# Prediction of Stock Market Prices using Machine Learning

Radhika Sreedharan, Archana Praveen Kumar

Manipal Institute of Technology Manipal Academy of Higher Education Manipal, Karnataka, India

Corresponding author: Radhika Sreedharan, Email: radhika.sreedharan@gmail.com

Stock prices prediction for the stock exchanges has become a very challenging task. The data for stock prediction was taken from history_60d data set where it has various stock listings belonging to NYSE, AMEX and NASDAQ. Certain stock listings of those stock exchanges are considered to predict stock prices for them. The foremost aim of this project is to envisage the stock prices formulated upon the given data set based on machine learning algorithms such as Support Vector Machine, Random Forest and Decision Tree Regression. These algorithms were used to predict future close prices making used of open prices. Also, they could predict low prices by utilizing high prices formulated upon the given data set. The predicted values were compared with the actual values. Accordingly, Mean Square Error, Root Mean Square Error and Mean Absolute error were computed using the actual and predicted rates of close and low prices. Machine learning algorithms like Support Vector Machine, Random Forest Regression and Decision tree regression couldn't predict values much closer to the actual values and they didn't get decimal values. These limitations were overcome by Linear Regression and Polynomial Regression. These algorithms could predict values closer to the actual values of close prices. Hence, mean square error, mean absolute error and root mean square error were less using the predicted values computed using Linear and Polynomial Regression. It is concluded that Linear Regression and Polynomial Regression could predict values closer to the actual stock prices compared to SVM, RF and DTR. The stock market project can be used for prediction of gold prices.

**Keywords**: Stock price, Prediction, Machine learning, Linear regression.

2021. In Raju Pal & Praveen K. Shukla (eds.), *SCRS Conference Proceedings on Intelligent Systems*, 339–347. Computing & Intelligent Systems, SCRS, India.
https://doi.org/10.52458/978-93-91842-08-6-34

# 1 Introduction

The main idea of this project is to compute stock prices of certain stock listings belonging to NYSE, NASDAQ and AMEX. Machine learning algorithms like Support Vector Machine, Random Forest Regression and Decision Tree Regression are employed to envisage close prices based upon the given data set. The obtained values will be compared with the actual values indicating how proximate the predicted values are to the actual values. The techniques like Linear Regression and Polynomial Regression are used to given closer predicted values to actual values compared to the these algorithms..The medium for dealings of a firm's bonds, dividends, shares and derivation at a pre-determined price is a stock market. It is oneof the acclivitous parts in any nation. Forecasting of stock prices becomes a demanding work because of active character as well as responsible to swift transformations in stock price. Values of stock prices are stipulated with regard to its 'closing price' as well as its 'adjusted closing price'. The 'raw' price which is the cash value of final accomplished price prior to closing of market is known as closing price. Adjusted closing price improves a closing price of stock for faultlessly considering the value of stock subsequent to examining for any corporate comportment. It is contemplated to be the true price of that stock and is generally utilized while enquiring factual earnings or determining a thorough analyzation of factual earnings. The adjusted closing price characterizes in whatsoever that may influence stock price at the close of market. A stock's price is usually contrived by supply and demand of market associates. However, some corporate activities, like stock splits, dividends/ distributions and rights offerings, contrive a stock's price and adjustments are required to appear at a practically authentic appearance of stock's true value. Shareholders should comprehend how corporate activities are considered for adjusted closing price of stock. Itis particularly purposeful while scrutinizing factual earnings asit delivers specialists a faultless illustration of equity value ofa firm. The closing price illustrates the recent estimated price of a security until trading initiates afresh on subsequent trading day. A number of monetary instruments are interchanged subsequent hours (even though with considerably less volume and liquidity levels), so the security's closing price may not match its subsequent hours price. Closing prices don't contemplate the subsequent hour's price or corporate activities; even though they still may act as functional markers for investors for estimating stock prices fluctuation over time- the closing price of one day can be distinguished with previous closing price for measuring market opinion for a given security over a trading day. The closing price is only purposeful at the time of phases when a company has not broadcasted any cash dividends or organized any corporate activities, like stock splits, reverse stock splits and stock dividends.

The foremost aim of the tasks is to compare the algorithms in terms of forecasting stock prices.

# 2 Background Theory

[1] Trippi & Turban, 1992; Walczak, 2001; Shadbolt & Taylor, 2002: The favored and very new procedure which includes scientific investigations for making forecasting in financial markets is Artificial Neural Network (ANN). This method incorporates a collection of baseline working. These services pointed at past background subsequent to joining one another with flexible weights and they are useful for making forecasting. [2]In 1995, Kua & Liu explored the out-of instance estimating capability of perennial as well as feed-forward neural networks on the basis of experiential foreign exchange rate data. [3]In 2017, Khasei and Rahimi assessed work of sequence and correlation approaches to regulate better one utilizing ARIMA (Auto regressive integrated moving average) and MLP (Multi-layer Perceptron).

[4]According to Samek and Varaccha, 2013, ANNs were utilized to clarify many issues because of their

adaptable type. [5]In 2012, Yodele et al. gave a hybridized way, that is, a consolidation of variables of basic and scientific investigation of stock exchange indexes for forecasting subsequent stock exchanges to boost current approaches. [6]In 2011, Kara & Boyacioglu considered stock exchange indicator action utilizing two models on the basis of Artificial Neural Network (ANN) and Support Vector Machine (SVM). They examined work of both models and accomplished standard work of ANN model was considerably of superior quality than SVM model.

[7]Qi & Zhang, 2008 examined the perfect trend time series modeling utilizing Neural Network and utilized four distinct methods that is, raw data, raw data with a time index, detrending and differencing for modeling variegated trend patterns and culminated neural network outputs of superior quality. [8]Huck (2009) showed a pre-requisite to subsequent applications of deep learning in finance. In his initial innovation, he utilized Elman perennial neural networks for working periodical forecasting for every components of S & P universe. A long-short portfolio comprising of stocks with maximum certainty produces enough returns of 0.8% per week at a 54% directional correctness. Huck (2010) adapts this method and presents multi-step forward estimates. [9]In 2014 Moritz and Zimmermann applied Random Forests to the U.S. stock universe from 1968 to 2012. They utilized rebound placed characteristics and 86 other firm-based features, as well as obtained hazard-adapted enough returns of 2.28% per month. [10] Feature-oriented approaches utilized methods based on Statistics like PCA (principal component analysis (Tsai and Hsiao, 2010) and information gain (Lee, 2009b) for selecting efficient characteristics. The work of a likely representative will be enhanced if the little pertinent characteristics in the input are distant. Many characteristics are actual values of an assured extent, and are weekly thoughtful in case of progression data. [11]Some considerations have accustomed fuzzy logic for processing feature data to acquire trend moving characteristics (Atsalakis and Valavanis, 2009b, Chang and Liu, 2008) and convert every characteristic into a probability distribution over numerous types; thereby enhancing characteristic's meaning for classification. [12]According to Li et al. 2016, the Extreme Learning Machine (ELM) is used to forecast the stock market. It can expedite training and enhance generality work with the help of randomly generated hidden layer units.

[13] Mani and Bloedorn (1997) introduced a method utilizing graph representation for extracting multi-document concept which is for connecting text's verbal facts to the model. [14] Choudhary and Bhattacharya (2002) introduced an approach for building the document feature vectors as stated by UML diagram. These approaches have rectified the issue of incorrectly stating passage arrangement and affinity to a convinced level, however these models are very complex, as well as approaches determining models' comparability are still inadequate. [15]Tsai and Wang (2016) utilized fiancé specialized opinion lexicon for inspecting relationships among monetary perception words as well as monetary hazard. The exploratory outcomes showed utilizing only monetary perception words gives rise to accomplishment corresponding to utilizing all texts.

[16] For stocks in TSEC (Taiwan Stock Exchange Corporation), Robert K. Lai et al. have accepted a financial time series-forecasting model by enlarging and congregating decision tree. The prediction model consolidated a data clustering technique, genetic algorithms (GA) and a fuzzy decision tree (FDT), and for constructing a decision-making system on the basis of past data and technological evidences. The previous data set can be split into k sub-clusters by using Kmeans algorithms. GA was employed for many fuzzy terms for every input index in FDT so the model's prediction correctness can be enhanced. [17] For getting progression expression characteristics, some reviews have accepted fuzzy logic for processing character data (Atsalakis and Valavanis, 2009b, Chang and Liu, 2008). They remodel every characteristic value into a probability distribution over numerous types; hence promoting the character's suggestive for systematization. [18] The model-oriented method targets on enhancing models' fitting capability. The support vector machines (SVM) and neural networks (NN) have proved to be very potent for stock exchange forecasting (Kara et al., 2011a). The SVM training algorithm has a cubic time complexity, and both NN and SVM are over-fitting without difficulty

because of enormous size of specification.

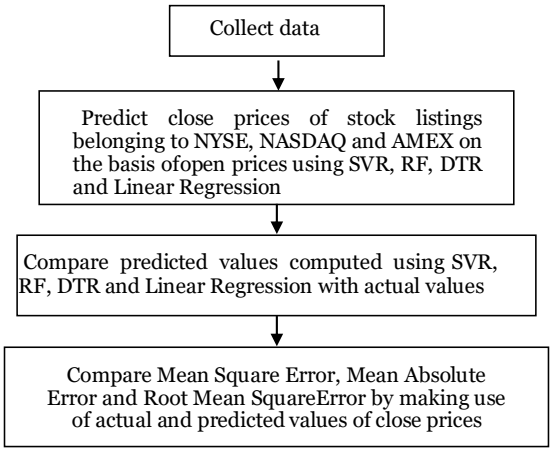
[19] The Extreme learning machine (ELM), which can accelerate training and promote conclusion, work across arbitrarily produced hidden layer units, is useful for predicting the stock market (Li et al., 2016).[20] Lodhi, Saunders, Shawe-Taylot, Cristianini and Watkins (2002) introduced a recent kernel for classifying verse evidence minus converting them to fixed-length, real-valued feature vectors.[21] Pfoh, Schnieder, and Eckert (2013) utilized a string kernel for utilizing subsequent information inbuilt in a system call trace for detecting malwares.

[22] Vishwanathan, Schraudolph, Kondor, and Borgwardt (2010) explored an idea known as graph kernels, a kernel task that calculated scalar multiplication on graphs. Random walk kernel is one of the majorities of spontaneous ideas. Zeng, Yi, and Liu (2010) apprehended spectral graph theory and random walk kernel in image denoising. [23] Li, Su and Wang (2012) apprehended purchaser communal power uniformities into a graph random walk kernel and make SVR models for forecasting buyer beliefs. [24] Lu (1990) introduced an abstract representation approach applied to area of data recovery. [25] Russell and Norvig (1995) presented the linguistic network illustration of awareness.

[26] Zhang et al. depicted an abridgement of current data in executions of neural networks for predicting. Notable utilization of ANNs is for time-series forecasting. ANNs have the convenience of precision prediction; their outcomes are conflicting in undoubted specific happenings. Dai and Zhang et. Al (2013) commenced the task with 3M stock data as the training data applied for their fieldwork. The prediction system was trained utilizing diverse algorithms. These algorithms comprehended a logistic regression, a quadratic analysis for differentiation and SVM. The discoveries were rationalized by Dai and Zhang (2013), which asserted that US equity, is semi-strongly systematic, inferring that no fundamental or practical evaluation could be used to make lowly gain. The length-term forecast system gave rise to powerful results, nonetheless proliferated when SVM manifested a high precision of 79.3%.

[27]Mackin et al (2018) put in the stacked LSTM neural network for jump detection as well. However, the approach is put on a high frequency intra-day limit order book depth data, rather than a time series.

# 3 Methodology



The formula for calculating price is

$$Price = (high\ price + lowprice + close\ price)\ 3$$

Root Mean Squared Error (RMSE): It isthe quality divergence of the forecasting errors. Residuals are estimation of how distant from the regression line data points are. It is an estimation of how unfold these residuals are:

Lesser the RMSE, better the model

$$RMSE = \sqrt{1\ n} \sum (yactual - ypredicted)2\ n\ i=1$$

Coefficient of determination: It is the portion of the variation in the relied variable that is foreseeable from the covariates. Higher the r2 score better the model.

Mean Absolute error: It is the mean ofabsolute value of the errors

$$MAE = 1 \sum i=1 |yactual - ypredicted|\ n$$

Mean Squared error: It is the average of squared inaccuracies

$$MSE = 1\ n \sum (yactual - ypredicted)2\ n\ i=1$$

The type of linear regression where link between covariate x and relied variable y is formed at the nth degree polynomial is PolynomialRegression. This form of regression suits an indiscriminate connection connecting the value of x and the commensurable conditional mean of y, denoted E(y|x).

R-Square value is a measure utilized to estimate correctness of a linear regressionmodel. It provides the goodness of fit of a regression model.

The ideal value of r-square is 1. If the value of r-square is closer to 1, the better is the model fitted.

R-square is a juxtaposition of residual sum of squares (SSres)with total sum of squares (SStot).

SStotal=SUM(yi-yavg)2

# 4 Result Analysis

Random Forest Regression could predict close price based on the values of open price.

Close price for Stock listing 'AB' could be predicted based on the value of Open price using Random Forest Regression, Support vector Regression and Decision Tree Regression and Regression plots were created.

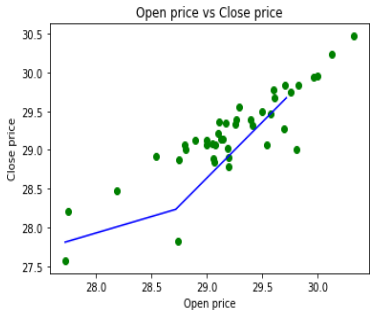



**Fig. 1.** Regression plot using Decision Tree Regression for 'AB' topredict close price


**Fig. 2.** Regression plot using Random Forest regressionfor 'AB'

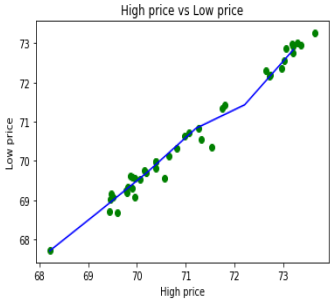



**Fig. 3.** Regression plot using SVM for 'AB'

The prices predicted were almost same for Stock listing'AB' using the above 3 algorithms. Linear Regression could predict close prices closer the actual value compared to DTR, RF and SVM. Low prices were computed for 'AAXJ' using DTR, RF and SVM.

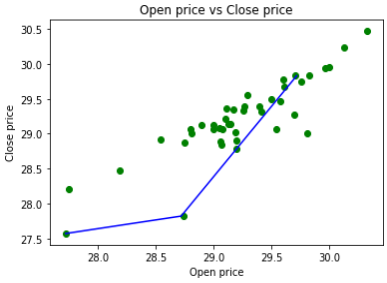



**Fig. 4.** Regression plot for 'AAXJ' using DTR topredict low price

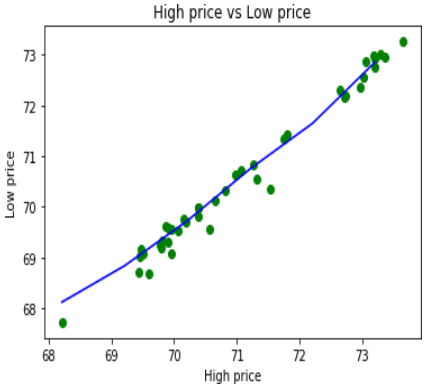



**Fig. 5.** Regression plot for 'AAXJ 'using RF topredict low price

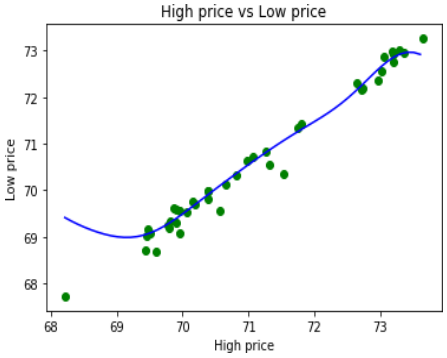



**Fig. 6**. Regression plot for 'AAXJ' using SVM topredict low price

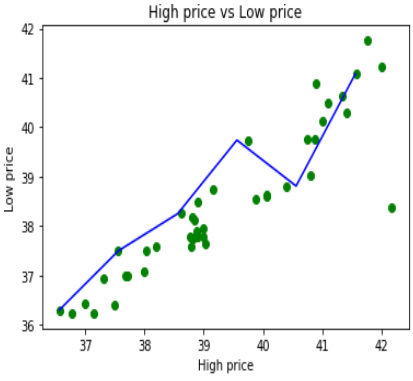



**Fig. 7.** Regression plot using 'DTR' to compute lowprices for 'AE'

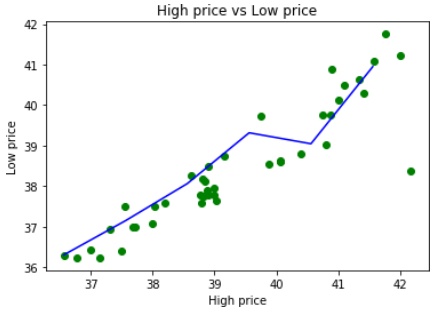



**Fig. 8.** Regression plot using 'RF' to compute lowprices for 'AE'

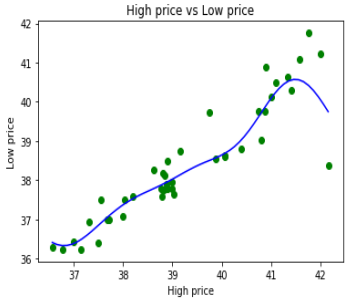



**Fig. 9.** Regression plot using SVM to compute lowprices for 'AE'

**Table 2.** Future low-price values using high prices for'AAXJ' based on given data set

| | High price | Actual lowprice | Predicted lowprice |
|---|---|---|---|
| DTR | 74.96 | 74.3 | 73 |
| RF | 74.96 | 74.3 | 73 |
| SVM | 74.96 | 74.3 | 71 |
| Linear Regression | 74.96 | 74.3 | 74.60206197 |
| Polynomial Regression | 74.96 | 74.3 | 75.07898214 |

**Table 3.** Future low-price values using high prices for 'AE'based on given data set

| | High price | Actual lowprice | Predicted lowprice |
|---|---|---|---|
| Linear Regression | 39.5 | 38.5 | 38.59659531 |
| Polynomial Regression | 39.5 | 38.5 | 38.45672283 |

| | | | |
|---|---|---|---|
| DTR | 39.5 | 38.5 | 39 |
| RF | 39.5 | 38.5 | 39 |
| SVM | 39.5 | 38.5 | 38 |

For 'AE' stock listing, polynomial and linear regression could predict low price values closer to actual low price values.

# 5 Conclusion

Close prices and low prices were predicted using Support Vector Machine, Random forest and Decision Tree Regression algorithms. With the given data set, the predicted stock prices were compared with the actual values of close and low prices. The stock prices predicted using Linear Regression and Polynomial Regression were closer to the actual values than the stock prices predicted using SVM, DRT and RF and it was applicable when variation in stock prices were less for stock listings. Polynomial regression couldn't predict closer values were there were more variations in stock prices. In such cases, linear regression could predict prices closer to actual values. Mean Square error, Root mean square error and Mean Absolute error were computed using actual and predicted values of close prices. Those close prices were predicted using SVM, RF, DTR, Linear Regression and Polynomial Regression. MSE, RMSE and MAE were less using predicted values obtained using Linear Regression. R square was almost approximately equal to one when predicted values were compared with actual close price values. R square indicated that model fitted in better.

The algorithms and techniques used in this algorithm can be used for gold prices prediction in future. Based on the current data set or the present-day scenario, future prices can be predicted using these algorithms and can be compared with actual values to find the difference between real values and future values.

# References

[1] Selvamuthi, D., Kumar. V. and Mishra. A. (2019). Indian stock market prediction using artificial neural networks on tick data. *Financial Innovation*, 5(1):16.

[2] Huck, N. (2019). Large data sets and machine learning: Applications to statistical arbitrage. *European Journal of Operational Research*, 278(1):330-342.

[3] Liu,G. and Wang, X. (2019). A new metric for individual stock trend prediction. *Engineering Applications of Artificial Intelligence*, 82:1-12.

[4] Long, W., Song, L. and Tian, Y. A. (2019). New graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity. *Expert Systems with Applications*, 118:411-424.

[5] Sharma, A., Bhuriya, D. and Singh, U. (2017). Survey of stock market prediction using machine learning approach. In *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology*, 506-509.

[6] Jyothirmayee, S. et al. (2019). Predicting stock exchange using supervised learning algorithms. *International Journal of Innovative Technology and Exploring Engineering*, 9(1): 4081-4090.

[7] Jay, F. K. et al. (2020). Jump Detection in Financial Time Series using Machine Learning Algorithms. *Soft Computing*, 24:1789-1801.