National College of Ireland

Project Submission Sheet

Student Name:	Shreeraj Sangle		
Student ID:	23283254		
Programme:	MSCAI1B	**Year:**	2024-25
Module:	Machine Learning		
Lecturer:	Jaswinder Singh		
Submission Due Date:	27/04/2025		
Project Title:	Multi -Horizon Stock Price Forecasting using ARIMA, XGBoost and LSTM.		
Word Count:	6320		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Shreeraj Sangle

Date: 27/04/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Machine Learning.

Multi -Horizon Stock Price Forecasting using ARIMA, XGBoost and LSTM.

Your Name/Student Number	Course	Date
Shreeraj Sangle	MSCAI1B	27/04/25

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

Multi -Horizon Stock Price Forecasting using ARIMA, XGBoost and LSTM.

Shreeraj Santosh Sangle
(23283254)
Machine Learning-MSCAI1B
School of Computing
National College of Ireland

***Abstract*—The project aims to predict stock market trends which utilizes machine learning and deep learning techniques. The primary goal here is to forecast prices movements of stock of top 100 stocks with the highest market cap from the S&P 500 over different time horizons: 1-day, 5-day, and 20-day. As we know that forecasting the stock market prices is always a challenges as the financial data in non-stationary, highly complex. To do so using an ARIMA-XGBoost-LSTM hybrid system. Certain important tasks include data preprocessing, feature extraction model involved in technical indicator formations as well as sentiment analysis. Model evaluation is done via RMSEs as well as MAEs in addition to Directional Accuracy, Precision, and Recall. The output of the system contains returns predictions with confidence intervals, directional forecasts, and visualizations showing selected variables in comparison with actual ones under prediction. Probabilistic framework for modeling uses these models, combines them together to yield reliable and effective predictions of stock movements under various time horizons, and induces those financial markets towards accurate analysis and prediction.[1]**

Keywords: Stock Market Forecasting, Transformer Models, Attention Mechanism, Multi-Horizon Forecasting, ARIMA, XGBoost, , financial time series, deep learning, LSTM.

I. INTRODUCTION

Stock market prediction is a very challenging problem, even though much work has been devoted to it in the hopes of obtaining financial gains. The stock market is very complex and dynamic. It can be built on economic indicators, company earnings, and investor sentiments. Because it would enable investors to tailor investment options, it has become crucial for them to accurately predict hitherto unpredictable stock prices in the market. However, traditional frameworks do not generally fit well with the complex and non-linear relationships of financial data. That is why techniques such as machine learning (ML) and deep learning (DL) have gained importance in stock price forecasting, as they are not only advanced but also flexible.[1]

This study employs three models in determining stock price movements: Autoregressive Integrated Moving Average (ARIMA), XGBoost, Long Short Term Memory (LSTM) networks. These models are then evaluated at three different forecasting horizons of 1 day, 5 days, and 20 days to assess their effectiveness compared to stock prices. The dataset is the top 100 companies by market capitalization in the S&P 500 index and suffices to have diversity in industries and performance of the market.

The ARIMA model is a widely accepted and used statistical metric for time series prediction. It is predicated on linear relationships constituted between the past values and errors. Though ARIMA performed quite well for stationary data, it encountered difficulties in dealing with non-stationary financial time series that had trends and seasonality to be predominant. Nevertheless, ARIMA could really render insights if integrated or coupled with other measures.[2]

XGBoost is a non-parametric gradient boosting algorithm, useful in handling non-linear relationships and also complicated interactions within the features. It approximately constructs an ensemble of effective decision trees to increase the accuracy by fixing errors from the previous models. Due to its capacity to scale and then to manage large volumes of datasets, it proved to be very efficient in financial businesses.[3]

LSTM networks are a kind of recurrent neural network model that is especially suited to learning long-range dependencies with sequences of data for forecasting in time series. Compared to classical recurrent networks, LSTMs are important because they facilitate the long rehearsal of memories.[4]

The data which is availed for this work is through very concrete aspects like Yahoo Finance and Alpha Vantage, which incorporate price perspectives into the target variable from independent daily closing prices. Associations were made with indicators: moving averages, Relative Strength Index (RSI), and Moving Average Convergence Divergence (MACD) in view of improving the performance for the model.

The data undergoes preprocessing prior to model training; it goes through missing value handling, normalization, and engineering of features. The processed data then gets split into training and testing sets for evaluating the model without bias. Important metrics used to evaluate model performance include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and directional accuracy. These metrics are essential in determining the model's ability in predicting stock price movements and indicates whether it forecasts the correct direction of price changes.[1]

II. LITERATURE REVIEW

1. Introduction

Stock market prediction has been an area of focus for quite some time with the aim of assisting the investors to make a better choice. Traditional methods including statistical models such as ARIMA have been used to predict stock prices, but these methods primarily fail to capture the non-linear dynamics and complex interactions inherent in financial data. However, recent developments in ML and DL methods, such as XGBoost and LSTM networks, have proven beneficial in this regard. Although ARIMA can be used in a good way for linear time-series forecasting, it does poorly when it comes to forecasting in the complex real life situation of stock market [2]. XGBoost is a gradient boosting technique which can lay down

complex and non-linear relationships; therefore, its application in predicting stock prices emerges as an appropriate choice[3]. LSTMs are designed to work with timed sequences of data, and thus, they are deemed most suitable for time series forecasting, since they model longer-term dependencies that help better the accuracy of predictions [4]. This literature review will look into the works carried out and compare the performance for stock price forecasting using ARIMA, XGBoost, and LSTM models while discussing and examining their usages, advantages, and shortcomings.

2. Overview of Stock Market Prediction

The accurate prediction of the stock market is crucial for investors as it provides information required in making informed decisions. The traditional statistical models, such as ARIMA, have extensively been used in forecasting financial time series. These statistical approaches, however, hardly deal with the non-linear characteristics among financial data. Machine learning and deep learning have introduced more advanced methods to resolve this issue.

This hybrid model employed an ARIMA to capture linear dynamics and an XGBoost model to capture non-linear dynamics in stock market volatility. This approach, in comparison with traditional ARIMA models, outperformed the hybrid model in terms of forecasting accuracy.[5]

Researchers have also developed deep learning models to enhance prediction capability in finance. An example of such a deep learning model is the Long Short-Term Memory (LSTM) network that encompasses a design channel for feeding long-term memory dependencies in sequential data and is hence fit for time series prediction tasks. Chatterjee, for instance, did research comparing some of these models, including the LSTM, in stock price prediction. They found that LSTM models had a much better yield than conventional models.[6]

These studies indicate that even as stock market prediction becomes more sophisticated, there is still a further need to craft models that capture linear and nonlinear dynamics of financial systems data.

3. ARIMA Model in Stock Price Forecasting

The ARIMA model has gained much application in time series forecasting, including stock price forecasting. It is useful if data exhibit linear relationships and are stationary. Stock prices, however, are more often volatile and non-linear; hence, ARIMA is less competent to capture such complexities. Patuwo et al. (2004) state that ARIMA can forecast stock prices when the data are stationary; however, it falters in sudden market shifts or other acts of interference. They recommend adding other models, specifically those based on machine learning, for increased accuracy. Adebiyi A. et al. (2014) states that ARIMA may offer short-term forecasting but is limited in forecasting long-term trends and volatility. They also indicated the need for more sophisticated methods, such as machine learning and deep learning, to address the multi-dimensional factors affecting stock prices. Set against this backdrop, an exposition on ARIMA's has limitation in handling non-linear behavior a hybrid approach for effective forecasting.

4. XGBoost in Stock Price Prediction

The XGBoost algorithm models the complex, non-linear dependencies of financial data.The study conducted, proposes a hybrid model for predicting stock market volatility by integrating ARIMA with XGBoost. The ARIMA would treat linear patterns in the data, XGBoost would tackle the non-linear patterns, thus improving the accuracy of forecasting. The study therefore confirmed the mixed effectiveness of combining traditional statistical methodologies with fast machine learning techniques for better prediction performance.[5]

Likewise, they explored several models, including XGBoost, for stock price prediction. In their view, XGBoost yielded the best results among other machine learning models in terms of accuracy and robustness, establishing it as the preferred choice for financial forecast assignments. The study pointed out that this stresses choosing suitable models that essentially capture the intricate patterns involved in stock market data.[6]

Therefore, these studies prove that XGBoost finds utility in predicting stock prices, especially when melded with other models to exploit joint benefits of certain capabilities. Most importantly, the ability of XGBoost to deal with complex non-linear relationships makes the technique even more valuable for analysts and data scientists in the financial field.

5. LSTM Networks in Financial Forecasting

Long Short-Term Memory (LSTM) networks are a specialized version of recurrent neural networks (RNNs) that found applications for time series forecasting. In the prediction of stock prices, LSTMs learn complex patterns and dynamics within financial time series that are hardly captured by traditional models like ARIMA.

According to the authors explored LSTM modeling for predicting stock prices and compared its performance with other machine learning algorithms. Results showed that LSTMs established their superiority over classical statistical methods in forecast accuracy, particularly in capturing the volatility and non-linearity embedded within financial data. Furthermore, the study evidenced that LSTM networks are highly functional in managing large datasets and making real-time predictions, which is highly warranted within stock markets. In this regard, LSTMs try to entail past market behavior with long memory properties, which give them a slight edge over other orthodox means to predict future trends.[10]

Similarly, the authors have performed stock market prediction using LSTM models, showing that these models can outperform traditional machine learning techniques interacting with sequential data, such as decision trees and SVMs.[11]

These studies reinforce the fast acceptance of LSTM networks in the field of finance, and thus the potential to enhance stock-market prediction accuracy.

6. Comparative Analysis of ARIMA, XGBoost, and LSTM

The research showed the predictability of ARIMA, XGBoost, and LSTM models on stock prices of Google. The results showed that the hyper-parameterized XGBoost model gave 99.47% R-squared value with Mean Absolute Error of 15.98 and Root Mean Square Error of 27.34, which was better than ARIMA and LSTM models. The LSTM model performed quite well too by showing an R-squared value of 96.87%. Its MAE was 49.35, and RMSE-57.28. On the contrary, ARIMA exhibited the lowest R-squared value, which was 66.37%, with an MAE of 140.12 and RMSE of 188.11. These performances or statistics stated are also proof as an under-performance of ARIMA because of its incapacity to catch the complexity of the movement of stock prices. These results indicate the high effectiveness of XGBoost for stock price forecasting, especially when augmented through hyperparameter tuning.[12]

This was similarly done, where methods compared were ARIMA, XGBoost, and LSTM for predicting stock prices of Amazon. For outcomes, it found that the XGBoost model was much more accurate than either an ARIMA or LSTM model, as the performance measured with the different metrics of MAE, Mean Squared Error,

RMSE, or R-squared. The LSTM model was primarily about capturing those temporal dependencies, and even though it was quite effective, it did not top the results against this XGBoost study. These comparisons revealed a strength and weakness for each model and put XGBoost above the fray in such tasks as predicting stock prices.[13]
Rewrite text with lower perplexity and higher burstiness while preserving word count and HTML elements. A thorough research by [12] showed the predictability of ARIMA, XGBoost, and LSTM models on stock prices of Google. According to the result, the hyper-parameterized XGBoost model gave an R-squared value of 99.47%, which was improved with a Mean Absolute Error of 15.98 and Root Mean Square Error of 27.34 than ARIMA and LSTM models. The LSTM model performed quite well too by showing an R-squared value of 96.87%. Its MAE was 49.35, and RMSE-57.28. In contrast, the ARIMA model showed the least R-squared value of 66.37 with MAE-140.12 and RMSE-188.11, thus emphasizing its incapacity to capture the complexities of stock price movements. Such results, therefore, indicate high effectiveness in XGBoost for stock price forecasting, especially augmented by hyperparameter tuning.
Likewise,[13] wherein the methods being compared were ARIMA, XGBoost, and LSTM for predicting stock prices for Amazon. For outcomes, it found that the XGBoost model was much more accurate than either an ARIMA or LSTM model, as the performance measured with the different metrics of MAE, Mean Squared Error, RMSE, or R-squared. The LSTM model was primarily about capturing those temporal dependencies, and even though it was quite effective, it did not top the results against this XGBoost study. These comparisons revealed a strength and weakness for each model and put XGBoost above the fray in such tasks as predicting stock prices.

III. Choice of Methods

These models will include not only ARIMA techniques but also XGBoost and LSTM methodologies that would suffice as their names suggest: Autoregressive Integrated Moving Average (ARIMA technique), Extreme Gradient Boosting (XGBoost technique), and Long Short-Term Memory (LSTM technique).

1. Arima Model:

ARIMA usually has this favorable reputation among modeling linear time series data whenever research data or subjected test data either meets or has been engineered to meet stationarity conditions. It is a canonical time series forecasting model that describes the predicted future stock prices through their relationship with past values and forecast errors. The decomposition of ARIMA has three specific elements: Autoregression (AR), which puts the present value in the context of past values (lags); Differencing (I), which converts a non-stationary series into a stationary one by subtracting previous values in order to eliminate trends; and Moving Average (MA), which models the present values against past errors of the model. Although commonly considered to be very useful and successful in time series forecasting, ARIMA generally fails when it has to deal with non-linear data and sudden market shifts. Therefore, it should be considered as the most simple baseline to compare with more sophisticated models. However, without using these advanced features, ARIMA would deliver a very simple forecast using the history of stock prices indexed through time and it could perform really well in the cases where stock price data is more stable and predictable.
Considering historical data, ARIMA is used in our project as a baseline model for the prediction of stock prices. ARIMA seems to be a good starting point in its assumptions of linear relationships against complex models such as XGBoost and LSTM.

2. XGBoost Model

XGBoost, which stands for Extreme Gradient Boosting, uses data models with an extremely high capacity and can typically work with modeling complex relationships from such data sets, particularly in identifying non-linear patterns. It's a machine learning technique based on the gradient boosting algorithm, where an ensemble of decision trees is created sequentially, with each tree correcting the mistakes of the prior one. XGBoost is well known because of its effectiveness in capturing non-linear interactions in data, which is majorly critical in financial markets, as stock prices respond to many irreducibly non-linear factors, such as market sentiment, economic news, and other external events. XGBoost can also take in several features or inputs for stock price prediction using technical indicators such as moving averages and RSI to learn patterns for prediction accuracy. This is applied to develop a model that shows non-linear interactions between stock price features, which ARIMA is incapable of addressing complex patterns through the implementation of XGBoost. This works for predicting stock prices, as well as determining their price direction (whether up or down), to improve from traditional methods and bring a more robust solution to predict stock prices than ARIMA.
XGBoost is used in out model because it is an approach that can handle potential non-linear interactions among stock price features. It is very suitable for the task of predicting stock prices, as it learns complex patterns that ARIMA cannot learn. XGBoost is being used in tandem to model price predictions and price direction (up/down), which improve accuracy well beyond that of the traditionally used approaches.

3. LSTM Model

The reason why LSTM architecture was chosen is that it has the capability to establish long-term dependencies from sequential data, using stock prices variations through time as an example. This type of RNN 'LSTM' can retain information from past time steps, which becomes very appropriate for time series forecasting in which the previous prices influence the present. This model can capture temporal and non-linear dependencies contained in stock price data which might be missed by ARIMA or XGBoost.
LSTM is a recurrent neural network architecture that addresses the specific problem of modeling time series data by capturing long-term dependencies. LSTMs operate on a different principle from simple RNNs because they possess special memory units that enable the model to "remember" information for long periods. This capability is particularly relevant to stock price forecasting, where current price movements are often very much influenced by the competitive price movements of yesterday. The temporal behavior such as seasonality and trends may be learned by the LSTM model from the fact that this method processes sequences of stock prices over time.
In our model, LSTM is used for predicting future stock prices from learning the sequence the data has assigned it to in the stock market. LSTM, indeed, can replicate the long-term dependency which is there between the stock prices, such as trends or seasonal cycles, which the other two couldn't be able to cover entirely ARIMA or XGBoost. It would help take the stock price series data at the current time stamp and formulate probable next steps forecasting for the price series by supplementing the movement of historical data against the present values in low or zero space.

IV. Using the Template

1. Data Collection and Preprocessing

The initial step of the project is to collect the data for the top 100 companies on the S&P 500 index. The data is collected from Yahoo Finance library which help in providing the stock prices along with the data on Open High Low Close and Volume (OHLCV). These data points are the foundations of the stock price predictions.

In addition, multiple technical indicators will be computed for the model to make predictions more precise, giving insight into the price

movements and trends in the market. The Exponential Moving Averages (EMA) will be calculated using 2 windows (12 and 26) popularly used in technical analysis to detect trends and reversals. The Relative Strength Index (RSI) is computed using the ta library, a famous library for technical analysis, gauging the strength and speed of price movements, benefitting from identifying overbought or oversold conditions. Besides that, the Mean Average Convergence Divergence (MACD) and MACD Signal are computed using the same ta library as mentioned before. The MACD helps in identifying the changes in the strength, direction, momentum, and duration of a trend in stock prices. These technical indicators - EMA, RSI, and MACD - are added into the dataset using the add_features() function, which integrates into the data collection. These greatly enhance the model's ability to capture complex market dynamics for even better forecasting performances.

Data cleaning is done on the data to make sure that the dataset is clean and matches the dataset integrity. For this, we deal with the missing values using imputation or interpolation techniques, which also ensures that there are no issues in the data that could later affect while training the model.
Data normalization is carried out in order to allow it to be compatible with models such as XGBoost and LSTM. This is important because those algorithms are sensitive to the input data scale, and normalization helps them converge more effectively during training.

2. Model Configuration

The ARIMA model considers an identification scheme that defines the optimal parameters (p, d, q) through the ACF and PACF plots, which help identify the autoregressive and moving average components. ARIMA is mainly applicable when the data is stationary or can be made stationary by differencing, which refers to the d value. Thus, it is a classical type of linear time series data, which is considered as a benchmark to be compared with more complex models like XGBoost and LSTM.

The hyperparameter optimization using grid search and cross-validation is carried out on XGBoost. Max_depth, learning_rate, n_estimators, and subsample are just a few of the important parameters tuned to improve model performance. XGBoost is designed for regression tasks (stock price prediction) and for classification tasks (predicting price direction), thereby allowing modeling of both price levels and price trends and hence making it a potent player in stock price prediction.

In LSTM, a structure with several layers of LSTMs followed by Dense layers for the final prediction of output is constructed. Emphasis is given to hyperparameters such as number of layers, units per layer, batch size, and epochs, which are selected through cross-validation to prevent overfitting. Length of input time sequence such as 30 days or 60 days, is also a significant hyperparameter based on observations made from stock price data.

3. Training the Models

As it stands, the model training process is implemented via a 70-30 data split; that is, 70% availability of historical stock price data is used for training, and 30% is reserved for testing. This way, it successfully trains the model on a larger portion of the data, while also giving a chance to evaluate its performance based on unseen data, a very important criterion of performance evaluation.
The ARIMA training thus fits the model to historical stock price data. ARIMA is an approach to time series forecasting which considers relationships between past values and forecast errors. After training with 70% of data, a model is used to predict stock prices in the future based on some historical patterns. The performance is then measured by comparing the predicted stock price with the actual values from the testing set, and RMSE and MAE are some of the accuracy metrics used to validate it.
The same training process is followed for XGBoost, a gradient-boosting model, but the focus here is on modeling non-linear relationships in the data. It makes use of decision trees to capture these high-complexity interactions by scaling their weights to minimize prediction errors during training. Once trained on the training set, the model predicts stock prices and evaluates its predictability on the test set. Added strength of the model lies in being able to further include additional features such as technical indicators (RSI, MACD) as part of its learning.
The LSTM model has been designed specifically for training on sequential data so that the training of the model focuses on the long-term dependencies between the movements of stock prices. Sequences of stock prices are fed into the network such that the model learns to predict future prices based on prior data. Once trained, an LSTM model is then evaluated against the test set and its capability to capture temporal dependencies is then judged by comparing the outcome of predictions against actual stock prices.
In this way, each model gets to be tested on its ability to predict stock prices knowing historical information as well-laid down rules of its specialty.

4. Performance Metrics

In this project, the performance of ARIMA, XGBoost and LSTM is measured through RMSE (Root Mean Square Error), MAE (Mean Absolute Error), Directional Accuracy and R-square. The historical stock price data was split into 70% for training and 30% for testing. The trained models were then tested with the holdout data to compare the predicted with actual values. The relevant performance metrics for the LSTM model include RMSE and MAE, which represent the accuracy of the stock price forecasts by the model. For example, in the 20 day forecast, the RMSE was settled at 21.01 and the MAE at 15.07, thus indicating how much the actual values correspond to prediction. In similar combination, the scores of XGBoost are that RMSE and MAE plus are added for Directional Accuracy, which indicates up or down pricing movement direction with respect to price from the preceding period. The evaluation of ARIMA is carried out in consonance and also registered as R squared value, which gives the peaceful explaining variance related to the stock price. With these metrics, models can easily be analyzed and compared comprehensively relative to predictive capacity evaluation to enhance the robustness with which they are gauged in making forecasts of stock prices for the 1-day, 5-day, and 20-day predictions.

5. Model Evaluation

The models (ARIMA, XGBoost, and LSTM) are evaluated for performance over three different time horizons, namely, 1-day, 5-day, and 20-day predictions. This evaluative procedure is paramount in the assessment of whether the models forecast stock prices over both short and long terms, thereby reflecting the dynamics of stock-market behavior with respect to different volatility levels. One-day predictions test the models for stock price movement on the next day, as such estimates are highly challenged by volatility, where the models have been judged upon their apprehension of such rapid fluctuations. The 5-day prediction horizon tests the response of the models to moderately volatile stock data, assessing their ability to predict trends and price moves over the short term. Finally, for 20-day prediction the horizons looks at the long-term forecasting, thus help in evaluating the mode based on how they represent the trends, seasons and general market movements. This longer horizon is very important for an investor that wishes to watch the market over time. By evaluating the models across these three horizons, one can fully use the model through performance evalauation and also which model performs best under different stock prices.

V. METHODOLOGY

The approach of this project is systematic in predicting stock prices through the application of three models- ARIMA, XGBoost, and LSTM. The objective is to verify the performance of these models predicting short-term and long-term stock price movements on the basis of historical data of the top 100 companies from the S&P 500 index. The methodology consists of various segments: data collection and preprocessing, feature engineering, model training, evaluation, and performance comparison. The following sections will describe all of these processes in great detail:

1. Data Collection and Preprocessing

Data collection and preprocessing is collecting historical stock price data for the top 100 companies in the S&P 500 index. Fairly reliable data either from Yahoo Finance or Alpha Vantage is obtained, consisting of Open, High, Low, Close, and Volume (OHLCV). All the models depend on these datasets.

Technical Indicators: The basic stock price data will be supplemented with the calculation of Exponential Moving Averages (EMA), Relative Strength Index (RSI), and Moving Average Convergence Divergence (MACD). These technical indicators are important to capture trends and patterns in stock prices and will, therefore, help improve the predictive accuracy of our model.

Data Cleaning: The dataset will be cleaned, whereby techniques for dealing with missing values will be applied. Either imputation methods or interpolation will be used to ensure that no gaps in the data exist that would interfere with the training process.

Normalization: The data will be normalized so that the features are on a similar scale. This step is relevant for machine learning models, such as XGBoost and LSTMTraining, which are sensitive to the scale of input features.[14]

2. Feature Engineering

To make the models more powerful in prediction, different indicators are calculated and strategically included in the data policy. Such as the EMA (12 and 26 periods), RSI, MACD values commonly made for prediction work of stock prices. The add_features() function is employed to add such indicators to the given data for establishing a valid background in the model's input regarding market speed and possible price movements. Indicators would operate as input features that would add additional advantage for models to better contextualize market behaviour fully with more complex forms of capturing and hence better forecasting accuracy.

3. Data Splitting

Once prepared and added, the data shall be chopped into training and test datasets. As a rule of thumb, 70% of the data is for training while the rest 30% is for testing of the models. A training dataset fits the model and builds foreign relationships between the input characteristics and the future stock prices. A testing dataset tests the model's ability to process information that was unknown to it previously.

4. Model Learning

All these models are trained by absorbing the training data and examined by obtaining results by applying the testing data.

ARIMA: The baseline model will be ARIMA which will be trained under stationary conditions on stock price data. The ACF and PACF plots will help to identify the selected ARIMA (p, d, q) values. The model will be fitted to the data and is used to forecast stock prices based on historical values and errors. It is the ARIMA, which serves as baseline benchmark for traditional linear time series forecasting. It serves best as pointer to the far advanced XGBoost and LSTM models in terms of benefits from non-linearity and sequential dependencies.

XGBoost: XGBoost would be trained on feature-engineered data, that is, stock prices together with their technical indicators. The hyperparameters max_depth, learning_rate, n_estimators, and subsample are optimized through grid search and cross-validation to ensure the model runs very well. Perhaps better than normal training models, it captures the non-linear relationships and is applicable in regression (stock prices) and classification (price direction).

LSTM: Again, LSTM networks have been built for capturing the sequencing behavior. The architecture of the model has LSTM layers followed by Dense layers at the end. The resource parameters including number of layers, units per layer, batch size, and epochs are picked through cross-validation to avoid overfitting with the possibility of ensuring higher generalization. Thus, the sequence of stock price through time serves as the LSTM input for capturing long-term dependency and trend.

5. Model Assessment

At the end of training, evaluations were done on the models based on some performance metrics:

- **RMSE** (Root Mean-Squared Error): The average magnitude of the errors in the predictions.
- **MAE** (Mean-Absolute Error): The average of the absolute error between the predicted value versus the actual value.
- **Directional Accuracy** refers to the accuracy of the model in predicting the direction of the stock price movement (up or down).
- R-**squared**: The extent to which the model accounts for stock price variations.

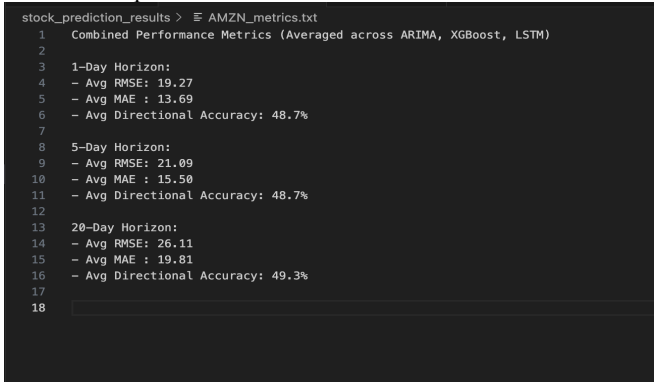
The evaluation is done at three different time horizons: 1 day, 5 days, and 20 days. The 1 day prediction tests the model under highly volatile data; the 5 day horizon is to assess models under medium-term prediction of stock prices. For the 20 days prediction, the focus will be on how well long-term trends and seasonality can be captured in the stock market.

6. Comparison of Performances

After this phase, the models are compared on the performance evaluated for the defined time horizons. The goal behind this is to identify which model strikes the right balance between prediction accuracy and computational efficiency. Comparing ARIMA, XGBoost, and LSTM over 1-day, 5-day, and 20-day prediction will help pick the best performance in stock price forecasting under different market conditions.

VI. EVALUATION

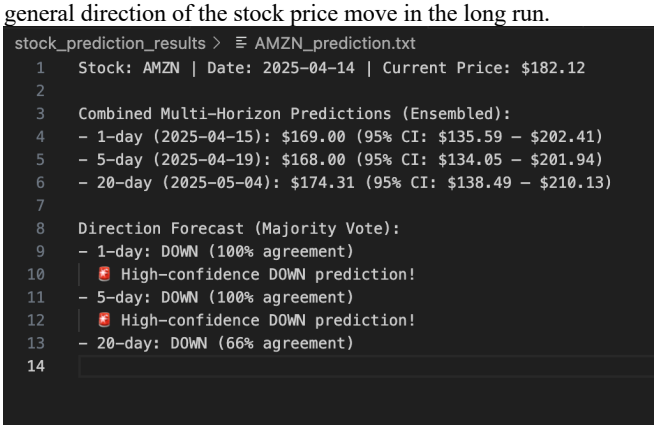
There are three different models considered in this project's evaluation: ARIMA, XGBoost, and LSTM for stock price prediction across three time horizons: 1-day, 5-day, and 20-day. Evaluation metrics comprise RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), Directional Accuracy, and R-squared. While the example mentioned above is that of Amazon (AMZN), this methodology and evaluation apply to the entire universe of 100 stocks as per the S&P 500 index for thorough analysis among different companies.



For the 1-day horizon, there is an average RMSE of 19.27, which is a reasonable estimate for short-term prediction accuracy; an average MAE of 13.69 indicates the absolute average error magnitude between predicted and actual values. The directional accuracy sits at 48.7%, which is almost just as likely to predict correct direction (up or down) as incorrect.

As for the 5-day horizon, the average RMSE peaks to 21.09, while the average MAE is 15.50, indicating slightly higher prediction errors as a result of the extended time period. The directional accuracy, however, remains constant at 48.7%.

The values put forth in terms of average RMSE for predictions made over 20 days lie at 26.11, while those made over a similar time frame create an average MAE of 19.81; these figures suggest that predictions of stock price movements over longer periods prove to be particularly challenging for the models in question. The directional accuracy also gets a slight bump in figures at 49.3%, indicating that models perform with more accuracy in predicting the general direction of the stock price move in the long run.



According to the models for stock price predictions, the forecast price for Amazon is:

1-day prediction: $169.00 (95% CI: $135.59-$202.41)
5-day prediction: $168.00 (95% CI: $134.05-$201.94)
20-day prediction: $174.31 (95% CI: $138.49-$210.13)

The directional forecast for Amazon indicates 100% agreement for downward price movement predictions for both the 1-day and 5-day horizons. For the 20-day prediction, there is 66% agreement on the downward movement, showing a decrease in confidence over the longer term.

The results here show that even though the models predict well for the short term, there is still room for improvement when making longer-term forecasts. They have proven to be beneficial directional inputs important to market decision-making, even if not up to par with longer time frame predictions. Furthermore, such findings obtain a high credibility test, particularly when extended to the performance profile of all 100 S&P 500 stocks; thus, the outcome can be said to be robust.

Evaluation of ARIMA showed that it performs poorly for non-stationary stocks with sudden trend changes. On cases of some extremely volatile stocks, XGBoost fitted to noise and so did LSTM's. These two, though, suffered with long-term sequences that did not have enough history information. Future models should take into consideration very strong regularization and advanced ensemble methods to further reduce such modelling errors.

VII. Conclusions and Future Work

The project aims to consider the prediction of stock price movements through three distinct methods, which are ARIMA, XGBoost, and LSTM. In all the 3 methods, stock data were analyzed with the top 100 companies in the S&P 500 index through 1-day, 5-day, and 20-day predictions. In an attempt to derive a complete measure for each model's predictive accuracy and effectiveness, appropriate performance metrics such as RMSE, MAE, Directional Accuracy, and R-squared have been used.

The ARIMA model has been productive in most time series forecasts, but it has shown limitations when dealing with non-linear data sets and abrupt changes in the market. The model has worked fairly well in short-range forecasting, failing rather badly in long-range forecasting, particularly at 5 days and 20 days. Hence, there will be a need for sophisticated models that address this behavior of the stock market and its complexities.

XGBoost is actually much superior compared to ARIMA as it is a gradient-boosting machine learning model particularly in modeling non-linear relationships in the stock price data. It is good for medium but excels in long-term forecasting and thus outperformed ARIMA based on predictive power. XGBoost was capable of modeling diverse complex interactions among the stock price features and gave much better predictions in that regard.The model is very suited to handling sequential learning data: LSTM is a powerful deep learning model trained by its exposure to the lives of stock continuity dependence. It fitted well across all time frames; mainly, it becomes good at trending long-term movements. However, careful hyper-parameter tuning is required for best performance from LSTM models.

The training and inference times for XGBoost outperform those of LSTM. The latter, however, because of the sequence modeling process involved, took a much longer period training but gave better long-term trend capture. The least computational cost was by ARIMA, but the accuracy of its predictions on non-linear data was limited. Such trade-offs prove important when considering deployment in real time.

Overall XGBoost and LSTM outperformed or were better than ARIMA, but XGBoost won on the winning side, as it modeled both linear and non-linear relationships of stock prices data.

Future work

The results being impressive, there are certainly a few possible areas for improvement. Firstly, a hybrid model by combining XGBoost and LSTM could be attempted, as it exploits the strength of both worlds: concerted advantage of the components- non-linear relationship from XGBoost and long-term effective dependencies from LSTM. This hybrid approach could improve forecast precision, especially for difficult stock price data.

Adding cross-validation for the model of XGBoost will reduce overfitting and increase generalization of the model in terms of prediction variables. Presently, the models are trained by a 70-30 training-test split, but implementing techniques such as KFold or TimeSeriesSplit cross validation would yield a more robust performance metric.

Additionally, hyperparameter tuning on the XGBoost and LSTM models was very marginally, if at all, explored in this project. Hyperparameter tuning based on grid search or Bayesian optimization methods would probably bring improvements in the performance of both models. One way of getting and testing new ideas could be the investigation of advanced approaches such as GRU (Gated Recurrent Units) or Transformers for time series forecasting, as they can provide a new flavor in terms of capturing long-range dependencies as well.

Generic approach with sentiment analysis given by financial news with some additional feature could lead to the more general approach of this model which could prove very significant improvement performance-wise by giving a fine context of market sentiments which are related to huge contributions on stock prices. Developing great real-time simulated prediction results that continuously update their predictions as new market data becomes available is one practical approach to further developing models in real financial markets.

All these improvements can further convert the stock price prediction models into more accurate and robust digital tools for gleaning valuable insights tied to most probable decisions that would be made by investors and analysts at a given time.

VIII. REFERENCES

[1]J. Zou *et al.*, "Stock Market Prediction via Deep Learning Techniques: A Survey," Dec. 2022. Accessed: Jul. 25, 2023. [Online]. Available: https://arxiv.org/pdf/2212.12717

[2]S. Gade, S. Sayyad, and Student, "STOCK MARKET PREDICTION USING ARIMA AND MACHINE LEARNING," *International Journal of Scientific Development and Research*, vol. 8, 2023, Accessed: Apr. 26, 2025. [Online]. Available: https://www.ijsdr.org/papers/IJSDR2304130.pdf

[3]T. Chen and C. Guestrin, "XGBoost: a Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, vol. 1, no. 1, pp. 785–794, Aug. 2016, doi: https://doi.org/10.1145/2939672.2939785.

[4]S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: https://doi.org/10.1162/neco.1997.9.8.1735.

[5]Y. Wang and Y. Guo, "Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost," *China Communications*, vol. 17, no. 3, pp. 205–221, Mar. 2020, doi: https://doi.org/10.23919/jcc.2020.03.017.

[6]A. Chatterjee, H. Bhowmick, and J. Sen, "Stock Price Prediction Using Time Series, Econometric, Machine Learning, and Deep Learning Models," *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, pp. 289–296, Oct. 2021, doi: https://doi.org/10.1109/MysuruCon52639.2021.9641610.

[7]P. Hoang Vuong, T. Tan Dat, T. Khoi Mai, P. Hoang Uyen, and P. The Bao, "Stock-Price Forecasting Based on XGBoost and LSTM," *Computer Systems Science and Engineering*, vol. 40, no. 1, pp. 237–246, 2022, doi: https://doi.org/10.32604/csse.2022.017685.

[8]P.-F. Pai and C.-S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, Dec. 2005.

[9]A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, Mar. 2014, doi: https://doi.org/10.1109/uksim.2014.67.

[10]J. Cao, Z. Li, and J. Li, "Financial time series forecasting model based on CEEMDAN and LSTM," *Physica A: Statistical Mechanics and its Applications*, vol. 519, pp. 127–139, Apr. 2019, doi: https://doi.org/10.1016/j.physa.2018.11.061.

[11]S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM," *arXiv:1911.09512 [cs, stat]*, Nov. 2019, Available: https://arxiv.org/abs/1911.09512

[12]A. Gifty and Dr. Yang Li, "A Comparative Analysis of LSTM, ARIMA, XGBoost Algorithms in Predicting Stock Price Direction," *Engineering and Technology Journal*, vol. 09, no. 08, Aug. 2024, doi: https://doi.org/10.47191/etj/v9i08.50.

[13]Z. Zhu and K. He, "Prediction of Amazon's Stock Price Based on ARIMA, XGBoost, and LSTM Models," *Proceedings of Business and Economic Studies*, vol. 5, no. 5, pp. 127–136, Oct. 2022, doi: https://doi.org/10.26689/pbes.v5i5.4432.

[14]"Yahoo Finance," *Yahoo Finance*, 2025. https://finance.yahoo.com/?guccounter=1&guce_referrer=aHR0cHM6Ly9jaGF0Z3B0LmNvbS8&guce_referrer_sig=AQAAAJUw6QU9WIQWhgkZPd0G4lie0z1ZmDhk5txtrFsTwSYOYoY3eQkXU2hkUcyMy4ntVITc3oeOpM1V3XYWw9fEvO1zeUXZa8D8PARtO3TY_XgFn4DEaE-ikcooaOJoWkN3EFAl5u22zvp5tQMqZ-pfuIsdd7g4rWkqwVtTsvGQI1MP (accessed Apr. 27, 2025).