



International Journal of Marketing Management

ISSN 2454 - 5007



www.ijmm.net

Email ID: editor@ijmm.net , ijmm.editor9@gmail.com

A Study on Machine Learning Algorithms for Indian Stock Market Predictions

¹ Dr. B. Nageshwar Rao, ² G. Meghana,

¹Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.

² MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract:

Because of its inherent complexity and variety, the stock market has always piqued the interest of those seeking methods to foretell its future movement. But this is made more challenging by the fact that financial markets throughout the world are quite volatile. While statistical approaches and modeling do a good job, they aren't foolproof when it comes to predicting the stock market's behavior. Addressing and solving the complex challenges in the stock market was beyond the capabilities of the old approaches utilized. Machine learning and artificial intelligence solutions simplify the handling of Big Data's complicated scenarios and difficulties. The goal of this study is to find a model that can predict outcomes that are close to the real ones by combining six distinct algorithms: Support Vector Machine, Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, and Gradient Boosted Trees. The BSE index data from April 2015 to March 31, 2020 was processed using these techniques, and the model with the least relative error was determined. Gradient Boosted Trees is selected as the most efficient model out of all the ones used since it has the lowest relative error and standard deviation. The outcomes are predicted using Additional Gradient Boosted Trees.

Keywords:

Machine learning, Random Forest, Support Vector Machine, Big Data, Artificial Intelligence, and the BSE Sensex

I. INTRODUCTION

The unpredictable and ever-changing character of the stock market has made price forecasting a daunting task in recent times. As far as the Indian stock market

is concerned, "Predicting the unpredictable" is currently the epicenter. The stock market, sometimes referred to as an economic barometer, is a chaotic system that is also loud, unpredictable, non-linear, and deterministic. [1] Political, cultural, and financial events, as well as the attitudes of investors and the decisions made by international institutional investors, are just a few of the many variables that influence the stock market and make it difficult to forecast financial time series values. The two most common methods for predicting stock market movements are Fundamental analysis based on the Theory of Intrinsic Value and technical analysis using the chartist approach. Efficient Market Hypothesis, which assumes a random walk, is largely at odds with these ideas. The idea that today's pricing and market conditions are not indicative of tomorrow's is fundamental to the logic of random walks [8]. The premise upon which EMH theory is built is that stock prices fully represent all relevant and up-to-date information, and therefore even the most ignorant investors may obtain a respectable rate of return when compared to financial specialists [13]. Academics and researchers from across the globe have been using machine learning algorithms in a variety of contexts as of late. Because machine learning can generalize patterns, it is mostly used to predict stock values across stock markets. Importantly, this article applies six forecasted the final values of the BSE SENSEX using methods for machine learning.

A. Vector Machine Support

Support Sparse kernel application, capacity control of the decision function, and sparsity of solution are the defining characteristics of support vector machine (SVM), a popular and enduring machine learning technique [5]. It has great potential for use in classification and regression tasks. The data points

that are near to the hyperplane and may affect its location and direction are called support vectors.

(B. SVMAlgorithm)

The support vector machine (SVM) algorithm aims to magnify the hyperplane and data point margins. The goal of applying a loss function, often called as hinge function, is to maximize the margin.

$$L(w) = \sum_{i=1} \underbrace{\max(0, 1 - y_i[w^T x_i + b])}_{\text{Loss function}} + \underbrace{\lambda \|w\|_2^2}_{\text{regularization}}$$

The regularization term prevents overfitting in the solution vector, whereas the loss function penalizes misclassifications. A random forest is a group of trees that may predict future outcomes; each tree's value is based on the size of an independent random vector sample [3]. Given the abundance of evidence suggesting that "a cluster of uncorrelated trees (models) operating on a combined basis outperform individual constituent model," the Random Forest Algorithm becomes operational. (D. Decision Trees) The decision-making process in machine learning's decision tree analysis follows the structure of a tree, including its branches and nodes. Choice trees can be based on classification or regression. Classification trees are used to identify and categorize problems, while Predicting a variable using a regression tree is like using a stock price. There is a goal at every node and a choice at every branch. Once the decision tree is set up, the error rate is calculated to prune it. [16] E. Trees with Gradient Boosting The area of machine learning known as gradient boosting has gained widespread acceptance. Along with regression difficulties, loss function may also handle classification challenges. [2] in F. Deep Learning Using the idea of neural networks, Deep Learning is essentially a subfield of machine learning. It employs the algorithms present in the neural network's other layers in response to the data provided as input. The equation $h_{t+1} = \delta(W h_t + b)$ represents a non-linear connection between the variables h_t and h_{t+1} , where δ is the activation function, W is the basic model parameters, and b is the value of h_t . Machine learning models for classification and regression issues are a collection of G. Generalized Linear Models. This diagram shows that the connection between the two variables, "independent" and "dependent," is linear.

Though most variables do not adhere to a normal distribution, the ordinary linear model performs well for those that do. Non-normal distributions are well-suited to GLM. The remaining article is organized as follows: Section II provides a general overview of literature research linked to machine learning algorithms; Section III lays out the research methods and data; Section IV is devoted to the analysis of the results; and Section V concludes the study.

II. REVIEW OF LITERATURE

The sentence compares four prediction models for the movement of share prices and finds that Random Forest (RF) performs better overall than the other three models, including Artificial Neural Network (ANN), Support Vector Machine (SVM), and naïve Bayes [15]. Nevertheless, a research that used several artificial neural networks such as MLP, CNN, and RNN to forecast stock price movements in the S&P 500 index found that convolutional neural networks (CNN) outperformed the other designs in predicting stock price changes [7]. Utilizing convolutional neural networks (CNNs) has resulted in a 65% accuracy rate for predicting prices one month from now and a 60% accuracy rate for predicting prices one week from now. Logistic regression and Support Vector Machines (SVM) have shown marginally superior performance, as previously reported [6]. According to research that compared two types of artificial neural networks (ANNs)—feedforward and recurrent—to predict the short-term values of ten NYSE equities, feedforward performed better than long short-term (LSTM) [10]. In order to forecast the movement of stock prices, the research has employed a Support Vector Regression (SVR) predictive model. We also compare this model to the Efficient Market Hypothesis's (EMH) Random Walk model. According to their research, the Support Vector Model is a powerful forecasting tool [9]. Investors may employ Artificial Intelligence (AI) and Machine Learning approaches to minimize multicollinearity difficulties and estimate share values, according to another research that has shed light on the fundamentals of machine learning [17]. Additionally,

one research examined the closing prices of exchange-traded funds using four different types of machine learning algorithms. Support vector regression in neural networks is ranked second, while random forest approaches are ranked third, all pointing to the fact that deep learning is the best method for predicting [14]. A hybrid strategy for predicting stock prices has been suggested in this research. They have optimized the stock price prediction using a hybrid strategy that combines the Support Vector Regression model with the Hodrick-Prescott filter. It has been determined showing the suggested approach outperforms the competition when it comes to forecasting stock prices [12]. In order to extract features from financial periods series sample data and forecast stock prices, another research on the idea has proposed a multifilter neural network (MFNN). The Chinese stock market index CSI300 was used to test this concept. Compared to convolutional and LSTM networks, as well as to statistical models and more conventional machine learning approaches, they find that their network is more profitable, accurate, and stable [11]. Even for shares that are difficult to arbitrage, the authors of this research claim that deep learning signals assist investors turn a profit. Nevertheless, when dealing with big trade volumes and considerable trading costs, machine learning performance suffers. Additionally, in the long term, the investor may make more money with the deep learning signals [4]. Machine learning models outperform statistical and economic models, according to the machine learning study. It is also claimed that ensemble machine learning models outperform individual machine learning models. The efficacy of ensemble machine learning models based on trees is examined in this work. For the purpose of predicting the future movement of the stock price, we compare the following classifiers: Random Forest (RF), XGBoostClassifier (XG), BaggingClassifier (BC), AdaBoost Classifier (Ada), Extra Trees Classifier (ET), and Voting Classifier (VC). The tree-based ML algorithms are ranked using the Kendall W test of concordance. The top model for predicting stock price movement was found to be the Extra Trees Classifier (ET) [2]. On the other hand, there is

research that focuses on time complexity; the authors of that study compute the algorithms' time complexities and compare pre- and post-implementation data from the Indian financial market [18].

III. RESEARCH PROBLEM

The purpose of this paper is twofold: first, to show how different machine learning algorithms can be used to predict the direction of the market index (Sensex); second, to compare and contrast the results of support vector machines (SVMs) and random forests (RFs) in order to forecast the direction of stock market movement; and finally, to use the selected model to determine the predicted values.

VI. DATA&RESEARCH METHODOLOGY

Information pertaining to the daily closing prices of the BSE SENSEX from April 1, 2015, to March 31, 2020, is presented in this article. The information is collected from the BSE website. There will be many steps to the data analysis process (Fig. 1):

- Data loading
- General data pre-processing
- The following steps are involved:
 - encoding the data
 - handling text columns
 - preparing the data for correlation calculations
 - removing columns with constant values
 - sampling data based on characteristics
 - creating an actual correlation matrix
 - defining the result name.

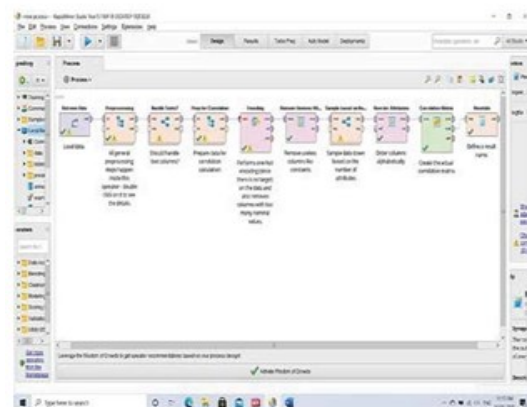


Fig.1.ProcessofApplicationofMachineLearning Algorithms

A. Visualizing Data BSE Sensex closing prices seem to be growing with some volatility over the specified term (2015–2020), as shown in Fig2. However, a decrease in price is noticeable in January and February 2020 as a result of COVID-19 and sluggish global markets.

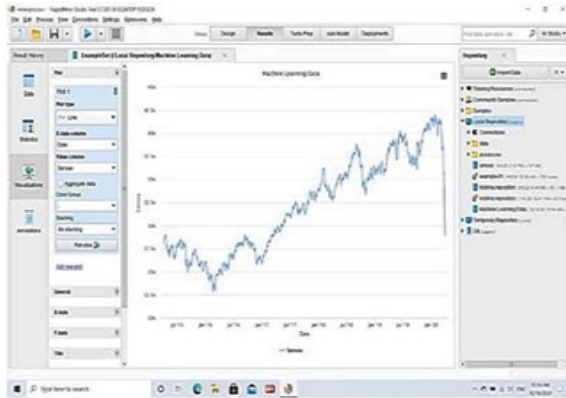


Fig.2.Data Visualization

V. APPLICATION OF MACHINE LEARNING ALGORITHMS

The BSE SENSEX closing prices for a specific period from 2015 to 2020 are predicted using the following algorithms (Fig3) • Deep Learning • Generalized Linear Model Gradient Boosted Trees, Decision Tree, Random Forest, and Support Vector Machines



Fig.3.SelectionofMachineLearningAlgorithms

Description of Relative Error & Standard Deviation of Algorithms.

Table1

Model	Relative Error	Standard Deviation
Generalized Linear Model	4.0%	0.1%
Deep Learning	3.2%	0.3%
Decision Tree	0.8%	0.1%
Random Forest	6.7%	0.1%
Gradient Boosted Trees	0.8%	0.0%
Support Vector Machines	2.3%	0.2%

When it came time to forecast the values of the Sensex, Decision Tree and Gradient Boosted Trees had the best relative errors at 0.8% (Table1), while Gradient Boosted Trees often had the lowest standard deviations at 0.0 %.A. A Linear Predictive Model Root mean squared error was 1634.094 and deviation was 132.703 when GLM was conducted. The GLM Prediction Chart (Fig. 5) shows the reported outliers and the deviations of the projected values from the actual values. The 4.0% relative error is in line with what GLM predicted.



Fig.4.GeneralizedLinear Model

Fig.7.DeepLearningModel-PredictionChart



Fig.5.GeneralizedLinearModel-PredictionChart

B. Learning Methods The closing prices were predicted by the Deep Learning model with a relative error of 3.2%; nevertheless, the projected values were consistently 0.3% off from the actual values (Fig7). A root-mean-squared error of 1296.5 and a deviation of 112.4 were given by the model's predicted values.



Fig.6.DeepLearning Model



The branches and nodes of the Decision Tree model are shown in Figure 8. The stated value predictions had a standard deviation of 0.1% and an error of 0.8%, the lowest of all the stall models.D. WildForest A Random Forest model was run with a standard deviation of 0.1% and a maximum relative error of 6.7%. The prediction chart (Fig. 9) also shows outliers.



Fig.8.RandomForestModel



Fig.9.RandomForestModel-PredictionChart

Grade-Boosted Model (E) In terms of predicting BSE closing prices, this model was the most effective, with a reported error of 0.8% and a minimal standard deviation of 0.0%. The reported root-mean-squared error is 372.4 with a variation of 43.6. After minimizing root-mean-squared error, the objective is to minimize relative error. Figure 11 shows the prediction chart, which clearly shows the value that is most closely matched to the actual value.



Fig.10.GradientBoostedTreesModel



Fig.11.GradientBoostedModel-Prediction Chart

F. Vector Machine Support The SVM Model is said to have a relative error of 2.3% with a standard deviation of 0.2% when used with the Kernel Model. The reported root mean squared error is 1239.8 with a deviation of 109.3. There are few outliers, but the prediction chart shows that the results are similar to the actual values (Fig13).



Fig.12.SupportVectorMachine Model

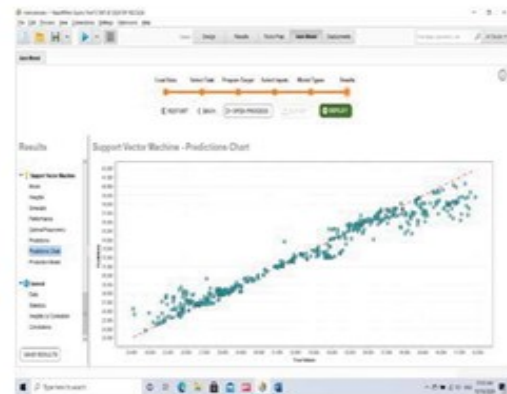


Fig.13.SupportVectorMachine-Prediction Chart

VI. DEPLOYMENT OF GRADIENT BOOSTED TREES

The deployment of Gradient Boosted Trees is seen in Figure 14. As the machine uses this prediction model, it becomes evident that tree-like models are used to estimate the values.

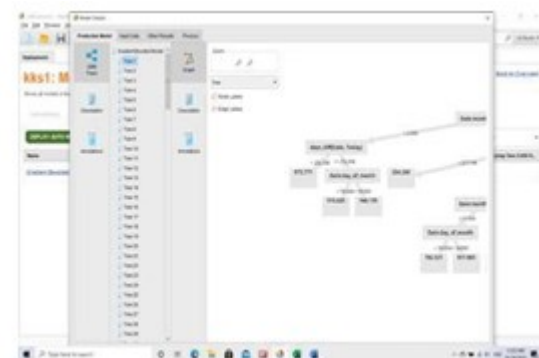


Fig.14.DeploymentofGradientBoostedTrees

After the Gradient Boosted Trees model was put into action, the results shown in Figure 15 show that the mean absolute error was 124.873276, the root mean squared log error was 0.00641, and the mean residual deviation was 43422.95. Additionally, it displays that there are a total of 150 trees.



Fig.15.Modelsummary-GradientBoosted Trees

The output of the Gradient Boosted Trees model's prediction is shown in Fig. 16. The date is shown in Column I, the Sensex values in Column II, and the predicted value in Column III of the figure. Between April 10 and April 14, 2015, the BSE Sensex closed at 28516.59, 28707.75, 28885.21, and 28879.38, whereas the corresponding projected values were 28259.885, 28238.69, 28247.723, and 28312.072.

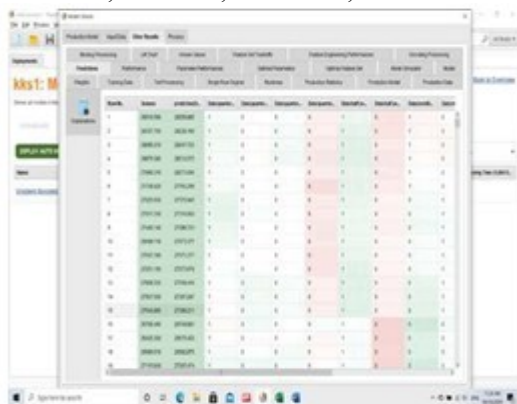


Fig.16.PredictedValuesasperGradientBoostedTrees

VII. RESULT ANALYSIS

The BSE Sensex was forecasted using six models in this study: Support Vector Machine, Gradient

Boosted Trees, Generalized Linear Model, Deep Learning Decision Tree, Random Forest. It was used to forecast the BSE Sensex because, as previously mentioned, Gradient Boosted Trees has a lower relative error and standard deviation ($\pm 0.0\%$ compared to the other five models). With the lowest mean absolute error of 124.873276 and the lowest residual deviation of 43422.95 (Fig. 16), the Gradient Boosted Model was chosen for deployment. It was discovered that the closest prediction is supplied exclusively by this model after the deployment of GradientBoostedModel. Some models have strayed a little farther than others.

VIII. CONCLUSION

We urgently want a durable stock market prediction pricing model. When formulating plans for the financial markets, it is difficult yet crucial. In this regard, a plethora of statistical, linear programming, and machine learning technologies are used. The BSE Sensex closing prices from April 2015 to March 31, 2020 were used to estimate future price movement in this research. The six models were Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine.

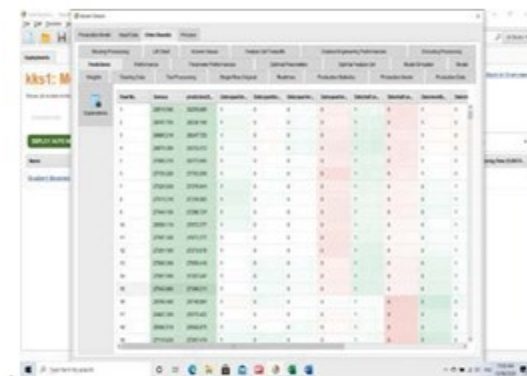


Fig.17.PredictedvaluesasperGradientBoostedTrees

On the first day, the Gradient Boosted Model predicted 28,259.885, which was 256.705 basis points below the actual 28,516.590. But the Deep Learning Model predicted 26598.2 bps and the Machine Learning Model 24643.4 bps for the same day. The Machine Learning Model has a latency of 3873.19 bps, in contrast to the Deep Learning

Model's 1918.39 bps. As a result, the GRADientBoostedModel was determined to have the best performance in forecasting the BSE Sensex price.

IX. REFERENCES

[1] Ahangar RG, Yahyazadehfar M, Pournaghshband H (2010) The comparison of methods artificial neural network with linear regression using specific variables for prediction of stock price in Tehran stock exchange. *Int J Comp Sci Informat Sec* 7(2):38–46

[2] Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement [Abstract]. *Information*, 11(6), 332. doi:10.3390/info11060332

[3] Avramov, Doron and Cheng, Si and Metzker, Lior, Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability (August 18, 2020). Available SSRN: <https://ssrn.com/abstract=3450322> or <http://dx.doi.org/10.2139/ssrn.3450322>

[4] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

[5] Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbour algorithm for stock market indices prediction. *Expert Systems with Applications*, 80, 340–355.

[6] Dingli, A., & Fournier, K. S. (2017). Financial Time Series Forecasting – A Deep Learning Approach. *International Journal of Machine Learning and Computing*, 7(5), 118–122. <https://doi.org/10.18178/ijmlc.2017.7.5.632>

[7] Di Persio, Luca & Honchar, Oleksandr. (2016). Artificial neural networks architectures for stock price prediction: Comparisons and applications. 10. 403–413.

[8] Fama, E. F. (1995). Random walks in stock market prices. *Financial analysts journal*, 51(1), 75–80.

[9] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of Finance and Data*

Science, 4(3), 183–201. <https://doi.org/10.1016/j.jfds.2018.04.003>

[10] K. Khare, O. Darekar, P. Gupta and V. Z. Attar, (2017) "Short term stock price prediction using deep learning," 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017, pp. 482–486, doi: 10.1109/RTEICT.2017.8256643.

[11] Long, W., Lu, Z., & Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, 164, 163–173. <https://doi.org/10.1016/j.knsys.2018.10.034>

[12] M. Ouahilal, M. El Mohajir, M. Chahhou and B. E. El Mohajir, (2016) "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, 2016, pp. 290–294, doi:10.1109/CIST.2016.7805059

[13] Malkiel, B. G. (1989). Efficient market hypothesis. In *Finance* (pp. 127–134). Palgrave Macmillan, London.

[14] Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*,

[15] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications*, 42(1), 259–268.