EFFIECIENT STOCK MARKET PRICE PREDICTION USING MACHINE LEARNING METHODS

Nuha Hurayein Laeeq
PG Student, CSE Department, Shadan women's College of Engineering and Technology,
Hyderabad, India
nuhahurayein@gmail.com

Dr. I. Samuel Peter James
Associate Professor, CSE Department, Shadan women's College of Engineering and Technology,
Hyderabad, India
drisamuelpeterjames@gmail.com

Dr. K. Palani
Professor and Principal, CSE Department, Shadan women's College of Engineering and Technology,
Hyderabad, India
principalswcet2020@gmail.com

ABSTRACT

Predicting stock market trends is a complex task due to the multitude of influencing factors, such as economic indicators, historical prices, and investor behavior. Traditional forecasting models often fail to effectively capture the intricacies and dynamic nature of financial markets. This study presents a robust approach to stock market prediction using advanced machine learning techniques, specifically CatBoostRegressor and XGBRegressor. These models are chosen for their superior performance in handling both numerical and categorical data, as well as their ability to uncover complex relationships within financial datasets. Utilizing historical stock data from Apple Inc. (AAPL), the models are trained and evaluated based on metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2).

The models' predictions were visualized against actual stock prices, providing a clear indication of their predictive power. The results highlight significant predictive capabilities, demonstrating the potential of machine learning in enhancing stock market forecasts. This research contributes to the field by advancing the application of machine learning in financial forecasting, offering a promising direction for future market analysis and investment strategies. The findings suggest that machine learning techniques, particularly gradient boosting algorithms, can provide substantial improvements over traditional methods, paving the way for more accurate and reliable stock market predictions.

Keywords: Machine Learning (ML), Stock market

I. INTRODUCTION

Predicting stock market trends is a complex and challenging task that has fascinated researchers, financial analysts, and investors for decades [1]. The ability to accurately forecast stock prices can lead to significant financial gains and better risk management. However, the stock market is influenced by a myriad of factors including economic indicators, market sentiment, political events, and individual company performance. These factors contribute to the market's volatility and unpredictability, making accurate predictions a difficult endeavor.

Traditional forecasting models, such as ARIMA (AutoRegressive Integrated Moving Average) and other statistical methods, have been commonly used to predict stock prices. While these models can capture certain patterns in historical price data, they often fall short in handling the complex, non-linear relationships inherent in financial markets. The limitations of these traditional approaches have led to the exploration of more advanced techniques, including machine learning (ML) methods. Machine learning offers a powerful alternative to traditional methods by leveraging large datasets and identifying complex patterns that may not be apparent through conventional analysis. ML models can learn from historical data and improve their predictions over time. Among the various ML techniques, ensemble learning methods, which combine the predictions of multiple models, have shown promise in enhancing prediction accuracy. This study focuses on the application of machine learning techniques, specifically using CatBoost and XGBoost algorithms, to predict stock prices. These algorithms are well-suited for handling the high-dimensional and non-linear nature of stock market data. By utilizing these advanced ML models, this research aims to improve the accuracy and robustness of stock market predictions compared to traditional methods.

By integrating modern machine learning algorithms with financial data, this research seeks to contribute to the growing body of knowledge on stock market prediction and provide a valuable tool for investors and financial analysts.

OBJECTIVE

The primary objective of this project is to develop an advanced machine learning-based system for predicting the stock prices of Apple Inc. (AAPL). Traditional stock market prediction models often struggle to capture the complex and dynamic nature of stock price movements [2]. To address these limitations, this project aims to enhance prediction accuracy by leveraging sophisticated machine learning algorithms, particularly CatBoostRegressor and XGBRegressor. The study focuses on utilizing historical stock market data to train and validate these models. The predictive performance of each model will be evaluated using key metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and the R-squared (R^2) value. By comparing the performance of CatBoost and XGBoost, this project will analyze the strengths and limitations of each model in the context of stock price prediction.

II. PROBLEM STATEMENT

Predicting stock prices, especially for companies like Apple Inc. (AAPL), is a complex challenge due to the market's volatility and the influence of numerous factors such as economic trends, global events, and investor behavior. Traditional models

often fail to capture the nonlinear patterns in stock data. Advanced machine learning models like CatBoost and XGBoost offer more effective solutions but require further exploration to maximize accuracy and efficiency. This project aims to apply these techniques to improve stock price prediction and assess their performance compared to traditional methods.

EXISTING SYSTEM

Current stock price prediction systems often rely on traditional models and methods, such as linear regression, ARIMA (AutoRegressive Integrated Moving Average), and basic time-series analysis. While these models have been commonly used in financial forecasting, they come with several limitations. First, they struggle to handle the complexity and non-linear patterns present in stock market data, as they typically assume linear relationships between variables that may not accurately capture the dynamic nature of financial markets. Additionally, conventional models often fail to fully utilize the vast amount of available financial data, overlooking crucial technical indicators and market signals, leading to less-than-optimal predictions. Furthermore, these methods may lack the adaptability required to respond to changing market conditions, making them less reliable in terms of prediction accuracy over time. As a result, many existing systems primarily depend on historical stock prices and basic statistical techniques, which restrict their ability to provide precise and actionable insights into market trends .

Disadvantage of Existing System

Existing stock prediction systems, such as linear regression and ARIMA, struggle with several limitations. They often assume a linear relationship between variables, failing to capture the non-linear complexities of financial markets. These traditional models rely heavily on historical data, overlooking key technical indicators and external factors, which leads to suboptimal predictions. Additionally, they lack the flexibility to adapt to rapid market changes, resulting in reduced accuracy and reliability, especially in volatile conditions. Overall, their inability to handle complex patterns and adapt to evolving trends limits their effectiveness in stock price forecasting.

PROPOSED SYSTEM

The proposed system aims to improve stock price prediction accuracy by addressing the limitations of traditional models through the use of advanced machine learning techniques, specifically XGBoost and CatBoost, along with the yfinance library for data acquisition.

Data Acquisition and Feature Engineering: The system integrates yfinance to retrieve historical stock data for Apple Inc., including stock prices, trading volumes, and other financial metrics. The data is further enriched with technical indicators such as moving averages (MA_10, MA_50, MA_100), volume averages (Volume_10), and the Relative Strength Index (RSI) to provide deeper insights into market trends and potential price movements.

Advanced Machine Learning Models: XGBoost and CatBoost, two powerful gradient boosting algorithms, are employed to handle complex data patterns. XGBoost is known for its speed and accuracy, while CatBoost excels in managing categorical data and offering reliable predictions [3].

Model Training and Evaluation: The models are trained on the prepared dataset and evaluated providing a comprehensive measure of their performance.

By integrating yfinance and leveraging advanced machine learning models, the proposed system offers more accurate and actionable stock price predictions, addressing the shortcomings of conventional methods and delivering enhanced forecasting for investors.

Advantages of Proposed System

The proposed system offers several key advantages over traditional stock prediction models. By using advanced machine learning techniques like XGBoost and CatBoost, it provides improved accuracy in predicting stock prices by capturing complex, non-linear relationships in the data. The system also incorporates additional technical indicators such as moving averages and the Relative Strength Index (RSI), enriching the dataset and enhancing prediction precision. The integration of the yfinance library ensures efficient data acquisition, while the adaptability of these algorithms allows the system to adjust to changing market conditions. This leads to more reliable, actionable stock forecasts for investors.

III. RELATED WORKS

Several studies have explored various methodologies for stock price prediction, leveraging different models and techniques.

An LSTM-Based Model for Stock Price Prediction by V. Sarika et al. (2023) focuses on the application of Long Short-Term Memory (LSTM) networks, The application of a sort of machine learning technique known as Long Short Term Memory, which is based on recurrent neural networks (RNNs), is the main topic of this study [4].

Stock Price Prediction Based on LSTM Neural Network: The Effectiveness of News Sentiment Analysis by Yuqiao Guo (2020) , this paper retrieves news articles from the New York Times and conducts sentiment analysis for news headline and text body, then combine quantitative sentiment score with stock historical stock basic features together, using LSTM neural network to predict both future stock close price and stock return [5].

Stock Trend Prediction Based on ARIMA-LightGBM Hybrid Model by Xiuyan Zheng et al. (2022) In the proposed paper, The ARIMA model was used for the six-month prediction of exogenous variables. Secondly, the LightGBM model is used to model the exogenous variables predicted by the ARIMA model to obtain the predicted stock trend in the next six months [6].

Stock Index Prediction Method Based on ARIMA-ELM

Combination Model by Yi Peng et al. (2021), This paper combines two models, differential integrated moving average autoregressive model(ARIMA) and the extreme learning machine model(ELM), to predict the closing price of stock indexes [7].

These works illustrate the diverse approaches and advancements in stock price prediction, highlighting the ongoing evolution of models and techniques to improve forecasting accuracy and reliability.

IV. METHODOLOGY OF PROJECT

The project's methodology involves several key steps to ensure effective stock price forecasting. Initially, data acquisition is carried out using yfinance to collect historical stock data for Apple Inc. (AAPL) from January 1, 2020, to January 1, 2023 [8]. This dataset includes crucial elements like stock prices and trading volumes, which are essential for analysis. Additional financial metrics such as moving averages and the Relative Strength Index (RSI) are derived from this raw data to enhance the predictive features of the model.

In the feature engineering phase, the raw data is refined by calculating technical indicators. These include Moving Averages (MA_10, MA_50, MA_100) to identify trends, Volume Averages (Volume_10) to assess market activity, and RSI to measure price movement momentum. These indicators are vital for capturing various market behaviors and boosting model performance.

During data preprocessing, features are normalized using StandardScaler to ensure consistent scaling, which is crucial for effective model training. The dataset is then divided into training and testing sets with an 80-20 split, allowing the model's performance to be evaluated on unseen data.

For model training, XGBoost and CatBoost algorithms are used, known for their ability to handle complex datasets and capture non-linear relationships. The models' performance is assessed using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2), providing a detailed evaluation of their accuracy, robustness, and predictive performance.

MODULES:

The project is structured into several key modules to ensure a comprehensive approach to stock price prediction:

1. Data Acquisition Module: This module focuses on gathering historical stock data for Apple Inc. (AAPL) using the yfinance library. It retrieves data from January 1, 2020, to January 1, 2023, including critical variables like stock prices and trading volumes. This data forms the foundation for further analysis and modeling.

2. Feature Engineering Module: In this module, raw stock data is enhanced through the calculation of various technical indicators. Moving Averages (MA_10, MA_50, MA_100) are computed to capture trends and smooth price fluctuations. Volume Averages (Volume_10) provide insights into trading activity, while the Relative Strength Index (RSI) helps identify overbought or oversold conditions. These features are essential for improving the model's predictive accuracy.

3. Data Preprocessing Module: This module handles the preparation of data for modeling. It includes normalization of features using StandardScaler to ensure that all input variables are on a comparable scale, which is crucial for effective machine learning performance. The data is then split into training and testing sets with an 80-20 ratio, allowing for robust model evaluation.

4. Model Training and Evaluation Module: The core of the project, this module involves training machine learning models using the XGBoost and CatBoost algorithms. These models are known for their high performance and ability to handle complex data patterns. The trained models are evaluated using performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2), which provide insights into their accuracy and reliability in predicting stock prices.

V. ALGORITHM USED IN PROJECT

The project utilizes two advanced machine learning algorithms: XGBoost and CatBoost. XGBoost, an implementation of gradient boosting. It is particularly effective at managing non-linear relationships within datasets, making it a powerful tool for stock price prediction. CatBoost, on the other hand, is designed to work exceptionally well with categorical features and provides robust predictions with minimal preprocessing. Both algorithms leverage the boosting technique, which builds multiple decision trees iteratively to improve accuracy. By using these models, the system aims to capture complex patterns in stock data and enhance prediction reliability.

Benefits

The proposed system offers several benefits over traditional stock prediction models. By utilizing advanced machine learning algorithms like XGBoost and CatBoost, the system is capable of capturing complex, non-linear relationships in stock market data, leading to more accurate predictions. These models are known for their high performance, adaptability, and ability to handle large datasets efficiently. Additionally, the integration of technical indicators such as moving averages and the Relative Strength Index (RSI) enhances the predictive power by providing deeper insights into market trends. The use of the yfinance library for real-time data acquisition ensures that the system remains updated with the latest stock information, further improving the reliability of predictions. Overall, the system is designed to offer investors and analysts more actionable and precise insights, helping them make informed decisions.

DATA FLOW DIAGRAM

Fig: 1 Flow diagram

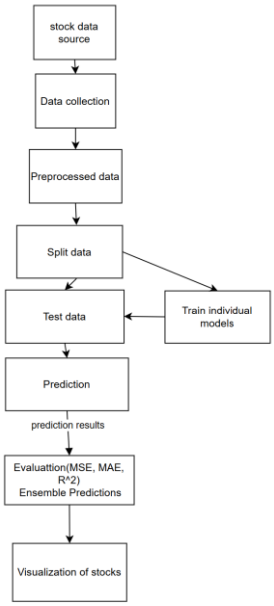

Fig: 2 Flow Diagrams

VI. SYSTEM ARCHITECTURE

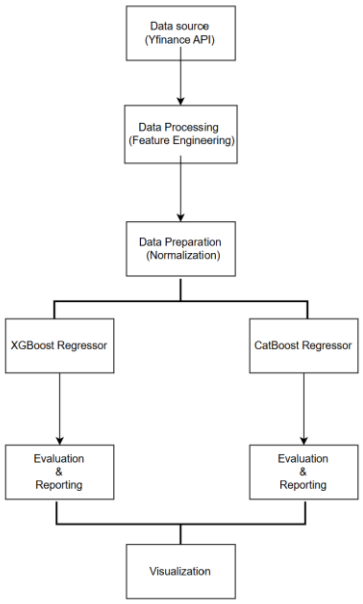

Fig: 3 ARCHITECTURE DIAGRAM OF SYSTEM

VII. CONCLUSION

This project demonstrates the potential of machine learning in financial forecasting, using advanced models like XGBoost and CatBoost to predict Apple Inc. (AAPL) stock prices. By incorporating historical data and technical indicators such as moving averages and RSI, the models were able to capture complex market dynamics, yielding accurate predictions. While the results are encouraging, future improvements could include adding macroeconomic data, sentiment analysis from news and social media, and exploring deeper learning techniques. Expanding the system for real-time predictions and applying it to other markets could further validate the models' effectiveness.

In conclusion, the use of XGBoost and CatBoost in this project highlights their value in stock prediction and opens the door for more comprehensive financial forecasting approaches.

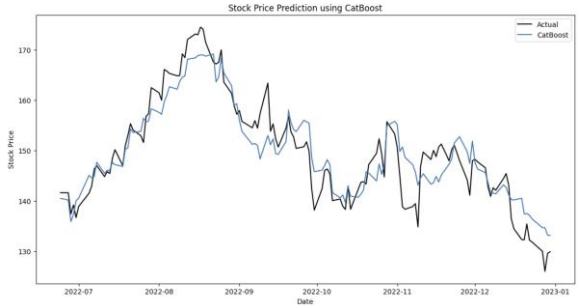

Fig: 4 Prediction using CatBoost

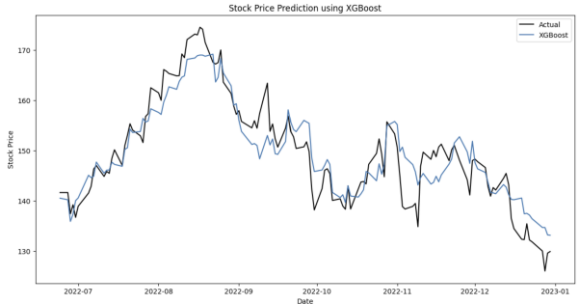

Fig: 5 Prediction using XGBoost

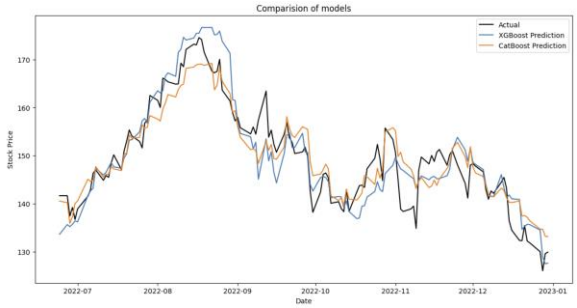

Fig: 6 Comparison of model

I. FUTURE ENHANCEMENT

The future scope of this project includes several potential enhancements to improve the model's predictive accuracy and broaden its application. Expanding the dataset to incorporate a wider variety of financial indicators, such as macroeconomic factors and alternative data sources like sentiment analysis from news and social media, could offer a more holistic perspective on market dynamics, leading to more precise predictions. Incorporating real-time data streaming would enable the model to dynamically adjust to current market conditions, providing timely and up-to-date forecasts.

Future work may also investigate advanced feature engineering methods and consider deep learning techniques for potentially superior performance. Additionally, adapting the model for different stocks and market sectors would increase its usefulness, making it applicable across various industries. Ongoing updates and validations with fresh market data will ensure the model remains relevant and reliable, maintaining its effectiveness in the continuously changing financial landscape.

II. REFERENCES

[1] Ankita Bharti, Shivani, Anushka Kumari, Jinisha Jain, B. Indra Thannaya, “Algorithmic Stock Market Prediction,” International Journal Of Engineering Research And Development e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com Volume 20, Issue 2 (February. 2024), PP. 43-49.

[2] Wuzhida Bao , Yuting Cao , Yin Yang , Hangj un Che , Junjian Huang , Shiping Wen , “Data-driven stock forecasting models based on neural networks: A review” https://www.sciencedirect.com/science/article/pii/S1566253524003944

[3] https://www.analyticsvidhya.com/blog/2020/02/4-boosting-algorithms-machine-learning/

[4] V. Sarika, G. V. S. Kamal, S. V. Pratham, N. V. S. S. Deepak and T. Veneela, "An LSTM-Based Model for Stock Price Prediction," *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, Kanjirapally, India, 2023, pp. 1-6, doi: 10.1109/AICERA/ICIS59538.2023.10420270.

[5] Y. Guo, "Stock Price Prediction Based on LSTM Neural Network: the Effectiveness of News Sentiment Analysis," *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, Chongqing, China, 2020, pp. 1018-1024, doi: 10.1109/ICEMME51517.2020.00206.

[6]] X. Zheng, J. Cai and G. Zhang, "Stock Trend Prediction Based on ARIMA-LightGBM Hybrid Model," *2022 3rd Information Communication Technologies Conference (ICTC)*, Nanjing, China, 2022, pp. 227-231, doi: 10.1109/ICTC55111.2022.9778304.

[7] Y. Peng, K. He and Q. Yu, "Stock Index Prediction Method based on ARIMA-ELM Combination Model," *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, Changsha, China, 2021, pp. 637-642, doi: 10.1109/AEMCSE51986.2021.00132.

[8] https://deepnote.com/guides/data-science-and-analytics/introduction-to-python-for-banking