Comparative Analysis of Machine Learning Algorithms to Forecast Indian Stock Market

G.Naga Chandrika[1*], *Sai Venkata Rohit Gumudavelli*[1], *Ashish Kalleru*[1], *Vijitha Kambhampati*[1], and *Pragnika Kandaggatlta*[1]

[1]Department of Information Technology, VNR Vignana Jyothi Institute Of Engineering and Technology, Hyderabad, Telangana, India

Abstract. Forecasting the stock market is a complex and challenging task, as it involves analyzing a vast amount of data and taking into account various economic, political, and social factors. This paper presents an overview of different approaches and techniques used for stock market forecasting, including fundamental analysis and machine learning. The study also highlights the different algorithms used and discusses their effectiveness in predicting the stock market. This research proposes to use five different algorithms as Decision Trees, Random forest, Generalized Linear model,Gradient boosted trees, and Support Vector Machines. This research identifies models that are close to real predictions. These algorithms are applied to BSE index data from November 2017 to February28, 2023.

1 Introduction

Forecasting the stock market is a complex and challenging task that involves analyzing various economic, financial, and geopolitical factors[1]. The stock market is a highly dynamic and volatile entity, influenced by a multitude of factors that can impact the prices of individual stocks, as well as the overall performance of the market.

Despite the challenges involved in forecasting the stock market, accurate predictions can be of great value to investors, financial institutions, and policymakers[2]. By identifying market trends and anticipating changes in economic conditions, forecasts can help investors make informed decisions about their investments and minimize their risk exposure[3].

This paper aims to explore the performance of machine learning models in forecasting the stock market in India. Specifically, we compare the standard deviation and error rate of machine learning algorithms like support vector machine, random forest, decision trees, Generalized Linear models and Gradient boosted trees in predicting the future trends of the Bombay Stock Exchange (BSE) index. Our analysis covers the period from 2017 to 2023, using daily closing prices, opening prices, and high and low values as input variables. Our results indicate that machine learning models can produce reliable forecasts of the BSE index, with Linear Regression showing the least error rate and higher standard deviation.

*Correspondingauthor:gnchandrika@gmail.com

The stock market is critical to the Indian economy because it provides a venue for businesses to raise capital and for investors to profit from their investments[4]. Over the years, the Indian stock market has seen significant growth and volatility, influenced by various economic, political, and social factors[5]. This has created a need for accurate forecasting of stock prices and market trends, to help investors make informed decisions and minimize their risk exposure[6].

Stock market forecasting in India has become increasingly important, given the significant impact of market movements on the country's economy. Accurate forecasting of stock prices can help investors identify opportunities for profitable investments, and avoid potential losses[7]. Moreover, it can also help policymakers and regulators make informed decisions about the economy, such as monetary and fiscal policies, based on the current market trends and future outlook[8].

Despite the growing importance of stock market forecasting in India, it remains a complex and challenging task, influenced by a multitude of factors that can affect the market's performance[9]. Therefore, accurate forecasting requires the use of advanced analytical tools and techniques, such as machine learning to capture the complex relationships among various variables and identify patterns in the data.

Overall, stock market forecasting in India is critical to the success of the Indian economy, and can potentially provide significant benefits to investors, policymakers, and the general public.

2 Methods used for Forecasting

2.1 Generalized Linear Model

The two key features of the generalized linear model are the ability to model response variables with non-normal distributions, and the use of a link function to connect the response variable's mean to the linear predictor [10]. In the context of stock forecasting, the GLM can be used to model the relationship between a set of predictor variables (such as economic indicators, company financials, or news sentiment) and a response variable (such as stock prices or returns) that is not normally distributed.

2.2 Decision Tree

To build a decision tree for stock forecasting, one could use historical data to identify the most important predictors of stock price movements, such as economic indicators, company financials, or news sentiment. Quinlan explains that the decision trees created using this algorithm are simple and are built in a top-down manner by testing a single attribute at each decision node[11]. These predictors would be used as the input variables for the decision tree. At the leaf node that corresponds to the last set of decision rules, the final prediction is formed. Because they are simple to understand and are capable of capturing non-linear correlations between predictors and outcomes, decision trees can be particularly helpful for stock forecasting.

2.3 Random Forest

Random forests can be a powerful tool for stock forecasting because they can capture complex non-linear relationships between predictor variables and outcomes, and are less prone to overfitting than individual decision trees. In Breiman's (2001) paper on random

forests, he describes the algorithm as a combination of decision trees, where the trees are built using a random subset of the features and training data[12].

2.4 Gradient Boosted Trees

Gradient Boosted Trees (GBTs) are a prominent machine-learning approach for stock market prediction. Boosting can produce good predictive accuracy by combining rough and moderate rules of thumb, but requires a large number of them to do so. Gradient boosting provides a general framework for incorporating many weak learning algorithms into a single model[13]. Decision trees are iteratively added to the model as part of the method, with each tree attempting to fix the flaws of the preceding tree. The process continues until the model has reached the optimal number of trees, or until a certain stopping criterion is met.

2.5 Support Vector Machine

Another well-liked machine learning approach for predicting the stock market is called support vector machines (SVMs). It can be applied for classification and regression applications. SVMs seek to find the hyperplane that maximizes the margin between two classes of data points, while also ensuring that no data points are misclassified[14]. After that, projections based on fresh data can be made using the hyperplane.

3 Methodology

3.1 Data Collection and Preprocessing

These are critical steps in any stock market forecasting study, The dependability and accuracy of the results are heavily influenced by the quality of the data used. In this sector, we will outline the data collection and preprocessing procedures used in this research on stock market forecasting in India.

3.1.1 Data Collection

In this research, the collected data on the Bombay Stock Exchange (BSE) index, which is one of the most widely used indicators of the Indian stock market's performance. We obtained the data from publicly available sources, including the BSE website.
The dataset is taken from the official bseindia.com website. The dataset consists of five attributes namely date, open, low, high, and close values. The dataset has been taken from November 2017 to February 2023 which is of 1322 records. This period was chosen to capture a range of market conditions and economic events that could influence the stock market's performance.

3.1.2 Data Preprocessing

We preprocessed the raw data after gathering it to remove any discrepancies, missing numbers, or outliers that might have impacted the precision of the research.
First, the data is checked for any missing values and replaced them with interpolated values using appropriate time-series techniques. We also normalized the data to ensure that all variables have the same scale and are comparable.

Finally, we used an 80:20 split to divide the data among training and testing sets. The machine learning model's performance was evaluated on the testing set after it had been trained on the training set.

3.2 Feature Selection and Engineering

Feature selection and engineering are critical steps in any machine learning-based stock market forecasting study, as they determine the input variables that will be used to train the models. In this section, we will outline the feature selection and engineering procedures used in this study on stock market forecasting in India. The workflow model of the proposed model is described in Fig1.

3.2.1 Feature Selection

Numerous variables, including financial news, world market conditions, and economic indicators have an impact on the BSE index. To identify the underlying patterns and connections in the data, it is crucial to choose the appropriate characteristics[15]. Open price, closing price, and high and low values, among other potential input variables, were all taken into account in this study. To find the most pertinent features, we combined statistical and machine learning-based techniques.
Our investigation led us to use the BSE index's daily closing price, opening price, and high and low values as the main features for our models. These factors are frequently employed in stock market forecasting research and are recognized to have a big effect on the performance of the stock market.

3.2.2 Feature Engineering

In addition to selecting the input variables, we also performed feature engineering to extract additional information from the raw data and improve the performance of our models.
Overall, the feature engineering process involved selecting the most relevant variables and extracting additional information from the raw data to improve the accuracy of our models.

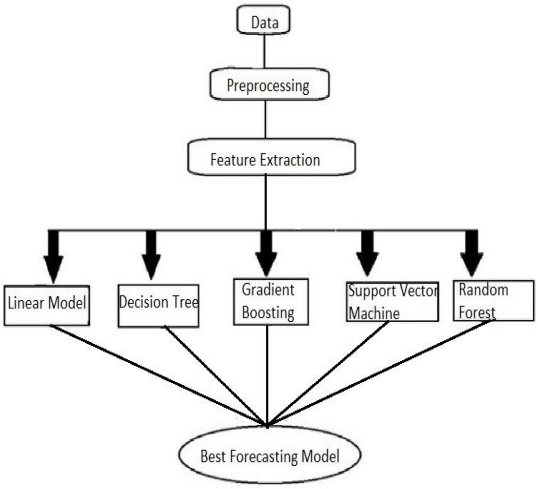

Fig. 1. Workflow

For each machine learning algorithm, the model gets trained based on the attributes such as Date, Open, Low, High, and Close. Finally, these models are ready to forecast and the comparison of all the algorithms with their respective error rate and standard deviation are presented.

4 Evaluation Metrics

4.1 Standard Deviation

A set of data points dispersion from the mean or average value is measured by their standard deviation. It is often used in stock market forecasting as a way to measure the volatility or risk of a particular stock or portfolio.
To calculate the standard deviation of stock returns, you would first need to gather data on the historical returns of the stock or portfolio. Then, you would use statistical software or a spreadsheet program to calculate the standard deviation of those returns.
The formula for calculating the standard deviation is

$$\sigma = sqrt(\Sigma(xi - \mu)2 / (n - 1))$$

where as,
σ- standard deviation
xi -i^{th} data point
μ- mean or average value of the data points
n - number of data points

After determining the standard deviation, you may use it to gauge the stock's or portfolio's level of risk. When comparing risk or volatility, Greater risk or volatility is indicated by a higher standard deviation, where as lesser risk or volatility is indicated by a lower standard deviation.

4.2 Error Rate

Calculating the error rate in stock market forecasting entails comparing the predicted returns of a stock or portfolio with the actual returns to assess how accurate a prediction model is. This is commonly done by calculating the error or residual, which is the difference between the expected and actual returns. The mean squared error calculates the average squared difference between the desired and actual returns, which can then be obtained by squaring the error and adding it up across all time periods. The root mean squared error (RMSE), which is a measurement of the average, is obtained by taking the square root of the MSE. Finally, dividing the RMSE by the average return of the stock or portfolio gives the percentage error rate.

$$\% \text{ Error Rate} = (\text{RMSE} / \text{Average Return}) * 100\%$$

where
RMSE is the root mean squared error.
Average Return is the mean return of the stock or portfolio over the same time period.

The percentage error rate gives a measure of the accuracy of the forecasting model, with a lower percentage indicating a more accurate prediction. It's critical to remember that this formula presumes that the forecasting model is linear and that the errors are normally distributed. If these hypotheses are violated, alternative measures of error may be needed.

5 Results

In the results section, we will present the forecasting results for the machine learning techniques utilized in this research on stock market forecasting in India. We have compared the standard deviation and error rating for all five machine learning algorithms. The error rate and standard deviation of all the five models on the train data are shown in Table1. The graph in Fig2. represents error rate and standard deviations of five methods on the train data.

Table 1. Summary of the methods on train set

Algorithm	Error rate	Standard Deviation
Random Forest	102.4574	1370.2039
SVM	1354.9043	31.91433
Decision Tree	115.3866	1384.2265
Gradient boosted trees	102.5381	1372.5882
Linear Regression	67.2372	1370.2906

Fig. 2. Error rate and Standard deviation of the methods on trainset

Table 2 . Summary of the methods on the test set

Algorithm	Error rate	Standard Deviation
Random Forest	13.6580	218.1191
SVM	178.4100	41.3563
Decision Tree	17.2597	213.4106
Gradient boosted trees	15.3967	218.9091
Linear Regression	7.9624	216.3163

The error rate and standard deviation of all the five models on the test data are shown in Table2. The graph in Fig3. represents error rate and standard deviations of five methods on the test data.

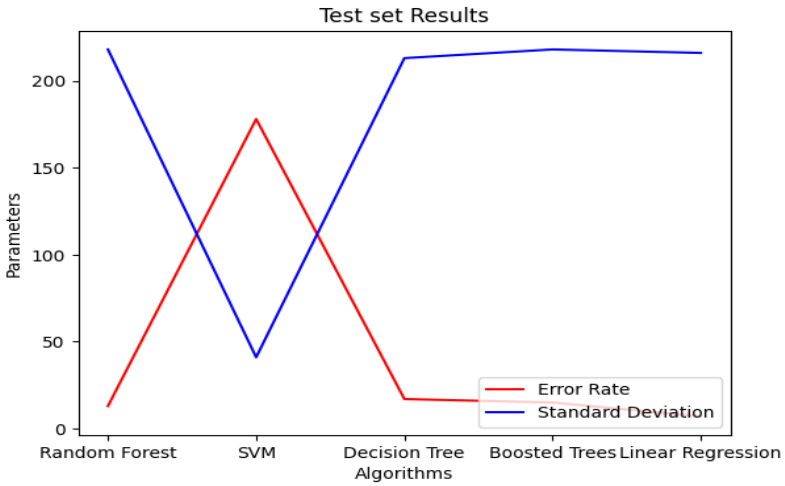

Fig. 3. Error rate and Standard deviation of the methods on test set

Overall, all five machine learning algorithms performed well in predicting the movements of the BSE index.

6 Conclusion

In this study, we explored the use of machine learning algorithms in India's stock market forecasting. We collected and preprocessed historical data on the Bombay Stock Exchange (BSE) index and used five different machine learning algorithms, namely linear regression, support vector machine, random forest, gradient boosted trees, and decision trees to develop forecasting models. We have compared the standard deviation and error rating for all five machine learning algorithms.

This study on stock market forecasting in India using machine learning algorithms yielded promising results, we would like to continue our work in the following areas. Real-time forecasting of stock market movements can provide valuable insights to investors and traders. Developing machine learning models that can process and analyze real-time data streams can be a valuable tool for financial analysts.

Stock market conditions are constantly changing, and historical patterns may not always be a reliable indicator of future movements. As a result, the forecasting models developed in this study may not be accurate in predicting stock market movements during periods of significant market volatility.

The models developed in this study were specific to the BSE index in India and may not apply to other stock markets or financial instruments. Furthermore, the performance of the models may vary depending on the specific features used and the modeling techniques applied.

References

1. L. Di Persio, O. Honchar, *Artificial neural networks architectures for stock price prediction: Comparisons and Applications*, International Journal of Circuits, Systems, and Signal Processing, **10**, 403-413, (2016).
2. E. F. Fama, *Random walks in stock market prices*, Financial Analysts Journal, **51(**1), 75-80, (1995).
3. B. G. Malkiel, *Efficient Market Hypothesis*, in Eatwell, J., Milgate, M., Newman, P. (eds) Finance. The New Palgrave. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-20213-3_13
4. E. K. Ampomah, Z. Qin, G. Nyame, *Evaluation of Tree-Based Ensemble Machine Learning Models* in Predicting Stock Price Direction of Movement, Information, **11**, no. 6,332, (2020).
5. B. M. Henrique, V. A. Sobreiro, H. Kimura, *Stock priceprediction using support vector regression on daily and up to theminute prices*, The Journal of Finance and Data Science, **4**, Issue 3, pp. 183-201, (2018).
6. M. Ouahilal, M. El Mohajir, M. Chahhou, B. E. El Mohajir, *Optimizing stock market price prediction using a hybrid approach based on HP filter and support vector regression*, 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, Morocco, 290-294, (2016) doi: 10.1109/CIST.2016.7805059.
7. Q. Qin, Q. G. Wang, J. Li, S. S. Ge, *Linear and Nonlinear Trading Models with Gradient Boosted Random Forests and Application to Singapore Stock Market*, Journal of Intelligent Learning Systems and Applications, **5**, 1-10, (2013).
8. D. Avramov, S. Cheng, L. Metzker, *Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability*, Management Science, (2021).
9. Y. Chen, Y. Hao, *A feature-weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction*, Expert Syst. Appl. **80**, C, 340–355, (2017). https://doi.org/10.1016/j.eswa.2017.02.044
10. J. A. Nelder, R. W. Wedderburn, *Generalized Linear Models*, Journal of the Royal Statistical Society. Series A (General), **135(3)**, 370-384, (1972).
11. J. R. Quinlan, *Induction of decision trees*, Machine Learning, **1(1)**, 81-106, (1986).
12. L. Breiman, *Random forests*, Machine learning, **45(1**), 5-32, (2001).
13. J. H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, Annals of Statistics, **29(5)**, 1189-1232, (1999).
14. V. N. Vapnik, A. YA. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability & Its Applications, **8(2)**, 264-280, (1963).
15. D. Srivastava, B. Lekha, *Data Classification Using Support Vector Machine*, Journal of Theoretical and Applied Information Technology, **12(1)**, 1-7, (2010).