

STOCK DATA ANALYSIS AND PRICE DETECTION

Ambika Mishra¹ Bharti Anisha² Abhithilak Srivastava³

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MADAN MOHAN MALAVIYA UNIVERSITY OF TECHNOLOGY GORAKHPUR

Abstract - The aim of the project is to examine a number of different forecasting techniques to predict future stock returns based on past returns and numerical news indicators to construct a portfolio of multiple stocks in order to diversify the risk. We do this by applying supervised learning methods for stock price forecasting by interpreting the seemingly chaotic market data. This research paper presents a machine learning-based approach for predicting stock price movements using historical market data. The goal of this study is to analyze how supervised learning models can detect patterns, reduce prediction errors, and improve decision-making for investors. The methodology includes data collection, preprocessing, feature scaling, and model training using algorithms such as Linear Regression, Random Forest, and LSTM. Experimental results show that machine learning models can capture both short-term and long-term trends with reasonable accuracy. This work highlights the importance of data quality, model selection, and evaluation metrics in building an effective stock prediction system .

I. INTRODUCTION

The fluctuation of stock market is violent and there are many complicated financial indicators. However, the advancement in technology, provides an opportunity to gain steady fortune from stock market and also can help experts to find out the most informative indicators to make better prediction. The prediction of the market value is of paramount importance to help in maximizing the profit of stock option purchase while keeping the risk low. The stock market is one of the most dynamic and unpredictable financial environments, influenced by economic conditions, global events, investor behavior, and numerous external factors. Traditional statistical methods often struggle to capture these complex patterns. With the rapid advancement of technology, Machine Learning (ML) has become a powerful tool for analyzing historical stock data and identifying hidden trends. ML algorithms can process large datasets, learn from patterns, and provide predictions that assist investors in making informed decisions. This research focuses on applying machine learning techniques to forecast stock price movements and evaluate their performance through standard metrics.

II. METHODOLOGY

The methodology followed in this research includes several systematic steps that ensure accurate and reliable stock price prediction. The first step involves data collection from publicly available stock market datasets. After collecting the data, preprocessing is performed to remove missing values, normalize the features, and convert the dataset into a machine-learning-friendly format. Feature engineering is applied to extract meaningful indicators such as moving averages, daily returns, and volume trends.

After preprocessing, multiple machine learning models—including Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) networks—are trained using the processed dataset. Each model is evaluated using performance metrics such as Mean Squared Error (MSE) and R^2 score. The final step involves comparing these models to identify which one provides the most accurate stock price prediction.

III. RESULTS AND DISCUSSION

The performance of each machine learning model was evaluated after training on historical stock data. Linear Regression demonstrated moderate accuracy but struggled to capture non-linear trends present in financial time series. The Random Forest model performed significantly better due to its ability to handle complex relationships and noise within the data. However, the most accurate predictions were obtained using the Long Short-Term Memory (LSTM) model, which is specifically designed for sequential data.

LSTM effectively captured long-term dependencies and provided smoother prediction curves. Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score indicated that LSTM outperformed the other models. The graphical comparison between actual and predicted stock prices further highlighted the superior performance of deep learning techniques. These results confirm that machine learning, especially neural networks, can be highly effective in forecasting stock market trends.

IV. CONCLUSION

This research demonstrates that machine learning provides an effective approach for predicting stock price movements using historical market data. Traditional models such as Linear Regression offer simple and fast predictions but lack the capability to capture non-linear behavior in financial time series. In contrast, more advanced models like Random Forest and LSTM show significantly improved accuracy and stability.

The results confirm that LSTM-based deep learning methods are well-suited for sequential data analysis and outperform conventional algorithms in most evaluation metrics. Although no model can perfectly predict the stock market due to its volatile and uncertain nature, machine learning techniques provide valuable insights that can support investor decision-making and risk management.

V. LITERATURE REVIEW

The application of machine learning in stock market prediction has gained significant attention due to its ability to identify non-linear patterns in large datasets. Early studies primarily relied on statistical approaches such as ARIMA and Moving Average Models, which performed well for linear and stable trends but failed in highly volatile markets. Later, researchers introduced Support Vector Machines (SVMs), Decision Trees, and k-Nearest Neighbors (k-NN) to capture complex relationships in financial time series.

Recent advancements highlight the effectiveness of ensemble techniques such as Random Forest and Gradient Boosting, which combine multiple weak learners to produce more reliable predictions. According to several empirical studies, ensemble models outperform traditional regression methods by reducing overfitting and improving generalization.

Deep learning models, especially Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, have further enhanced predictive accuracy. Researchers have demonstrated that LSTM captures dependencies across longer time intervals and adapts better to noisy financial data. Modern studies also explore hybrid architectures such as CNN–LSTM, which extract spatial features before temporal modeling.

Overall, existing literature suggests that machine learning, particularly deep learning, is highly suitable for stock price forecasting. This research builds upon previous work by comparing multiple ML models and evaluating their performance using real-world stock market data.



Fig. RMSE value calculation

VI. SYSTEM ARCHITECTURE

The system architecture used in this research consists of five major components that work together to achieve accurate stock price prediction.

The first module is **Data Collection**, where historical open, high, low, close (OHLC) values and volume data are gathered from reliable stock market sources. The second module is **Data Preprocessing**, involving removal of missing values, normalization, and transformation of date-time formats into machine-readable structures.

The next component is **Feature Engineering**, where indicators such as Moving Average (MA), Exponential Moving Average (EMA), Relative Strength Index (RSI), and price momentum features are generated. These features enrich the dataset and help the model interpret underlying trends.

The fourth component is **Model Training**, where algorithms like Linear Regression, Random Forest, and LSTM are trained on the processed dataset. Each model undergoes hyperparameter tuning to achieve maximum accuracy.

Finally, the **Prediction Module** generates future stock prices using the best-performing model. The modular design of the architecture ensures flexibility, scalability, and easy integration with real-world applications.

VII. LIMITATIONS

Although machine learning improves the accuracy of stock market forecasting, several limitations must be considered. Stock prices are affected not only by historical data but also by unpredictable global events, political tensions, economic announcements, and investor sentiment. Machine learning models may fail to capture sudden market changes caused by such external factors.

Additionally, complex models like LSTM require large datasets and high computational power. Training deep learning models without proper hardware can be time-consuming.

Another major limitation is model generalization. A model that performs well on past data may not always predict future patterns accurately. Overfitting is a common issue when dealing with financial data. Therefore, predictions should be interpreted cautiously and used for supporting decisions rather than replacing human judgement.

Algorithm	RMSE Value	R-squared Value
Random Regressor	1.4325434e-07	0.956669
Bagging Regressor	1.329966e-07	0.959771
Adaboost Regressor	2.9882972e-07	0.909611
KNeighbours Regressor	0.00039015	-117.01176
Gradient Boosting Regressor	1.274547e-07	0.961448

VIII. FUTURE WORK

Future work can focus on incorporating additional datasets such as real-time news headlines, social media sentiment, and macroeconomic indicators to improve prediction accuracy. Integrating Natural Language Processing (NLP) techniques can help capture the effect of public opinion on stock movements.

Further research can also explore advanced deep learning architectures such as Transformer Models and attention-based LSTM networks, which have shown excellent performance in sequential data tasks.

Real-world deployment of the model as a web application or mobile app is another promising direction. This would allow users to access live forecasts and trend analysis in real time. Expanding the model to include multiple stocks or entire indices could make the system more robust and widely usable.

IX. RESULTS

Based on the results obtained, it is found that Gradient Boosting Regressor consistently performs the best. This is followed by Bagging Regressor, Random Forest Regressor, Adaboost Regressor and by K Neighbour Regressor. Bagging Regressor is found to perform good as Bagging (Bootstrap sampling) relies on the fact that combination of many independent base learners will significantly decrease the error. Therefore we want to produce as many independent base learners as possible. Each base learner is generated by sampling the original data set with replacement. From the results, it is safe to say that additional hidden layer(s) improve upon the score of the models. Random Forest is an extension of bagging where the major difference is the incorporation of randomized feature selection.

X. ACKNOWLEDGMENT

We would like to thank Dr. Vimal Kumar, our course instructor for Statistical Methods in AI, and clearing basic concepts required as part of the Project.

XI. REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, 2015.
- [2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.
- [3] C. Krauss, X. Do, and N. Huck, "Deep neural networks for stock market prediction," *European Journal of Operational Research*, 2017.
- [4] F. Chollet, "Deep Learning with Python," Manning Publications, 2017.

[5] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, 2009.

[6] Y. Kim, "Financial market prediction using machine learning," IEEE Conference on Big Data, 2020