Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 20s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Stocks Analysis and Prediction Using Big Data Analytics

Gollapalli Bhargav Sai[1], Dr. Dara Vikram[2]

[1]Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

Email ID: bhargavsaigollapalli@gmail.com

[2]Dept. of CSE, Professor, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

Email ID: daravikram@kluniversity.in

Cite this paper as: Gollapalli Bhargav Sai, Dr. Dara Vikram, (2025) Stocks Analysis and Prediction Using Big Data Analytics. *Journal of Neonatal Surgery*, 14 (20s), 21-29.

ABSTRACT

Big data analytics plays a crucial role in various sectors, enabling the accurate prediction and analysis of large datasets. This approach focuses on stock market prediction, where large volumes of stock data are processed to predict daily gains or losses. By utilizing big data tools, such as the PySpark API, streaming or batch data is processed to generate predictions based on historical stock information. The goal is to identify patterns in stock price movements and predict future trends with a high level of accuracy. Performance is evaluated using R-squared metrics, ensuring that the most effective model is selected. Among the models tested, the Long Short-Term Memory (LSTM) algorithm demonstrates the highest predictive accuracy, achieving an R-squared value of 0.97%. This result highlights LSTM's capability to closely match predicted stock prices with actual test data, offering a reliable solution for stock market forecasting. The approach showcases the potential of big data and machine learning in financial analysis, helping investors make informed decisions based on historical trends and future predictions.

Keywords: *Stock Analysis, Machine Learning, Stock Prediction, Big Data, LSTM, Stock Price Prediction*.

1. INTRODUCTION

Big data has gained significant attention across various sectors due to its potential to drive innovation, enhance decision-making, and provide valuable insights. In the business world, organizations increasingly rely on big data to transform raw information into meaningful business intelligence, allowing them to streamline operations, predict market trends, and improve customer experiences. With the ability to process and analyze vast amounts of structured and unstructured data, companies can identify patterns, optimize strategies, and gain a competitive edge. The application of big data tools enables businesses to forecast demand, understand consumer behavior, and implement personalized marketing strategies [1].

In the healthcare sector, big data plays a pivotal role in improving medical outcomes, enhancing patient care, and optimizing healthcare services. By analyzing patient records, medical histories, and treatment outcomes, healthcare providers can identify trends, predict diseases, and provide more accurate diagnoses. The integration of big data with machine learning algorithms also assists in predicting health trends, reducing healthcare costs, and supporting personalized treatment plans [2]. Healthcare organizations are increasingly utilizing big data analytics to address challenges such as early disease detection, treatment optimization, and operational efficiencies, making it an essential tool for advancing modern healthcare systems [3].

Moreover, the information technology (IT) and cloud computing sectors have embraced big data to enhance system performance, security, and service delivery. The rise of cloud computing has facilitated the storage and processing of enormous volumes of data, making it more accessible and scalable. Cloud platforms provide the infrastructure necessary to support big data analytics, enabling organizations to deploy applications that can analyze data in real time. With the growing complexity of systems, big data helps IT companies improve data management, ensure seamless integration, and automate decision-making processes [4]. This dynamic combination of big data and cloud computing has reshaped how IT systems function, providing businesses with the agility to adapt to changing demands [5].

The integration of big data analytics in these sectors has not only enhanced operational efficiencies but also paved the way for new business models and innovative technologies. As industries continue to explore and harness the power of big data, its transformative impact on global economies becomes increasingly evident [6]. Therefore, big data remains an indispensable resource that drives technological advancements, boosts productivity, and contributes to the growth of various industries worldwide.

2. LITERATURE SURVEY

The field of stock market prediction has gained significant traction due to the increasing availability of vast amounts of data and the growing importance of big data analytics in making informed decisions. Several studies have explored various methods and techniques to predict stock price trends, leveraging big data processing environments, machine learning algorithms, and advanced analytical models. One of the earliest works in this domain by Zhao and Wang [1] introduces the use of outlier data mining algorithms for predicting stock price trends. They propose a methodology where outlier data points are identified and excluded from analysis to improve the accuracy of stock market predictions. Their work highlights the importance of data quality in prediction models, emphasizing the role of anomaly detection in enhancing forecasting performance.

Jaweed and Jebathangam [2] further explore the integration of big data processing environments in stock market analysis. Their study investigates how big data tools can be utilized to handle massive volumes of stock market data. They suggest that using a big data framework, such as Hadoop or Spark, allows for the efficient processing and analysis of diverse data sources, including historical stock prices, financial news, and social media sentiment. This approach helps uncover hidden patterns and trends that can be crucial for accurate stock price prediction. They argue that big data not only facilitates real-time data processing but also enables more comprehensive and reliable stock market analysis by incorporating multiple data streams.

Tiwari, Bharadwaj, and Gupta [3] focus on the use of data analytics for stock price prediction, specifically employing machine learning techniques. They propose the application of algorithms such as regression analysis, decision trees, and neural networks to predict stock prices. Their study shows how these algorithms can be trained on historical data to develop predictive models, enabling investors to forecast future stock movements based on past patterns. The authors suggest that by combining multiple machine learning algorithms, it is possible to enhance the accuracy of predictions and make more informed investment decisions.

Singh and Thakral [4] take a statistical approach to stock market analysis, focusing on the analysis of stock market indexes and their constituents. Their research investigates how statistical techniques such as time series analysis and correlation studies can be used to analyze the movement of stock prices and their dependencies on various market factors. By studying the relationship between different stock indices and their individual components, they propose that investors can gain deeper insights into the dynamics of the stock market. This approach aims to identify factors that influence stock price movements and predict future trends with a higher degree of confidence.

Peng [5] introduces a comprehensive study on stock analysis and prediction using big data analytics. His work proposes the use of advanced machine learning algorithms, such as support vector machines (SVM) and random forests, to predict stock market behavior. Peng emphasizes the importance of data preprocessing, including normalization and feature selection, to improve the performance of predictive models. By applying these techniques to historical stock data, his approach aims to improve the accuracy of stock predictions and help investors make more reliable decisions. Additionally, Peng's work incorporates sentiment analysis of news articles and social media, demonstrating the value of unstructured data in predicting stock market trends.

Attigeri, Pai, and Nayak [6] focus on a big data approach to stock market prediction, employing machine learning algorithms such as decision trees and support vector machines. They highlight the potential of big data in predicting stock market trends by incorporating diverse data sources, including economic indicators, stock prices, and social media sentiment. Their proposed system uses a hybrid approach that combines technical analysis and fundamental analysis, integrating data from various sources to improve prediction accuracy. This approach aims to help investors identify profitable trading opportunities by leveraging a wide range of data types and advanced algorithms.

Jeon, Hong, Kim, and Lee [8] take a different approach by proposing the use of pattern graph analysis to predict stock prices. Their study examines the role of pattern recognition techniques in stock market forecasting, where patterns in historical stock data are analyzed to identify recurring trends and predict future price movements. They argue that pattern graphs, combined with big data tools, can provide a more accurate representation of stock price trends and help investors make better trading decisions. Their approach suggests that pattern graph analysis, when applied to large datasets, can uncover complex relationships between stock prices and external factors, improving prediction reliability.

Choudhry and Garg [9] propose a hybrid machine learning system for stock market forecasting, combining multiple machine learning models to improve prediction accuracy. Their system integrates neural networks, decision trees, and support vector machines to create a robust forecasting model. By combining the strengths of different algorithms, their approach aims to improve the generalization ability of the model and reduce prediction errors. The authors suggest that hybrid models can outperform single algorithms by capturing various aspects of stock market behavior and providing more accurate predictions.

3. MATERIALS AND METHODS

The proposed system aims to leverage big data analytics and machine learning techniques for accurate stock market

prediction. By processing large volumes of historical stock data using big data tools such as the PySpark API, the system will utilize both streaming and batch data processing methods. The core prediction models include Long Short-Term Memory (LSTM) [12], AutoRegressive Integrated Moving Average (ARIMA) [11], and XGBoost [10]. LSTM will be employed for its ability to capture long-term dependencies and patterns in sequential stock price data, while ARIMA will model time-series data to forecast future stock prices based on past observations. XGBoost, a powerful gradient boosting algorithm, will be used for its high predictive performance in tabular data. The system will analyze and predict daily stock gains or losses, with performance evaluated using R-squared metrics to ensure model accuracy. This approach will provide a reliable framework for making informed investment decisions based on historical data and predictive trends.

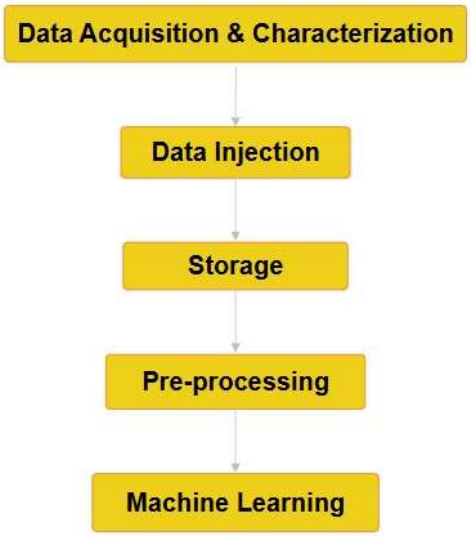

Fig.1 System Architecture

The system architecture (fig.1) begins with Dataset Acquisition & Characterization, where relevant stock data is gathered and analyzed for key features. This data is then sent through Data Injection into the processing pipeline. The data is stored in a scalable database for easy retrieval. Pre-processing cleans and transforms the data for consistency. Finally, Machine Learning models such as LSTM, ARIMA, and XGBoost are applied to predict stock trends and provide insights.

i) Dataset Collection:

The dataset for this work consists of historical stock market data, including daily stock prices, volume, open, high, low, and close values for various companies. These datasets are sourced from financial platforms or APIs like Yahoo Finance, Alpha Vantage, or Kaggle. Each file includes time-series data with timestamps and numerical values, stored in CSV format. Additional features such as moving averages, RSI, and MACD can be generated for analysis. Large datasets simulate real-time streaming and batch processing for scalability.

ii) Modules:

a) Read Data using PySpark: This module reads stock data using PySpark. It initializes a Spark session and sets up a Spark Streaming Context for processing streaming or batch data. The dataset is loaded using PySpark's readStream or read methods, supporting CSV or other formats. Once loaded, the data is displayed to verify correct loading. The module ensures efficient handling of large stock datasets for further machine-learning-based prediction and analysis.

b) Data Normalization: This module focuses on data normalization to standardize the scale of stock dataset features. It applies normalization techniques, such as Min-Max scaling or StandardScaler from PySpark MLlib, ensuring that all features are within a uniform range. Normalization eliminates discrepancies caused by varying units or ranges of features, making the data consistent and suitable for machine learning algorithms. This step is essential to enhance the effectiveness and accuracy of predictive models.

c) Feature Engineering with XGBoost: This module leverages the XGBoost algorithm for feature engineering to identify and select the most relevant features from the stock dataset. Irrelevant or redundant features are excluded, streamlining the dataset for improved model performance. XGBoost [10] evaluates feature importance during training, focusing on impactful features that enhance prediction accuracy. This step ensures that ARIMA and LSTM algorithms receive optimized inputs, resulting in more efficient and accurate stock gain/loss predictions.

d) Training ARIMA Model: This module trains an ARIMA model on the preprocessed stock dataset. It specifies the model's

order parameters (p, d, q) to capture temporal dependencies in the time-series data effectively. The ARIMA model identifies patterns like trends and seasonality, making it suitable for predicting stock prices. Training involves fitting the model to historical data, and its performance is evaluated to understand its predictive accuracy before being used for future predictions.

e) Evaluate ARIMA Model: This module evaluates the performance of the ARIMA model by applying it to the test dataset. The predicted stock prices are compared with actual values, and the R-squared metric is calculated to assess how well the model explains variance in the test data. A higher R-squared indicates better model performance. This evaluation helps determine the ARIMA [11] model's effectiveness and suitability for accurate stock price prediction in real-world scenarios.

f) Training LSTM Model: This module trains an LSTM model using the preprocessed stock dataset. A Sequential model is created with Keras, incorporating LSTM [12] layers to capture long-term dependencies in the time-series data. The model is configured with suitable hyperparameters, such as the number of neurons and dropout rates, and trained using historical stock prices. LSTM's ability to learn sequential patterns makes it highly effective for predicting future stock movements accurately.

g) Evaluate LSTM Model: This module evaluates the LSTM model's performance by applying it to the test dataset. Predicted stock prices are compared with actual values, and the R-squared metric is calculated to assess the model's accuracy. The R-squared value indicates how effectively the model captures variance in the data. Finally, the LSTM's performance is compared to the ARIMA model, highlighting the model better suited for accurate stock price prediction.

iii) Algorithms:

XGBoost is employed for feature selection and optimization. It enhances prediction models by identifying impactful features while excluding irrelevant ones. This gradient-boosting [10] algorithm is efficient and scalable, making it ideal for processing large datasets, improving performance by focusing only on the most significant inputs.

ARIMA models time-series data by capturing trends, seasonality, and temporal dependencies. It predicts future values based on historical patterns and residuals. ARIMA [11] is effective for datasets with linear temporal relationships and is often used for forecasting tasks, such as stock price prediction.

LSTM excels in analyzing sequential data, capturing both short-term and long-term dependencies. It is highly effective for time-series predictions, as its memory cells retain essential information over time. LSTM [12] is commonly applied to stock price forecasting, recognizing patterns in historical data to predict future trends.

4. RESULTS

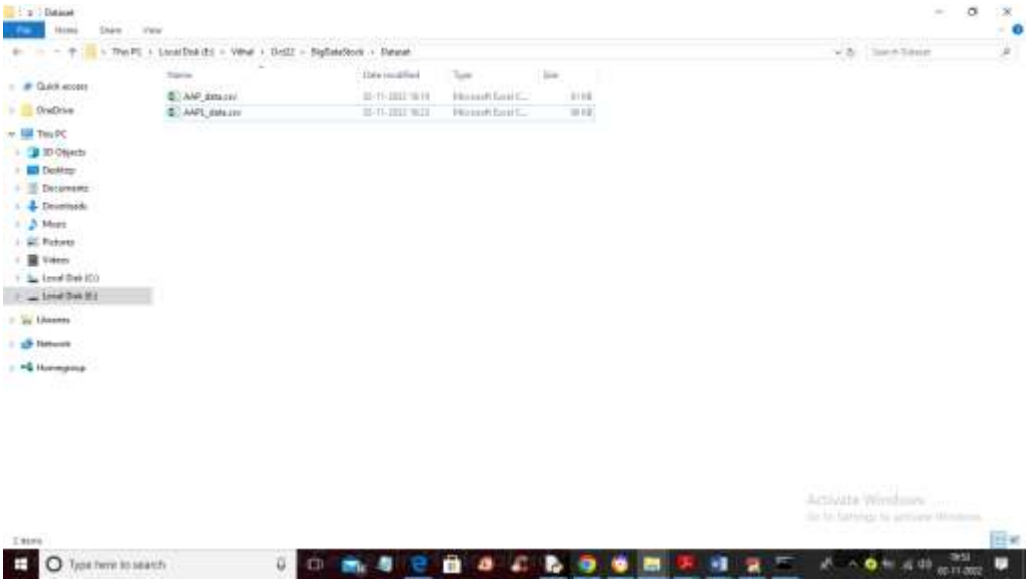

Fig.2 Stock dataset files

In above screen showing stock data files used in this work. Below screen showing JUPYTER code and output with comments

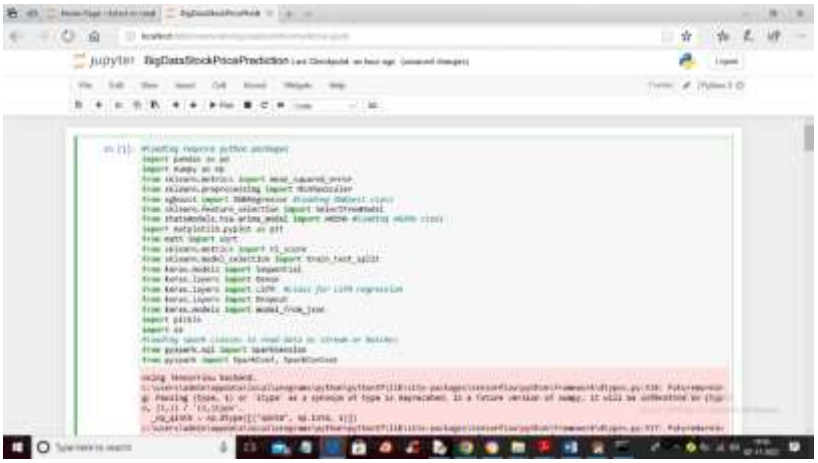

Fig.3 Packages Imported

In above screen we are importing all the classes required to implement this work

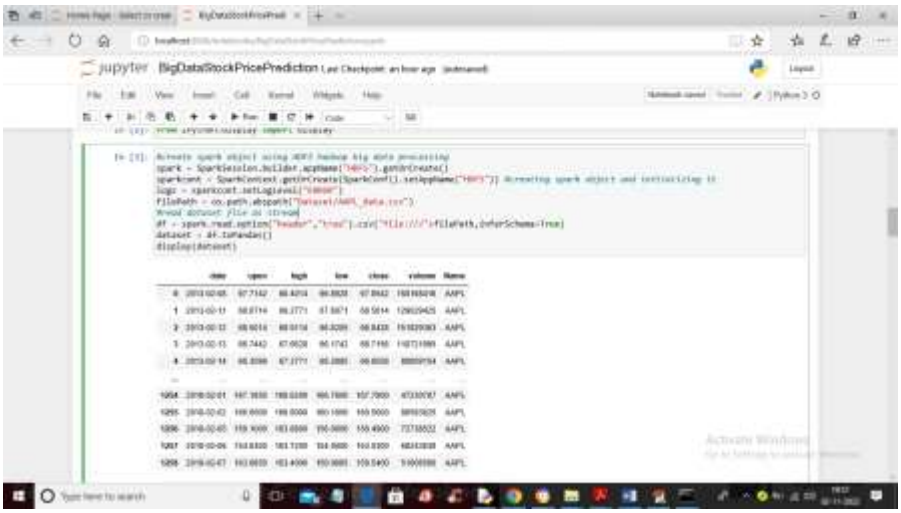

Fig.4 Dataset

In above screen using SPARK classes we are reading dataset as stream, here we don't have any online streaming so we are reading data from file and then displaying

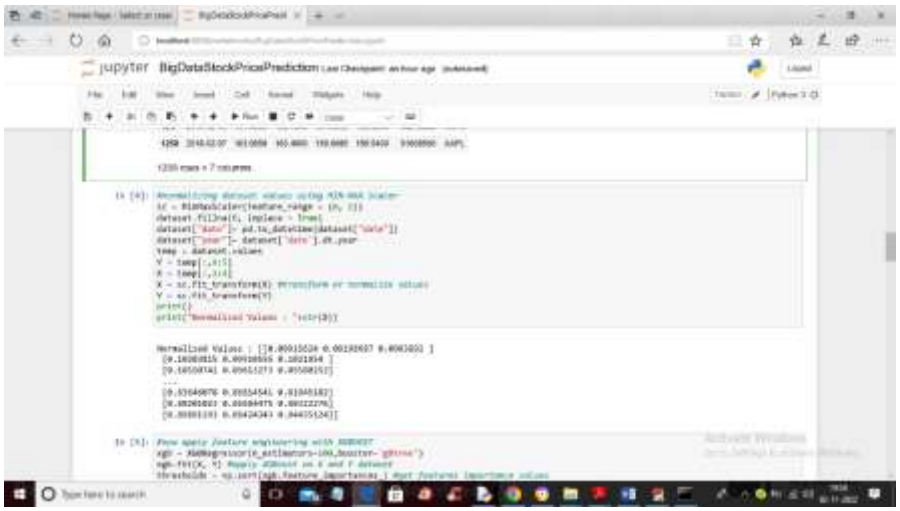

Fig.5 Normalization

Using above screen code we are normalizing dataset values

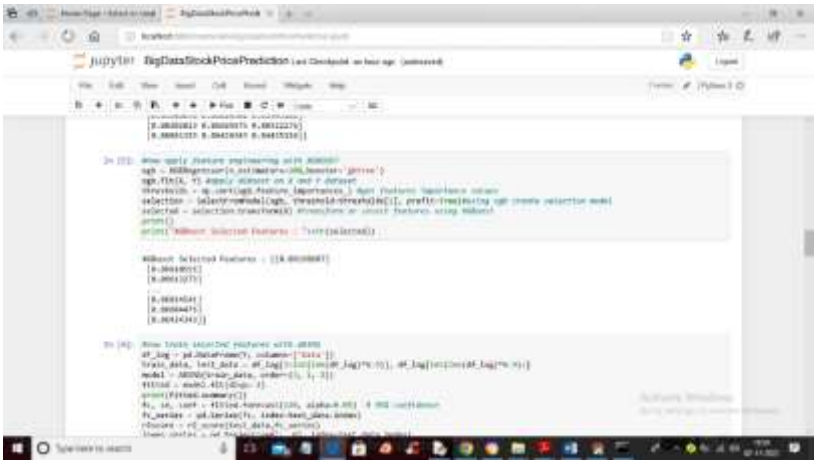

Fig.6 XGBoost

In above screen we are applying XGBOOST algorithm to apply features engineering and then displaying selected values from dataset. Now processed data will be input to both algorithm

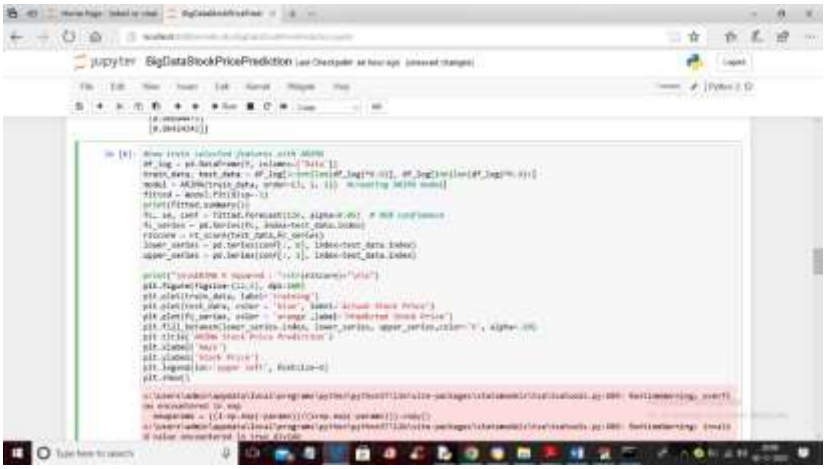

Fig.7 Training Dataset with ARIMA

Using above screen code we are training dataset with ARIMA class and below is the ARIMA stock prediction output

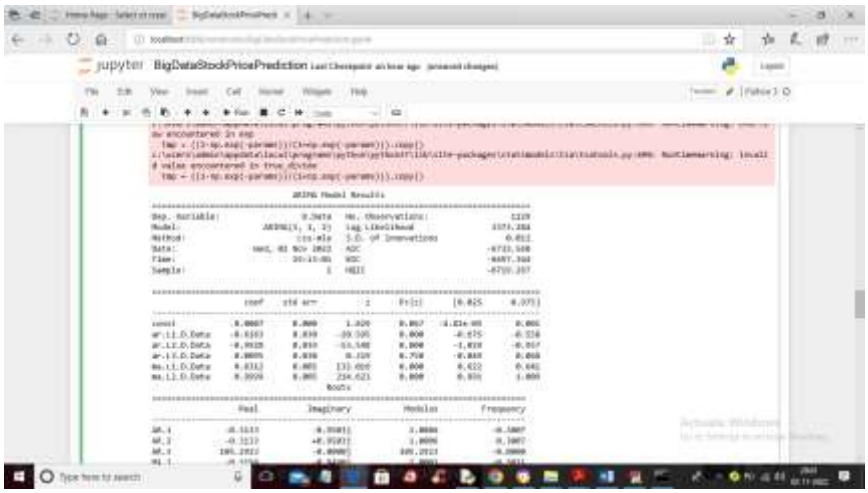

Fig.8 ARIMA

In above screen ARIMA starts building model and will get below output

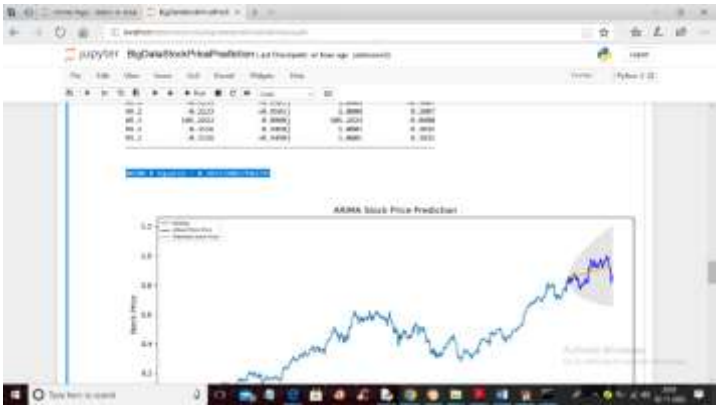

Fig.9 ARIMA R^2

In above screen in blue color text we can see ARIMA R SQUARED as 0.38 and below is the graph

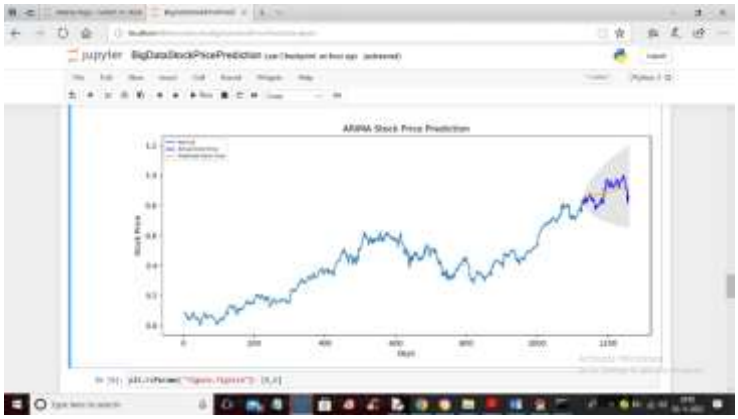

Fig.10 ARIMA Stock Price Prediction Graph

In above graph x-axis represents training stock days and y-axis represents stock prices and big light blue line represents training stock prices and dark blue line in the last is the TEST prices and orange line is the predicted prices and orange line is not as zigzag as train and test data so its prediction is not accurate and due to that reason we got its R SQUARED as 0.38%. In below screen we are showing LSTM code

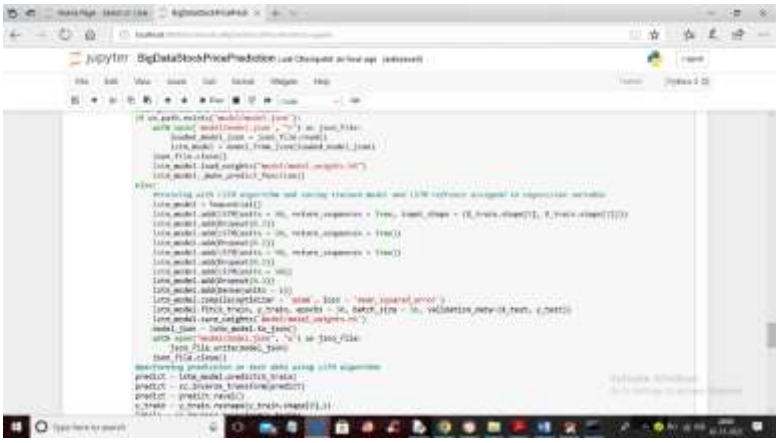

Fig.11 Training Dataset with LSTM

In above screen we are training dataset with LSTM algorithm and below is the prediction output

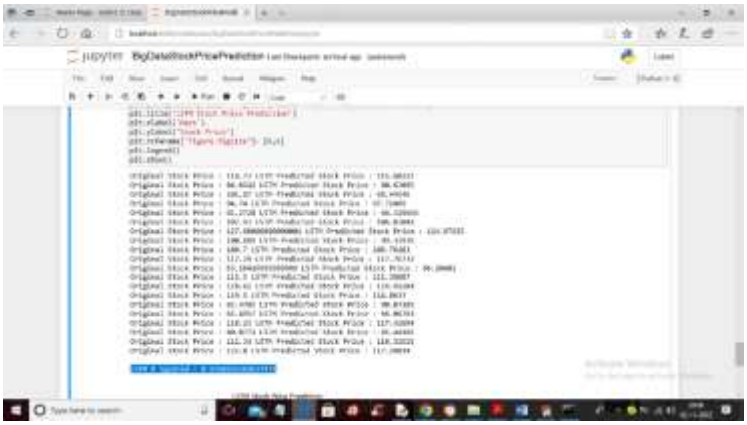

Fig.12 LSTM R^2

In above screen first we can see Original Test data stock price and then we can see LSTM predicted prices and we are displaying few records where you can see there is very close difference between original test prices and predicted prices and in blue color text we can see LSTM R SQUARED as 0.97%. so we can say LSTM prediction is accurate and in below screen we can see LSTM graph

Fig.13 LSTM Stock Price Prediction Graph

In above graph x-axis represents Number of Days and y-axis represents STOCK PRICES and red line represents original TEST stock prices and green line represents LSTM predicted prices and we can see both lines are fully overlapping so test prices and predicted prices are very close.

So from above result we can say LSTM is accurate.

5. CONCLUSION

In conclusion, big data analytics plays a pivotal role in the realm of stock market prediction [11] by processing vast amounts of historical data to forecast future trends. By utilizing powerful tools like the PySpark API, which allows for efficient batch or streaming data processing, it is possible to uncover patterns in stock price movements. The use of machine learning algorithms such as Long Short-Term Memory (LSTM) proves particularly effective in this context. With an impressive R-squared value of 0.97%, LSTM demonstrates its capability to closely match predicted stock prices with actual test data, indicating its strong predictive accuracy. This approach showcases the potential of integrating big data and machine learning for financial analysis, providing investors with a reliable tool to make informed decisions. The success of LSTM in this study underscores the significance of combining historical data insights with advanced predictive models, offering a robust solution for stock market forecasting [12]. Ultimately, this methodology serves as an effective means to assess market trends and support data-driven investment strategies.

The ***future scope*** of this approach lies in further enhancing prediction accuracy by incorporating additional data sources, such as real-time news sentiment analysis, social media trends, and macroeconomic indicators. Moreover, exploring hybrid models that combine LSTM with other advanced machine learning algorithms, like reinforcement learning, could lead to even more robust stock market forecasts. Continuous refinement of models and techniques can improve decision-making for investors in increasingly dynamic and volatile financial markets.

REFERENCES

[1] L. Zhao and L. Wang, “Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm,” in 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, Dalian, China, 2015, pp. 93–98.

[2] M.D. Jaweed and J. Jebathangam, “Analysis of stock market by using Big Data Processing Environment” in International Journal of Pure and Applied Mathematics, Volume 119

[3] S. Tiwari, A. Bharadwaj, and S. Gupta, “Stock price prediction using data analytics,” in 2017 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, 2017, pp. 1–5

[4] P. Singh and A. Thakral, “Stock market: Statistical analysis of its indexes and its constituents,” in 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bangalore, 2017, pp. 962–966.

[5] Z. Peng, “Stocks Analysis and Prediction Using Big Data Analytics,” in 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Changsha, China, 2019, pp. 309–312.

[6] G. V. Attigeri, Manohara Pai M M, R. M. Pai, and A. Nayak, “Stock market prediction: A big data approach,” in TENCON 2015 - 2015 IEEE Region 10 Conference, Macao, 2015, pp. 1–5.

[7] W.-Y. Huang, A.-P. Chen, Y.-H. Hsu, H.-Y. Chang, and M.-W. Tsai, “Applying Market Profile Theory to Analyze Financial Big Data and Discover Financial Market Trading Behavior - A Case Study of Taiwan Futures Market,” in 2016 7th International Conference on Cloud Computing and Big Data (CCBD), Macau, China, 2016, pp. 166–169.

[8] S. Jeon, B. Hong, J. Kim, and H. Lee, “Stock Price Prediction based on Stock Big Data and Pattern Graph Analysis:,” in Proceedings of the International Conference on Internet of Things and Big Data, Rome, Italy, 2016, pp. 223–231.

[9] R. Choudhry and K. Garg, “A Hybrid Machine Learning System for Stock Market Forecasting,” vol. 2, no. 3, p. 4, 2008.

[10] K. Kim, “Financial time series forecasting using support vector machines,” Neurocomputing, vol. 55, no. 1–2, pp. 307–319, Sep. 2003.

[11] M. Makrehchi, S. Shah, and W. Liao, “Stock Prediction Using EventBased Sentiment Analysis,” in 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 2013, pp. 337–342.

[12] H. Pouransari and H. Chalabi, “Event-based stock market prediction,” p. 5.

.