2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)
Department of Electrical & Electronics Engineering, Galgotias College of Engineering and Technology, Gr. Noida, India

2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) | 978-1-7281-7741-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICACITE51222.2021.9404642

Comparative Analysis of Machine Learning Algorithm to Forecast Indian Stock Market

Mr. Sachin Rohatgi
Assistant Professor
Amity Global Business School, Noida,
mr.rohatgi@gmail.com

Dr. Krishna Kumar Singh
Associate Professor
Symbiosis Centre for Information Technology, Pune,
krishnakumar@scit.edu

Deepmala Jasuja
Assistant Professor
Amity Global Business School, Noida,
deepmalajasuja13@gmail.com

Abstract: **Complexity and diversity of the stock market has always attracted the researchers since ages to find out the ways to predict the future movement. However, the volatility across the globe in different stock markets makes this task more difficult. Statistical methods and modelling are effective but can't address the vast range of problems encountered in the prediction of movement of the stock market. The traditional methods used were not able to address and give solution to the complex problems prevailing in the stock market. The machine learning and artificial intelligence tools are used to solve the complex situations and problems of Big Data with ease. In this paper authors propose to use six different algorithms i.e., Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees and Support Vector Machine and identify model which predict near to actuals. These algorithms were applied on the BSE index data from April 2015 to 31st March 2020 and the model with least relative error is identified. Amongst all the models applied, Gradient Boosted Trees is chosen to be the efficient one, as it has least relative error and standard deviation. Further Gradient Boosted Trees is used to forecast the results.**

Keywords: BSE SENSEX, Big Data, Predictive algorithms, Artificial Intelligence, Random Forest, SVM, Gradient Boosted Trees, Machine Learning

I. INTRODUCTION

These days, forecasting the movement of prices in the stock market has become a challenge due to unforeseen events and its dynamic nature." Predicting the unpredictable" has now become the Centre of Indian Stock Market movement predictions. Stock Market being called as the barometer of the economy which drives the economic growth is random, noisy, unpredictable, non-linear and deterministic chaotic system [1] There are many factors which drives the stock market including political, cultural and financial events along with investors' sentiments, foreign institutional investors decision which make the prices of financial time series prediction challenging. Conventionally, there are two approaches to predict the stock market behaviour namely Chartist approach (Technical analysis) & Theory of Intrinsic Value (Fundamental Analysis). Majorly, these theories contradict Efficient Market Hypothesis which follows random walk. Logic of Random Walk is based on the premise that successive price fluctuations are not dependent on yesterday's price and market condition [8]. Following on the same lines, EMH theory is based on the idea that stock prices completely reflects all available and current information and even the uninformed investors can achieve reasonable rate of return as compared to financial experts [13]. Recently, Researchers and Academicians all over the world are applying machine learning algorithms in various fields. Machine learning is predominantly being used in the forecasting of stock prices across stock markets due to the techniques it applies to generalize the patterns. This article applies six prominently used machine learning algorithms to predict BSE SENSEX closing prices.

A. Support Vector Machine

Support Vector machine (SVM) is a classic and widely used machine learning algorithm featured by capacity control of the decision function, application of Kernel function and sparsity of solution [5]. It can very well be utilized for classification and regression assignments. Support Vectors the data points close to hyperplane and can leave an impact on the site and orientation of hyperplane.

B. SVM Algorithm

In SVM Algorithm, we attempt to amplify the margin of data points and hyperplane. There is a loss function also referred to as hinge function which is applied to maximise the margin.

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

Loss function will penalize misclassifications and regularization term will avoid overfitting in the solution vector.

C. Random Forest

Random Forests are tree predictors bunch, wherein value of individual tree is dependent on the magnitude of random vector independent sample [3]. Random Forest Algorithm works on the plethora that," A cluster of uncorrelated trees (models) operating on a combined basis outperform individual constituent model".

D. Decision Trees

In decision tree analysis of machine learning, decision is taken through the way of tree structure, its branches and its nodes. Decision trees can be classification or regression based. Classification trees are to classify a problem and find out the solution whereas Regression trees predict a variable simply like a stock price. Every node presents a target and each branch presenting a decision. After the establishment of decision tree, error rate is computed to prune the decision tree [16]

E. Gradient Boosted Trees

Gradient Boosted is one of the commonly accepted field of machine learning. Using loss function, classification problems can also be catered to, along with regression problems [2]

F. Deep Learning

Deep Learning can be primarily called as a field of machine learning applying the concept of Neural Networks. It

uses the layers of neural network algorithms at other layers given the raw input data. It is a non-linear relationship between two variables ht and ht+1 and is expressed as

$$ht+1 = \delta (Wht+ b)$$

Where, δ = Activation Function, W & b are model parameters.

G. Generalized Linear Model

Generalized Learning models are bunch of machine learning models for classification and regression problems. It represents a linear relationship between independent and dependent variable. Ordinary linear model works best for normal distribution, but practically most variables follow arbitrary distributions. GLM suits for non-normal distributions.

The remaining article is assembled as follows, Section II broadly covers related literature research to Machine Learning Algorithms followed by Section III constituting data and research methodology and Section IV focussed on Result Analysis with Section V concluding the research study.

II. Review of Literature

Compares four prediction models, Artificial Neural Network (ANN), Support Vector Machine (SVM), Random forest (RF) and naïve-Bayes for the share price movement and concludes that the Random forest (RF) outperform all the other three prediction models on overall performance [15].However, a study used different artificial neural networks like MLP, CNN & RNN to predict the stock price movements in S&P500 index and concluded that the convolutional neural networks (CNN) is better than the other architectures in the prediction of stock price movement [7].The use of convolutional neural networks (CNN) have given a 65% accuracy while forecasting next month price prediction and 60% accuracy while forecasting next week price prediction. Further it is confirmed that logistic regression and Support Vector Machines (SVM) have achieved slightly better result [6]. The study which has used the two different ANN i.e. feed forward neural network and recurrent neural networks to forecast ephemeral prices of 10 stocks on NYSE have suggested that the feed forwards performs better than the long short term (LSTM) in forecasting the share prices [10].The study have used Support Vector Regression (SVR) predictive model in the prediction of stock price movement. Further this model is compared with the Random Walk model given by Efficient Market hypothesis (EMH). Their study concludes that the Support vector Model has the predictive power[9].There is another study which has given an insight into the basic concepts of machine learning and offer a solution of how an investor can use Artificial Intelligence (AI) and Machine learning techniques to project the share prices by minimizing the multicollinearity problems [17]. Further there was study in which authors have studied the closing price data of exchange traded fund through four models of machine learning algorithms. They conclude that the deep learning method is better in forecasting than the other methods and they have put support vector regression method in neural networks at second rank and random forest methods at third rank [14]. This study has proposed a hybrid approach for the stock price prediction. In this hybrid approach they have joined Support Vector Regression model and Hodrick-Prescott filter for optimizing the stock price prediction. They have concluded that the proposed model is more accurate in predicting the stock prices [12]. There is another study done on the concept which has proposes a multifilter neural network (MFNN) for feature extraction on financial times series sample data and stock price prediction. This model was applied on Chinese stock market index CSI300.They conclude that their network performs better in terms of profitability, accuracy and stability as compared to the traditional machine learning models, statistical models and convolutional and LSTM networks [11]. In this article authors states that the deep learning signals helps investor to give profitability in the case of difficult to arbitrage shares also. However, the machine learning performance reduces in the presence of substantial trading cost due to higher trading volumes. Further the deep learning signals are more profitable to the investor in the long run [4]. The Study on machine learning states that Machine learning models perform better than the statistical models and econometric models. Further it is stated that the ensemble machine learning models are better than the single machine learning models. In this study the effectiveness of tree -based ensemble machine learning models Random forest (RF), XGBoost Classifier (XG), Bagging Classifier (BC), AdaBoost Classifier (Ada), Extra Trees Classifier (ET), and Voting Classifier (VC) are compared in forecasting the stock price movement. The Kendall W test of concordance is used to rank the performance of the tree-based ML algorithms. It was concluded that the Extra Trees Classifier (ET) is the best model in the prediction of stock price movement [2]. However, there is another study which emphasis on time complexities and authors in their study calculates the time complexities of the algorithms and stimulates the Indian financial market data before and after the implementation of the algorithms [18].

III. Research Problem

This paper attempts to demonstrate and verify the forecast of market index (Sensex) direction by applying various machine learning algorithms and secondly, to contrast the output of SVM and random forest to foresee the direction of the stock market motion and deploying the selected model to find out the predicted values.

IV. Data & Research Methodology

Research Data in this paper pertains to daily closing prices of BSE SENSEX from 1st April 2015 to 31st March 2020.Data is retrieved from official website of BSE. Data Analysis will be in following phases (Fig1):

- Loading the data
- General Pre-processing of Data
- Handling of Text Columns
- Preparing data for correlation calculation
- Encoding the data
- Removing Columns with constant
- Sample data down based on attributes
- Creating actual correlation matrix
- Define a result name.

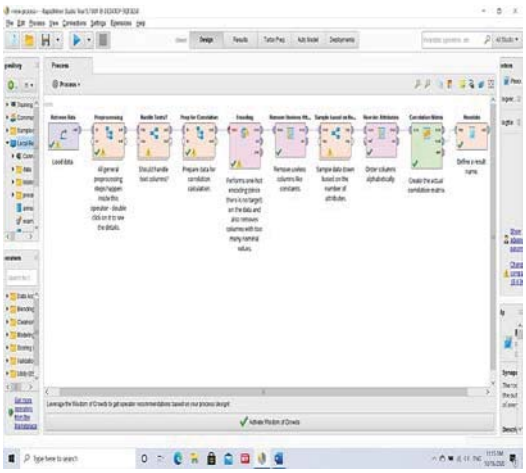

Fig.1. Process of Application of Machine Learning Algorithms

A. Data Visualization

Fig2 depicts the stock price movement for 5 years (2015-2020), BSE Sensex closing prices seems to be rising with some volatility during the said duration, but a dip in prices is visible in January & February 2020 due to Covid19 and weakened global markets.

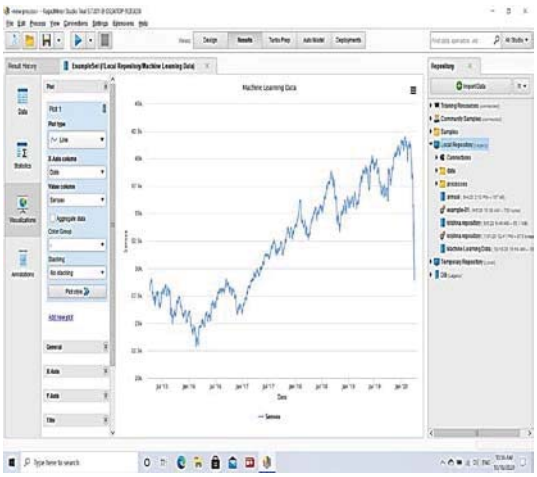

Fig.2. Data Visualization

V. APPLICATION OF MACHINE LEARNING ALGORITHMS

Following algorithms are applied to predict the BSE SENSEX closing prices of a given duration of 2015-2020 (Fig3)

- Generalized Linear Model
- Deep Learning
- Decision Tree
- Random Forest
- Gradient Boosted Trees
- Support Vector Machines

Fig.3. Selection of Machine Learning Algorithms

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%

Following with algorithms run, Decision Tree and Gradient Boosted Trees reported the lowest relative error 0.8% (Table1) in predicting the prices of Sensex, but Gradient Boosted Trees tends to report the lowest standard deviation (0.0%)

A. Generalized Linear Model

GLM was run with root mean squared error of 1634.094 along deviation of 132.703. GLM Prediction Chart (Fig5) depicts deviations of predicted values from the actual values and outliers being reported in the chart. This is consistent with GLM prediction with a relative error of 4.0%.

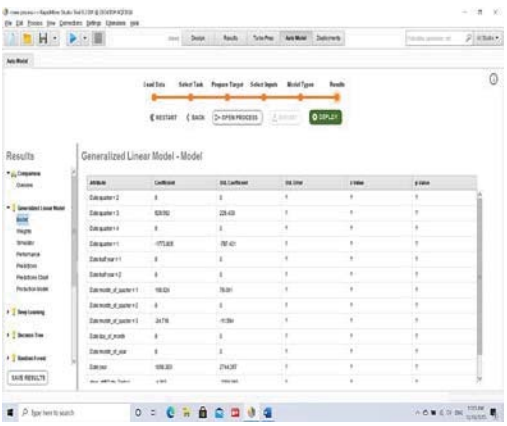

Fig.4. Generalized Linear Model

Fig.5. Generalized Linear Model-Prediction Chart

B. Deep Learning

Deep Learning model predicted the closing prices with a relative error of 3.2% but the predicted values are relatively deviated from actual values as consistent with standard deviation of 0.3% (Fig7). Model prediction values reported root mean squared error of 1296.5 with a deviation of 112.4.

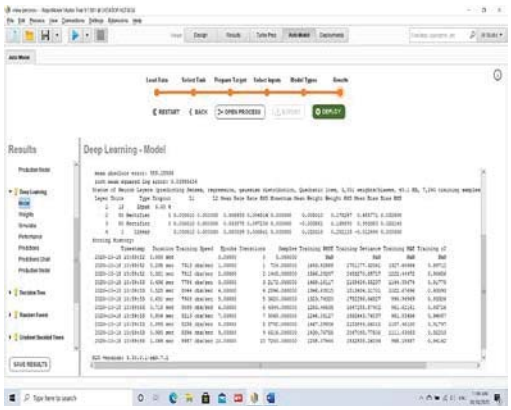

Fig.6. Deep Learning Model

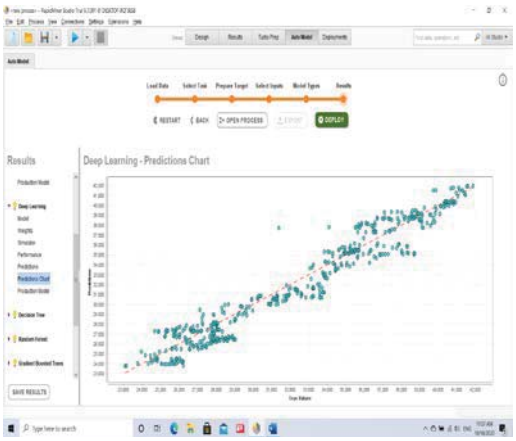

Fig.7. Deep Learning Model-Prediction Chart

C. Decision Tree

Decision Tree model is presented (Fig8) with its branches and nodes. Prediction of values are reported with an error of 0.8% which is relatively lowest amongst all models along with standard deviation of 0.1%.

D. Random Forest

Random Forest model performed with the highest relative error of 6.7% and standard deviation of 0.1%. Outliers are also visible in the prediction chart (Fig9)

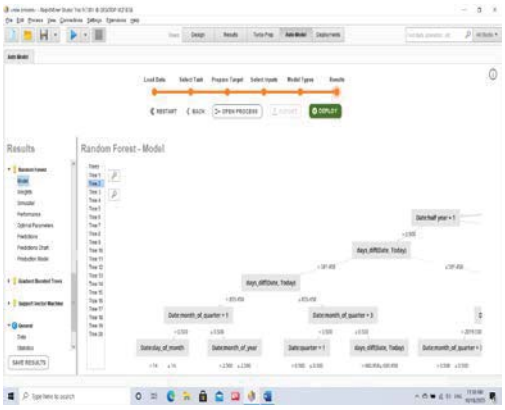

Fig.8. Random Forest Model

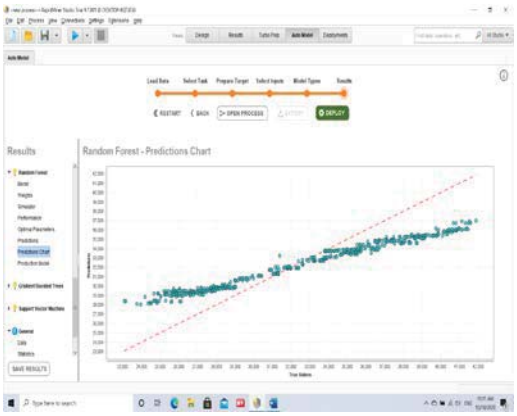

Fig.9. Random Forest Model-Prediction Chart

E. Gradient Boosted Model

This model proved to be the most efficient to predict the closing prices of BSE as it reported the lowest error of 0.8% and minimum standard deviation of 0.0%. Root mean squared error is reported to be the lowest 372.4 with deviation of 43.6. Lowest root mean squared error is followed by least relative error. Prediction chart clearly depicts the closest match between actual and predicted values (Fig11)

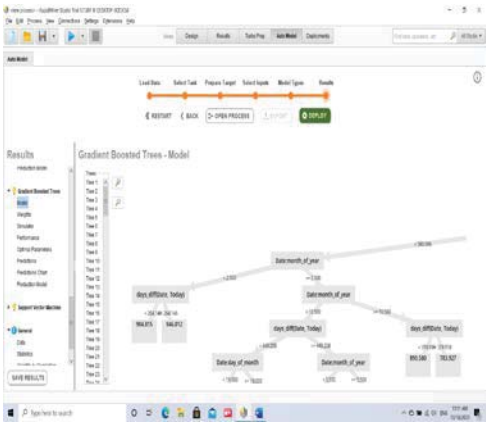

Fig.10. Gradient Boosted Trees Model

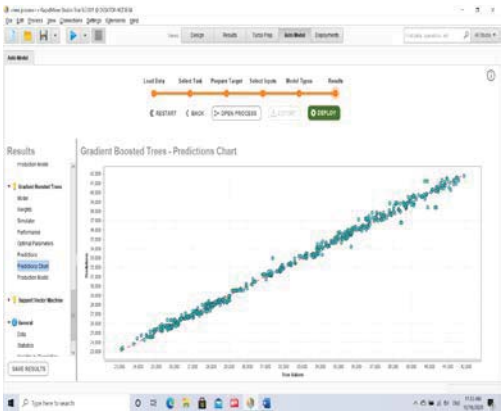

Fig.11. Gradient Boosted Model-Prediction Chart

F. *Support Vector Machine*

SVM Model is applied with Kernel Model and relative error is said to be 2.3% with standard deviation of 0.2%. Root mean squared error is reported to be 1239.8 with a deviation of 109.3. Prediction chart clearly states that values are close to actual values but there are some outliers (Fig13)

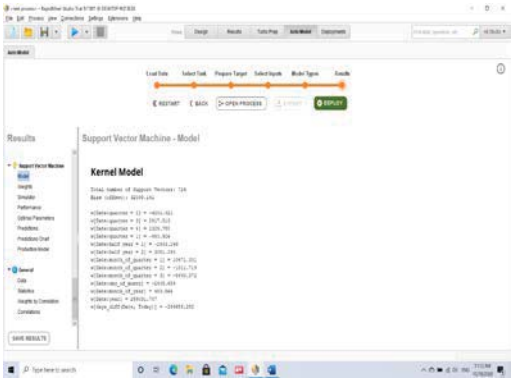

Fig.12. Support Vector Machine Model

Fig.13. Support Vector Machine-Prediction Chart

VI. DEPLOYMENT OF GRADIENT BOOSTED TREES:

Gradient Boosted Trees is deployed as shown in the Fig14.The machine is deploying this prediction model and it is clear that the values are estimated by making tree like models.

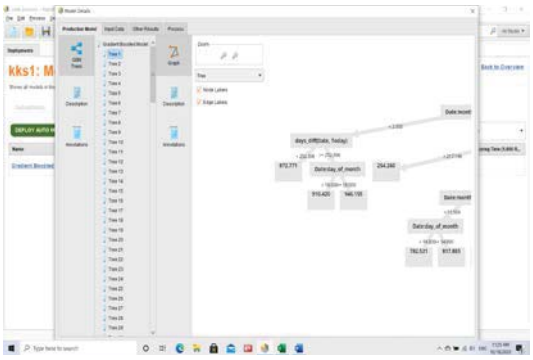

Fig.14. Deployment of Gradient Boosted Trees

After the Gradient Boosted Trees model deployed its model summary Fig15 has given the mean residual deviance 43422.95, mean absolute error 124.873276, root mean squared log error 0. 00641.It also shows the total number of trees are 150.

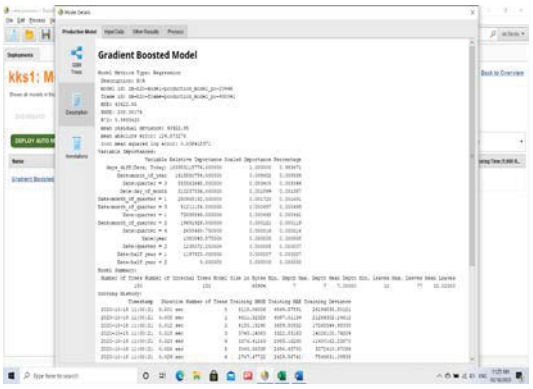

Fig.15. Model summary- Gradient Boosted Trees

The following Fig16 shows the final result of the prediction given by Gradient Boosted Trees model. In the figure, Column I denote the date and Column II denotes Sensex values and the third column denotes the prediction value. The BSE Sensex values from 10th April 2015 to 14th April 2015 were 28516.59, 28707.75, 28885.21, 28879.38 respectively and the predicted value for these dates were 28259.885, 28238.69, 28247.723,28312.072 respectively.

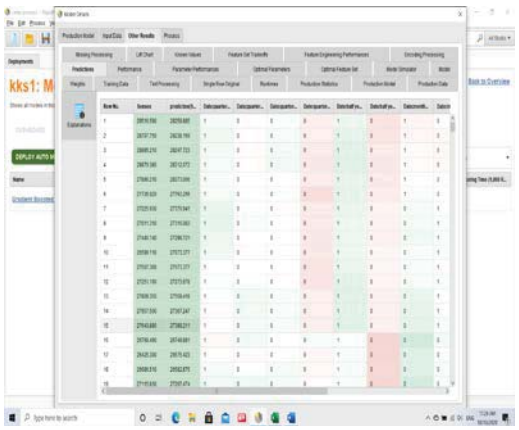

Fig.16. Predicted Values as per Gradient Boosted Trees

VII. RESULT ANALYSIS

In this study the 6 models were used to predict the BSE Sensex namely., Generalized Linear Model, Deep Learning,

Decision Tree, Random Forest, Gradient Boosted Trees and Support Vector Machine. As it is already explained that because the relative error and standard deviation of Gradient Boosted Trees is least i.e., 0.8% and ± 0.0 % as compared to the other 5 models therefore it was deployed to predict the BSE Sensex. As the Gradient Boosted Model has the lowest residual deviance 43422.95 (Fig16) and lowest mean absolute error 124.873276 this model was deployed. After the deployment of Gradient Boosted Model, it was found that the nearest prediction is given by this model only. Other models have deviated relatively higher.

VIII. CONCLUSION

Finding a sustainable predictive price model for the stock market is the need of the hour. It is challenging though but important while framing the financial market strategies. Many statistical tools, liner programming tools and machine learning tools are used in this direction. In this study 6 models i.e., Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees and Support Vector Machine were used on the BSE Sensex closing prices from April 2015 till 31st March 2020 to predict the price movement.

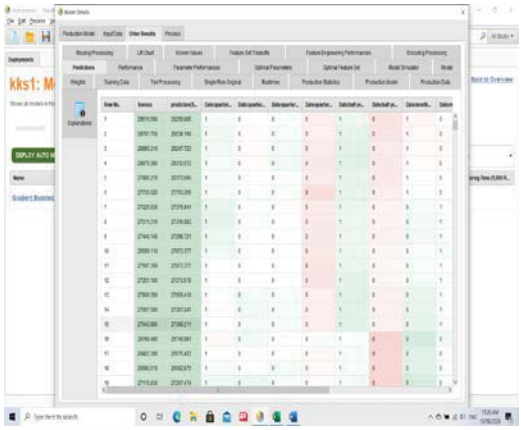

Fig.17. Predicted values as per Gradient Boosted Trees

The prediction given by Gradient Boosted Model on the 1st day was 28259.885, which was lagging from the actual 28,516.590 by 256.705 bps. However, the prediction for the same date given by the Deep Learning Model and Machine learning Model were 26598.2 bps and 24643.4 bps respectively. Deep Learning Model gives a lag of 1918.39 bps whereas the Machine Learning Model has a lag of 3873.19 bps. Therefore, it was concluded that Gradient Boosted Model has outperformed all the other models in predicting 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 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.knosys.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, 26(4), 164–174. https://doi.org/10.1002/isaf.1459

[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. https://doi.org/10.1016/j.eswa.2014.07.040

[16] Qin, Q., Wang, Q. G., Li, J., & Ge, S. S. (2013). Linear and nonlinear trading models with gradient boosted random forests and application to Singapore stock market

[17] Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine Learning for Stock Selection. Financial Analysts Journal, 75(3),70-88. ttps://doi.org/10.1080/0015198x.2019.1596678

[18] Singh K.K., Rohatgi S. (2021) Green Algorithmic Impact of Computing on Indian Financial Market. In: Goyal D., Bălaş V.E., Mukherjee A., Hugo C. de Albuquerque V., Gupta A.K. (eds) Information Management and Machine Intelligence. ICIMMI 2019. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-4936-6_12