Stock Market Price Analysis and Prediction: LSTM-Based Approach

Paridhi Arya (S20210020305), Pratik Agrawal (S20210020312), and Vaibhav Prajapati (S20210020328)

***Abstract*—Time Series data plays a crucial role in financial markets, offering insights into the dynamics of stock prices. This project focuses on employing Long Short Term Memory (LSTM) networks for the statistical analysis and prediction of stock price movements, with a specific emphasis on technology stocks such as Apple, Amazon, Google, and Microsoft. The project follows a comprehensive methodology, encompassing data retrieval, descriptive statistics, technical analysis, and LSTM-based predictive modeling.**

This project employs a comprehensive approach to analyze and predict stock price movements, focusing on technology stocks such as Apple, Google, Microsoft, and Amazon. Utilizing Python libraries, the study explores descriptive statistics, moving averages, daily returns, and correlation between stock prices. Furthermore, the implementation of Long Short-Term Memory (LSTM) models for stock price prediction is showcased. The findings reveal insights into risk assessment, technical indicators, and predictive modeling, offering valuable contributions to the field of stock market analysis.

***Index Terms*—Stock Price Prediction, Feature Extraction, Long Short-Term Memory (LSTM), Predictive Modeling, etc.**

I. Introduction

A. Motivation

IN the realm of financial markets, particularly concerning technology stocks, the motivation for delving into the realm of stock price prediction is grounded in the intricate dynamics and inherent volatility characterizing these markets. The quest for reliable tools capable of deciphering and forecasting market trends is a perpetual endeavor for investors and traders alike. Conventional financial models often grapple with the multifaceted nature of stock movements, necessitating a more adaptive and nuanced approach. This introduction sets the stage for the utilization of Long Short-Term Memory (LSTM) networks, a variant of recurrent neural networks (RNN), implemented through a Python script. By embracing the capabilities of deep learning, the objective is to offer a refined and more precise predictive modeling framework, adept at navigating the intricate relationships and swift fluctuations emblematic of the technology stock sector.

B. The Classification Problem

In tackling the formidable challenge of accurately predicting technology stock prices, this study delves into the realm of advanced machine learning techniques, with a particular focus on Long Short-Term Memory (LSTM) networks. The intricacies and volatility inherent in the technology stock market pose a significant hurdle for conventional financial analysis methods. Recognizing the limitations of traditional approaches, the research turns to LSTM networks for their proven efficacy in modeling temporal dependencies, making them well-suited for capturing the nuanced dynamics of stock price movements.

The motivation behind this departure from conventional methods lies in the rapid pace of technological advancements and their profound impact on stock markets. Technological shifts introduce a level of complexity that demands a more sophisticated and adaptable predictive model. By harnessing the capabilities of LSTM networks, the study aspires to develop a robust forecasting model capable of discerning subtle patterns within historical stock data. This approach aims to provide a more nuanced and accurate understanding of the intricate relationships and trends that govern the technology stock market, ultimately contributing to more informed and precise predictions in this dynamic financial landscape.

II. Methodology

A. Dataset Description

The dataset utilized in this study was obtained through the `yfinance` library, capturing historical financial data for the selected financial instrument. The dataset includes the following key columns:

1) **Date:** The date corresponding to each trading day.
2) **Open:** The opening price of the financial instrument for the given trading day.
3) **High:** The highest price reached by the financial instrument during the trading day.
4) **Low:** The lowest price reached by the financial instrument during the trading day.
5) **Close:** The closing price of the financial instrument for the given trading day.
6) **Adj Close:** The adjusted closing price, incorporating corporate actions like dividends and stock splits.
7) **Volume:** The total number of units of the financial instrument traded on the specific trading day.

The table below illustrates a snippet of the dataset:

Date	Open	High	Low	Close	Adj Close	Volume
2023-01-01	150.00	155.00	148.50	153.00	152.50	100,000
2023-01-02	153.20	158.00	152.10	155.80	155.00	120,000
...	...	...	...	...	...	...

This dataset serves as the foundation for subsequent analyses, allowing for the exploration of patterns, trends, and the impact of selected features on the predictive models employed

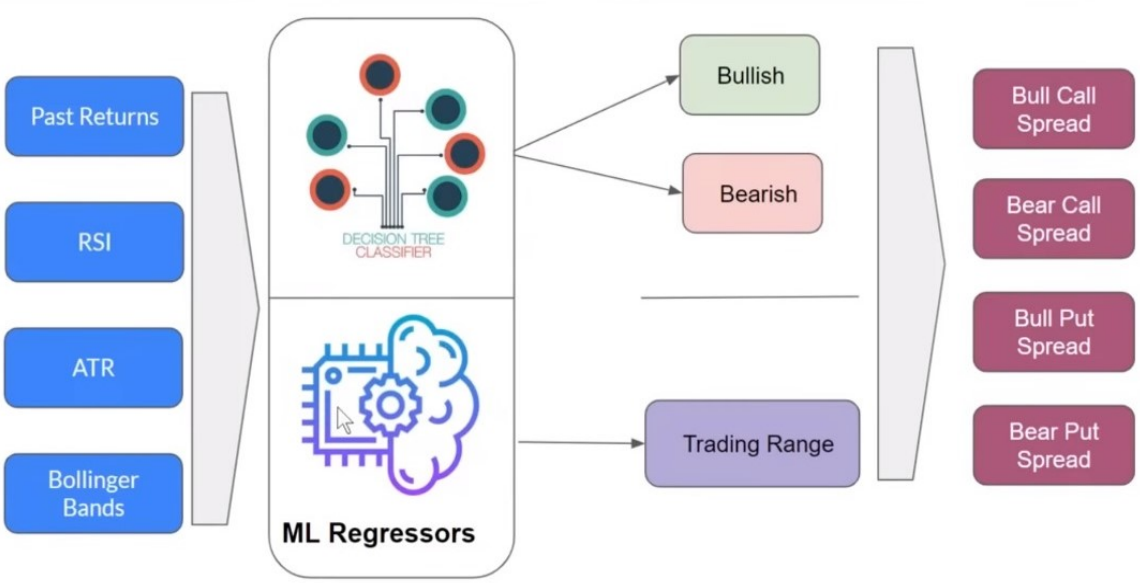

Fig. 1: Block Diagram of Stock Market Closing Price Prediction using LSTM

in this study. The next sections detail the preprocessing steps, feature extraction, and visualization techniques applied to enhance the dataset's suitability for the predictive models.

B. Additional Features Derivation

We derive additional features from the raw data to enhance the predictive power of our model. Let C_t represent the closing price at time t, H_t and L_t denote the high and low prices, and V_t is the volume. The following features are derived:

- **Momentum_1D**: $C_t - C_{t-1}$, capturing the change in closing price over one day.
- **RSI_14D**: The 14-day Relative Strength Index, RSI_{14D}, is calculated using the formula:

$$RSI_{14D} = 100 - \frac{100}{1 + \frac{AverageGain_{14}}{AverageLoss_{14}}}$$

  where $AverageGain_{14}$ and $AverageLoss_{14}$ are the average gains and losses over the 14-day period.
- **Volume_plain**: Raw volume data without any transformations, denoted as V_t.
- **Bollinger Bands**: Comprising the Middle Band (`BB_Middle_Band`) and Upper/Lower Bands (`BB_Upper_Band`, `BB_Lower_Band`), calculated using the formula:

$$MiddleBand = SMA_{20}(C_t)$$

$$UpperBand =$$

$$MiddleBand + numsd \times StandardDeviation_{20}(C_t)$$

$$LowerBand =$$

$$MiddleBand - numsd \times StandardDeviation_{20}(C_t)$$

- **VWAP**: Volume Weighted Average Price, calculated as:

$$VWAP_t = \frac{\sum_{i=1}^{t} V_i \times \frac{H_i + L_i}{2}}{\sum_{i=1}^{t} V_i}$$

- **High-Low Range (HL)**: The difference between the high and low prices, $HL_t = H_t - L_t$.
- **absHC** and **absLC**: Absolute differences between the high/low of the current day and the previous day's closing price.
- **True Range (TR)**: $TR_t = \max(HL_t, absHC_t, absLC_t)$.
- **ATR**: The 14-day Average True Range, calculated as the mean of TR_t over a 14-day window.
- **NATR**: Normalized Average True Range, defined as $\frac{ATR_t}{C_t} \times 100$.

C. Data Pre-Processing

In the context of our stock market analysis project, data pre-processing plays a crucial role in enhancing the quality and relevance of our analytical results. Our initial step involves obtaining time series data for technology stocks (Apple, Google, Microsoft, and Amazon) from Yahoo Finance using the yfinance library. This raw data, indexed by date, includes essential features such as the adjusted closing prices, trading volumes, and other financial metrics. To ensure consistency and accuracy in our analysis, we employ various pre-processing techniques.

We handle missing weekend records, perform descriptive statistical analyses using methods like `.describe()` and `.info()` to gain insights into the data structure, and visualize trends over time through matplotlib and seaborn. Furthermore, we engineer additional features, such as moving averages, daily returns, and technical indicators like the Relative

Strength Index (RSI) and Bollinger Bands. Standardization and normalization techniques are applied to scale the data appropriately.

To assess the normality of the daily returns, we conduct the Shapiro-Wilk test, a statistical test used to determine whether a sample comes from a normal distribution. This test provides valuable insights into the distributional properties of the data, helping us make informed decisions about the appropriateness of certain statistical analyses.

The resulting pre-processed dataset becomes a foundation for our subsequent exploratory data analysis and predictive modeling, enabling a more informed and meaningful interpretation of stock market trends and behaviors.

D. Feature extraction and visualization

In the context of feature extraction for our predictive modeling utilizing Decision Tree Regression, a meticulous process was initiated to curate a set of pertinent features for analysis. The selected features encompassed metrics such as Normalized Average True Range (NATR), Volume Weighted Average Price (VWAP), the Middle Band of Bollinger Bands (BB_Middle_Band), the 14-day Relative Strength Index (RSI_14D), unprocessed trading volume (Volume_plain), one-day Momentum (Momentum_1D), and adjusted closing prices (Adj Close). To ensure uniformity and scalability in our analysis, Min-Max scaling was applied to normalize these features. Subsequently, the target variable, represented by the adjusted closing prices, was temporally shifted by one day to facilitate the predictive modeling approach.

To gauge the significance of each feature within our predictive model, a Decision Tree Regressor was employed, and feature importances were assessed. The analysis revealed VWAP as the most influential feature, contributing approximately 80% to the predictive efficacy of our model. Conversely, the adjusted closing prices (Adj Close) exhibited a lesser, albeit still noteworthy, influence, contributing approximately 20%. These findings offer valuable insights into the critical determinants influencing the prediction of closing prices, guiding the selection of features that contribute most substantially to the accuracy and reliability of our model.

The visual representation of feature importance through bar plots further underscores the prominence of `VWAP`, emphasizing its pivotal role in predicting stock closing prices. This insight into feature importance serves as a foundation for the refinement of our predictive model, ensuring optimal performance by leveraging the most impactful features in a principled and methodical manner.

III. Results

A. Performance Without Feature Extraction

In this subsection, we assess the predictive performance of our model without employing feature extraction techniques. The methodology involves training a Long Short-Term Memory (LSTM) neural network on scaled data, incorporating a sequence of historical closing prices. The evaluation is based on root mean squared error (RMSE), providing insights into the accuracy of the model's predictions.

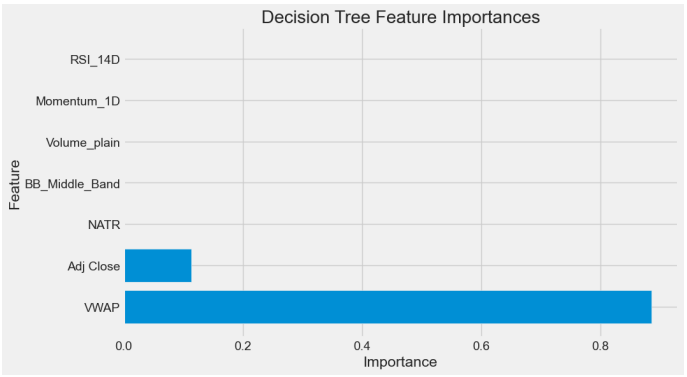

Fig. 2: Decision Tree Feature Importance

In the absence of feature extraction, our predictive modeling approach utilizes a Long Short-Term Memory (LSTM) neural network trained on scaled data. The model incorporates a historical sequence of closing prices to predict future values. The performance evaluation, quantified by the root mean squared error (RMSE), provides valuable insights into the accuracy of the model's predictions. The obtained RMSE before feature extraction is 7.54.

B. Performance With Feature Extraction

In this subsection, we enhance our predictive modeling by incorporating feature extraction, particularly leveraging the Decision Tree method to identify the most influential feature. The identified feature, VWAP (Volume Weighted Average Price), is integrated into the LSTM neural network model. The evaluation metrics, including RMSE, illustrate the impact of feature extraction on predictive accuracy.

Building upon our predictive modeling framework, we introduce feature extraction to identify the most impactful feature using the Decision Tree method. VWAP (Volume Weighted Average Price) emerges as the key feature, significantly enhancing our LSTM neural network model. The evaluation metrics, including RMSE, provide a comparative analysis of the model's performance before and after incorporating feature extraction. The obtained RMSE after feature extraction is 3.49, representing a notable improvement. The improvement percentage is 53.73%, demonstrating the effectiveness of leveraging feature importance in our predictive modeling approach.

Metric	Value
RMSE before feature extraction	7.54
RMSE after feature extraction	3.49
Improvement Percentage	53.73%

TABLE I: Performance Metrics

IV. Conclusion

In conclusion, this research endeavor focused on refining stock market prediction models by integrating advanced feature extraction techniques into LSTM models. The comprehensive journey from literature review and guidance through real-time demonstrations to data pre-processing, LSTM modeling,

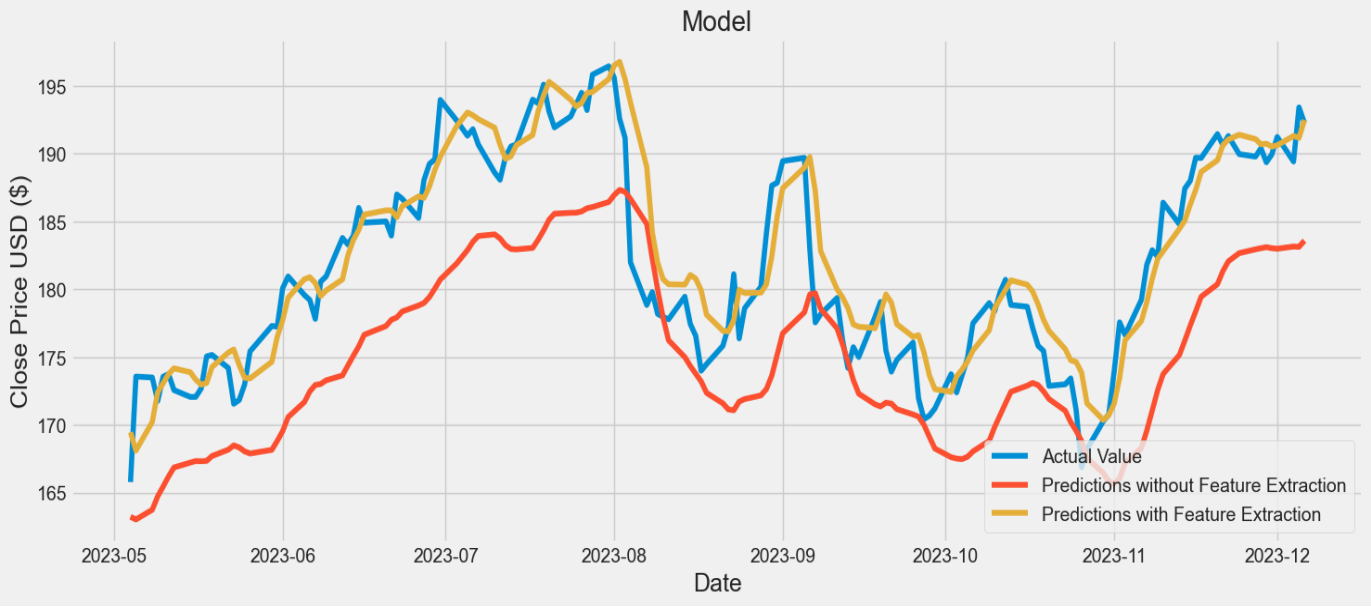

Fig. 3: Stock Market Closing Price Prediction (Using Both Cases)

and comparative analysis has yielded insightful findings and notable enhancements in predictive accuracy.

The development of LSTM models incorporating time series analysis laid the foundation for predictive modeling. Further advancements in feature extraction techniques greatly enhanced the models by deriving additional features crucial for improving predictive performance. The pivotal comparative study between models with and without the derived features demonstrated a clear and significant improvement in RMSE by 53.73% when incorporating the extracted features. This comparison underscored the importance of feature extraction in augmenting the models' predictive capabilities.

V. References

[1] S. B. Islam, M. M. Hasan, M. M. Khan, "Prediction of Stock Market Using Recurrent Neural Network," in *IEEE - IEMCON*, https://doi.org/10.1109/IEMCON53756.2021.9623206.

[2] P. S. Sisodia, A. Gupta, Y. Kumar, G. K. Ameta, "Stock Market Analysis and Prediction for Nifty50 using LSTM Deep Learning Approach," in *IEEE - ICIPTM*, https://doi.org/10.1109/ICIPTM54933.2022.9754148.

[3] X. Weng, X. Lin, S. Zhao, "Stock Price Prediction Based On LSTM And Bert," in *IEEE - ICMLC*, https://doi.org/10.1109/ICMLC56445.2022.9941293.

[4] K. J, H. E, M. S. Jacob, D. R, "Stock Price Prediction Based on LSTM Deep Learning Model," in *IEEE - ICSCAN*, https://doi.org/10.1109/ICSCAN53069.2021.9526491.

Appendix A
Individual Contributions

This project commenced with a comprehensive literature review, supplemented by guidance and real-time demonstrations from the team lead and subject matter expert, Pratik Agrawal, enabling a profound understanding of stock market dynamics. The project's pivotal aspect involved comparing model performance with and without feature extraction, revealing substantial improvements in predictive accuracy, and affirming the pivotal role of advanced feature extraction in refining stock market prediction models.

- Paridhi Arya:
  - Led the efforts in time series analysis, LSTM model building, testing, and subsequent improvements.
  - Spearheaded the development of models and executed rigorous testing protocols for enhancement.
- Pratik Agrawal
  - Engaged in feature extraction techniques on the time series data and led the comprehensive analysis of results.
  - Executed feature extraction methodologies and conducted an in-depth analysis of the obtained results.
- Vaibhav Prajapati:
  - Focused on exploratory data analysis (EDA) related to the stock market and utilized models for predictive analysis.
  - Conducted in-depth EDA on stock market data and applied various models for time series data prediction purposes.

Appendix B
Github Repository

Code for Stock Market Analysis and Prediction ('main.ipynb' in linked Github Repository)