Stock Prediction Using Machine Learning Algorithms

Pahul Preet Singh Kohli, Seerat Zargar, Shriya Arora and Parimal Gupta

Abstract Market systems are so complex that they overwhelm the ability of any individual to predict. But it is crucial for the investors to predict stock market price to generate notable profit. The ultimate aim of this project is to predict the behavior of Bombay Stock Exchange (BSE). We have taken into factors such as Commodity Prices (crude oil, gold, silver), Market History, and Foreign exchange rate (FEX) that influence the stock trend, as input attributes for various machine learning models to predict the behavior of Bombay Stock Exchange (BSE). The performances of the models are then compared against other benchmarks. A structured relationship was also determined among the different attributes used. The gold price attribute was found to have the highest positive correlation with market performance. The AdaBoost algorithm performed best as compared to other techniques.

Keywords Stock prediction · BSE index · Machine learning algorithms · Stock prediction classification

1 Introduction

Historically, high market prices often make the investors despondent from investing, while low market prices represent an opportunity. Predicting stock market price, therefore, becomes imperative for investors to yield a significant profit. Though predicting the financial markets and the stock movements is onerous [1], many

P. P. S. Kohli (✉) · S. Zargar · S. Arora · P. Gupta
Bharati Vidyapeeth's College of Engineering, New Delhi, India
e-mail: pahulpreet86@gmail.com

S. Zargar
e-mail: seeratzargar1996@gmail.com

S. Arora
e-mail: shriyaarora080696@gmail.com

P. Gupta
e-mail: guptaparimal1996@gmail.com

H. Malik et al. (eds.), *Applications of Artificial Intelligence Techniques in Engineering*, Advances in Intelligent Systems and Computing 698,
https://doi.org/10.1007/978-981-13-1819-1_38

researchers from different fields have scrutinized and used many algorithms and different combination of attributes to predict the market movements. But these algorithms are all on the basis of stock price itself which has random property.

In this project, we have proposed the use of macroeconomic factors such as commodity price, market history, and foreign exchange rate to predict the Bombay Stock exchange (BSE). These are some of the vital factors [2, 3] that predict whether BSE will increase or decrease on a particular day. The project is implemented in Ipython Notebook.

The rest of the paper is organized as follows. Section 2 gives a brief overview of machine learning algorithms. Section 3 describes the method that is proposed for the implementation of models on the stock market data. Section 4 presents the simulation and test results of the paper. Section 5 concludes this paper.

2 Methodology

In this research, four machine learning algorithms are used and compared on the basis of their training accuracy. These models are as follows.

2.1 Support Vector Machines

SVM algorithm is based on statistical learning hypothesis. Both regression and classification can be done using this supervised machine learning algorithm. SVM can also be used for outlier's detection [2, 3]. The data points are plotted in the n-dimensional space. SVM performs classification by hyperplane which is a boundary constructed over the dataset. The hyperplane separates the cases of different class labels. This hyperplane is constructed in the multidimensional space as given in Fig. 1. To find the right hyperplane in the case of classification, the distance between the nearest data points called support vectors and the decision boundary is maximized.

Hence, minimization of the norm of the vector w is needed; where w define the separating decision boundary. This is analogous to maximizing the margin between the two classes [4]. Considering the above figure, if we assume u to be some unspecified data point and was a vector which is perpendicular to the hyperplane, then the decision rule in SVM is given by:

$$\vec{\omega} \cdot \vec{u} + b \geq 0 \tag{1}$$

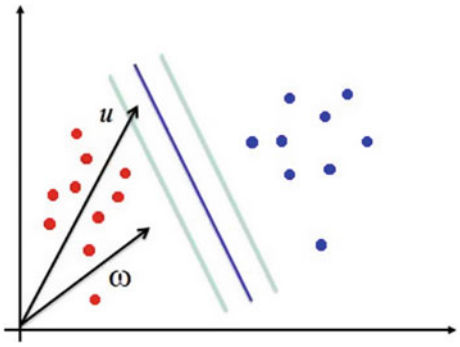

Fig. 1 Decision boundary in support vector machine

The maximization of width of the hyperplane is required for the expansion of the spread

$$W = \left[\frac{2}{||\omega||}\right] \tag{2}$$

$$W = max\left[\frac{2}{||\omega||}\right] \tag{3}$$

2.2 Random Forest

Random forest [5] is a machine learning algorithm which uses ensemble method for classification and regression problems. It uses the benefits of both Decision Trees and Bagging (Bootstrap Aggregation). Thus, it overcomes the problem of overfitting in decision trees. Bagging reduces the variance of high variance algorithms like decision trees (CART). Random forest is a group of unpruned classification and regression trees that are obtained from the subsamples of the training dataset.

In this algorithm, the process is as follows:

- The subsamples from the training dataset are created.
- CART model is applied to each of the subsamples to obtain a predicted output depending on the model.
- The bagging process is applied to get the ensemble of the predicted outputs of each model.
- Thus by this process, the predicted output has a lesser prediction error than each of the individual model.

2.3 Gradient Boosting

Gradient boosting is a kind of a boosting algorithm that trains many models sequentially. The loss function is gradually minimized by each new model. In Gradient Boosting, we assume a uniform distribution say D1 which is 1/*n* for all n observations. Then the algorithm progress according to the following steps:

- We assume an α_t.
- Calculate a weak classifier $h(t)$.
- Update the population distribution for the next step.
- The new population distribution is used again to find the next learner.
- Iterate Step 1–Step 4 until no hypothesis is found which can further improve the accuracy.
- Take the weighted average of the frontier using all the learners used till now.

2.4 Adaptive Boosting (AdaBoost)

Adaptive Boosting being [6] one of the first successful boosting algorithms can be used for both classification and regression processes. It involves ensemble of weak classifiers to build a strong one. Adaptive Boosting focuses on predicting current data set by giving equal weight to each attribute. If the prediction is incorrect, then it gives higher weight to the incorrect observation [7]. The iteration continues till almost no error is received. The basic prediction is made using a basic algorithm or decision stumps. The final output is predicted by calculating the weighted average of each weak classifier. Since the weights are calculated on the basis of the false predictions and changed accordingly, this algorithm is adaptive in nature, hence the name Adaboost.

3 Proposed Model

The idea implemented in this research is to use different attributes such as commodity prices, market history and foreign exchange rate as input attribute to predict the Bombay Stock Exchange (BSE) [8]. These input attributes were continuous numeric value of varied range, so in order to classify them, they were normalized as $[-1, 1]$. This is because all the input attributes can have either positive or negative values. The paper compares the outputs of all the four machine learning algorithms used. The output of each model is either 1 or -1 to describe either positive or negative impact on the market, respectively. All the factors are discussed separately (Fig. 2).

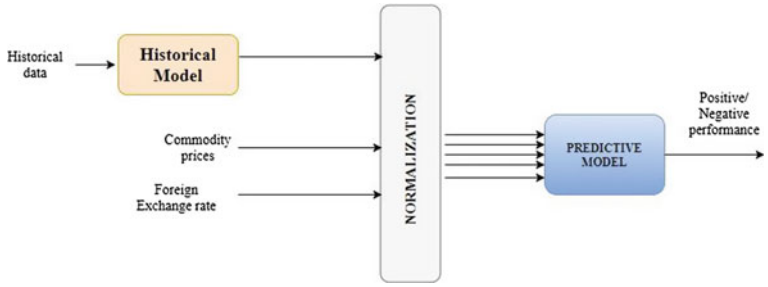

Fig. 2 Proposed model

3.1 Factors

The factors that affect the stock market prices are as follows:

- **Market History**
  A historical data of 2 years from January 2015 to December 2016 was collected. This historical data was not applied to the model directly but after applying a model based on historical data which was training on factors such as change in opening price, low price, change in volume of stocks, and high price.
- **Commodity Price**
  The price change of various different commodities (attributes) such as gold, silver, and crude oil has an impact on the overall change in the stock prices of the BSE.
- **Foreign Exchange**
  The foreign exchange rate change is known to play a vital role in the market performance. Foreign exchange rate between the INR (Indian National Rupee) and USD (United States dollar) is used as an input attribute in the model.

Based on the above factors, a total of 5 attributes are used as an input to the model and output as +1 (positive market) or −1 (negative market) is found.

3.2 Scope

A data spread of over 9 months from January 2017 to September 2017 is used in the research. All the data is taken from website https://www.investing.com (Fig. 3).

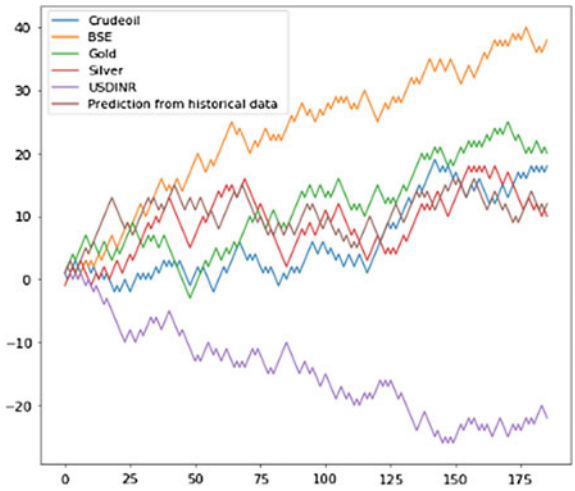

Fig. 3 Variation in input and output attributes

4 Simulation and Results

The data of different attributes was extracted and manipulated to convert it into a form that can be used as an input in model. The output of this model is +1 to represents positive market which describes the increase in BSE prices and −1 to represent negative market which describes decrease in BSE prices. The stimulation was performed using four different algorithms and accuracy for each was calculated from train/test split method. In this method, the data is split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on.

4.1 Implementation of Different Algorithms

The abovementioned machine learning algorithms were separately and successfully implemented on the dataset in python and accuracy for each model was calculated using different values of training and test dataset size. It observed an accuracy of 68.4% for 90% training data and 73% for 70% training data for Random Forest. For SVM, the accuracy was 78.95% for 90% training data and 73% for 70% training data. Accuracy for Gradient Boosting being equal to 78.95 for 90% training data and 73.2% for 70% training data. For AdaBoost, the accuracy is 78.95% for 90% training data and 77% for 70% training. Figure 4 shows the comparison between accuracy of different algorithms in the form of bar graph.

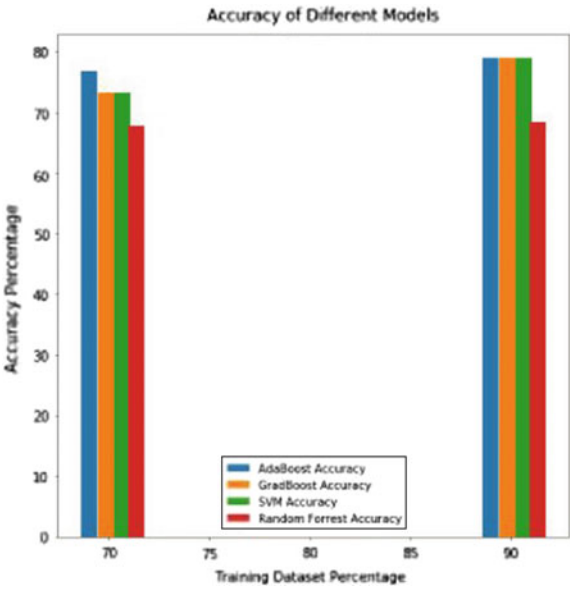

Fig. 4 Accuracy comparison of different machine learning algorithms

4.2 Dependency of Market Performance on the Attributes

Correlation and covariance between various attributes and market performance were calculated. The resulting dependency of the attributes in the form of correlation and covariance is described in Table 1.

It is evident from Table 1 that the attribute "Gold Rate" shows maximum positive correlation with market that is one variable decreases as the other variable decreases, or one variable increases while the other increases. Also the attribute "Foreign Exchange rate" shows maximum negative correlation with market that is when one variable increases as the other decreases and vice versa. The attribute that has the least impact was found to be "Silver Rate". The following Fig. 5 shows scatter plot between different attributes for spotting structured relationship between different attributes and market performance.

Table 1 Correlation and covariance between different attributes and market performance

S. No	Attributes	Correlation	Covariance
1	Oil rate	−0.7451	−102579.41
2	Gold rate	0.75293	16781.41
3	Silver rate	−0.2866	−114.30
4	Foreign exchange rate	−0.8961	−698.39

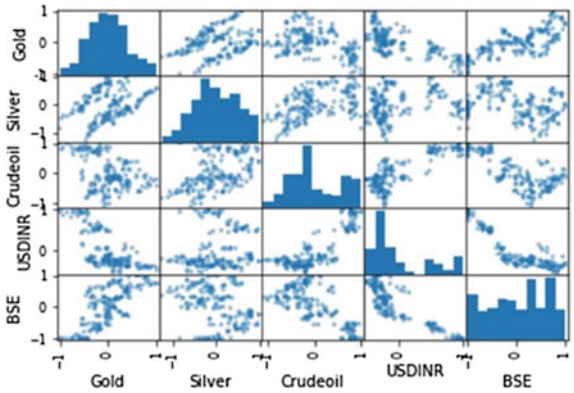

Fig. 5 Scatter plot between market performance and different attributes

4.3 Evaluation of Machine Learning Algorithms

The accuracy of all four machine learning algorithms over different values of training dataset and test dataset are compared in Table 2. From the given table, it is evident that the accuracy of AdaBoost is the highest in all of the four algorithms.

The AdaBoost was selected as the best predicting algorithm and was tested on untrained dataset and accuracy was found to be 75%. Figure 6 shows the scatter plot of actual and predicted market performance by AdaBoost. The number of instances is represented by the x-axis and the y-axis represents the class.

Table 2 Accuracy of different predictive models

Machine learning algorithms	Dataset used for verification		
	Training set (%)	Test set (%)	Accuracy (%)
AdaBoost	90	10	78.95
	70	30	76.79
Gradient boosting	90	10	78.95
	70	30	73.21
SVM	90	10	78.95
	70	30	73.00
Random forest	90	10	68.4
	70	30	67.8

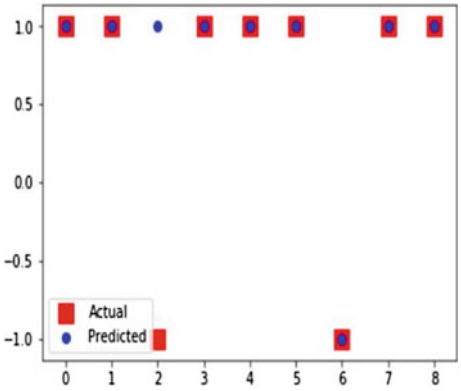

Fig. 6 Actual and predicted market performance by AdaBoost

5 Conclusion and Future Scope

The outcome of this research concludes that the machine learning algorithms can be used to predict the increase or decrease in the stock market performance. It verifies the dependency of BSE on the factors taken in the study. Our findings confirm that the dependency of BSE is highest on the gold rate, since the correlation factor is highest. Also, the correlation factor is lowest for silver rate, showing least dependency of BSE on it. Of all the machine learning algorithms used, AdaBoost shows the highest accuracy of 76.79% for 70% training data and 75% for untrained data. There is still a scope of improvement in this project. The project can be further extended to include additional variables such as interest policy, political, and economic reforms to get more accurate results.

References

1. A. Nayak, M. Pai, R. Pai, Prediction models for Indian stock market. Proced. Comput. Sci. **89**, 441–449 (2016)
2. Z. Hu, J. Zhu, K. Tse, Stocks market prediction using support vector machine, in *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering* (2013)
3. Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang, Forecasting stock market movement direction with support vector machine. Comput. Oper. Res. **32**(10), 2513–2522 (2005)
4. C. Cortes, V. Vapnik, Support vector networks. Mach. Learn. **20**, 273–297 (1995)
5. J. Ali, R. Khan, N. Ahmad, I. Maqsood, Random forests and decision trees. Int. J. Comput. Sci. Issues **9**(5), 272–278 (2012)
6. S. Yutong, H. Zhao, Stock selection model based on advanced AdaBoost algorithm, in *2015 7th International Conference on Modelling, Identification and Control (ICMIC)* (2015)
7. P. Wu, H. Zhao, Some analysis and research of the AdaBoost algorithm, in *Communications in Computer and Information Science* (2011), pp. 1–5
8. L. Zhao, L. Wang, Price trend prediction of stock market using outlier data mining algorithm, in *2015 IEEE Fifth International Conference on Big Data and Cloud Computing* (2015)

9. M. Usmani, S. Hasan Adil, K. Raza, S. Ali, Stock market prediction using machine learning techniques, *ICCOINS* (2016)
10. J. Ali, R. Khan, N. Ahmad, I. Maqsood, Random forest and decision trees. Int. J. Comput. Sci. Issues **9**(5), 272–278 (2012)
11. A. Bhargava, A. Bhargava, S. Jain, Factors affecting stock prices in India: a time series analysis. IOSR J. Econ. Finance **07**(04), 68–71 (2016)
12. N. Pahwa, N. Khalfay, V. Soni, D. Vora, Stock prediction using machine learning a review paper. Int. J. Comput. Appl. (0975–8887) **163**(5), 36–43 (2017)