Stock Market Analysis

Swagata Malik
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22252@bl.students.amrita.edu

Anaswara Reji
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham,India
bl.en.u4cse22273@bl.students.amrita.edu

Ruchi Chaurasiya
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, Ind
bl.en.u4cse22278@bl.students.amrita.edu

Shakshi Yadav
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham,India
bl.en.u4cse22286@bl.students.amrita.edu

Sangita Khare
Department of Computer Science and Engineering
Amrita School of Computing,Bengaluru
Amrita Vishwa Vidyapeetham,India
k_sangita@blr.amrita.edu

***Abstract*—Due to complexity and volume of data in different sectors have been increasing rapidly, big data analysis has become an important advancement for processing large and complex data in less time. It is becoming more trustworthy by processing real time data and providing security to then without doing bias based on the data structure .The characteristics of big data is enhancing it use in different fields for different propose. This paper is mainly focusing on the prediction of stock market by using the data set of different companies. Stock market prediction is to analyze the market trend or companies financial position in share market. We are overcoming the old and traditional method of prediction for stock market by using databrick and Apache spark framework.we are also using ML models for the prediction and selecting the best model based on high accuracy.**

***Index Terms*—Stock, Apache Spark,Finance, Databrick**

I. Introduction

Big data has become important for rapid development of a lot of different sectors .It has been widely used by corporate entities to codify critical business perception .It process and analyze the large volume of data. Due to the enhancement in the technology, each and every date more than petabyte of data is being generated. It is becoming tough and bulky for human to handle and work on them as of their high speed. As several tools and techniques are used in data processing. So by using big data analytics ,it become easier to store data, provide security to data ,can be analyzed any time we want and also customers acan get various other services which can satisfy them. Big data helps to collect data , interpret them an which makes the working of business easier.

There are various challenge that big data faces such as very few people have idea of using big data techniques for their business and research. The quality of data stored may affected the result. Even though the 7 different V 's of big data is ruling the data analysis world. Among 7 ,the 3 most important fundamental characteristics of big data are:

Volume: This refers to large amount of data generated and stored. It can range from terabytes to petabyte and even more.

Velocity: This refers to speed at which data is generated and being processed .It handles the real time data from sources like social media ,IoT etc and help to enable real time decisions.

Variety: This refers to the various type of data like structured, semi-structured and unstructured. Big data analytics integrate this various formats of data by providing various flexible tools can be process diverse data source.

In this paper ,we are mainly focusing on stock market analysis using big data approaches. Stock market prediction has gained high important in the field of business professionals as well as educational studies .It mainly represent the market trends and flow of different business in the market. It evaluates company based on various factors like management, economics status and financial statements. Due volatility in the market trends and change in the price value every seconds ,traditional methods is no more appropriate to work for the prediction of share market price. Nowdays ,the finance and banking sectors are monitoring the market activity, detecting financial fraud ,preventing illegal trading etc all with the help of big data analytics.

We are primarily concentrating on building a model to predict the future price of different companies .For that we are using the historic data of different companies and comparing their price so that investors can get maximum benefit from the share market of this companies. we are using a big data framework known as Apache spark inbulit with python I.e. pyspark.Apache spark is a high speed, versatile processing

engine which is prior and best choice for many financial organizations. A top build component of Apche spark , databricks has been for distributed computing of data on cloud. Different machine learning algorithms have also been used for prediction .

We are using big data ,to achieve more accurate predictions by properly preprocessing the data.

II. Literature Survey

Prit.Modi et al. [1] explicated Cloud era-Hadoop based pipe lined architecture predicting daily stock returns based on the analysis of real time US stock data taken from Yahoo Finance. Based on this paper market prices , there existed some level of predictability although they are not entirely random but rather a dynamic nonlinear system. ANN model was developed that generated historical repeating stock price daily models by feature extraction where the cases controlled were 85% training data and 15% test data. Results show that under the regularization terms in logistic regression the accuracy improves while under optimal SVMs there is higher accuracy.

A.Ashok et al. [2] emphasized the importance of accurate stock market prediction for financial decision-making. ARIMA (AutoRegressive Integrated Moving Average) was employed for its efficiency in modeling linear patterns, while LSTM (Long Short-Term Memory networks) handled the complexity of non-linear time series data. By merging these approaches, the hybrid model effectively captured both linear trends and non-linear dependencies in stock prices.

Mazhar Javed Awan et al. [3] proposed both established and novel machine learning classification techniques on modern platforms like Databricks and PySpark using Python.They worked on on historical data for 10 companies and applied machine learning libraries where the results was showing 80% accuracy on logistics regression compared to navies bayes.

S.Jaya Amruth et al. [4] highlighted how these models extract meaningful patterns from extensive datasets, leveraging data mining and heuristic approaches for real-time analysis. The results demonstrated that ensemble methods like Random Forest and XGBoost outperform standalone algorithms, offering higher prediction accuracy and better generalization to unseen data. p Nidhi Sharma et al. [5] analyzed to perform a comparative assessment of Hadoop MapReduce and Apache Spark frameworks on particular aspects in order to identify which would be a better option for analysis of stock exchange data. Because of some restrictions on hadoop, it is not suitable for active data processing. To evaluate the effect of covid-19, they employ nifty-50 data. They also remarked that processing of Apache spark can be completed very quickly as well owing to its processing capabilities.

Subhadra Kompella et al. [6] proposed that sentiment analysis is employed to calculate polarity scores, which are then used to determine whether articles have a positive or negative impact on stocks, aiding in further analysis.They have used a historic data where data calculation is displayed to the users as graph. The paper purposed mean squared error and mean squared log error scores of the Random Forest model output from those of the Logistic Regression model.

Shivani Shah et al. [7] explored that explore Big Data processing and visualization for NSE data using the capabilities of Azure, Databricks, and Power BI.Uses interactive dashboard allows users to utilize their data for insightful analysis and decision-making, facilitating more informed investment strategies.It also helped the users to analyze the historic and current stock data conveniently.

Zhihao Peng [8] proposed that a robust Cloudera-Hadoop-based data pipeline for analyzing various data types, focusing on selected U.S. stocks to predict daily gains using real-time data from Yahoo Finance.Mapreduce and HDFS has been used for data analysis of 13 oil stocks from SP500 stocks where they got the MSE 1.97% by using machine learning model I.e. linear regression.

Baldeep Singh et al. [9] proposed that variety of databased can be evaluated on a hybrid cloud environment .They have used Mongodb for for storing OHLC i.e. Open High Low Close data based on standard and custom workload experiments as the execution time in mongodb is less and also is cost effective.

Arunkumar.R et al. [10] designed Azure Databricks Delta-Lake with Azure DataLake Storage Generation 2 architecture and explained how it can be used for Fibonacci retracement analysis, which helps them analyze stocks and forecast market prices to make smarter investment decisions. Incorporating Machine Learning (ML) and Deep Learning (DL) methods aids in accurate prediction of the forecasted close price.

Yuxi Liu [11] investigated that streaming data analysis under Apache spark and sentiment analysis for stock data based on Apache spark investigated.This paper shows how forcasting financial time series enhances the realibilty and precision of stock market analysis by using different machine learning models and comparing their results.

I. J. Xianya et al. [12] proposed that method to Build a spark cluster,and gradually reduce the number of hosts in the cluster .They also uses the function Vector Assembler in pyspark.ml. feature to merge the feature attribute columns.They endeavored to perform a quantitative analysis how web is impacting on stock market from the behavior of investor.

Hind Daori et al. [13] proposed they used various NLP techniques to convert textual information into a machine-friendly for Al Rajhi Bank stock price analysis.This paper explained how map reduce programming model has been usedto process unstructured data and get efficient results.

Hansol Lee et al. [14] provided a comparative study of small vendor and big vendor but this time employing big data analysis techniques. Their results shows firm's spending on big data solution brings about favorable response in stock market. Concerning their forecast on the basis of the data, its shows dimensions of firm and vendor size make unique influence on the stock market trend and price.

Qing Li et al. [15] proposed that big data analytics also helps to catch illegal trading in the financial markets and helps to find fraud.They build a model using hadoop mapreduce to find

out the maximum and minimum volatility in stock and pyspark to predict the closing price of those stock using different ML models and also by sentiment analysis.At last they conclude that a lower volatility means that a collateral value does not fluctuate dramatically.

Bharathi Mohan et al. [16] studied the past data of NIFTY-50 portfolio constituents with a purpose to find the trends and patterns that are related to stock price movements. An important part of this research was focus on sentiment analysis which assessed the influence of news, events, and general market on the changes in stock prices.

E.Annapoorna et al. [17] speaks on predictions of this kind and necessity of tackling difficulties like market uncertainty and data. The Facebook Prophet is utilized by these authors to perform forecasting due to its time series decomposition and seasonality effects. Its automations and handling of missing values are especially relevant to financial forecasts.

III. System Design

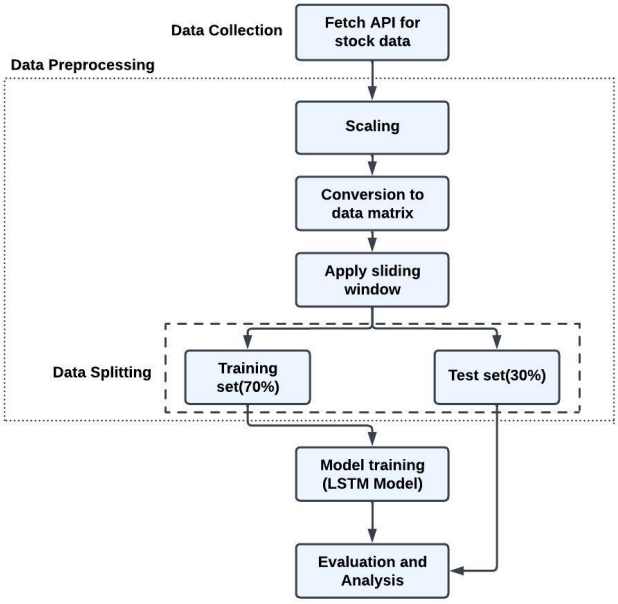

Fig. 1. Enter Caption

This infographic exemplifies the actions which are taken to construct a stock price prediction model with the assistance of Long Short Term Memory (LSTM) networks. It starts at the beginning that is data collection from an API, where the historical stock data is fetched. That data is further pre-processed so that the values can also be scaled as well as transformed into an appropriate data matrix format. Next up is the sliding window technique, to create input-output pairings for an LSTM model. The data is then divided into 70% training set and 30% test set. This LSTM model is trained on the training set and evaluated on the test set. In the end, by checking the model for observations, it identifies what can be improved in the predictions potentially and whether the stock prices are going to increase or decrease in upcoming days.

IV. Implementation

The implementation part consists of training the model for stock market prediction and to integrate it with GUI for easy user interaction.

A. Data Collection

The process starts with the data collection phase where stock prices are fetched through the yfinance library using Yahoo Finance. Users have to provide the ticker ID for the company (for example, "AAPL" for Apple) and the time frame for collecting historical stock prices, like 2010-2019. Adjusted closing prices will be loaded into Spark DataFrames, and therefore they will take advantage of distributed computing for scalability and efficiency in handling large datasets.

B. Spark Environment

Once stock price data has been collected, Spark is then applied to efficiently handle and preprocess the dataset. Transforming the historical stock data into a Spark DataFrame will make it possible to utilize Spark's distributed computing capabilities for effectively fast and reliable processing of huge datasets. Spark makes it easy to carry out the relevant data manipulations, including column selection, treatment of missing values, and scaling features in preparation for machine learning. Particularly, this ensures that everything is in place for seamless processing, even with very large historic data sets or in support of different stock tickers-an ideal scenario for large investments in financial data research.

C. Data Preprocessing

In this phase, the stock data is cleaned and prepared for analysis. Here, irrelevant columns are removed and missing values are deleted to ensure that the dataset is consistent. Thereafter, the adjusted closing prices are normalized by using MinMaxScaler to transform the data within the 0-1 range. This is necessary because it helps stabilize and improve the neural network model performance. This have an effect that will make sure that the training of the model is not skewed by and is normally distributed due to differences in magnitudes of the data points.

D. Model training and testing

As for training the models with scaled data, a sliding-window method is followed to develop input-output pairs. Every input contains a fixed amount of past stock prices (e.g., 100 time steps) whereas the corresponding output consists of next stock price in the series. Model is LSTM neural networks having multiple layers and hence they are trained to learn temporal patterns from sequential data. The model compilation takes place using Adam optimizer and Mean Squared Error as the loss function which is given in equation (1) and it was trained for a number of epochs (e.g., 50) to optimize model performance. The model was saved to use it for different companies.

Equation for loss function that is used to train model (Mean Squared Error)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where:

- y_i is the actual stock price,
- $\hat{y}_i$ is the predicted stock price, and
- n is the total number of predictions.

In the second phase which is testing and prediction, performance evaluation for the model is conducted using data that was not displayed to the model during training. The input data to be predicted is prepared by applying the sliding-window technique once again. The model gives stock price predictions which are then rescaled back to the original range by applying inverse transformation of MinMaxScaler. Comparison of these predictions and actual prices takes place to give accurateness assessments to this model.

E. Visualization

The application therefore includes visualization in order to make the results interpretable. A line graph is produced by Matplotlib whereby actual stock prices are plotted against predicted prices. This gives a clear cut view of the actual price and the predicted value for a given time thus making it easier for a user to evaluate the model performance.

The implementation finally ends with the creation of a Graphical User Interface (GUI) to Streamlit. This allows the users to dose by entering the stock ticker ID along with a date range against which the stock is to be analyzed. The LSTM model saved earlier now loads for predictions on the chosen stock, and the results will be immediately displayed including a line graph of both actual and predicted prices. With this, the application is rendered user-friendly for the smooth process of trend analysis and prediction evaluation on stocks.

V. Result and Analysis

The model consists of several LSTM layers to learn long term dependencies in the sequential data more accurately. There are Dropout layers after each LSTM layer to avoid overfitting. The last dense layer outputs a single value predicted at the next time step. **Model Summary:**

The architecture of the Long Short-Term Memory (LSTM) model used for stock price prediction is summarized below:

TABLE I
LSTM Model Architecture Summary

Layer (Type)	Output Shape	Parameters
LSTM	(None, 100, 50)	10,400
Dropout	(None, 100, 50)	0
LSTM	(None, 100, 60)	26,640
Dropout	(None, 100, 60)	0
LSTM	(None, 100, 80)	45,120
Dropout	(None, 100, 80)	0
LSTM	(None, 120)	96,480
Dropout	(None, 120)	0
Dense	(None, 1)	121

Total Parameters: 536,285
Trainable Parameters: 178,761
Non-trainable Parameters: 0
Optimizer params: 357,524

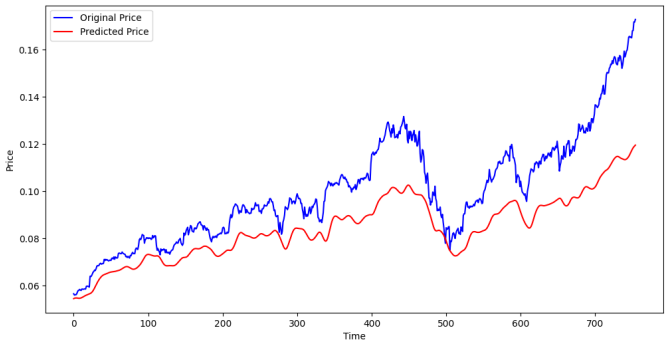

Fig. 2. Visualization of actual v/s predicted stock prices

Fig 2 is a graphical representation of the predictability of the LSTM model through the comparison of actual stock prices (indicated by blue line) against predicted stock prices (indicated by red line). Overall, the model captures the trend within the stock price market, albeit not entirely without some deviation from actual values predicted. Such deviations are common in financial time series data, generally quite noisy and volatile in nature.

The ability of the model to predict stock prices was analyzed on the test set.And then after prediction generation, values were converted back into original scale by inverse scaling using MSE (Mean Squared Error) as the performance measuring unit, defined as with respect to average squared differences between actual price values and predicted price values. Overall, the model is robust enough to be used for forecasting stock price trends.

VI. Conclusion

In this study, a Long Short-Term Memory (LSTM) neural network model was built to predict the stock prices. The model architecture comprises different layers of LSTM units placed in such a way that dropout layers are inserted to avoid overfitting and promote generalization better. Once trained on historical stock data, the model's capability was validated using unseen test data, denoting the ability to forecast future stock prices. The results indicate that the model is able to learn

and shed light on the temporal patterns using the study of comparison for predicted and actual prices. To evaluate model prediction accuracy, the Mean Square Error (MSE) metric was utilized, while the overall architecture of the model, with a parameter count of 536285, illustrates the complexity of the model as well as its potentiality for sequential data handling.

VII. Future Scope

However, in the future, its prediction accuracy can be improved by fine-tuning hyperparameters through grid search and random search. More diverse features such as technical indicators, market sentiment, and economic data would give a better understanding of the stock price movements. Finally, one may further consider very advanced neural network architectures like attention mechanisms or Transformers to capture the complex patterns still hidden in time series data. The model can also be extended for real-time predictions through a dynamic data pipeline, and it could be deployed as a web service for real-world applications such as automated trading or financial analysis. Another possibility would be to improve the model interpretability with techniques like SHAP or LIME to enhance user trust and offer insights into underlying decision-making.

References

[1] P. Modi, S. Shah, and H. Shah, "Big data analysis in stock market prediction," *International Journal of Engineering Research & Technology (IJERT)*, vol. 8, no. 10, 2019.

[2] A. Ashok and C. Prathibhamol, "Improved analysis of stock market prediction: (arima-lstm-smp)," in *2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)*, no. 2, 2021, pp. 1–5.

[3] M. Javed Awan, M. S. Mohd Rahim, H. Nobanee, A. Munawar, A. Yasin, and A. M. Zain, "Social media and stock market prediction: a big data approach," *MJ Awan, M. Shafry, H. Nobanee, A. Munawar, A. Yasin et al.," Social media and stock market prediction: a big data approach," Computers, Materials & Continua*, vol. 67, no. 2, pp. 2569–2583, 2021.

[4] S. J. Amruth, T. Nigelesh, V. S. Shruthik, V. S. Reddy, and M. Venugopalan, "Time-series-based stock market analysis using machine learning," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, no. 4, 2024, pp. 1–7.

[5] Y. K. Gupta and N. Sharma, "Propositional aspect between apache spark and hadoop map-reduce for stock market data," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2020, pp. 479–483.

[6] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–24, 2022.

[7] S. Sanap and K. Syed, "Real-time data visualization in pwas using power bi embedded," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–7.

[8] S. KOVVUR and L. FERREIRA, "Exploring the use of big data in auditing: A study of the indian context."

[9] B. Singh, R. Martyr, T. Medland, J. Astin, G. Hunter, and J.-C. Nebel, "Cloud based evaluation of databases for stock market data," *Journal of Cloud Computing*, vol. 11, no. 1, p. 53, 2022.

[10] S. Kamalakkannan, A. Yasmin, P. Kavitha *et al.*, "A model for the analytical performance of data lake in stock market analysis with databricks delta lake," in *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*. IEEE, 2023, pp. 1065–1071.

[11] Y. Liu, "The investigation of the application of apache spark in stock analysis."

[12] J. Xianya, H. Mo, and L. Haifeng, "Stock classification prediction based on spark," *Procedia Computer Science*, vol. 162, pp. 243–250, 2019.

[13] H. Daori, G. Alzahrani, A. Alanazi, M. ALHARTHI *et al.*, "Big data analytics by using spark of alrajhi stock," 2022.

[14] H. Lee, E. Kweon, M. Kim, and S. Chai, "Does implementation of big data analytics improve firms' market value? investors' reaction in stock market," *Sustainability*, vol. 9, no. 6, p. 978, 2017.

[15] Q. Li, Y. Chen, J. Wang, Y. Chen, and H. Chen, "Web media and stock markets: A survey and future directions from a big data perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 2, pp. 381–399, 2017.

[16] B. Mohan, A. Yadav, J. Toleti, D. S. Sruthik Reddy, R. K. Ahmed, and G. Koushitha, "Sentiment-driven predictive models for nifty 50 stock market fluctuations: A time series analysis perspective," in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, no. 16, June 2024, pp. 1–8.

[17] E. Annapoorna, S. V. Sujil, S. S, S. Abhishek, and A. T, "Revolutionizing stock price prediction with automated facebook prophet analysis," in *2024 International Conference on Inventive Computation Technologies (ICICT)*, no. 17, 2024, pp. 1307–1314.