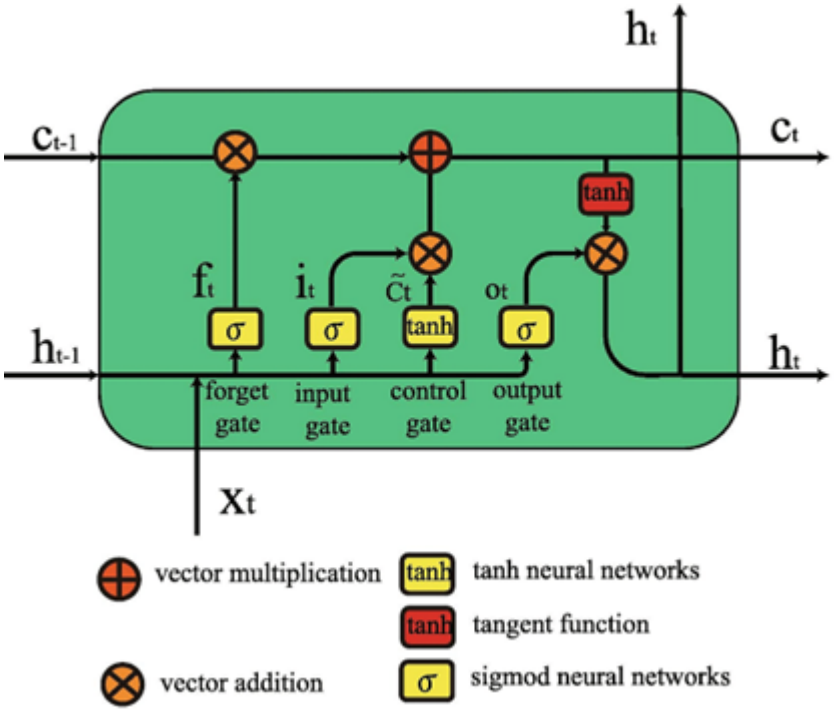


fig1. Representation of LSTM

The forget gate $(f)_t$ decides whether or not to retain some information from the previous cell state (C_{t-1}). It applies a sigmoid activation function to the current input (x_t) and the prior hidden state (h_{t-1}) as inputs, and outputs values between 0 and 1 for each component of the cell state. One signifies "keep this information," whereas zero denotes "discard this information." It is mathematically represented as

$$f_t = (W_i h_{t-1} + U_f x_t + b_f) \quad (1)$$

The input gate makes the decision of which new data should be added to the cell state. It functions similarly to the forget gate in that it accepts the current input (x_t) and the prior hidden state (h_{t-1}) and uses a sigmoid activation function to create values between 0 and 1. These numbers indicate how significant fresh information is for each component of the cell state.

$$i_t = (W_i h_{t-1} + U_i x_t + b_i) \quad (2)$$

The new data that can be included into the cell state is the candidate cell state. It is calculated by using a weighted sum of the prior hidden state (h_{t-1}) and the current input (x_t) and the hyperbolic tangent (tanh) activation function. The resultant numbers, which range from -1 to 1, reflect potential modifications to the cell state.

$$\hat{C}_t = \tanh (W_c h_{t-1} + U_c x_t + b_c) \quad (3)$$

The new candidate cell state ($\hat{C}_t$) and the old cell state (C_{t-1}) combine to create the updated cell state (C_t) in a given system. The input gate (i_t) regulates how much of the candidate cell state is added, while the forget gate (f_t) controls how much of the previous cell state is kept.

$$C_t = (C_{t-1} \otimes f_t) \oplus (\hat{C}_t \otimes i_t) \quad (4)$$

The hidden state (h_t) is the output of the LSTM, and the output gate decides what data from the cell state should be utilised to compute it. It uses a sigmoid activation function on the prior hidden state (h_{t-1}) and the current input (x_t) to obtain values between 0 and 1, which signify the importance of the cell state for the current output.

$$O_t = (W_o h_{t-1} + U_o x_t + b_o) \quad (5)$$

The LSTM cell's output for the current time step is the hidden state (h_t). It is calculated by putting the cell state (C_t) through a hyperbolic tangent activation function (tanh) before applying the output gate (O_t) to it. The numbers between -1 and 1 are scaled with the tanh function.

$$h_t = O_t \ .\tanh(C_t) \quad (6)$$

These equations describe how data moves through an LSTM cell as a whole. The input gate decides what fresh information should be added, the output gate decides what information should be used for the present forecast, and the forget gate governs what information should be forgotten from the past. The values in the cell state and concealed state are controlled by the tanh activation function.[12]

3.2. Linear Regression

One of the fundamental models in statistics and machine learning, noted for its clarity and simplicity, is linear regression. It is predicated on the notion that the characteristics, which are independent variables, and the target, which is the dependent variable, have a linear connection. The association between historical stock prices and pertinent variables, such as trade volume, macroeconomic indices, or company-specific data, is frequently modelled using linear regression in stock market forecasting. In order to predict a continuous target variable based on one or more input data, linear regression is a straightforward yet effective machine learning technique. Linear regression may be used to forecast stock prices in the context of stock market forecasting based on historical data or other pertinent considerations.[11]

One input feature serves as the input in a basic linear regression, which has one target variable as the dependent variable. A linear equation is used to illustrate the connection between the target variable (Y) and the input feature (X):

$Y=\beta_o+\beta_1 X+\epsilon$ (7)

The projected value of Y when X is 0, denoted by the y-intercept (constant) term β_o.

The slope coefficient, β_1, denotes the change in Y for a change in X of one unit.

The error term ϵ denotes the discrepancy between the actual and anticipated values of Y.

The objective of training a linear regression model is to identify the values of β_1 and β_0 that reduce the error term. For training,

First, we calculate the mean of X and Y

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad (8)(9)$$

Then calculate the slope and finally calculate the intercept (β_o)

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})}$$

$$\beta_0 = \overline{Y} - \beta_1.\overline{X} \qquad (10)(11)$$

We may use the trained linear regression model to predict values for additional data points after we know values for β_o and β_1 and the model has been trained. The model may be used to forecast future stock prices once it has been trained.

$Y_{pred} = \beta_o + \beta_1 . X$ (12)

Metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared are used to assess linear regression models. These metrics evaluate the model's ability to predict outcomes accurately and how well it matches the data.

4. Implementation

Important libraries like Tensorflow, Keras, and scikit-learn will be used for the stock prediction. Then the data is loaded directly from the website using the datareader. We may view the daily stock data by going to (https://uk.finance.yahoo.com/) and selecting "Historical Data" from the menu.Historical data of Adidas, Puma and Nike were taken from this website every month between January 2020 and July 2023.. Close prices are utilised as a general metric for the study in this work to create uniformity.

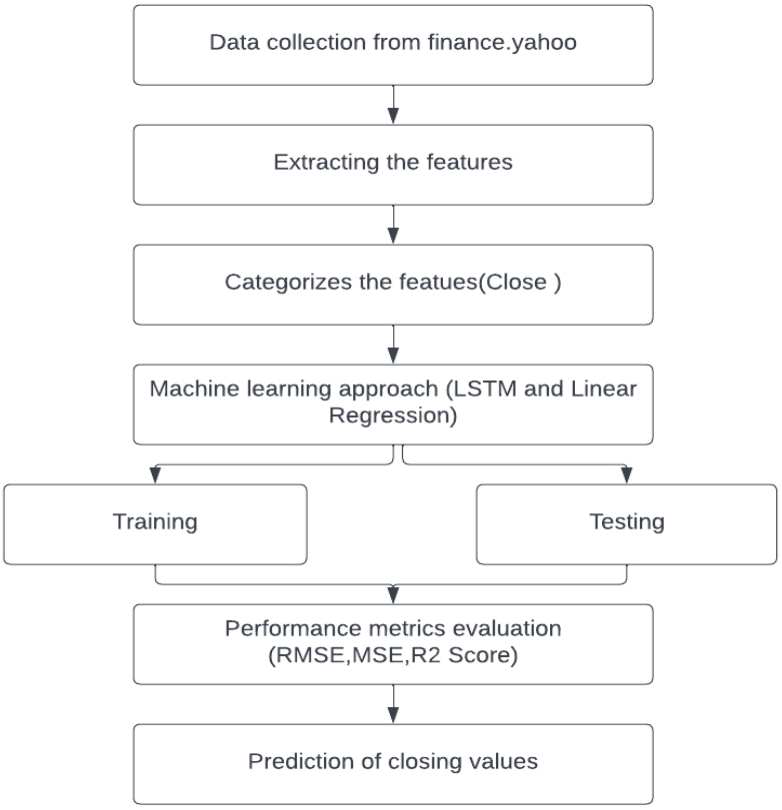


fig2. Steps and Methods for predicting stock market

Using matplotlib for some basic visualisations. Fig3, fig4 and fig5 are showing the close price history of the business firms for the past 3 years. Fig6, fig7, fig8 showing the close price history with the high and low points. Fig9, fig10 and fig11 showing the open and close price of each of the companies.

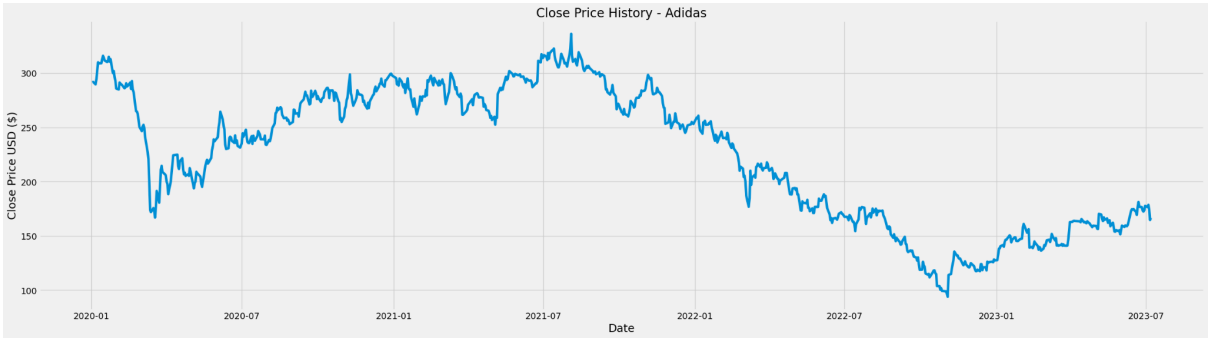


fig3

fig4

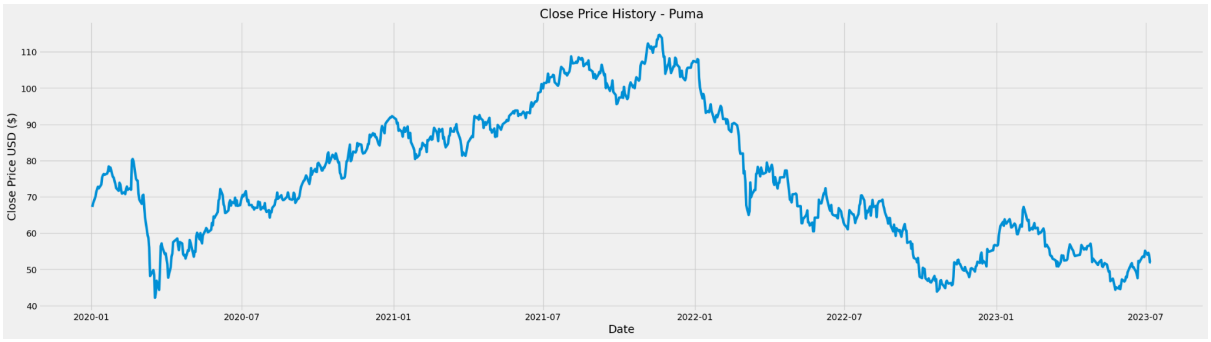


fig5

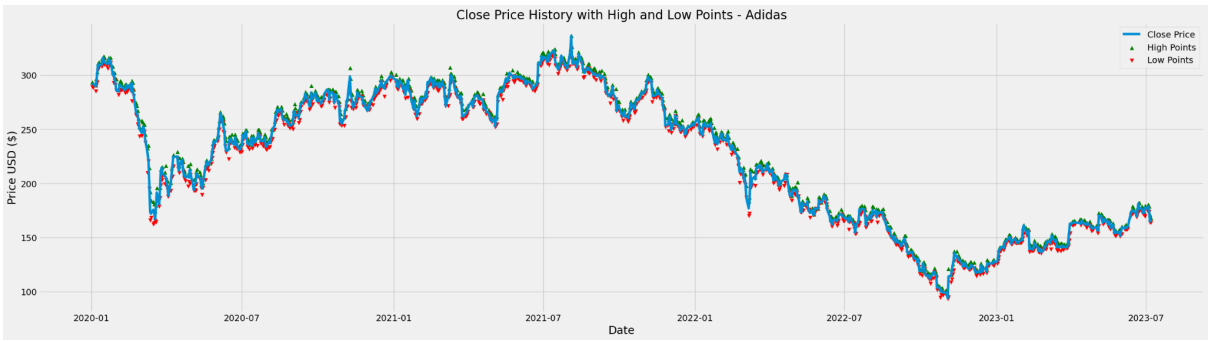


fig6

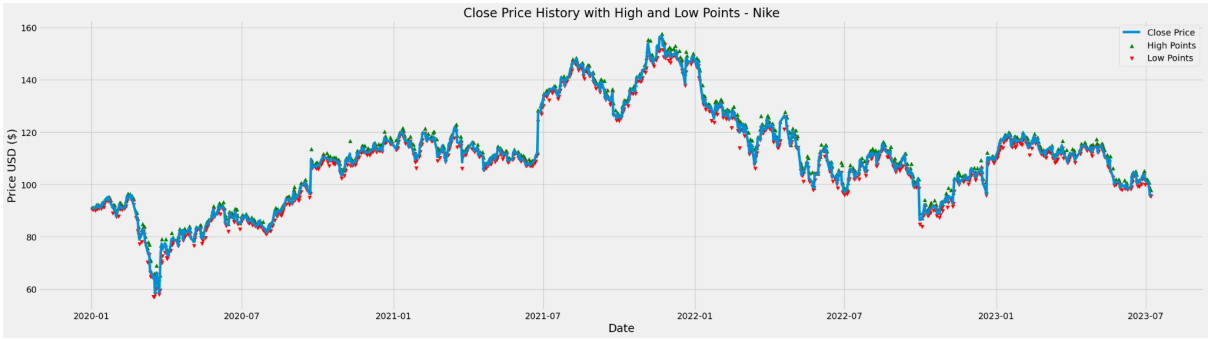


fig7

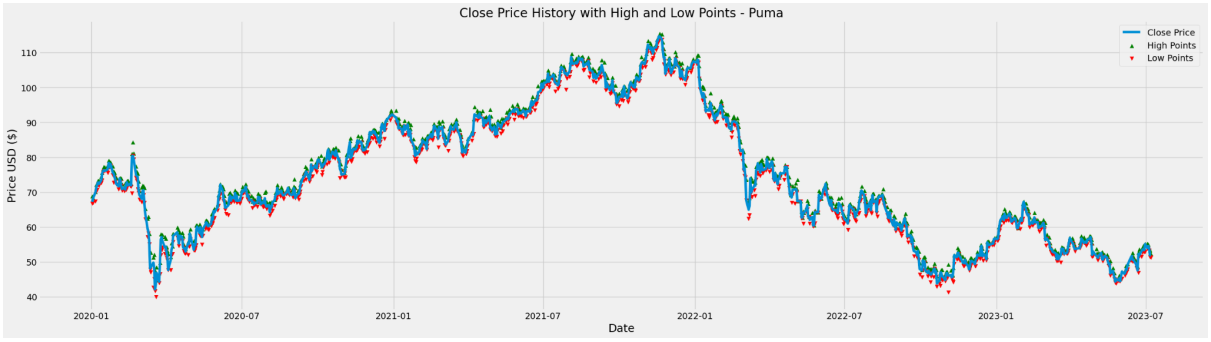


fig8

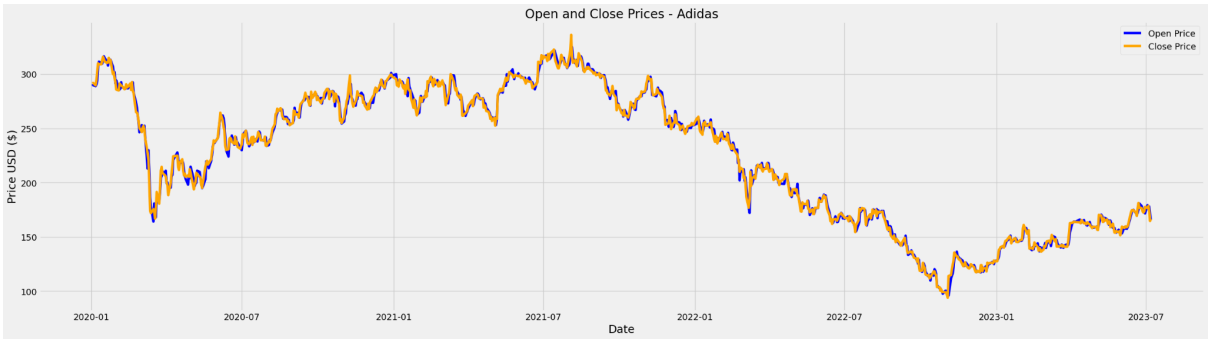


fig9

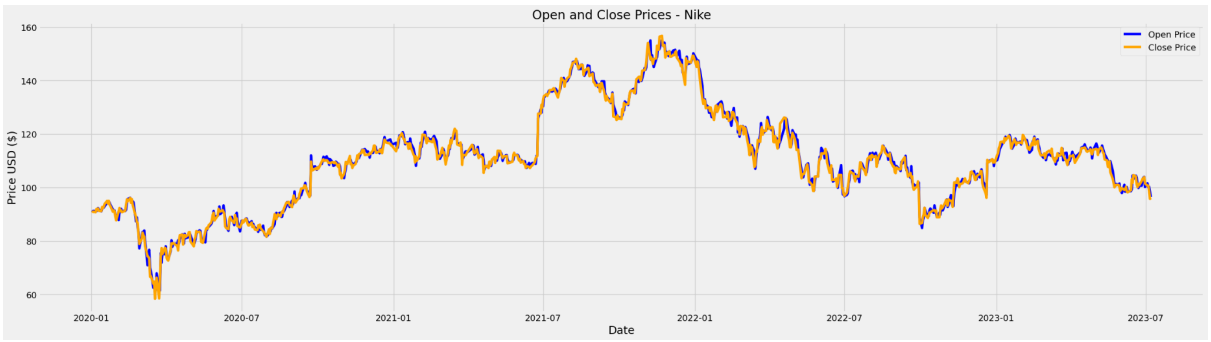


fig10

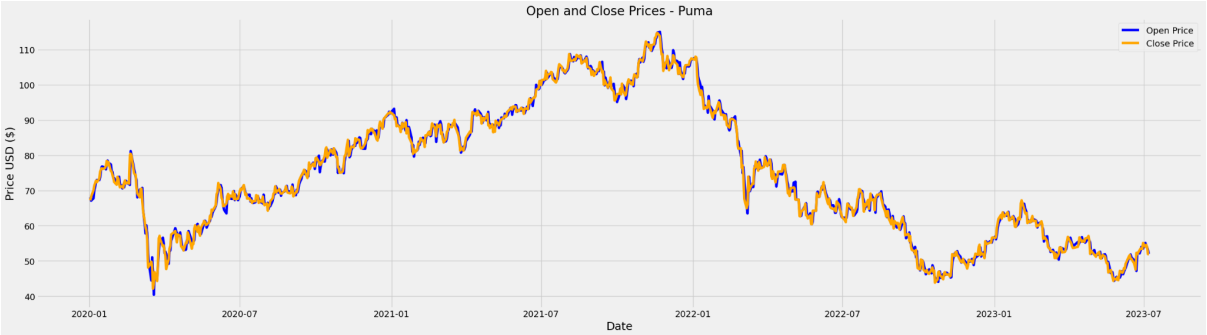


fig11

4.1 Pre-processing

Before utilising the historical stock price data to train and test the machine learning models, a number of preparation processes on the data were done in this code.
For normalisation on the closing price data, Min-Max scaling is used. The data are transformed via min-max scaling into a range between 0 and 1. In order to promote convergence during training, this scaling is crucial for neural network models like the LSTM.

The data is then separated into two parts, training and testing data. To confirm the results of predictions using test data, this fragmentation is required.The testing data is used to assess the performance of machine learning models, while the training data is used to train them.80% of the data were utilised for training, and just 20% for testing. Depending on particular needs, this split ratio may change.

Next step is to produce data sequences for the LSTM model input. Every group of numbers indicates a window of past closing prices.The LSTM model's input sequences (X) and associated target values (y) are created in this stage. Choosing a sequence length (eg., 60 days) to capture past trends in the data.

The input data is reshaped to fit the LSTM model's predicted input form. The reshaping produced a three-dimensional array with dimensions for the features, sample count, and length of the sequence.

After utilising the trained models to make predictions, inverse transformations were used to scale the predicted values back to their original range (in USD). To produce interpretable and significant forecasts in the original data units, this is crucial.

Reorganise the training and test data for the linear regression model to match the input structure. A simpler structure picked for linear regression because it often doesn't need sequence data as LSTM does.

These preparation procedures are necessary to get the data into a format that can be used to train machine learning models. Scaling ensures that all input characteristics are scaled consistently, and LSTM models only use sequence construction and reshaping to identify temporal connections in the data. In order to make predictions understandable in the context of the original data, inverse transformations are used.

4.2 Model Building

Utilising the Keras library to build a sequential model that allows for layer stacking. Two LSTM layers are added to this model.The “return_sequences” option was set to True in the first LSTM layer's code, which had 50 units. With 50 units and “return_sequences" set to False, the second LSTM layer should only provide one output, according to the configuration. The model is then given two Dense layers after the LSTM layers.The first Dense layer included 25 units, whereas the last Dense layer contained just one unit. in which the regression task's output layer is located. The model is then created using the Adam Optimizer and the loss function, mean squared error (MSE), as the issue is a regression one. The model is subsequently trained for 10 epochs using the training data.The model learns to generate predictions based on the input sequences during training.

Fig.12, fig13 and fig14 shows how the training loss of your model changes over training cycles.The goal is to observe if the loss decreases over time, indicating that the model is learning and improving its predictions. Ideally, the loss should decrease and stabilise, but this can vary depending on the complexity of the data and the architecture of the model.

Now we used the LinearRegression class from scikit-learn to build a linear regression model.After reshaping the data is trained for linear regression model.

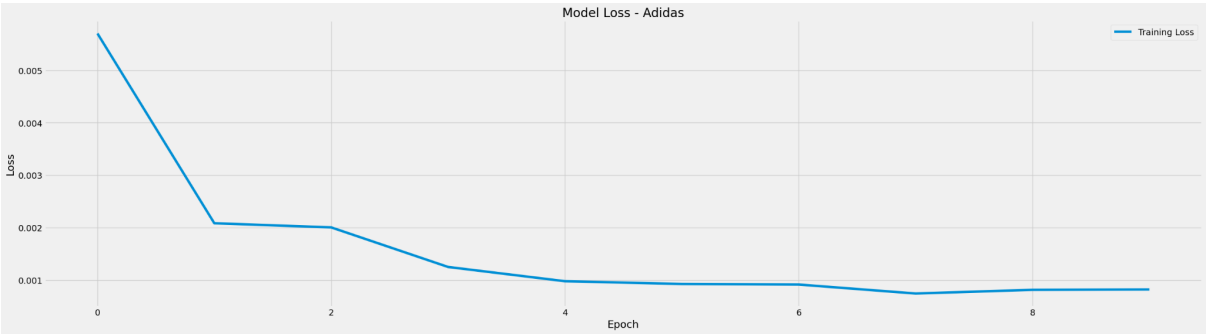


fig12

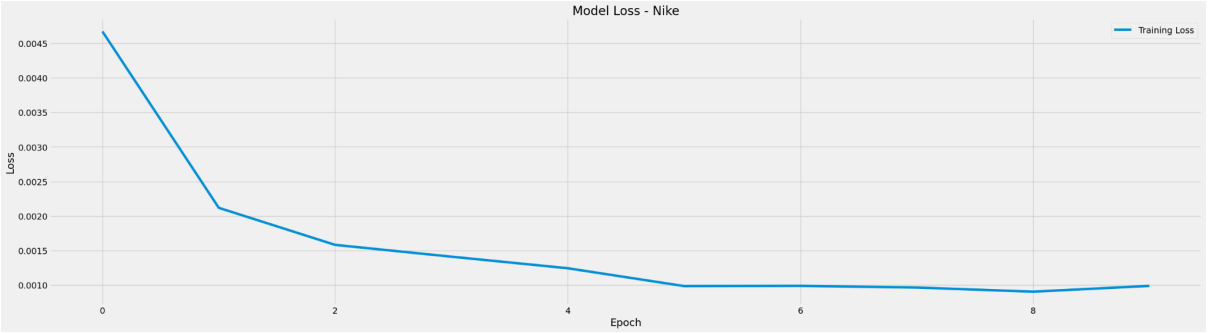


fig13

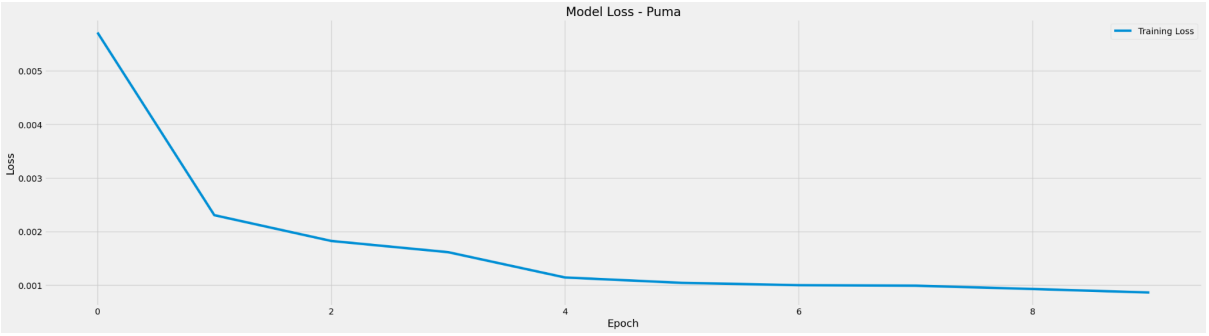


fig14

4.3 Model Evaluation

After creating any model, assessing the model is the most crucial step. For assessing the model, we have many evaluation matrices. However, whether we are addressing a regression problem, a classification problem, or any other sort of problem, the evaluation matrix to be used for the model assessment depends on the problem type. Here both LSTM and Linear Regression model performance are evaluated using the matrices MSE, RMSE and R2-score.

4.3.1 MSE

The Mean Squared Error (MSE) is likely the most straightforward and widely used loss function, and it is frequently included in beginning machine learning classes. The difference between your model's predictions and the actual data is squared, averaged throughout the whole dataset, and then used to determine the MSE. As long as we keep squaring the mistakes, the MSE will never be negative. The equation below serves as the official definition of the MSE:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2 \quad (13)$$

N is the number of samples we are comparing our results against.

Y_i is the actual target value and $\overline{Y}_i$ is the predicted target value.

4.3.2 RMSE

One of the methods most frequently used to assess the accuracy of predictions is root mean square error, also known as root mean square deviation. It illustrates the Euclidean distance between measured true values and predictions.

Calculate the residual (difference between predicted and truth) for each data point, together with its norm, mean, and square root in order to determine the root-mean-square error (RMSE). Due to the fact that it requires and utilises real measurements at each projected data point, RMSE is frequently used in supervised learning applications.

RMSE = $\sqrt{}$ MSE

$$=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2} \quad (14)$$

A lower MSE and RMSE number indicates more accuracy. They impose harsher penalties on bigger mistakes.The RMSE measure may be more useful if you wish to express the size of prediction mistakes in terms that accurately reflect the scale of the target variable.

4.3.3 R2_Score

R-squared is a metric that gauges how well a regression model fits the data. R-square ranges from 0 to 1. R-square is equal to 1 when the model accurately predicts the value and the actual value are identical. However, when the model does not learn any relationships between the dependent and independent variables and does not anticipate any variability in the model, we obtain R-square equal to 0.

$$R2 = 1 - \frac{SSR}{SST} \quad (15)$$

SSR stands for squared residual sum, which represents the variance between observed and anticipated values.

SST, often known as the sum of squares, represents the target variable's overall variance. Where,

$$SST = \sum_{i=1}^{N} (Y_i - \overline{Y}_i)^2 \quad \text{and} \quad SSR = \sum_{i=1}^{N} (Y_{pred} - \overline{Y}_i)^2$$

(16)(17)

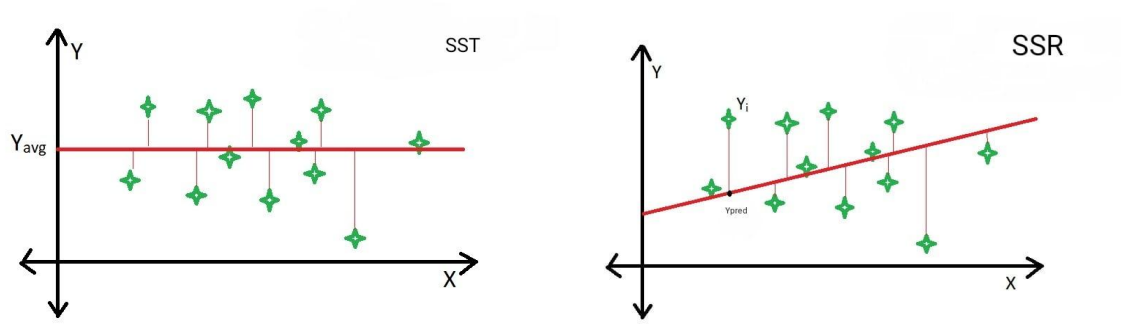


Fig15 :Representation of SST and SSR

A better match between the model and the data is shown by a higher R2 value. Lower MSE and RMSE values denote that the model's forecasts are more closely aligned with the measured data. A greater R2 rating indicates that a greater proportion of the variance in stock prices is explained by the model. The effectiveness of the Linear Regression model for forecasting stock prices may be measured using these measures.

5. Results

The following findings are drawn from the comparison of LSTM and linear regression models for forecasting the stock prices of Adidas, Nike, and Puma:

	LSTM		Linear Regression	
	MSE	**RMSE**	**MSE**	**RMSE**
Adidas	25.114285416252	5.011415510237801	18.217460009134477	4.268191655623547
Nike	4.11655327538944	2.0289290957028143	4.055150528714927	2.0137404323087242
Puma	2.008726721120592	1.4172955659002788	2.138554802766799	1.4623798421637242

Table 1: comparison of machine learning models

For predicting the Adidas stock, the Linear Regression model performs better than the LSTM model. With a Mean Squared Error (MSE) of 18.22 and a matching Root Mean Squared Error (RMSE) of 4.27, it performs more accurately than the LSTM, which has an MSE of 25.11 and an RMSE of 5.01.

In terms of predicting Nike stock values, LSTM and linear regression models both perform equally. The MSE and RMSE values are identical, coming in at 4.11 and 2.03, respectively, indicating similar prediction accuracy.

In terms of Puma stock predictions, the LSTM model has a tiny edge. Compared to the MSE of 2.14 and RMSE of 1.46 of Linear Regression, it obtains a lower MSE of 2.01 and RMSE of 1.42.

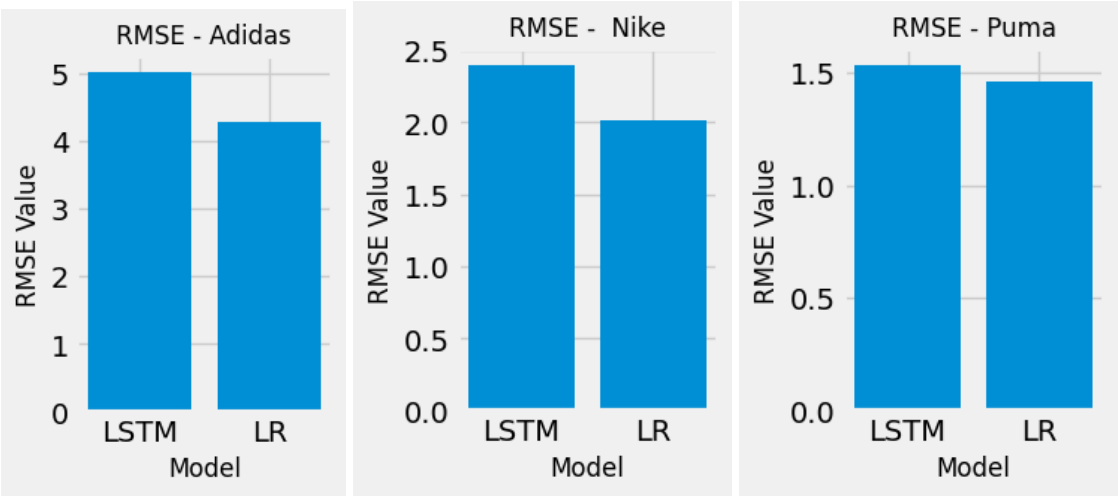


fig16: Comparison of RMSE for the machine learning models.

Typically, there is no clear pattern showing that one model regularly beats the other across all equities. Consider aspects like data qualities and model complexity while deciding between LSTM and linear regression.

The R2-score for both models is about equal, showing that they consistently account for a sizable percentage of the variation. When working with real-world data, it is uncommon for both LSTM and linear regression models to have exactly the same R2 score. The goodness of fit is measured by the R2 score, which evaluates how well the model accounts for variance in the target variable. It's conceivable that the features collected or the data utilised for both models have traits that will cause them to perform similarly. Or both models may provide identical R2 ratings if they are equally effective at detecting trends and patterns in the data. Also getting the same R2 score might be coincidental because randomness has the potential to affect it.

Predicted almost the same closing values using both the linear regression model and the LSTM. That is both the LSTM and Linear Regression models are capable of accurately capturing the underlying patterns and trends in the stock prices, as seen by the similarity in projected closing values between the two models.

The timeliness and accuracy of the models' predictions are displayed in the following graphs(fig17, fig18, fig19) As a result, it may be possible to gain understanding of the models' functionality and capacity for change.

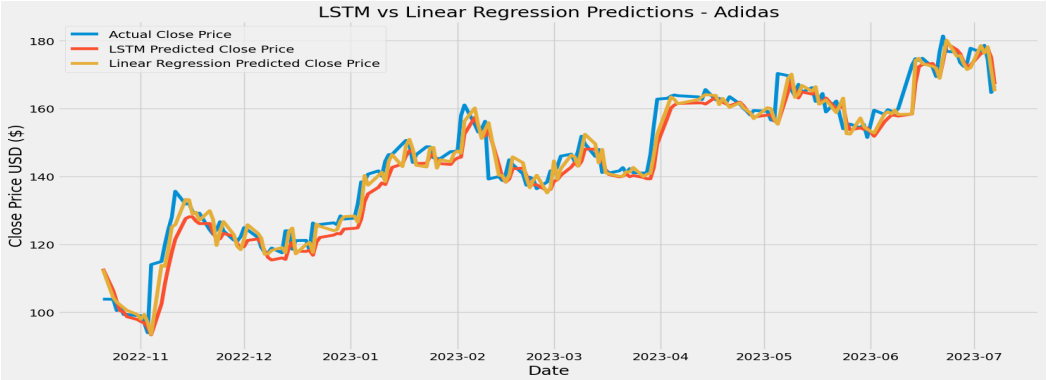


fig17

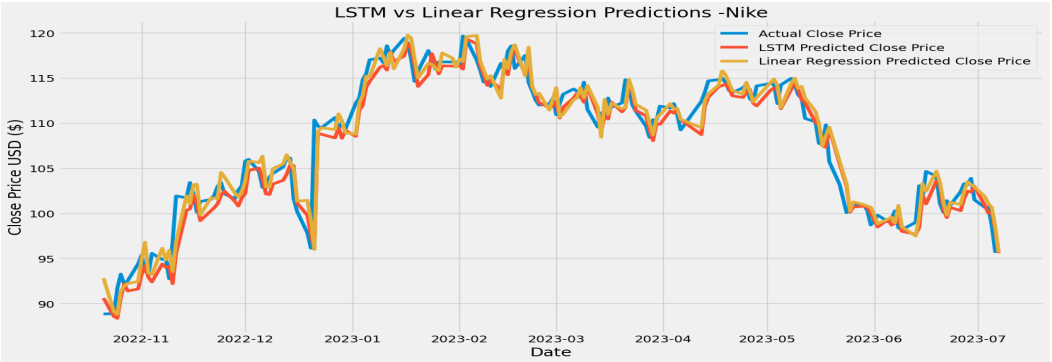


fig18

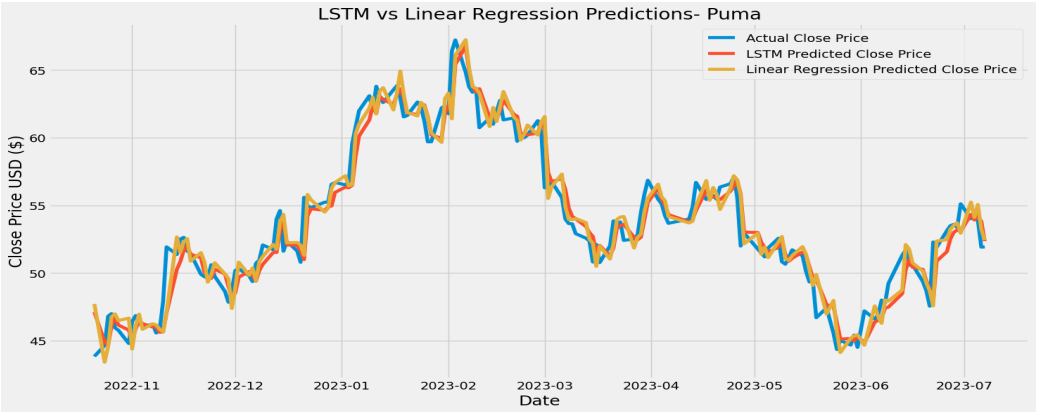


fig19

6. CONCLUSION

It can be difficult to forecast stock market returns since stock values are subject to frequent fluctuations and depend on a variety of factors that create intricate patterns. Only a few parameters, such as high, low, open, close, and adjacent close values of stock prices, as well as volume of shares traded, are included in the historical dataset that is available on the company's website. These features are insufficient. The decision to use LSTM or linear regression should take into account a number of variables, such as the particular stock being forecasted, the features of the data, and the desired balance between model complexity and interpretability. The MSE and RMSE numbers show that both models are capable of making accurate predictions, but the decision may change based on the stock and the particular use case. While LSTM is preferred for Puma stock, Linear Regression performs marginally better for Adidas and Nike.Although there are some little performance variations, both models deliver excellent outcomes with high R2-scores. It may be because of the same data split for both the models. In conclusion, even though both LSTM and Linear Regression models offer reliable and comparable forecasts for the closing prices of the stocks.

6.1.Reflection

This project gave me a great opportunity to practise time series analysis and machine learning. I received knowledge about data preparation, model creation, and performance assessment. The outcomes demonstrated that LSTM and linear regression models can both accurately forecast stock price. Linear Regression fared remarkably well, especially for Adidas and Nike stocks, although the LSTM model showed the potential to catch complex patterns. The project's objectives for creating and analysing stock market forecasting models were met with success. The models' advantages and disadvantages were brought out via comparison. Dealing with time series data that might have noisy and erratic stock market behaviour was one problem. To obtain best performance, fine-tuning and model hyperparameter adjustments were also necessary.

6.2.Future Work

Several directions might be investigated in future research to improve stock market prediction models. First off, using supplemental data sources like macroeconomic indicators and social media sentiment research might give a more complete picture of market dynamics. Second, intricate temporal correlations in stock price data may be captured via the development of more sophisticated deep learning architectures, such as attention-based

models like Transformers. Additionally, improving model resilience would involve adjusting model assessment measures to correspond with particular financial objectives and putting in place a rolling window strategy for continual learning and adaptability to shifting market conditions. For practical implementation in financial markets, real-time trading application deployment considerations and the incorporation of risk management measures are crucial.

A classification problem might be included to improve stock market forecasting and provide insightful information. Classification entails labelling stocks as "buy," "sell," or "hold" rather than attempting to forecast ongoing stock price changes. Given that it offers crystal-clear trading signals, this strategy may help investors make more sensible choices. Future research might investigate the use of feature engineering and sentiment analysis to improve categorization models, supporting traders in determining the best entry and exit positions. A comprehensive trading strategy that optimises both stock selection and timing for portfolio management may also be available by integrating classification and regression algorithms.

7. REFERENCE

[1]. F. Ferdiansyah, S. H. Othman, R. Zahilah Raja Md Radzi, D. Stiawan, Y. Sazaki and U. Ependi, "A LSTM-Method for Bitcoin Price Prediction: A Case Study Yahoo Finance Stock Market," 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), Batam, Indonesia, 2019, pp. 206-210, doi: 10.1109/ICECOS47637.2019.8984499.

[2]. Shen, S., Jiang, H. and Zhang, T., 2012. Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, pp.1-5.

[3]. "Stock market trend prediction with sentiment analysis based on lstm, neural network" by Xu jiawei, Tomohiro Murata

[4]. P. D. Yoo, M. H. Kim and T. Jan, "Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation," International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), Vienna, Austria, 2005, pp. 835-841, doi: 10.1109/CIMCA.2005.1631572.

[5]. Hegazy, O., Soliman, O.S. and Salam, M.A., 2014. A machine learning model for stock market prediction. *arXiv preprint arXiv:1402.7351*.

[6]. Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A. and Salwana, E., 2020. Deep learning for stock market prediction. *Entropy*, *22*(8), p.840.

[7]. "Stock market prediction with high accuracy using machine learning techniques " by Malti Bansal, Apoorva Goyal, Apoorva Choudhary.

[8]. International Conference on Computational Intelligence and Data Science (ICCIDS 2019) “Stock Closing Price Prediction using Machine Learning Techniques”
Mehar Vijha, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar

[9]. J. Jagwani, M. Gupta, H. Sachdeva and A. Singhal, "Stock Price Forecasting Using Data from Yahoo Finance and Analysing Seasonal and Nonseasonal Trend," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 462-467, doi: 10.1109/ICCONS.2018.8663035.

[10]. S. Addagalla, S. Koppuravuri, R. Krosuri, M. S. Kunapareddy, S. Reddy Mallu and M. Rashmi, "Stock Market Price Prediction Using Machine Learning Techniques," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-6, doi: 10.1109/INCET57972.2023.10170222.

[11]. Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. (2017). Stock market prediction using a linear regression. Paper presented at the 2017 international conference of electronics, communication and aerospace technology (ICECA).
[12]. “Predicting Stock Market using Natural Language Preprocessing”by Karlo Puh and Marina Bagic(2023)

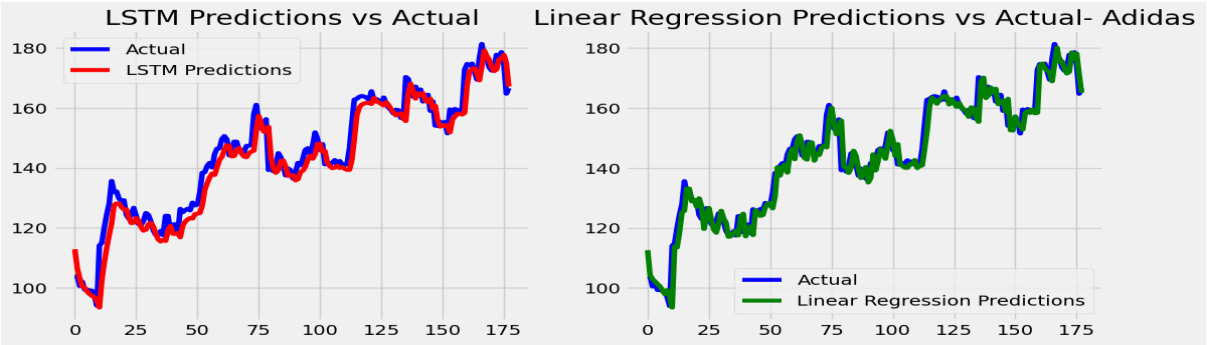
8. Appendices

(a) Predicted Closing Prices for LSTM and Linear Regression Model and their visualisation

1. Adidas

Date	Close	predictions_linear
2022-10-21	103.860001	112.621052
2022-10-24	103.820000	104.285409
2022-10-25	100.500000	102.963901
2022-10-26	102.000000	102.150348
2022-10-27	99.339996	101.304233
...	...	...
2023-07-03	176.619995	178.554706
2023-07-04	178.639999	176.377847

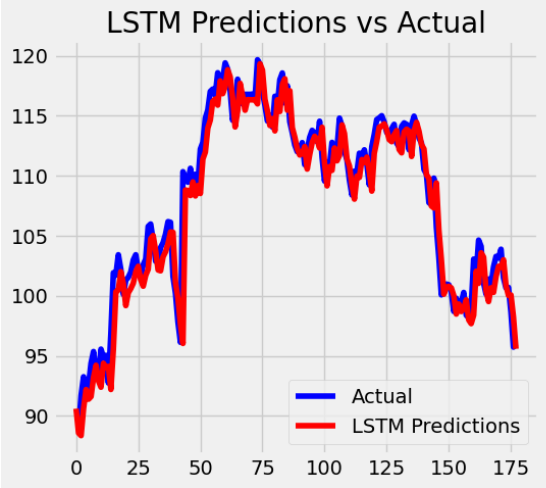
Date	Close	Predictions
2022-10-21	103.860001	112.823608
2022-10-24	103.820000	106.370705
2022-10-25	100.500000	103.162186
2022-10-26	102.000000	100.412125
2022-10-27	99.339996	100.021431
...	...	...
2023-07-03	176.619995	176.226868
2023-07-04	178.639999	176.471008

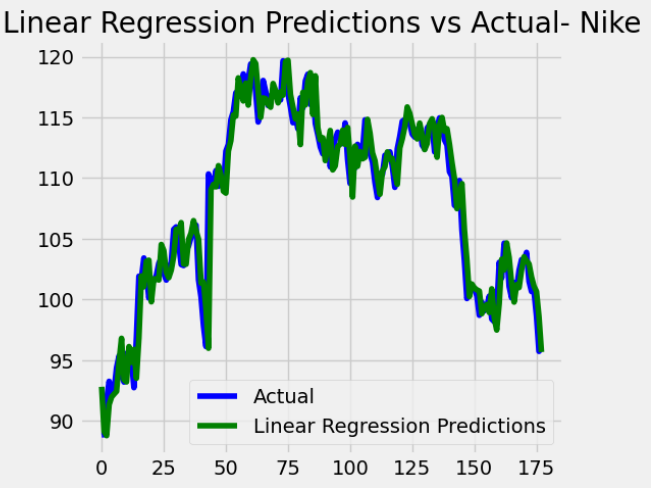


2. Nike

Date	Close	Predictions
2022-10-21	88.849998	90.602486
2022-10-24	88.870003	88.577682
2022-10-25	91.690002	88.355476
2022-10-26	93.260002	90.633430
2022-10-27	92.139999	92.195312
...	...	...
2023-07-03	100.660004	100.863708
2023-07-04	100.699997	100.038902

Date	Close	predictions_linear
2022-10-21	88.849998	92.811320
2022-10-24	88.870003	89.124904
2022-10-25	91.690002	88.787422
2022-10-26	93.260002	91.348183
2022-10-27	92.139999	91.973206
...	...	...
2023-07-03	100.660004	101.861036
2023-07-04	100.699997	101.091432

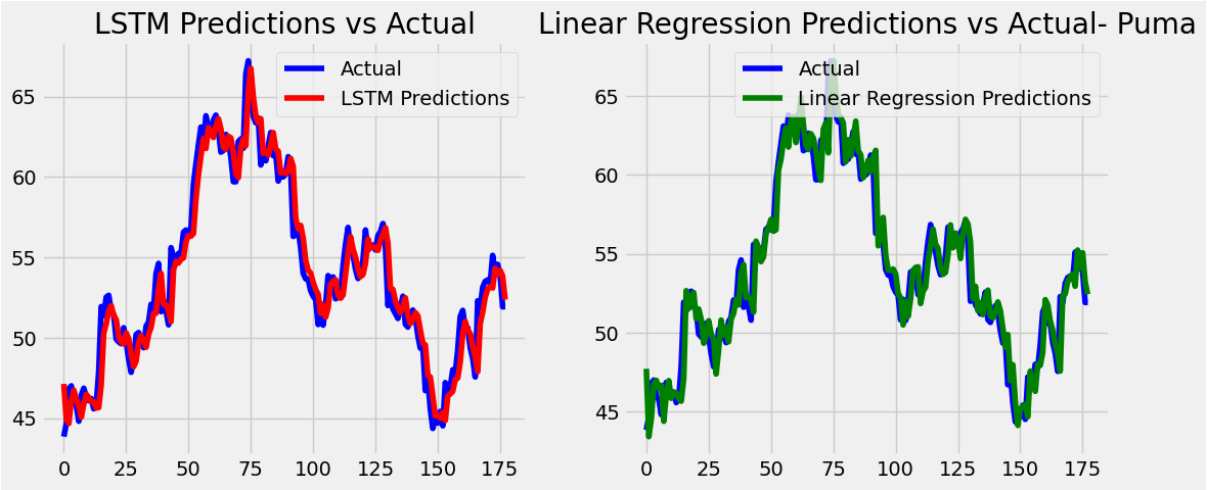




3. Puma

Date	Close	Predictions
2022-10-21	43.849998	47.129089
2022-10-24	44.660000	44.904518
2022-10-25	46.820000	44.672882
2022-10-26	47.009998	46.110310
2022-10-27	45.959999	46.735985
...	...	...
2023-07-03	54.020000	54.329323
2023-07-04	54.560001	53.924725

Date	Close	predictions_linear
2022-10-21	43.849998	47.721759
2022-10-24	44.660000	43.424920
2022-10-25	46.820000	44.580411
2022-10-26	47.009998	46.766609
2022-10-27	45.959999	46.989506
...	...	...
2023-07-03	54.020000	55.246512
2023-07-04	54.560001	54.119887



(b)Project Management link

(c) GitHub

(d) Google collab Adidas

(e) Google collab Nike

(f) Google collab Puma

(e) Google Drive