Analysis of Standard and Poor's (S&P) 500 Index Data And forecasting of the index's closing price

Abeba Bade
Texas A&M University
abeba.bade@tamu.edu

Tobi Oladunjoye
Texas A&M University
oooladu2@tamu.edu

Otis Koranteng Akrasi
Texas A&M University
otis.akrasi@tamu.edu

Abstract:

Predicting stock price market trends, especially the closing price is a challenge even though it has potential rewards [1]. According to the Efficient Market Hypothesis (EMH), markets are efficient and therefore unpredictable. Despite this assumption, researchers point out that the stock market can be predicted accurately with an accurate model and well-chosen variables [2]. In this work we aim to predict the closing price of the Standard Poor (S&P) 500 index utilizing opening prices, closing prices, high and low prices of the index. Using Python and R programming languages, and data analysis libraries we performed a systematic analysis involving data mining , visualization, statistical analysis, model selection and then modeling to choose the best performing model in predicting the closing price of the index. The overall goal is to accurately predict the closing price of the S&P 500 index to provide valuable insights for interested investors and analysts. After examining Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), and linear regression models over three years and 60 days of S&P 500 data with forecast windows of 3, 7, and 14 days, it was found that with a narrow forecast window of 60 days used, it gives a better prediction of closing price. Among the models evaluated, LSTM emerged as the most effective, exhibiting lower MAE, MAPE, and RMSE metrics compared to ARIMA and linear regression

1. Introduction

Forecasting is the process of making predictions of the future based on past and present data. Accurate forecasting of stock market price plays a crucial role in decision-making by providing organizations and investors with valuable insights into future trends and potential outcomes. It is also essential for investors to make informed investment decisions based on the past and present movement of the stock price. The investors are more likely to buy the stocks whose value is expected to increase in the future, and refrain from buying a stock whose value is expected to fall in the future [4]. Hence, an accurate prediction of the stock market index is essential for investors or traders to make profit. However, it is hard to predict the future price of stock exactly and properly due to the highly volatile and nonstationary nature of the market [2]. Hence, investors

have tried to predict stock price movement using technical and quantitative approaches. These methods involve identifying a relevant trend in market data and timing the best time to make an investing choice

Prediction of stock market trends is considered as an important task and is of great attention as predicting stock prices successfully may lead to attractive profits by making proper decisions [1]. The stock market therefore constitutes a pivotal element of the global economic infrastructure, exerting a profound influence on macroeconomic conditions as well as the financial prosperity of individuals and businesses. Conceived by Eugene Fama in the 1960s, the Efficient Market Hypothesis (EMH) contends that financial markets are efficient, and therefore investors cannot consistently outperform the market on a risk-adjusted basis, as all pertinent data is already priced in. Despite EMH proponents arguing against the predictability of stock prices, research shows that precise models and carefully selected variables can indeed forecast stock movements accurately [1].

Forecasting stock market indexes is an ongoing difficulty due to its unpredictable and nonlinear characteristics. The incorporation of machine learning and deep learning techniques has greatly improved the accuracy and efficiency of financial forecasting, surpassing the traditional methods such as linear regression and ARIMA that were previously considered fundamental [3]

Linear regression is a basic method of quantitative analysis used in stock market prediction. It is commonly used as a benchmark for comparing performance. ARIMA is highly regarded for its ability to accurately describe time series data that is inherently non-stationary. The efficacy of ARIMA is assessed by Mian [6] in comparison to sophisticated methods such as LSTM

LSTM models have become popular for their capacity to capture long-term dependencies in time-series data, which is essential for making precise predictions on stock prices. In their study, Paul, and Das [7] investigate the effectiveness of LSTM in comparison to other deep learning models like RNN and BiLSTM. They highlight LSTM's ability to accurately capture the temporal patterns of the NIFTY 50 index data spanning a period of 27 years. Khan et al. (2023) present a thorough performance comparison of various machine learning models, including LSTM. They also propose innovative investment strategies to improve prediction efficiency. Their findings demonstrate the superiority of LSTM and other machine learning models over traditional statistical models, especially when utilizing innovative training techniques that integrate real-time data and market sentiment.

Our objective is to predict the closing price of the S&P 500 index by using various features driven from the S&P 500 Index over three years and 60 days data with forecast windows of 3, 7, and 14

days. We also explored companies making up the S&P 500 index, using the data provided from Yahoo's financial report. Therefore, this project presents a structured analysis of the Standard and Poor's (S&P) 500 Index, which is a widely used indicator that serves as a benchmark to gauge a market performance, focusing on fundamental financial metrics.

Utilizing Python and R programming languages alongside data analysis libraries like Pandas, Matplotlib, and NumPy, we outline a systematic approach that includes data exploration, visualization, statistical analysis, and model selection. Model selection and modeling efforts extend to forecasting future trends in the S&P 500 Index, utilizing the best selected model to project future closing prices. These forecasts offer valuable insights for investors and analysts alike, serving as a guide for informed decision-making.

Problem Description

Throughout the project, we address questions regarding data quality, methodological rigor, interpretation of findings, limitations of models explored and future research directions. To achieve our goal, we begin with the datasets to identify features which potentially influence the S&P 500 stock closing price. Key questions which guided our research include:

1. What features can be extracted from the S&P 500 index dataset, S&P 500 companies' information, and other relevant datasets to predict the closing price of the S&P 500 index?
2. How can historical stock prices of S&P 500 companies be utilized as features in predicting the S&P 500 index closing price?
3. How accurate are the explored models in the predictions of the S&P 500 Index closing price?
4. What kinds of models are best in stock prediction?

To ensure the reliability and efficacy of each model examined, we addressed model evaluation and validation techniques. By presenting our research as a framework, and addressing the above questions, we aim to provide practical guidance for stock market analysis and investment decision-making, offering actionable insights and recommendations for stakeholders in the financial markets.

2. Exploratory Data Analysis

A. Exploration of S&P 500 Constituent Companies and Industries

The S&P 500 publishes a list of its constituent companies on Financial News and Reports such as on Yahoo finance, to get insights into top performers shaping the index's performance. The financial report also cutlets and reports into top-performing S&P 500 companies based on various

metrics. Therefore, the information regarding the constituent companies was derived from financial reports.

Data exploration ensures data quality, visualizations, including time-series plots, uncovering historical trends in the S&P 500 Index's closing prices over time for deeper insights. Our exploration involves dissecting the constituent companies and the top 10 industries within the S&P 500 index, primarily based on Market Capitalization (Market Cap), to unravel their compositions, characteristics, and significance within the stock market. Market Cap, representing the total value of a publicly traded company's outstanding shares of stock, serves as a pivotal metric guiding our analysis

Our initial study focuses on the sectors integrated into the S&P 500 index as shown in Fig 1, alongside spotlighting the top 10 industries within it based on Market Cap. This preliminary step sets the stage for understanding their compositions, inherent characteristics, and broader significance in the stock market ecosystem.

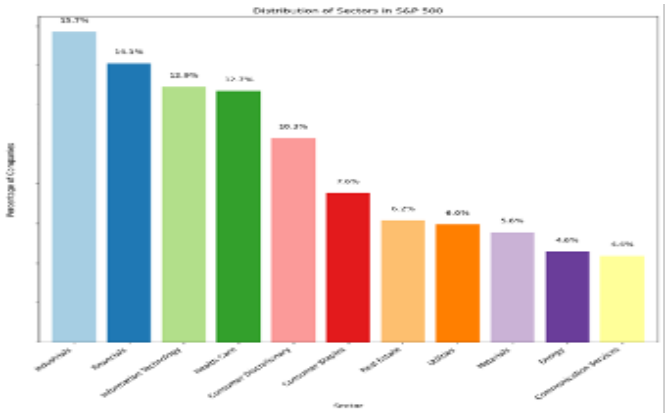


Fig 1: S&P 500 Sector Composition

To gauge the workforce scale and evolving trends, we further explored the average number of employees, in millions, within the top 10 industries across each sector (Fig 2). Moreover, analysis of the distribution of Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA) across these industries (Fig 3) aids in assessing profitability and overall financial performance.

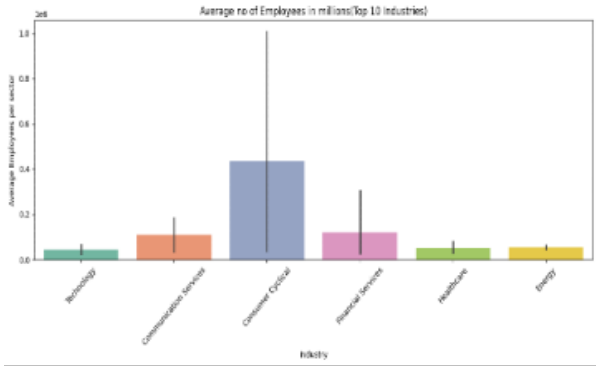


Fig 2: Average number of employees in million

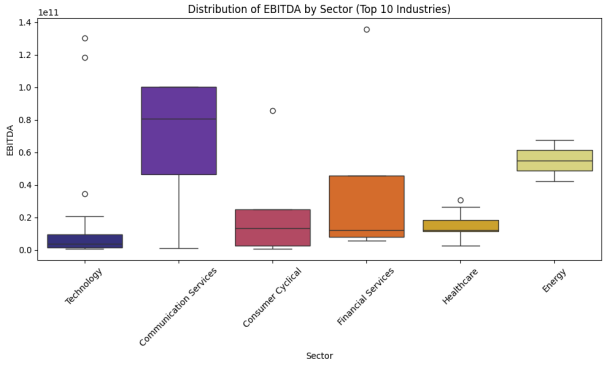
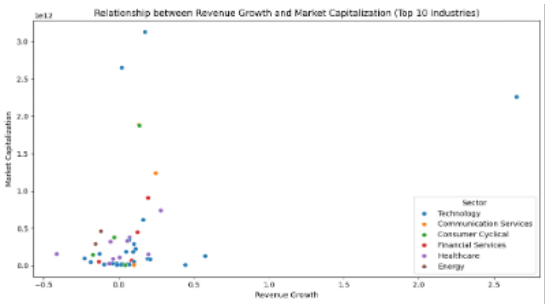


Fig 3: Distribution of EBITDA by sector

Our exploration extends to investigating the relationship between revenue growth and market capitalization (Fig 4), showing how market valuation correlates with revenue generation. Fig. 5

illustrates our examination of the distribution of market capitalization across industries within each sector, offering insights into the varying importance of different industries



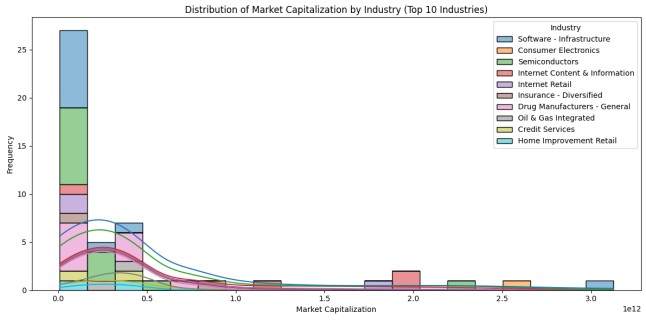


Fig 4: Revenue Growth vs Market Cap

Fig 5: Distribution of Market Cap by Industry

Additionally, we calculated the mean market capitalization for each industry within the top 10 industries of every sector (Fig 6), providing insights into industry-level valuation metrics. Finally, we determined the number of companies operating within each industry and sector (Fig 7), enabling a comprehensive assessment of industry representation and diversity.

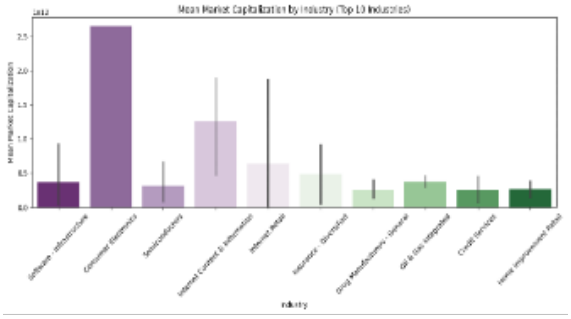


Fig 6: Mean market cap by industry

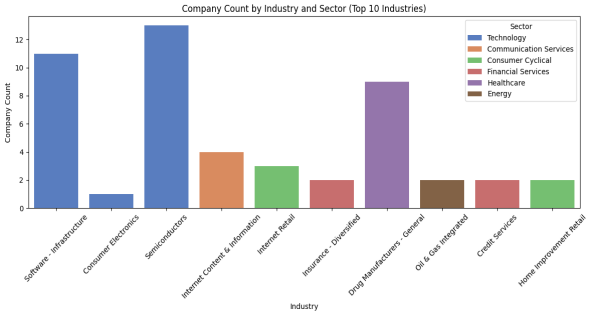


Fig 7: Company Count by Industry and Sector

Based on the Exploration of S&P 500 Constituent Companies and Industries and financial reports, the index consists of about 505 largest US based firms based on the market capitalizations, spanning 126 industries, which are grouped into 11 sectors, making up the index. This composition highlights the index's broad market representation as shown in the cart on the left. Moreover, the S&P 500 comprises approximately 80% of the total market capitalization of the U.S. stock market, largely due to its inclusion of the largest and most influential publicly traded companies across various sectors.

B. Exploration of S&P 500 stocks

The exploration of S&P 500 stocks involved the utilization of a dataset sourced from Yahoo Finance, a reputable financial data provider known for its comprehensive and up-to-date market information. The dataset encompassed stock prices for the S&P 500 index spanning from January 2021 to March 2024. Additionally, stock prices for the last 60 days leading up to March 2024 were extracted for analysis, offering a more in- depth view of recent market dynamics

.

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
2021-01-04 00:00:00-05:00	3764.610107	3769.989990	3662.709961	3700.649902	5015000000	0.0	0.0
2021-01-05 00:00:00-05:00	3698.020020	3737.830078	3695.070068	3726.860107	4591020000	0.0	0.0
2021-01-06 00:00:00-05:00	3712.199951	3783.040039	3705.340088	3748.139893	6064110000	0.0	0.0
2021-01-07 00:00:00-05:00	3764.709961	3811.550049	3764.709961	3803.790039	5099160000	0.0	0.0
2021-01-08 00:00:00-05:00	3815.050049	3826.689941	3783.600098	3824.679932	4773040000	0.0	0.0
...	...	...	...	...	...	...	...
2024-03-22 00:00:00-04:00	5242.479980	5246.089844	5229.870117	5234.180176	3374700000	0.0	0.0
2024-03-25 00:00:00-04:00	5219.520020	5229.089844	5216.089844	5218.189941	3331360000	0.0	0.0
2024-03-26 00:00:00-04:00	5228.850098	5235.160156	5203.419922	5203.580078	3871790000	0.0	0.0
2024-03-27 00:00:00-04:00	5226.310059	5249.259766	5213.919922	5248.490234	3850500000	0.0	0.0
2024-03-28 00:00:00-04:00	5248.029785	5264.850098	5245.819824	5254.350098	3998270000	0.0	0.0

814 rows x 7 columns

Table 1: The stock prices for the S&P 500 index spanning from January 2021 to March 2024

Columns Included: The dataset comprised essential columns necessary for thorough analysis, defined as follow:

Date: The date corresponding to each recorded stock price for trend analysis

Open: The opening price of the S&P 500 index at the beginning of the trading session

high: The highest price reached by the S&P 500 index during the trading session

Low: The lowest price reached by the S&P 500 index during the trading session

Close: The closing price of the S&P 500 index at the end of the trading session

Volume: The total number of shares traded for a particular stock on a given day

Scope of Analysis:

The dataset was analyzed to uncover patterns, trends, and potential insights into the behavior of S&P 500 stocks over the specified timeframe. By examining historical price data and recent trends, valuable information was gleaned to predict the closing price of the index; identify potential opportunities or risks and provide insight into long term trends and patterns.

Fig 8 below displays the closing prices of the S&P 500 Index from January 2021 to March 2024. The plot shows an overall bullish upward trend, indicating stock market growth amid volatility. Key periods include an initial pandemic recovery, a mid-2021 to early 2022 drop due to inflation concerns and policy tightening, a late 2022 to early 2023 rebound supported by economic data, and a steep late 2023 to March 2024 climb to new highs around 5,250. Fluctuations from factors like geopolitical tensions, monetary policy shifts, and recession fears, requiring investor attention despites optimism and caution

Fig 8: The S&P 500 Performance from January 2021 to March 2024

3. Methods : Model Selection

3.1 MLR : Brief overview

Multiple linear regression (MLR) expands upon the concept of simple linear regression by including many independent variables. This approach represents the dependent variable as a linear combination of the independent variables, together with an error term. The multiple linear regression equation is typically expressed in the following general form:

$$\gamma = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Using MLR provides a foundational framework for stock market trend prediction. We first performed a correlation matrix to examine the relationships between the different market variables. The matrix revealed high correlations between the Open, High, Low, and Close prices, suggesting these variables share a significant linear relationship. However, volume showed a negative correlation with other market variables, indicating less predictability associated with trading volume. Therefore, a linear regression model was fitted using the Open, High, and Low prices as predictors for the Closing price for both datasets (3-years and 60-days)

3.1.1 Results with Performance Analysis

A. Performance Analysis on 3-year data

Visualizing the model's performance, this graphic (refer Fig 9) showed that the model tracked index movements, despite occasional deviations during extreme volatility

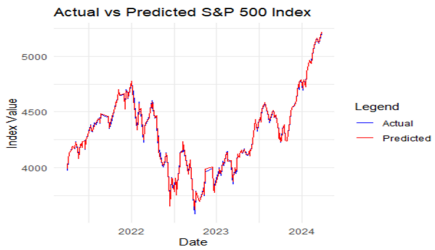


Fig 9: Actual vs Predicted S&P 500 on the 3-years data.

B. Performance Analysis on 60-days data

The graph of real against predicted results (Fig 10) shows that the model closely follows the trend. The graph shows both lines moving together, especially the January–March increasing trend.

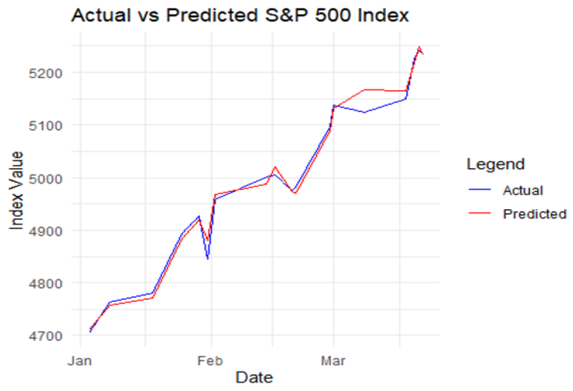


Fig 10: Actual vs Predicted S&P 500 on the 60-days data

C. MLR's Performance Metrics on 3-years data

The model displayed an extremely high coefficient of determination (R-squared value of 0.997), indicating that the model explains nearly all the variability of the response data around its mean. However, the presence of multicollinearity, as evidenced by high Variance Inflation Factors (VIFs), suggests that the predictors are highly correlated, which might inflate the variance of the estimated regression coefficients. VIF values greater than 10 are indicative of severe multicollinearity, and values in the hundreds (Open~290.34, High~243.70 and Low~201.55) suggest extremely high correlation that could be distorting the regression coefficients and their standard errors.

RMSE	MAE	MAPE	R^2
14.4391906	10.7951199	0.2552822	0.9977941

Table 2 : MLR's Performance Metrics on 3-years data

D. MLR Performance Metrics on 60-days data.

The model displayed an extremely high coefficient of determination (R-squared value of 0.996), indicating that the model explains nearly all the variability of the response data around its mean. However, the presence of multicollinearity, as evidenced by high Variance Inflation Factors (VIFs), suggests that the predictors are highly correlated, which might inflate the variance of the estimated regression coefficients. VIF values greater than 10 are indicative of severe multicollinearity, and values in the hundreds (Open~169.99, High~198.86 and Low~197.86), complicating the interpretation of model coefficients and potentially inflating the standard errors

RMSE	MAE	MAPE	R^2
15.7225951	11.7099446	0.2345173	0.9965935

Table 3: MLR's Performance Metrics for 60-days dataset.

E. Forecast Error Metric Evaluation

The model was used to conduct rolling forecasts using different time windows, including 3-day, 10-day, 1-month, and 3-month forecasts. The observed pattern aligns with the difficulties of making predictions for longer time frames, as the model encounters greater levels of uncertainty and variability in the available data. The error metrics for each horizon indicate that the model's accuracy decreases as the forecasting interval increases.

Forecast Window	RMSE	MAE	MAPE
3-Day	17.37826	11.94093	0.2820909
10-Day	26.33685	13.13877	0.3106544
1-Month	28.16177	13.22901	0.3124792
3-Month	28.02570	13.92528	0.3271620

Table 4 : Forecast Error Metric Evaluation for 60-days dataset.

The findings suggest that a substantial increase in prediction error occurs as the forecast horizon lengthens, with the 14-day forecast metrics exhibiting this trend most notably. The increasing errors indicate that although the model demonstrates considerable accuracy in short-term predictions, its dependability decreases as the duration increases

Forecast Window	RMSE	MAE	MAPE
3-Day	41.63886	19.26812	0.3943650
7-Day	31.93433	18.20989	0.3709526
14-Day	165.99080	54.57670	1.1060989

Table 5 : Forecast Error Metric Evaluation for 60-days dataset.

While multiple linear regression provides a simple and understandable framework for modeling association between variables, its use in our work was plagued with difficulties. With Variance Inflation Factors (VIFs) much above the widely used threshold of 10, indicating highly strong correlation among the independent variables (Open, High, Low), the study revealed significant multicollinearity among the predictors. Both models (3-years and 60-days) had a very high R^2 value, which means it fit the training data well. However, this can sometimes be a sign of overfitting. As a result, more advanced methods algorithms, which can handle the unique

characteristics and complexities of financial time series data are mostly preferred. In our case, we used ARIMA and LSTM.

3.2 ARIMA : Brief overview

ARIMA model is a model which stands for Autoregressive Integrated Moving Average is a model that is typically used for forecasting future values based on historical data. It combines three components: autoregression (AR), differencing (I), and moving average (MA). ARIMA modeling involves analyzing a time series dataset to identify patterns and relationships between observations. The Autoregressive component aims to capture the relationship between an observation and a linear combination of its lagged observations. The AR component can be modeled by the equation:

$$Y_T = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots \phi_p Y_{t-p}$$

where ϕ_1, ϕ_2 , …., ϕ_p are the coefficients of the model, Y_{t-1},Y_{t-2} , …, Y_{t-p} are the lagged observations and c is a constant term.

The Integrated component involves making the data more stationary. ARIMA models have an assumption that the data must be stationary and thus differencing provides a way to make it stationary. Differencing is the process of computing the differences between consecutive observations. It helps remove trends and seasonality, making the data suitable for modeling with ARIMA. Below is an example of a stationary data

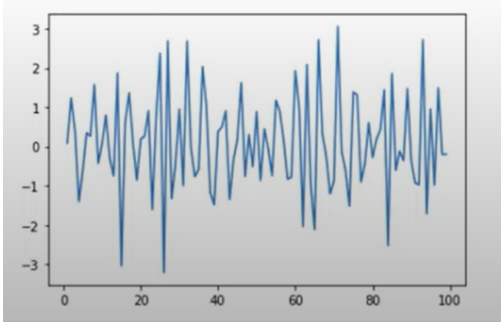


Fig 11: Example of stationary Data

The MA components models represents the relationship between an observation and a linear combination of past error terms and can be expressed as the equation. The equation of the Moving Average component of ARIMA is given by:

$$Y_T = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots \phi_q \epsilon_q + \epsilon_t$$

where Θ_1, Θ_2,... Θ_p are the coefficients and ε_1 , ε_2 , … ε_p are the error terms

Hyperparameter Selection

As ARIMA is a one feature model, the closing index price was just selected in order to train the model. Parameters p,d, and q which represent the three components of the ARIMA model respectively. Parameters p and q are obtained by PACF(Partial Autocorrelation Function) and ACF(Autocorrelation Function) plots which determine the lag order of the observations.

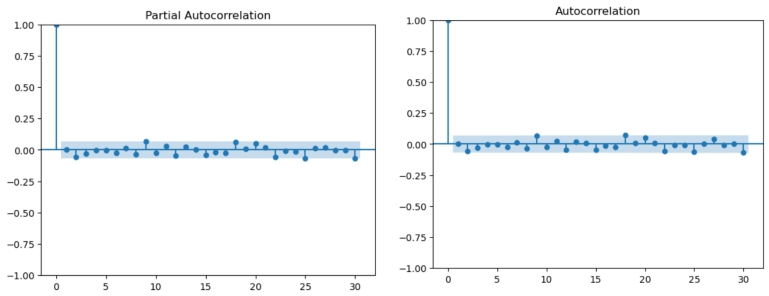


Fig 12: PACF (Left) and ACF (Right) plots

As from the plots we can see that lag order of 0 falls over the confidence interval so that was used as our optimal p and q however, as lag order 0 falls outside the confidence interval plot, we also used that as a lag order to test the results of the model. To get the parameter d and ADF test was used to check the stationarity of the data. The test uses hypothesis testing to check whether the data is stationary. On face value stock index data is not usually stationary cause it usually follows trends, thus we had to do some differencing to make it stationary. The number of times your data is differenced represents the integer value for parameter d. For our data, we only needed to difference once and do an ADF test which then told us to reject the null hypothesis that assumes stationarity of the data because of its high p value.

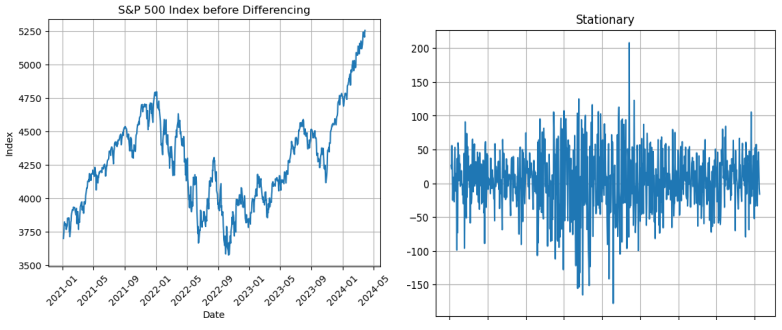


Fig 13: Data before differencing (Left) and after differencing (Right)

Thus, for the model an optimal hyper parameter selection of ARIMA (0,1,0) for the 3-year data and ARIMA(0,1,2) for the 60-day data. However, an ARIMA (2,1,0) was also used on both sets of data for the purpose of model comparison

3.2.1 Results with Model Performance Analysis

The two different hyperparameter selections were tested on 2 different sets of data - a 60-day range of data and a 3-year range of data. These 2 ranges were selected to see the effectiveness of ARIMA modeling on different types of historical data. For the 3-year data, three different forecasting windows (which serve as a testing dataset) were used for our prediction likewise for the 60-day data. The 3-year window uses forecasting windows of 3 days, 20 days, a month, and 3 months while the 60-day data uses forecasting windows of 3 days, 7 days, and 14 days.

For the ARIMA(0,1,0) model for the 3-year data it follows the same pattern as the Multiple linear Regression model as with increase in forecasting window the errors increase which is a common phenomenon with ARIMA as the Model tends to predict based on past behavior so if will be very easy to go way beyond the actual stock value. The 60-day model follows a similar trend to the 3-years data however the error values are smaller which highlights another limitation to the ARIMA model as it is suitable for short term data. The ARIMA (2,1,0) model seems to do better with the 3-year data with lower though not negligible error metrics.

Forecast Window	RMSE	MAE	MAPE
3-Day	28.5136	27.023	0.005154
10-Day	71.0587	63.406	0.012
1-Month	149.70614	132.0963	0.02555
3-Month	430.5713	371.5964	0.0743

Table 6: Error metrics for ARIMA (0,1,0) on 3-years data

Forecast Window	RMSE	MAE	MAPE
3-Day	27.6935	26.3674	0.00503
10-Day	70.0850	62.62515	0.01198
1-Month	148.7075	131.0306	0.02534
3-Month	430.973414	372.060	0.07440

Table7: Error metrics for ARIMA (2,1,0) on the 3-years data

Forecast Window	RMSE	MAE	MAPE
3-Day	15.93390	12.70069	0.002433
7-Day	23.57408	18.85923	0.003609
14-Day	38.36258	31.3416	0.00605

Table 8: Error metrics for ARIMA (0,1,2) on the 60-Days data

3.3 Long Short-Term Memory : Brief overview

The LSTM model is a type of recurrent neural network (RNN) designed to overcome the limitations of traditional RNNs in capturing long-term dependencies in sequential data. In the context of financial time series prediction, LSTM models have shown effectiveness due to their ability to retain information over extended time periods.

A. Data Preparation and Normalization

The primary dataset used in this study was the S&P 500 index, a popular US stock market index, used for the model prediction. We were interested in testing how good each model was by having two sets of datasets thus a three-year and 60-day dataset. This analysis helps in understanding the model's utility in practical trading scenarios. Initial data preparation involved for the 3-year dataset involved cleaning missing values and filtering the dataset to focus exclusively on the last three years of data, thus ensuring that the model's training phase is grounded in contemporary market conditions. The 60-days dataset focused on recent data by filtering for the last 60 days and a specific period from 2024-01-10 to 2024-04-05.

From this stage on going, the normalization and pre-processing were the same for both the 3-year and 60-days datasets respectively except for a minor change in the batch size used. The dataset was then divided into a 70-30 split for training and testing, enabling effective model validation. If the range of one feature varies more widely than the others, most ML algorithms might not perform well and as such to prepare the input data for the LSTM model, a normalization was applied using the min-max scaling technique, which rescaled the S&P 500 index values to a range from 0 to 1. This normalization is crucial for neural network models as it ensures all input features contribute equally to model learning, preventing gradients from vanishing or exploding during training.

B. Model Structure and Hyperparameter Tuning

The normalized data was then segmented into sequences that serve as inputs for the LSTM, facilitating the model's ability to learn from past information. The LSTM model was structured with two LSTM layers followed by a dense output layer. The first LSTM layer with 128 units returned sequences to provide temporal inputs for the second layer, which included 64 units. This configuration is designed to capture complex patterns in time series data effectively. The model was compiled with the Adam optimizer and mean squared error loss, a popular choice for these applications due to its adaptive learning rate capabilities, facilitating quicker convergence.

Parameters adjusted included the number of LSTM units, batch size, learning rate of the optimizer, and the number of training epochs. We experimented with various configurations, starting from simpler models with fewer units to more complex ones with a higher number of units. Testing configurations such as 64-32, 128-64, and 256-128 units, we sought a balance where the model had enough capacity to learn significant patterns without becoming computationally prohibitive or prone to overfitting. The final selection of 128-64 units was driven by this model's ability to efficiently process the inherent complexities of the S&P 500 index data without overfitting, as evidenced by the stability and generalization performance on the validation set.

The batch size, which determines how many examples the model sees before updating the weights, significantly impacts the training dynamics. We tested batch sizes of 20, 40, and 80, finding that a batch size of 40 offered the best compromise between efficient learning and model stability, given our computational resources and dataset size. However, a different batch size of 235 yielded the best results for the 60-days dataset. The learning rate for the Adam optimizer was another critical factor adjusted. Adam, known for its adaptive learning rate adjustments, was chosen to optimize the LSTM model. Other optimization functions were explored (such as SGD, Adagrad, AdaDelta and RMSProp) but Adam optimizer gave the best output.

Initial tests with learning rates of 0.001, 0.01, and 0.1 provided insights into how quickly the model converged to a local or global minimum. A learning rate of 0.01 was optimal in our case, as it helped the model converge quickly enough without overshooting the minimum of the loss landscape, which can occur with higher rates. We experimented with 50, 100, and 200 epochs. The model's performance on training and validation losses indicated that 100 epochs were sufficient

to achieve convergence without significant overfitting, as further increases in epochs did not yield improvements in validation loss.

3.3.1 LSTM Results with Model Performance Analysis

A. Model Performance Analysis for 3-year dataset.

The LSTM model's predictive accuracy is captured in the visualization of Actual vs Predicted S&P 500 Index." Figure x illustrates the model's ability to trace the general trend of the actual index values over time, despite some areas of deviation. Particularly, the model exhibits a commendable fit during the upward and steady-state trends but shows divergence in areas of high volatility.

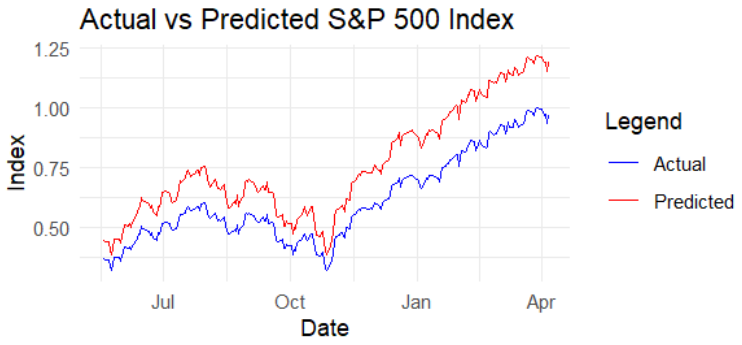


Fig 14: Actual vs Predicted S&P 500 Index on the 3-years data

B. Performance Metrics for 3-year dataset.

We analyzed the performance of the model by employing MAE, MAPE, and RMSE for the model's overall performance (Table 1). The MAPE of 16.54% signifies that, on average, the model's predictions deviate from the actual values by this percentage. While this is a moderate error rate, it highlights the challenges in achieving high precision in market forecasting. The RMSE of 0.11 reflects the model's average error in predicting the normalized S&P 500 values, and alongside the MAE of 0.10, indicates that the model generally maintains a close trajectory to the actual index values.

RMSE	MAE	MAPE
16.5495108	0.1128397	0.1041448

Table 9: Performance scores for the LSTM on the 3-years data

C. Model Performance Analysis for 60-days dataset

The graph below (Fig 15) indicates a strong alignment between the two sets of values thus predicted and actual data, particularly in capturing the trend and direction of the index movements. Although there are points where the predicted values do not perfectly match the actual data, the overall pattern suggests that the model can forecast the general behavior of the index with a notable degree of accuracy.

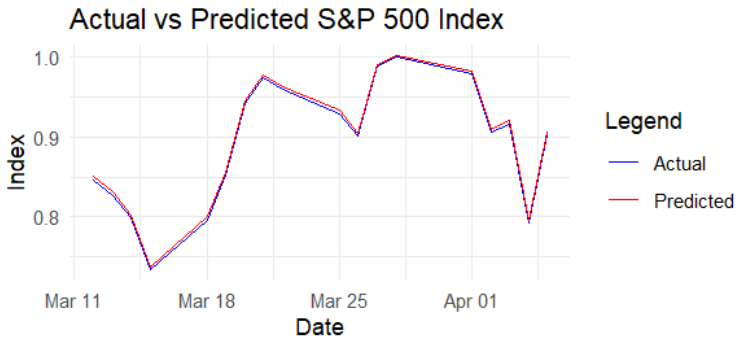


Fig 15: Actual vs Predicted S&P 500 Index om the 60-days data

D. Performance Metrics on 60-days data

We analyzed the performance of the model by employing MAE, MAPE, and RMSE for the model's overall performance (Table 10). An MAPE of 0.16% is remarkably low, indicating the model's strong alignment with the actual index values. RMSE and MAE metrics reinforce this precision, suggesting the model's reliability in the short term. The graphical depiction of predicted versus actual values offers a visual affirmation, showing the model's accuracy in capturing the market trend within the 60-day window (Fig 16)

RMSE	MAE	MAPE
0.1669636	0.0016294	0.0014124

Table 10 . Performance scores for the LSTM model on 60-days data

E. Forecast Error Metric Evaluation for 3-year dataset.

Forecast error metrics for various forecasting windows were calculated to assess the model's performance over different intervals (Table 11). This analysis helps in understanding the model's utility in practical trading scenarios, where prediction for various time horizons might be necessary. The implementation of a forecast function was used to assess the LSTM model’s accuracy by comparing actual stock index values against predictions shifted forward by specified

forecast windows. The Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) are tabulated for forecast windows of 3, 10, 20, and 60 days. There is an observable trend of increasing error values with the expansion of the forecast window. This finding is expected, as longer-term predictions are typically subject to greater uncertainty due to the compounding effects of unforeseen variables and market dynamics. The lowest errors are noted for the 3-day window, indicating the model's strength in short-term predictions.

Forecast Window	MAE	MAPE	RMSE
3	0.1086971	0.1712737	0.1198438
10	0.1297084	0.2023051	0.1445536
20	0.1535950	0.2312278	0.1745832
60	0.2382165	0.3327688	0.2799005

Table 11. Forecast error metrics for different time windows on the 3-year data

F. Forecast Error Metric Evaluation for 60-days dataset.

Due to the limited data range for this dataset, we computed forecast window analysis for only 3-, 7- and 14-days window size respectively (Table 12). The 3-day forecast window showcases the LSTM's strengths, suggesting its application for short-term trading could be highly beneficial. The errors for the 7-day and 14-day windows reveal the model's increasing uncertainty in prediction as time progresses. The RMSE value's growth from 0.08 in the 3-day window to 0.10 in the 14-day window is indicative of the challenges posed by extending the forecast horizon. It points to the potential exponential increase in unpredictability and risk in the financial markets as the prediction window widens.

Forecast Window	MAE	MAPE	RMSE
3	0.0690218	0.0777453	0.0811831
7	0.0964247	0.1038306	0.1110050
14	0.0967481	0.1061895	0.1004430

Table 12. Forecast error metrics for different time windows for 60-days dataset.

4. Conclusion

In the project the closing price of the S&P 500 index is predicted by drawing on the index data set and on information about the companies that make it up. Factors such as historical stock prices and trading volumes were used to forecast the closing price of the index. As explained in part A of data Exploratory analysis, stock prices of historical S&P 500 companies can provide valuable insights into market trends and correlations with the broader index (refer fig 1 – 9). Machine learning models including linear regression, ARIMA, and LSTM networks were used for forecasting, with metrics such as MAE, MSE, or RMSE evaluating the accuracy of the model

After exploring ARIMA, LSTM, and linear regression models on three years and 60 days of S&P 500 data with forecast windows of 3, 7, and 14 days, it was found that utilizing a narrower forecast window of 60 days yielded superior predictions for the closing price. Among the models evaluated, LSTM emerged as the most effective, exhibiting lower MAE, MAPE, and RMSE metrics compared to ARIMA and linear regression.

Optimizing the forecast window to a shorter time frame matches with the dynamic and rapidly changing nature of stock market data, allowing the models to adapt quickly to evolving trends and make more accurate forecasts. This adjustment shows the importance of tailoring model parameters to the specific characteristics of the data being analyzed.

Also, it is important to emphasize that the choice of data set should match the timing and schedule of the trader or investor. For daily traders considering weekly metrics, using data sets closer to the desired time frame, such as 60-day, 30-day, or even weekly data, provides an indication of short-term movement and an accurate prediction, as our findings show. In contrast, for long-term investors, considering data sets spanning years or decades gives a sense of market trends and stability broad, making it easier to make informed long-term investment decisions

Going forward, analyzing the relationship between international market movements and the S&P 500 index will offer a promising approach to analyze the integration of global markets and their impact on price movements on price dynamics. Additionally, considering other economic factors such as inflation, GDP, interest rates, and geopolitical events into the analysis could further enhance the understanding of long-term market dynamics and assist in making informed investment decisions. Therefore, incorporating dataset selection strategies, economic indicators, policy and analyzing the impact of global stock market movements, traders and investors can optimize the accuracy of their predictions and enhance their ability to navigate the complexities of the stock market effectively.

5. Reference

[1] Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. *Computer Science Review*, *34*, 100190. https://doi.org/10.1016/j.cosrev.2019.08.001

[2] Mehtab, S., Sen, J., & Dutta, A. (2021). Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models. In S. M. Thampi, S. Piramuthu, K.-C. Li, S. Berretti, M. Wozniak, & D. Singh (Eds.), *Machine Learning and Metaheuristics Algorithms, and Applications* (pp. 88–106). Springer. https://doi.org/10.1007/978-981-16-0419-5_8

[3] Hongying Zheng, Zhiqiang Zhou, Jianyong Chen, "RLSTM: A New Framework of Stock Prediction by Using Random Noise for Overfitting Prevention", *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 8865816, 14 pages, 2021. https://doi.org/10.1155/2021/8865816

[4] R. M. I. Kusuma, T.-T. Ho, W.-C. Kao, Y.-Y. Ou, and K.-L. Hua, "Using deep learning neural networks and candlestick chart representation to predict stock market," 2019, https://arxiv.org/abs/1903.12258.View at: Google Scholar

[5] K. Adam, A. Marcet, and J. P. Nicolini, "Stock market volatility and learning," *The Journal of Finance*, vol. 71, no. 1, pp. 33–82, 2016. View at: Publisher Site | Google Scholar

[6] Mian, T. S. (2023). Evaluation of Stock Closing Prices using Transformer Learning. *Engineering, Technology & Applied Science Research*, *13*(5), 11635-11642.

[7] Paul, M. K., & Das, P. (2023). A Comparative Study of Deep Learning Algorithms for Forecasting Indian Stock Market Trends. *International Journal of Advanced Computer Science and Applications*, *14*(10)

[8] Khan, A. H., Shah, A., Ali, A., Shahid, R., Zahid, Z. U., Sharif, M. U., ... & Zafar, M. H. (2023). A performance comparison of machine learning models for stock market prediction with novel investment strategy. *Plos one*, *18*(9), e0286362.

Dataset

1. https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks?select=sp500_companies.csv
2. https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks?select=sp500_index.csv
3. https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks?select=sp500_stocks.csv
4. https://finance.yahoo.com/quote/%5EGSPC/history