IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STOCKS ANALYSIS AND PREDICTION USING BIG DATA ANALYTICS AND MACHINE LEARNING

[1] Dr. B. Santhosh Kumar, [2] Parupalli Sai Likhitha, [3]Renangi Manoj Chandra, [4]Shaik Sana , [5]Vemula Navya

[1]Professor and Head of Department - Computer Science and Engineering (CSE), [2]Student, [3]Student, [4]Student, [5]Student

[1] Department of Computer Science and Engineering (CSE)

[1]Guru Nanak Institute of Technology, Ibrahimpatnam, India

ABSTRACT: Big data analytics are utilized in various sectors to produce accurate forecasts and analyze massive volumes of data. They reveal essential facts in large collections. This article proposes a powerful Cloudera-Hadoop data pipeline for any amount or kind of data. US stocks are predicted daily using Yahoo Finance real-time data. Distributing storage and processing allows Apache Hadoop to manage enormous data collections. The Spark machine learning module uses the US stock market and determines which stocks have the most value each day by dividing their daily values into training and test data.

Keywords *– Arima Model, LSTM, K-means and recurrent neural networks (RNN).*

I. INTRODUCTION

People have given big data a lot of value for the growth of many different fields. A lot of business groups have used it to make important business thoughts and information official. It has also been used by the healthcare industry to find important trends and information that can be used to make current healthcare systems better. Also, big data is very important for the fields of IT, cloud computing, and information technology. Recently, banks and finance companies employed "big data" to monitor financial markets. Big data and network analysis expose illegal traders. Large amounts of data are used by traders, major banks, financial institutions and trading companies to create advanced trading systems. Big data analytics has also uncovered fraud and refunds. The project plans to develop a system that uses real-time Yahoo financial data to forecast daily stock returns for US oil stocks. Approximately 13 companies in the US Treasury were selected and their daily returns were divided into training and testing so that Spark's ML could predict significant daily benefits. Based on our research, we recommend using the powerful Cloudera and Hadoop databases to perform such searches for all data types and volumes. Data on US oil prices will show how the US Oil Index affects other stocks on the US Oil Fund exchange. Among other things, it helps us determine which stocks will make money for buyers and the U.S. Oil Stocks Trading Group.

The goal is to make the most of big data by building strong ML models, especially ARIMA and LSTM algorithms. With the main goal of predicting how stocks will move, these models are meant to look at huge amounts of data. The project wants to help stock buyers find useful information and make decisions based on data by highlighting the importance of big data processing. Putting together PYSPARK, a Python API for Apache Spark, is the goal so that a lot of stock data can be processed quickly. Taking into account that this data is constantly changing, the project wants to make sure it can handle large datasets and stay quick to changes in stock prices in real time. The goal of the project is to use the XGBOOST method to improve the accuracy and usefulness of stock data features. The goal is to get the machine learning models to make more accurate guesses by getting rid of factors that aren't important. This makes the results more accurate, which is in line with the main goal of giving stock buyers useful information. The project's goal is to use the R SQUARED measure to compare how well the ARIMA and LSTM algorithms work. This goal gives a number value to how well the models guess how the stock will move. A higher R SQUARED number means that the model can make more accurate predictions. The goal is to help choose the best method for a full stock study.

The goal of this project is to use big data analytics to create a stock analysis forecast system. Network and big data analytics revealed financial market fraud. Traders, huge banks, corporations, and other financial entities created high-frequency trading systems using "big data".

II. LITERATURE REVIEW

Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm:

In this study, we introduce a new method for using data mining to predict the long-term performance of stocks. Older methods that used time series algorithms or price series volatility models to try to predict different patterns have proven inadequate. We propose a new outlier mining method that can find unusual things in a set of high-frequency signal data in the stock market. These unusual stock market developments constantly affect stock prices. Our technique correctly predicts stock market trends utilizing cluster information of these anomalies. Our testing show that our strategy earns money on the Chinese stock market, particularly over time.

Stock price prediction using data analytics:

Investors are very interested in making accurate predictions about the future of money. This essay suggests that data analytics could be used to help investors make accurate predictions about the future of money, which would help them make better business decisions. The work is done on two platforms: Python and R. The initial index price in R is predicted using Arima, Holt winters, neural networks (Feed forward and multi-layer perceptron), linear regression, and time series. Python uses multi-layer perceptron and support vector regression to forecast Nifty 50 stock prices. Latest stock tweets were utilized for mood analysis. Methods employ Data from the Nifty 50 (NSEI) stock index. The data is 9 years old. Comparing expected two- to three-year stock values to actual prices showed accuracy. Two months of Python and R forecasted pricing were used. The mean squared error and various error factors for each prediction system were obtained. The feed forward network predicts stock beginning prices with the lowest error, 1.81598342%.

Stock market: Statistical analysis of its indexes and its constituents:

The world of the stock market is always changing, and it's always doing well as things are changed. Because of this, getting money from it is hard and needs a lot of planning. Since this is the case, Stock Market research should be the first thing you do before investing any money. Taking into account the fact that stock prices tend to go up and down without warning, this creates an unpredictable situation. But a detailed and constant analysis is the most popular and tried-and-true way to gain understanding, intelligent humor, and smarts to get the best. The goal of this paper is to find the best stocks to buy in based on their high performance and good results compared to a given average. Using data from the past, we were able to find the best stocks to buy in. To be sure of our results, we also looked at modern statistics in the same way and found that the performance and returns of these stocks were still good, even though they were volatile.

"Stocks Analysis and Prediction Using Big Data Analytics:

A new buzzword in the business world right now is "big data." The stock market is a niche that is always changing, volatile, unsure, and full of interesting possibilities. It is an important part of predicting business and financial growth. For the stock market to work and come to useful conclusions, it has to deal with a lot of different and large amounts of data. Trends in the stock market are mostly determined by two types of analysis: mechanical and basic. Trends from the past and market values are used in technical analysis. On the other hand, basic analysis looks at how people feel, what they value, and the facts and answers they give on social media. We use big data analysis to help us make accurate predictions and make smart business choices and profitable investments because the data is so big, difficult, and growing at an exponential rate.

Stock market prediction: A big data approach:

There The stock market is unpredictable and subject to various factors. Thus, stock market projections are crucial to business and economics. Technical analysis and basic analysis are the two types of analysis that can be used to make predictions. There is talk about both basic and fundamental research in this study. ML is used to look at past data on stock prices for technical analysis, and mood analysis is used to look at data from social media for basic analysis. Today, social media data is more important than ever; it can help people guess what the stock market will do next. The method involves gathering information from news sources and social media and figuring out how people feel about things. After that, the link between the feelings and the stock prices is studied. The model that was learned can then be used to guess what stock prices will be in the future. It has been shown that this method can predict both how people feel and how well a stock will do. It has also been shown that current news and social data are closely linked.

III. METHODOLOGY

Recently, banks and finance companies employed "big data" to monitor financial markets. Big data and network analytics found unlawful stock traders. huge data was exploited by traders, huge banks, financial institutions, and enterprises to create high-frequency trading systems. Big data analytics also detected frauds and money transfers.

Disadvantages of existing system:

1. There is no correct estimate in data in the tools that are already in place

2. Systems that are already in place can't analyze the big data sets

Proposed System:

This project aims to construct a system that forecasts daily stock gains for US oil equities using real-time Yahoo Finance data. About all 13 companies in the US oil fund are selected, and their daily gain data are separated into training and test data sets so Spark's ML can predict significant daily gains. built on our research, we suggest a strong data flow built on Cloudera and Hadoop that can be used to do these kinds of studies on any kind and amount of data.

Advantages of proposed system:

1. It can handle all kinds of complicated research

2. more quickly

3. Because there are so many machine learning (ML) tools out there, it's been hard to choose the one that can do the best job of analyzing data and putting ML methods into practice.

4. Offers an open base for putting

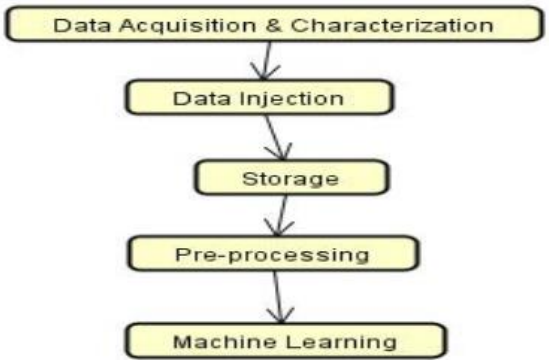

Fig.3.1: System architecture

MODULES:

1. Read Data using PySpark:

- Read stock data from a dataset file using PySpark.
- Initialize Spark and set up a Spark Streaming Context.
- Create a Spark session.
- Read the dataset as a stream or in batches using PySpark classes.
- Display the dataset to ensure it has been loaded correctly.

2. Data Normalization:

- Normalize dataset values to ensure consistency in scale.
- Apply data normalization techniques to scale the values of the dataset.
- Normalization is crucial when dealing with features that may have different units or ranges, ensuring that the algorithms can learn effectively from the data.

3. Feature Engineering with XGBoost:

- Apply XGBoost algorithm to perform feature engineering and select relevant features. while excluding irrelevant ones.
- XGBoost can help enhance the model's ability to make accurate predictions by focusing on the most impactful features.

4. Training ARIMA Model:

- Train the ARIMA model on the information that has already been cleaned up.
- Use the preprocessed dataset to train an ARIMA model, making sure to tell it which order to use (p, d, q).
- The ARIMA model is a time series model that shows how data changes over time.

5. Evaluate ARIMA Model:

- Evaluate the performance of the ARIMA model.
- Apply the trained ARIMA model to the test dataset.
- Evaluate the model's performance using metric R-squared.
- R-squared measures how well the model explains the variance in the test data.

6. Training LSTM Model:

- Train the LSTM model on the preprocessed dataset.
- Set up a Sequential model using Keras with LSTM layers.
- Train the LSTM model on the preprocessed dataset.
- LSTM models are effective for capturing long-term dependencies in sequential data.

7. Evaluate LSTM Model:

- Evaluate the performance of the LSTM model.
- Apply the trained LSTM model to the test dataset.
- Evaluate the model's performance using metrics like R-squared.
- Compare the R-squared value with the ARIMA model to determine which model performs better.

IV. IMPLEMENTATION

ALGORITHMS:

LSTM (Long Short-Term Memory):

Definition: The LSTM architecture is a type of recurrent neural network (RNN) that is meant to fix the problem of disappearing gradients that happens in regular RNNs. It is very good at finding patterns in sequential data over long periods of time because it keeps a cell state that can be carefully changed. This makes it easier to keep track of long-term relationships.

Why Used in the Project: LSTM is employed in the project for its capability to analyze and learn patterns in time-series data, which is crucial for predicting stock movements. Its ability to capture long-term dependencies in the sequential nature of stock prices makes it well-suited for forecasting stock

ARIMA (AutoRegressive Integrated Moving Average):

Definition: The ARIMA method is a scientific way to look at time series and make predictions. AutoRegressive (AR), Integrated (I), and Moving Average (MA) are the three key parts that make it up. Time-series data can be used with ARIMA models to find trends and patterns.

Why Used in the Project: ARIMA is applied in the project due to its effectiveness in modeling time-series data, making it suitable for predicting stock price movements over time. Its ability to account for trends and seasonality in the data complements the project's goal of forecasting daily gains or losses in the stock market.

XGBOOST (Extreme Gradient Boosting):

Definition: A strong ML method called XGBOOST is based on gradient boosting structures. It makes a bunch of decision trees and then puts together their estimates to get a single result. XGBOOST is famous for how fast, efficient, and well it works in many data science tasks.

Why Used in the Project: XGBOOST is utilized in the project for feature refinement within the stock data. Its capability to handle large datasets and prioritize the most influential features makes it suitable for improving the accuracy of predictions. By removing irrelevant variables, XGBOOST contributes to enhancing the overall performance of the ML models in the project.

V. EXPERIMENTAL RESULTS

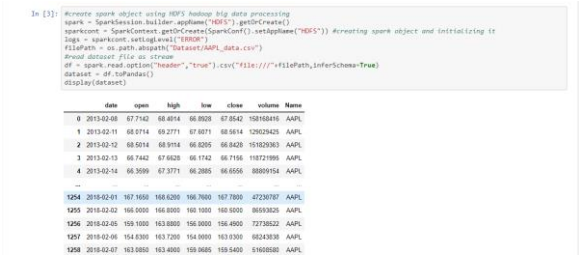

Fig.5.1: Spark class Dataset

We read the dataset as a stream in the screen above using SPARK classes. Since we don't have any online streaming here, we read the data from a file and then showed it.

Fig.5.2: Dataset Numbers adjustment

We are adjusting dataset numbers with the code shown above.

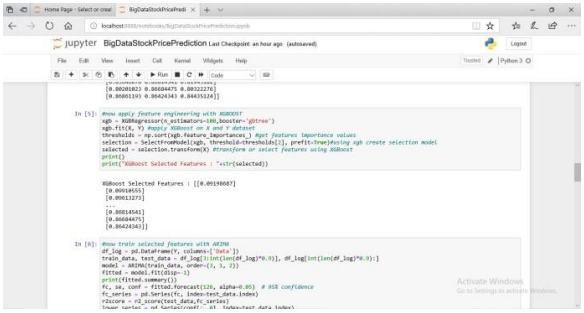

Fig.5.3: Feature extraction using XGBOOST method

On the screen above, we're using the XGBOOST method to do features engineering and then showing some numbers from the dataset. Now, data that has been handled will be fed into both algorithms.

Fig.5.4: ARIMA Class Usage

Using the code above, we are training a dataset with the ARIMA class. The result of the ARIMA stock forecast is shown below.

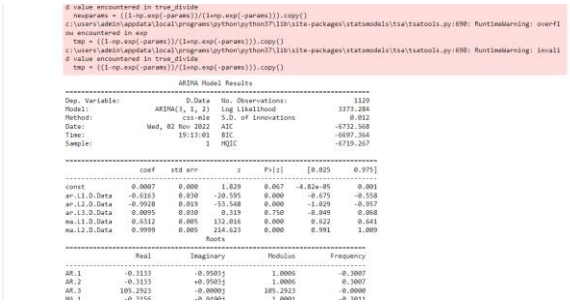

Fig.5.5: ARIMA Model Building

In the picture above, ARIMA starts to build the model, and the result below is what it shows.

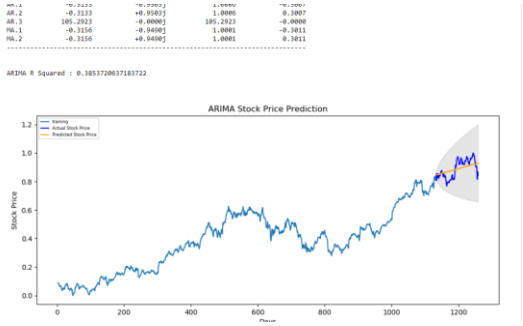

Fig.5.6: ARIMA Model Results

In blue words on the screen above, we can see that ARIMA R SQUARED is 0.38. The graph is below.

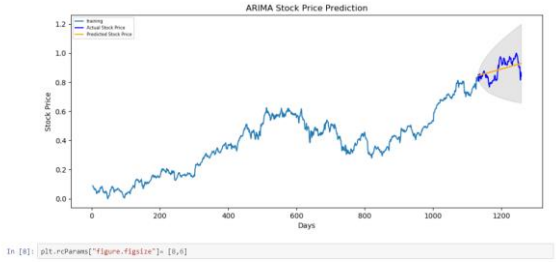

Fig.5.7: ARIMA Stock Prediction Result Graph

The x-axis represents the training days, the y-axis represents the stock price, the big blue line represents the training price, the dark blue line represents the testing price, and the orange line represents the estimated value. The estimated values are less variable than the study data and test data, resulting in an R-squared of 0.38%. Below we introduce LSTM codes.

Fig.5.8: LSTM Algorithm Usage

In the picture above, the LSTM algorithm is being used to learn a dataset. The forecast result is shown below.

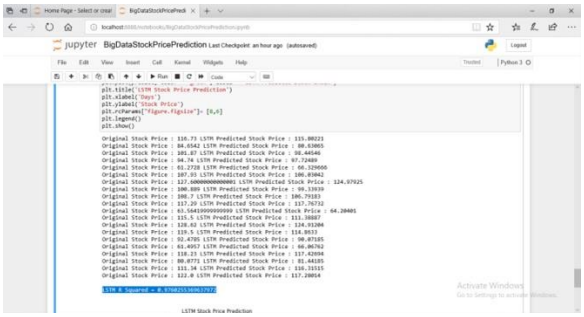

Fig.5.9: LSTM Model Building and Results

We see the Original Test data stock price first, then the LSTM forecasted pricing. In a few examples, the Original Test prices and anticipated prices are fairly close. R SQUARED LSTM is 0.97%. See the LSTM curve below.

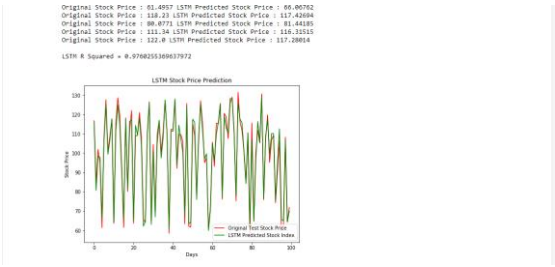

Fig.5.10: LSTM Stock Prediction Result Graph

The x-axis represents the date and the y-axis represents the stock price. The red line represents the original TEST stock price, the green line represents the LSTM estimated price. The two lines overlap, indicating that the test value and the predicted value are close.

Based on the above findings, we can say that LSTM works well.

VI. CONCLUSION

Big data analytics are used to quickly look at and guess what will happen in the stock market in this paper. In general, the stock market is a place where unpredictability and not being able to correctly predict the value of stocks can lead to huge losses in money. We were able to come up with a way to find stocks with positive everyday return rates through our work. These stocks could be good for better trade. This method will work like a Hadoop-based system to learn from past data and decide which US stocks are worth trading based on real-time changes. We also try to think of ways that our work could be better in the future. We plan to continue

our research by using the schedule tool to automate the analysis steps and then getting regular advice on how to trade US stocks. To get a good idea of what the US stock prices will do, we also want to try using Neural Network model-based learning instead of linear regression.

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 Event Based 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.