**Stock Market Prediction Using Machine Learning: A Comparative Study of Supervised Algorithms**

Rajeshwari Golande

PRN: 122B1D021

rajeshwari.golande22@pccoepune.org

Yash Jahagirdar

PRN: 122B1D027

yash.jahagirdar22@pccoepune.org

Rohit Kandelkar

PRN: 122B1D031

rohit.kandelkar22@pccoepune.org

**ABSTRACT**

This Research paper presents a comprehensive study on stock market prediction using machine learning algorithms. Stock market data is known for its non-linearity, volatility, and high dimensionality, which present unique challenges. In this study, we utilize supervised learning techniques such as Linear Regression, Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) to predict stock prices. The dataset used includes historical price data of various stocks, enriched with technical indicators and sentiment analysis from financial news. Results indicate that the machine learning models can provide significant predictive power, with Random Forest showing the highest accuracy. Furthermore, this study evaluates the performance of each model based on evaluation metrics, such as Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The implications of this research extent to both investors and analyst aiming to leverage machine learning for more informed decision-making in stock trading.

**Keywords**

Stock Market Prediction, Machine Learning, Supervised Algorithms, Random Forest, Linear Regression, Technical Indicators, Sentiment Analysis.

# INTRODUCTION

# The stock market is a key part of the global economy. It’s where people can buy and sell shares of companies. This helps businesses get the money they need to grow and innovate. For investors, it means owning a part of a company and potentially earning money from it.

# The stock market also helps keep the economy stable. It shows the real-time value of companies based on how much people are willing to pay for their shares. This transparency ensures that resources are used effectively. Additionally, by letting investors spread their investments across different companies, the stock market helps reduce the risk of big losses and supports overall financial health.

# METHODOLOGY

This section outlines the approach used in developing and evaluating machine learning models for stock market prediction. It details the algorithms used, data collection process and the preprocessing steps that were taken to ensure reliable and insightful predictions.

## MACHINE LEARNING ALGORITHMS

## Five models were used to analyze and predict data effectively:

1. **Linear Regression:**

Linear regression is a fundamental statistical modelling tool widely used in predictive analytics. This algorithm works by establishing a linear relationship between the dependent variable (such as stock prices) and one or more independent variables (features). By fitting a straight line to the data, it simplifies the interpretation of how different features impact the outcome. Linear regression serves as an essential benchmark for evaluating the performance of more sophisticated algorithms, due to its simplicity and ease of understanding.[7]

1. **Adaptive Boosting (AdaBoost):**

AdaBoost is an ensemble technique designed to enhance the accuracy of weak classifiers. It operates by sequentially adding classifiers to the model, each focusing on the errors made by the previous ones. By assigning higher weights to misclassified instances, AdaBoost improves the model's robustness and performance. This method is particularly effective for scenarios where boosting the accuracy of simpler models is essential, making it valuable for applications requiring precise classification.[8]

1. **Random Forest:**

Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions for more accurate and stable results. During training, each tree is constructed from a random subset of the data, reducing the risk of overfitting and enhancing the model's generalization capabilities. This method excels in capturing complex interactions within the data, making it suitable for tasks such as stock price prediction where intricate relationships between variables exist.[9][10][11]

1. **Support Vector Machines (SVM):**

Support Vector Machines are powerful supervised learning algorithms used for classification tasks, especially in high-dimensional spaces. SVM works by finding the optimal hyperplane that best separates different classes in the feature space. The use of kernel functions allows SVM to handle non-linear relationships by transforming data into higher dimensions. This characteristic makes SVM highly effective in distinguishing between complex and overlapping categories, ensuring robust classification performance. The outcomes of the study by *Naliniprava Tripathi* indicate that the average prediction accuracy is 60.2% [11][12][13]

1. **XGBoost**:

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of gradient boosting algorithms. It is designed for both speed and performance, employing techniques such as regularization to prevent overfitting and parallel processing for faster computation. XGBoost's ability to handle large datasets and deliver state-of-the-art predictive accuracy makes it a preferred choice in many competitive machine learning tasks. Its applications range from finance and healthcare to marketing analytics, demonstrating its versatility and effectiveness in various domains. XGBoost, as a regression model, builds an ensemble of trees where each new tree corrects the errors of the previous ones. It starts with an initial prediction and iteratively adds trees, each focusing on the residuals. This process involves splitting features to optimize predictions. The final prediction combines the outputs of all trees, resulting in a robust and accurate model for regression tasks. [14]

## MODEL TRAINING AND EVALUATION

### DATA COLLECTION

Historical stock market data was sourced from NSE (National Stock Exchange). The data has the following attributes:

1. **Date**: The date of the stock market activity (in "01-Jan-14" format). It serves as the timeline for tracking daily stock prices and trends.
2. **OPEN**: The stock price at the beginning of the trading day. This is the first price a stock trades at once the market opens.
3. **HIGH**: The highest price at which the stock traded during the day. This reflects the maximum value the stock reached.
4. **LOW**: The lowest price at which the stock traded during the day. This shows the minimum value the stock dropped to.
5. **PREV. CLOSE**: The stock's closing price on the previous trading day. This is useful for calculating price changes or percentage growth from the prior day.
6. **ltp (Last Traded Price)**: The price at which the stock was last traded on that particular day. It's the most recent price before the market closes.
7. **close**: The stock’s closing price for that day, which is the final price at which the stock was traded during the regular market hours.
8. **VWAP (Volume-Weighted Average Price)**: A trading benchmark that gives the average price the stock has traded at throughout the day, based on both volume and price. It's useful for understanding the overall price trend.
9. **52W H (52-Week High)**: The highest price the stock has reached in the last 52 weeks (1 year). It's used to gauge long-term performance.
10. **52W L (52-Week Low)**: The lowest price the stock has hit in the last 52 weeks. It's another long-term performance metric.
11. **VOLUME**: The number of shares traded during the day. It indicates the liquidity and trading activity for the stock.
12. **VALUE**: The total value of the trades made during the day. It can help assess the overall interest and movement in the stock.
13. **No of trades**: The total number of transactions or trades that occurred during the day for that stock.

### PREPROCESSING

Data preprocessing involved handling missing values, sorting according to dates, and converting numerical values to suitable form, such as removing commas from numbers and converting strings to numerical.

The dataset was split into training (80%) and testing (20%) subsets. Models were evaluated using Root Mean Squared Error (RMSE) as it measures the average prediction error in the same units as the target variable. It is sensitive to large errors, making it effective for identifying models that perform poorly on significant deviations from actual values.

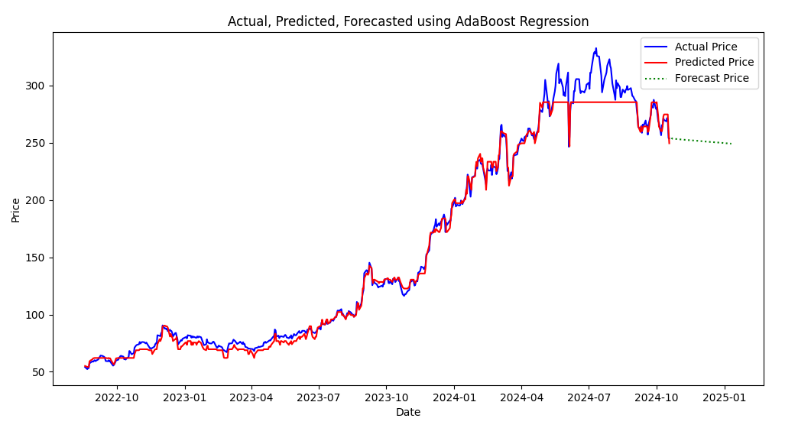
# RESULTS & DISCUSSION

The performance of the models varied significantly. Support Vector Machine produced the highest error rate. XGBoost and AdaBoost significantly performed better but suffered from overfitting. Linear Regression achieved the best accuracy, with highest RMSE of 13.34 and MAPE of 0.72% which is significantly lower than other models, indicating its effectiveness in capturing complex patterns. Random Forest, showed an accuracy more than SVM but was incomparable to XGBoost and AdaBoost models.

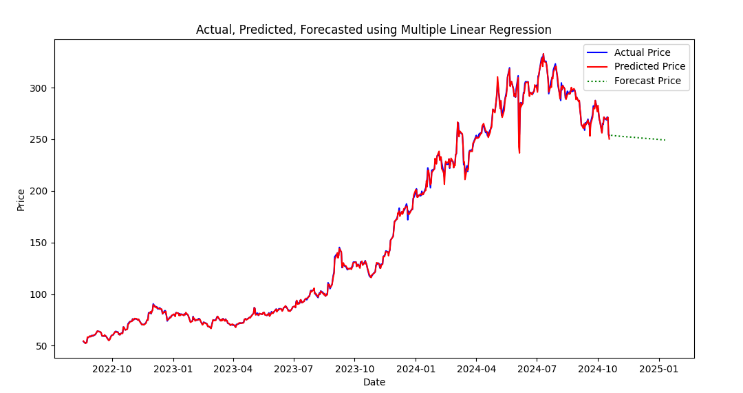
In terms of practical implications, Ensemble techniques and Boosting Algorithms provides a more reliable and accurate model for regression and classification task

**Comparison of various models using BHEL stock as example:**

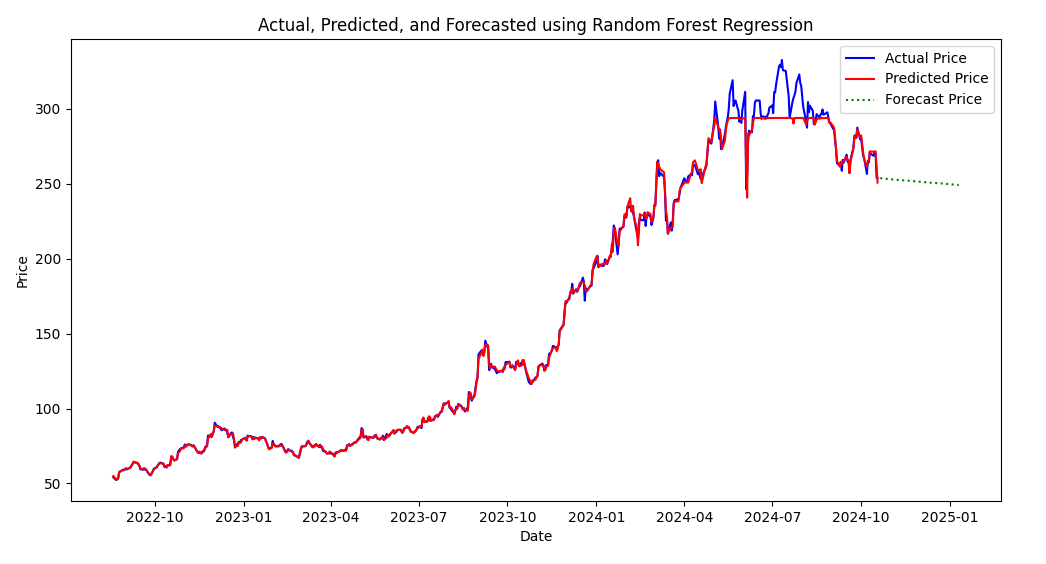
1. Using Adaptive Boosting:



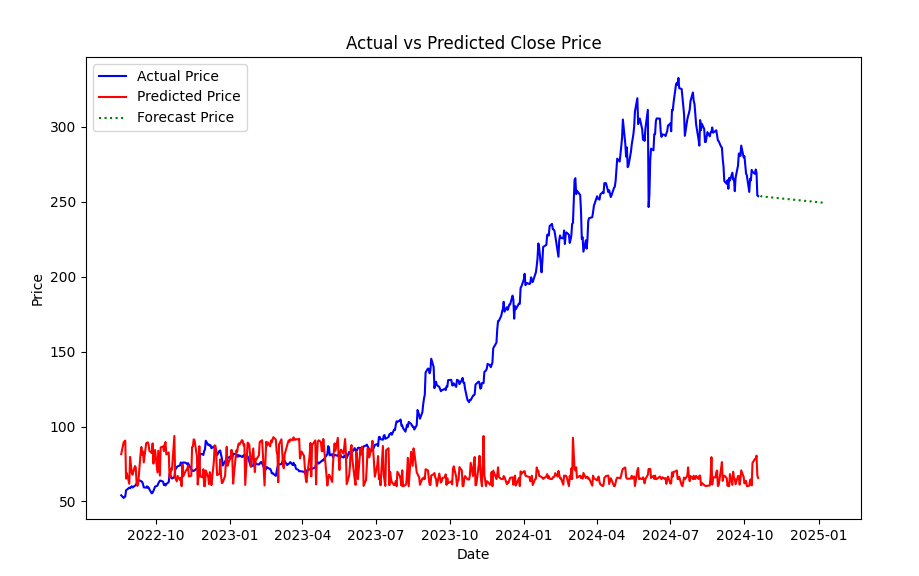
1. Using Linear Regression:



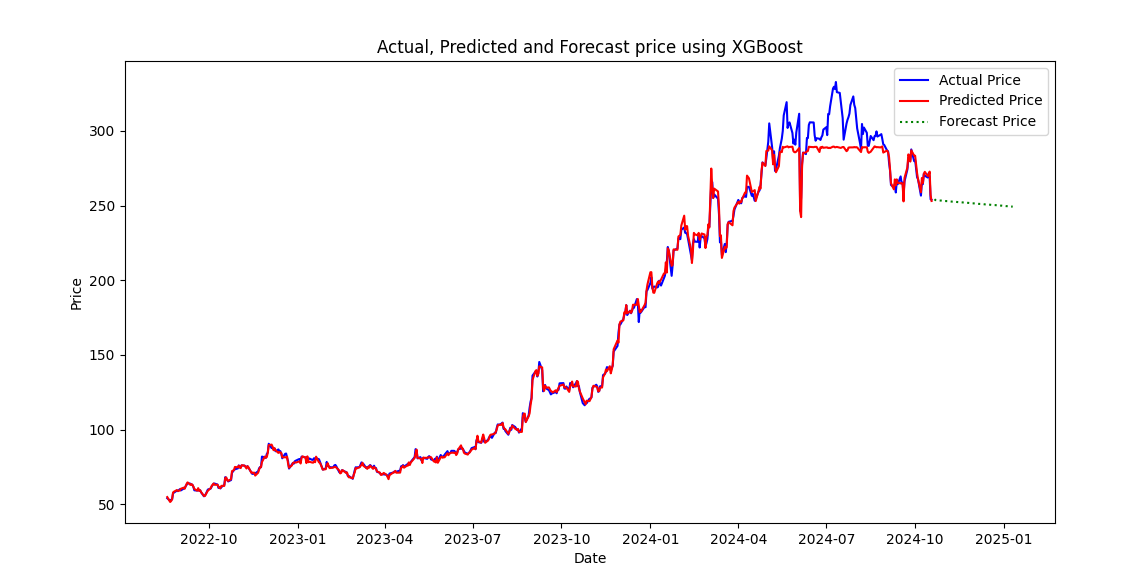
1. Using Random Forest Regression:



1. Using Support Vector Regression:



1. Using XGBoost Regression:



Even visually, it can be seen that Linear Regression offers the highest accuracy.

RMSE values

|  | AdaBoost | Linear Regression | Random Forest | SVM | XGBoost |
| --- | --- | --- | --- | --- | --- |
| BHEL | 8.56 | 1.82 | 5.65 | 16663.4 | 7.03 |
| OIL | 18.72 | 4.19 | 11.82 | 45859.37 | 17.36 |
| TataMotors | 246 | 4.01 | 233.83 | 183404.61 | 242.97 |
| TCS | 165.74 | 13.34 | 139.74 | 1357884.62 | 146.65 |
| Zomato | 69.57 | 2.21 | 68.74 | 16984.72 | 68.97 |

MAPE values (in percentage):

|  | AdaBoost | Linear Regression | Random Forest | SVM | XGBoost |
| --- | --- | --- | --- | --- | --- |
| BHEL | 3.61 | 0.65 | 1.13 | 45.3 | 1.55 |
| OIL | 2.57 | 0.61 | 1.08 | 32.76 | 1.54 |
| TataMotors | 17.43 | 0.38 | 15.92 | 42.72 | 17.04 |
| TCS | 2.02 | 0.27 | 1.49 | 29.68 | 1.68 |
| Zomato | 23.79 | 0.72 | 23.29 | 56.47 | 23.49 |

# CONCLUSION

This study demonstrates that machine learning techniques, particularly Linear Regression, can significantly enhance the accuracy of stock predictions. While models like SVM are limited in handling the complexities of stock data, advanced algorithms such as AdaBoost and XGBoost are better equipped to capture intricate relationships. Future research should focus on exploring deep learning models, such as Long Short-Term Memory (LSTM) networks, for more robust predictions. Machine Learning plays a vital role in financial markets, providing tools that enhance decision-making and potentially lead to higher returns on investment.

# ACKNOWLEDGMENTS

We extend our gratitude to the National Stock Exchange (NSE) repositories for providing historical stock data. Special thanks to our academic advisor, Mrs. Sujata Kolhe ma’am for their valuable insights into the stock market domain.

# REFERENCES

1. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The journal of Finance, 25(2), 383-417
2. Ghosh, R, & Chakrabarti, A. (2019). Stock market prediction using machine learning algorithms: A comparative study. International Journal of Advanced Research in Computer Science, 10(5), 75-82.
3. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
4. Shiller, R. J. (2003). From efficient markets theory to behavioral finance. Journal of Economic Perspectives, 17(1), 83-104.
5. Huang, W., Nakamori, Y., & Wang, S. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10), 2513-2522.
6. Li, F., & Yu, L. (2019). Application of machine learning in stock market prediction. Journal of Financial Data Science, 1(2), 17-28.
7. Stock Price Prediction Website Using Linear Regression - A Machine Learning Algorithm, ITM Web of Conference (2023)
8. An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification, Yi Ding, Hongyang Zhu, Ruyun Chen, Ronghui Li, MDPI (2022)
9. The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance, Jiajian Zheng, Duan Xin, Qishuo Cheng, Miao Tian, Le Yang
10. Predicting Stock Market Trends using Random Forest, K. Lavingia et al., IEEE Xplore (2022)
11. Stock Market Prediction using Random Forest and Support Vector Machine, R S Abirami, K.Varalakshmi, Maddika Jaswanth Reddy, Kota Venkata Madhava Reddy, Chittipi Reddy Akash, IJCRT (2022)
12. Stock Market Prediction using Support Vector Machine, Zhen Hu, Jie Zhu, Ken Tse, IEEE Xplore (2013)
13. Stock Market Prediction using Support Vector Machine Approach, Naliniprava Tripathi, International Academic Conference on Management and Economics (2019)
14. Stock Price Prediction Based on XGBoost and LightGBM YueYang, Yang Wu, Peikun Wang, Xujiali, E3S Web of Conferences (2021