International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Indian Stock Price Prediction Using Machine Learning Algorithm and Sentiment Analysis

Harshwardhan Patil[1], Rahul Patil[2]

[1]Master of Technology, Computer Engineering, Pimpri Chinchwad College of Engineering, Maharashtra, India

[2]Professor, Computer Engineering, Pimpri Chinchwad College of Engineering, Maharashtra, India

Abstract: ***Stock market prediction is a major exertion in the field of finance and establishing businesses. Stock market is totally uncertain as the prices of stocks keep fluctuating on a daily basis because of numerous factors that influence it. One of the traditional ways of predicting stock prices was by using only historical data. But with time it was observed that other factors such as peoples' sentiments and other news events occurring in and around the country affect the stock market, for e. g., national elections, natural calamity etc. Investors in the stock market seek to maximize their profits for which they require tools to analyze the prices and trend of various stocks. Machine learning algorithms have been used to devise new techniques to build prediction models that can forecast the prices of stock and tell about the market trend with good accuracy. Many prediction models have been proposed to incorporate all the major factors affecting the price of stocks. This project focuses random forest and sentiment-based prediction to decide buy or sell of the stock.***

Keywords: Stock Market Prediction; Machine Learning Algorithm

1. Introduction

The Stock market plays a vital role in the country's economic growth as well as the individual economy to a large extent. Finding the right time to buy and sell the shares is dependent on predicting the trends in the stock market. The technique for most accurate prediction is to learn from past instances and design a model to do this by using traditional & machine learning algorithms. Predicting stock and stock price index is difficult due to uncertainties involved. There are two types of analysis which investors perform before investing in a stock. First is the fundamental analysis. In this, investors look at intrinsic value of stocks, performance of the industry and economy, political climate etc. to decide whether to invest or not. On the other hand, technical analysis is the evaluation of stocks by means of studying statistics generated by market activity, such as past prices and volumes. The current trend of Algorithmic trading is taking boom in Stock Market World. Technical analysts do not attempt to measure a security' s intrinsic value but instead use stock charts to identify patterns and trends that may suggest how a stock will behave in the future. The Stock market trend varies due to several factors such as political, economics, environment, society, etc. Since years, many techniques have been developed to predict stock trends. Initially classical regression methods were used to predict stock trends. Since stock data can be categorized as non-stationary time series data, non-linear machine learning techniques have also been used. Artificial Neural Networks (ANN) and Support Vector Machine (SVM) are two machine learning algorithms which are most widely used for predicting stock and stock price index movement. Each algorithm has its own way to learn patterns. ANN emulates functioning of our brain to learn by creating network of neurons.

The major aim of the paper can be summarized as following:

- To use the Indian Stock Market API for fetching the live DataStream
- To use data from Stock Economics related keywords for sentiment analysis
- To use Random Forest Classifier for classification as it has more accuracy of prediction
- To Merge the data from various data sources.
- Generate Algorithmic Trading Model

2. Related Work

Efficient market hypothesis by Malkiel and Fama (1970) states that prices of stocks are informationally efficient which means that it is possible to predict stock prices based on the trading data. This is quite logical as many uncertain factors like political scenario of country, public image of the company will start reflecting in the stock prices. So, if the information obtained from stock prices is pre-processed efficiently and appropriate algorithms are applied then trend of stock or stock price index may be predicted

Hassan, Nath, and Kirley (2007) proposed and implemented a fusion model by combining the Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and Genetic Algorithms (GA) to forecast financial market behavior. Using ANN, the daily stock prices are transformed to independent sets of values that become input to HMM. Wang and Leu (1996) developed a prediction system useful in forecasting mid-term price trend in Taiwan stock market. Their system was based on a recurrent neural network trained by using features extracted from ARIMA analyses. Empirical results showed that the networks trained using 4-year weekly data was capable of predicting up to 6 weeks market trend with acceptable accuracy. Hybridized soft computing techniques for automated stock market forecasting and trend analysis was introduced by Abraham, Nath, and Mahanti (2001). They used Nasdaq-100 index of Nasdaq stock market with neural network for one day ahead stock forecasting and a neuro-fuzzy system for analysing the

trend of the predicted stock values. The forecasting and trend prediction results using the proposed hybrid system were promising. Chen, Leung, and Daouk (2003) investigated the probabilistic neural network (PNN) to forecast the direction of index after it was trained by historical data. Empirical results showed that the PNN-based investment strategies obtained higher returns than other investment strategies examined in the study like the buy-and-hold strategy as well as the investment strategies guided by forecasts estimated by the random walk model and the parametric GMM models. A very well-known SVM algorithm developed by Vapnik (1999) searches for a hyper plane in higher dimension to separate classes. Support vector machine (SVM) is a very specific type of learning algorithms characterized by the capacity control of the decision function, the use of the kernel functions and the scarcity of the solution. Huang, Nakamori, and Wang (2005) investigated the predictability of financial movement direction with SVM by forecasting the weekly movement direction of NIKKEI 225 index. They compared SVM with Linear Discriminant Analysis, Quadratic Discriminant Analysis and Elman Backpropagation Neural Networks. The experiment results showed that SVM outperformed the other classification methods. SVM was used by Kim (2003) to predict the direction of daily stock price change in the Korea composite stock price index (KOSPI). Twelve technical indicators were selected to make up the initial attributes. This study compared SVM with back-propagation neural network (BPN) and case-based reasoning (CBR). It was evident from the experimental results that SVM outperformed BPN and CBR. Random forest creates n classification trees using sample with replacement and predicts class based on what majority of trees predict. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to over-fit the training data more than a single model would, but in practice, some ensemble techniques (especially bagging) tend to reduce problems related to overfitting of the training data. Tsai, Lin, Yen, and Chen (2011) investigated the prediction performance that utilizes the classifier ensembles method to analyze stock returns. The hybrid methods of majority voting and bagging were considered. Moreover, performance using two types of classifier ensembles were compared with those using single baseline classifiers (i. e., neural networks, decision trees, and logistic regression). The results indicated that multiple classifiers outperform single classifiers in terms of prediction accuracy and returns on investment. Sun and Li (2012) proposed new financial distress prediction (FDP) method based on SVM ensemble. The algorithm for selecting SVM ensemble' s base classifiers from candidate ones was designed by considering both individual performance and diversity analysis. Experimental results indicated that SVM ensemble was significantly superior to individual SVM classifier. Ou and Wang (2009) used total ten data mining techniques to predict price movement of Hang Seng index of Hong Kong stock market. The approaches include Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), K-nearest neighbor classification, Naive Bayes based on kernel estimation, Logit model, Tree based classification, neural network, Bayesian classification with Gaussian process, Support vector machine (SVM) and Least squares support vector machine (LS-SVM). Experimental results showed that the SVM and LS-SVM generated superior predictive performance among the other models. It is evident from the above discussions that each of the algorithms in its own way can tackle this problem. It is also to be noticed that each of the algorithm has its own limitations. The final prediction outcome not only depends on the prediction algorithm used but is also influenced by the representation of the input. Identifying important features and using only them as the input rather than all the features may improve the prediction accuracy of the prediction models. A two-stage architecture was developed by Hsu, Hsieh, Chih, and Hsu (2009). They integrated self-organizing map and support vector regression for stock price prediction. They examined seven major stock market indices. Specifically, the self-organizing map (SOM) was first used to decompose the whole input space into regions where data points with similar statistical distributions were grouped together, so as to contain and capture the non-stationary property of financial series. After decomposing heterogeneous data points into several homogenous regions, support vector regression (SVR) was applied to forecast financial indices. The results suggested that the two-stage architecture provided a promising alternative for stock price prediction. Genetic programming (GP) and its variants have been extensively applied for modeling of the stock markets. To improve the generalization ability of the model, GP have been hybridized with its own variants (gene expression programming (GEP), multi expression programming (MEP)) or with the other methods such as neural networks and boosting. The generalization ability of the GP model can also be improved by an appropriate choice of model selection criterion. Garg, Sriram, and Tai (2013) worked to analyze the effect of three model selection criteria across two data transformations on the performance of GP while modeling the stock indexed in the New York Stock Exchange (NYSE). It was found that FPE criteria have shown a better fit for the GP model on both data transformations as compared to other model selection criteria. Nair et al. (2011) predicted the next day' s closing value of five international stock indices using an adaptive Artificial Neural Network based system. The system adapted itself to the changing market dynamics with the help of genetic algorithm which tunes the parameters of the neural network at the end of each trading session. The study by Ahmed (2008) investigated the nature of the causal relationships between stock prices and the key macro-economic variables representing real and financial sector of the Indian economy for the period March, 1995–2007 using quarterly data. The study revealed that the movement of stock prices was not only the outcome of behavior of key macro-economic variables but it was also one of the causes of movement in other macro dimension in the economy. Mantri, Gahan, and Nayak (2010) calculated the volatilities of Indian stock markets using GARCH, EGARCH, GJR-GARCH, IGARCH & ANN models. This study used Fourteen years of data of BSE Sensex & NSE Nifty to calculate the volatilities. It was concluded that there is no difference in the volatilities of Sensex, & Nifty estimated

under the GARCH, EGARCH, GJR GARCH, IGARCH & ANN models. Mishra, Sehgal, and Bhanu Murthy (2011) tested for the presence of nonlinear dependence and deterministic chaos in the rate of returns series for six Indian stock market indices. The result of analysis suggested that the returns series did not follow a random walk process. Rather it appeared that the daily increments in stock returns were serially correlated and the estimated Hurst exponents were indicative of marginal persistence in equity returns. Liu and Wang (2012) investigated and forecast the price fluctuation by an improved Legendre neural network by assuming that the investors decided their investing positions by analyzing the historical data on the stock market. They also introduced a random time strength function in the forecasting model. The Morphological Rank Linear Forecasting (EMRLF) method was proposed by Araújo and Ferreira (2013

System Architecture

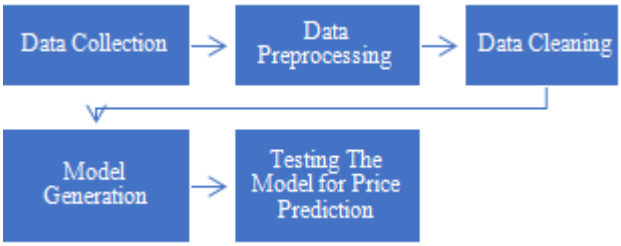


Figure 1: System Architecture

3. Proposed Methodology

In this paper, we explored, Machine Learning technique and Sentiment Analysis for prediction of futuristic stock prices. Usingthese techniques, it is beneficial to the traders of Indian stocks, to generate more profit and minimize their losses in trading world.

This research work possesses using the Indian Stock Market API for fetching the live DataStream of Indian stock prices. Data from Stock Economics related keywords use for sentiment analysis. Random Forest Classifier for classification as it has more accuracy of prediction. Merge the data from various data sources to get more information and improve the result of prediction. Finally from this we are able to generate prediction trading model

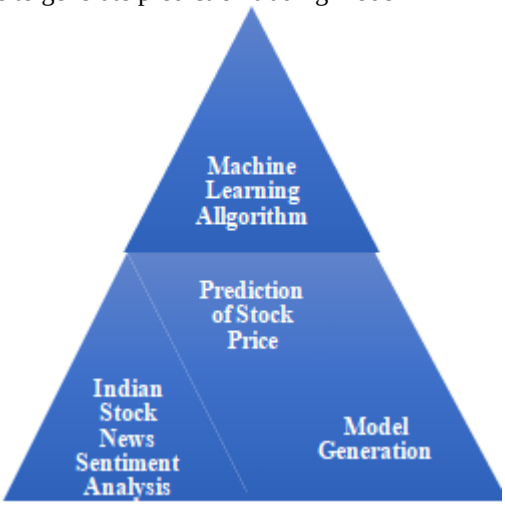


Figure 2: Proposed Methodology

The above figure depicts the overall scenario of proposed methodology, We will provide Quant data of NSE listed Indian Stock to ML algorithm. Next, we do Indian stock sentiment analysis we generate model.

Table 1: Performance Metrics

Performance Metric	Value
Mean Absolute Error	0.7445293661973035
Mean Squared Error	1.2783847671784652
Root Mean Squared Error	1.1306567857570506

Table 2: Accuracy

Mean Absolute Error	0.74 degrees
Accuracy	99.91 %

Implementation

In this paper, the effort is to build the system that is able to predict the futuristic price movement of the Stock on its analysis by means of Sentiment Analysis of news-based articles and actual historical data of stock using Machine Learning Algorithm.

System that is able to fetch data from News API news sources

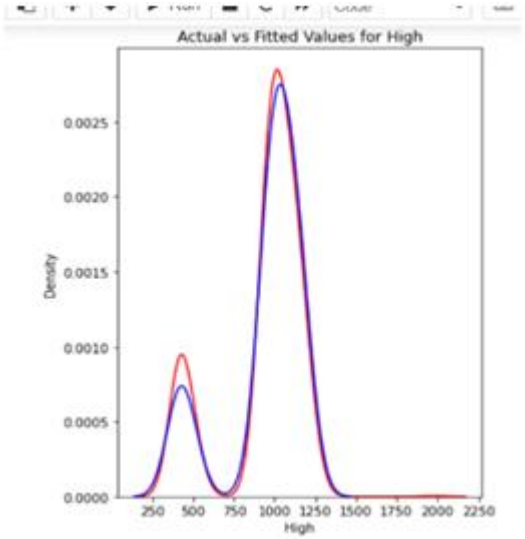
Use Random Forest Classifier for classification and more accuracy of prediction

Merge the data from various data sources.

Generate Price Prediction Model

4. Experimental Results Analysis

In the fig below the actual vs predicted price of stock is generated using random forest algorithm. From the accuracy of random forest is 99.91%



	Actual	Predicted
406	1140.95	1140.13100
546	1181.95	1182.97105
689	971.20	971.79510
14	409.63	409.78374
261	1080.98	1080.81506
...	...	...
549	1191.00	1190.83800
71	413.75	414.46631
49	387.53	388.33643
474	1123.40	1123.45279
521	1196.00	1196.40360

284 rows × 2 columns

5. Conclusion and Future Work

In this paper, Stock market is a long-time attractive topic for researcher and investors from its existence The Stock prices are dynamic day by day, so it is hard to decide what is the best time to buy and sell stocks. Currently we have trained the model using Random Algorithm. It is giving accuracy with 99.91%.

The results will improve once News API news sources sentiment analysis-based model will be merged with the Random Forest Algorithm. The system can be further improved further by adding artificial intelligence system components to facilitate the doctors and the patients. The data, consisting medical history of many patients' parameters and corresponding results, can be explored using data mining, in search of consistent patterns and systematic relationships in the disease. For instance, if a patient's health parameters are changing in the same pattern as those of a previous patient in the database, the consequences can also be estimated. If the similar patterns are found repeatedly, it would be easier for the doctors and medical researchers to find a remedy for the problem.

References

[1] Dharmaraja Selvamuthu, Vineet Kumar & Abhishek Mishra Indian stock market prediction using artificial neural networks on tick dataFinancial Innovation volume 5, Article number: 16 (2019)
[2] Kunal Pahwa, Neha Agrawal. Stock Market Analysis using Supervised Machine Learning. In *2019 IEEE International Conference on Machine Learning, BigData, Cloud and Parallel Computing (FiCloudW (Com-IT-Con), India, 14th-16th Feb 2019.*
[3] Sukhman Singh, Tarun Kumar Madan. Stock Market Forecasting using Machine Learning: Today and Tomorrow2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)
[4] Indian stock market prediction using artificial neural networks on tick data. DharmarajaSelvamuthu, etlhttps: //jfinswufe. springeropen. com/articles/10.1186/s40854-019-0131-7
[5] Stock Prediction with Random and Long Short-term Memory. Shangxua
[6] Research on the text sentiment classification about the social hot events on Weibo Fulian Yin etl
[7] Analysis of various machine learning algorithm and hybrid model for stock market prediction using python. Sahil Vazirani; Abhishek Sharma; Pavika Sharma
[8] Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks. Sai Vikram Kolasani1
[9] Stock Market Analysis using Supervised Machine Learning. KunalPahwa; Neha Agarwal
[10] Word Embeddings and Their Use In Sentence Classification Tasks
[11] Adi Shalev etlStock Market Forecasting using Machine Learning: Today and Tomorrow
[12] Sukhman Singh; TarunKumarMadan; Jitendra Kumar; Ashutosh Kumar Singh