A Survey of Machine Learning Techniques Used on Indian Stock Market

Archit A. Vohra* and Dr. Paresh J. Tanna
School of Engineering, RK University, Rajkot, India.

*architvohra@hotmail.com

Abstract. Stock market provides nonlinear time series data that has been difficult for making any kind of predictions. Since the advent of various Machine Learning techniques lot of research is covered in this area. Researches have mostly been carried out using Supervised Machine Learning algorithms on predicting the price of a particular stock and also on predicting the stock index movement. The studies also cover various unsupervised machine learning algorithms as well as some statistical methods used in the stock price prediction. Study has been covered on generating portfolio of stocks that maximizes the profit with minimum risk. Social Media today generates a large amount of data and has a significant impact on the buy/sell decision made by experts. Specifically, research has been carried out on how tweets and news have an impact on stock price movement. Recent researches focus more on neural networks, deep learning and hybrid models. This study tries to identify the different approaches used by researchers in order to improve the accuracy of machine learning algorithms on the India Stock Market. The survey emphasizes that all aspects related to stock market need to be considered before making any kind of prediction.

1. Introduction

The Indian stock market is the ninth largest in the world with a market capitalization of US $2.02 trillion as of December 2019. The market capitalization is likely to cross US $ 6 trillion by 2027, thus providing a huge opportunity for investors to grow. Investment in stock market is done mainly for two reasons - first to beat the inflation rate and second to make quick money. Investment in stocks after proper study helps in maximizing the profit and minimizing the risk.

The Indian stock market majorly comprises of two big stock exchanges the Bombay Stock Exchange (BSE) and the National Stock Exchange (NSE). The Bombay Stock Exchange has been in existence since 1875 and is Asia's oldest stock exchange. Approximately 5400 companies are listed. However only 500 companies from those listed form about 90 % off market capitalization. The BSE has a stock index known by SENSEX that is Stock Exchange Sensitive Index and lists 30 companies which are some of the largest and most actively traded. The NSE was founded in in 1992 and it started trading in 1994. Almost 1900 companies are listed. The NSE came up with an equity benchmark index called NIFTY - National Stock Exchange 50. NIFTY represents the weighted average of 50 Indian company stocks in various sectors.

2. Machine Learning Techniques

A computer scientist expert in artificial intelligence, Arthur Lee Samuel first coined the term Machine Learning way back in 1959. Machine Learning has gained popularity since last decade because of its
applicability in the fields like – prediction, image recognition, speech recognition, medical diagnosis, financial industry and trading. The main reasons behind the development of Machine Learning include – increase in computational power and increased amount of data generation by various digital devices. Machine Learning is a set of algorithms that allows soft-ware applications to learn from input data and improve their accuracy with time without the need of being explicitly programmed. Main types of Machine Learning algorithms include Supervised Learning, Unsupervised Learning and Reinforcement learning.

2.1. Supervised Machine Learning
Here the system is trained to build a model with a set of input and its corresponding output. Once the model is built, it will be tested for another set of input to check the output. Finally, the system will be able to find the output for any corresponding live input data.

2.2. Unsupervised Machine Learning
Here the system is trained to build a model only with the set of input data without the corresponding output. Here the model works on its own to discover correct output.

2.3. Reinforcement Learning
Here the system is trained to build a model using the concept of reward and punishment. The system learns by getting rewards for correct performances and punishments (penalties) for in-correct performances. Thus, develops itself by maximizing rewards and minimizing punishments.

The most common ML algorithms mostly researched the most are briefly discussed next.

**K-Nearest Neighbours (KNN):** KNN is one of the simplest supervised classification algorithms that can sometimes be used for regression as well. KNN is based on a simple assumption that similar data exist nearer to each other. Given an input it tries to identify the nearest neighbour mostly using Euclidean distance and hence classifies the data. KNN can be used in stock price prediction as well as in portfolio creation.

**Naïve Bayes Classifiers:** These are a collection of probabilistic classification algorithms based on Bayes’ Theorem. An important assumption is that every pair of attributes being classified is assumed to be independent of each other. Naïve Bayes classifier is highly scalable and takes minimum computational time and processor cycles. Due to its speed and ability of classifying multi class problems they are useful in real time stock market prediction. They are also useful in text classification and sentiment analysis.

**Random Forest:** Random Forest is a supervised ML algorithm, mostly used for classification but can be used for regression also. Multiple decision trees are generated based on sample data. Each tree is generated using different sample values. The result that is common among multiple decision trees is considered as the final decision. Random forest tries to overcome the problem of overfitting but requires a lot of computational power. Random forest is found to be very accurate on very large amount of data and hence can be utilized on stock market data.

**Multiple Linear Regression (MLR):** MLR also known simply as Multiple Regression (MR) is a statistical method used to model the linear relationship between more than one independent variable and one dependent variable. Hence MLR can be used to predict the value of dependent variable given the values of independent variable. Thus, it can be useful in predicting the stock market trend as well as stock prices.

**Support Vector Machine (SVM):** SVM is a supervised ML algorithm that helps in classifying the data points in an N-Dimensional space through a hyperplane. There can be multiple hyperplane’s that classify the given data points, but SVM identifies one hyperplane that maximizes the distance between the data points in the n-dimensional space. SVM is also useful in regression analysis and
hence making it a powerful technique. SVM can capture complex non-linear relationship between data points and thus very useful in predicting the stock index movement and stock prices.

**Neural Network (NN):** These algorithms are modeled based on neurons of human brain. Neural Networks also known as ANN – Artificial Neural Network consist of multiple layers of neurons. The first layer is called the input layer and the last one is known as output layer. There may be one or more hidden layers between the input and output layers. Each layer may consist of multiple neurons, which are connected to neurons of adjacent layers. In a Feed For-forward Neural Network (FFNN), the information moves only in forward direction from the input layer to the hidden layers and finally to output layers. In FFNN with backpropagation, if the output pattern does not match the output layer, the error signal is back propagated from the output layer to the input layer. A special type of FFNN is Radial Basis FFNN, where Radial Basis functions are used as activation functions. Neural Networks are good for time series and that is the reason that they are used a lot in stock market predictions.

**K-Means Clustering:** It is one of the simplest unsupervised ML methods to group or cluster the data with similar characteristics. K-Means Clustering creates K clusters from the set of unlabelled data with most similar characteristics. This method is significantly useful in creating and optimizing stock portfolio. Moreover, it can also be useful in sentiment analysis.

3. Related Work – Approaches Used
This survey has tried to get an overview about the different research works that are conducted on Indian Stock Exchanges i.e. BSE and NSE. Most of the studies have used historical data from BSE or NSE sources directly or from websites like Yahoo Finance. Web scraping is generally used to fetch data from Yahoo Finance or some similar sites providing stock data. Some of the common approaches used to predict stock price or stock index trend are discussed next.

![Figure 1. Single Layer approach using basic indicators](image)

Initial studies on stock market implemented a basic approach as shown in figure 1, implementing any of the ML algorithms using basic indicators on the stock market dataset. Mostly supervised machine learning algorithms use basic parameters like high price, low price, close price and volume of stock for the day to indicate the trend of stock price or predict the stock price. This type of approach does not consider other fundamental indicators, sentiments or other economic variables. Hence it just assists in giving an idea about the stock price trend and not very efficient while predicting stock price.

The research covered in [13] implements the above approach by using Neural Network algorithm and considers only the Close price of the day. The study analysed two implementations of NN – Feed Forward NN with back propagation and Radial Basis NN. The objective here was to predict the stock price and trend of stock price. The data set consisted of 9 stock closing prices for four consecutive days from NSE. The FFNN gave better results while predicting the trend of stock price. Similarly, while predicting the stock price the RBNN gave better results. Further the study concluded that if number of significant inputs to the ANN is increased and some economic or fundamental factors are included the results can still improve.

The approach shown in figure 1 is again used by [9], where Neural Net-work algorithm is applied on the basic parameters. The novelty in the research is the use of Binarized Genetic Algorithm for optimizing the weights while training and hence the model name BGANN – Binarized Genetic
Artificial Neural Network. In this study in order to improve the accuracy, Min-Max Normalization is carried out while data pre-processing. The experiment of using Binarized Genetic Algorithm during ANN training is carried out using R programming. Dataset considered here is bank data for three years from 2014, collected through NSEINDIA website. Further it is found that the error rate of BGANN is least when compared to SVM, NN and ARIMA. The study intends to include fundamental analysis in future.

Technical indicators of stock like Simple Moving Average (SMA), Exponential Moving Average (EMA), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Commodity Channel Index (CCI) etc. are calculated from the basic indicators figure 2. Specifically, for SVM and NN these indicators are used to improve the accuracy of prediction. However, in this approach also other factors affecting the stock market are not considered.

Cuckoo Optimized SVM for Stock Market Prediction is researched in [19]. Cuckoo search is a technique for optimizing the parameters of SVM. Thus, the study aims at improving the accuracy of SVM classifier. The CS-SVM has better prediction accuracy when compared to SVM and ANN. The system is based on the basic indicators – open, high, low and volume shown in figure 1 as well as 5 technical indicators – EMA, RSI, MACD, MFI and SO included in figure 2. Yahoo finance is used to collect daily closing prices. The study suggested that incorporating the effects of social media sentiments and considering the impact of crude, oil and dollar price may further improve the prediction accuracy.

The study [1] uses first the basic stock parameters and then it also uses few technical parameters and hence follows both figure 1 and figure 2. The CLFANN (Computationally Efficient Functional Link Artificial Neural Network) uses both ADE (Adaptive Differential Evolution) and LMS (Least Mean Square) to train the weights of the network. MLP suffers from multiple drawbacks like slow convergence, local minimum, overfitting and generalization. Moreover, the number of neurons in the hidden layer are chosen by trial and error. The performance of weights optimized using ADE is found to be much better than that of improved LMS and it also overcomes the limitations of MLP. The dataset consists of 3400 records of Reliance (BSE India) and IBM, Oracle (DWSG-US). MAPE, RMSE and MAE are used as error measures.

Increase in the use of digital devices and social media has a major impact on the stock market. Sentiment analysis is another area of research that has evolved relying on twitter data and news data. It is observed that apart from other factors, global political news has a significant impact on the stock market. Political tensions between America and Iran resulted in the stock markets sliding down by almost 1.5% to 2%. The indications of peace by America than got the market moving up almost by 1-1.5%. At present the public sentiments have a major impact on the stock market. Figure 3 tries to depict the effect of the same on stock market, however to improve the accuracy other factors must also be considered.
The study of news sentiment effects on stock market is studied in [11]. The research considered pharmaceutical stocks and developed a dictionary to further develop a sentiment analysis model. A directional accuracy of 70.59% was obtained while predicting trends in stock price movement. Only news sentiments were considered excluding the effects of other parameters on the stock market. The reliable moneycontrol.com was used to retrieve news articles. To transform text into numerical vectors, python library – ‘pattern’ was used. The study concluded that in sentiment analysis a word weighing approach may be useful. Moreover, a hybrid approach that considers ARIMA or LSTM along with sentiment analysis may improve the directional accuracy.

Approach shown in figure 4 is a hybrid approach that considers the effect of both ML algorithms and sentiment analysis. This approach tends to better results as it considers multiple factors affecting the stock market. However fundamental analysis and many other factors that have an impact on the stock market are still missing.

The research described in [16] uses the approach shown in figure 4. The study first compares the performance of three ML algorithms- ANN, SVM and MLR. The result depicts that ANN performs better than SVM and MLR. Hence ANN is combined with sentiment analysis to build a prediction model that finally gives the stock price. The stock market data is collected from Yahoo finance, news related data from economic times and reviews from twitter. The prediction model will be based on monthly prediction and daily prediction to forecast the next day open price.

The model proposed in [6] combines NN with sentiment analysis resembling the approach shown in figure 4. The paper concludes that there is significant correlation between actual stock prices and those predicted by proposed model. Here the input layer consists of 5 neurons, 4 indicating previous 4 days close price and the 5th indicating the sentiment score of previous day. The study suggested using deep multilayer neural networks with more than two hidden layers for deter-mining text sentiments. The research has collected news data from moneycontrol and historical stock data from yahoo finance.

Two different models – one for daily trend prediction and the other for monthly prediction are developed in [4]. The daily prediction model combines historical prices with sentiments, again depicting the approach shown in figure 4. Historical data is trained using supervised machine learning algorithms. Decision Boosted Tree performs better than SVM and Logistic Regression. The daily prediction model results in an accuracy of almost 70%. The monthly prediction model indicates that the trend of previous month and current month are least correlated. For sentiment analysis the data is collected from twitter and 2 different news websites while yahoo finance is used to collect stock data.
Only SVM classification method is used by the study [20] and the approach used here is as per figure 4 as technical indicators are used to predict CNX NIFTY index value. However, this study is an in-depth study and instead of doing research on multiple different ML algorithms, it has focused on four different Kernels of SVM classification model. The work has considered only 1046 data samples of CNX NIFTY and only basic parameters (open, high, low & close) are fetched from Yahoo Finance. From the 4 basic indicators, 20 other technical features are calculated and considered. Since there are many features that are considered, the SVM kernel functions can solve the problem easily with high optimization. Here 5-fold cross validation is used and hence the processing time slightly increases. Orange data mining tool is used to study all the four kernel functions i.e. Linear, Radial Basis, Polynomial and Hyperbolic Tangent. The study reveals that Polynomial Kernel function takes longer training time without good accuracy. The Linear and Radial Basis Kernel result in good accuracy with linear kernel being slightly better of the two. The research states that connections between number of features, data parts and accuracy can be further explored.

One more research [15] utilizes the approach shown in figure 4 to predict the stock price. The study uses a combination of Neural Network and Genetic Algorithm. Here a soft computing approach is used on stock market related news and social media data. This method was tested on only one stock CUB collected from NSE website. Neither the accuracy of the approach is measured, nor the method is applied to multiple different stocks.

A novel Multi-Layer approach using a combination of multiple ML algorithms is depicted in figure 5. The study [8] focused on the indices S&P BSE & CNX Nifty of India. It proposed two stage fusion approach similar to the one shown in figure 5 involving SVR (Support Vector Regression) in the first stage. ANN (Artificial Neural Network), RF (Random Forest) and SVR (Support Vector Regression) are used in the second stage resulting into SVR--ANN--RF & SVR--SVR fusion model to predict stock market index. Ten technical indicators are selected as input to these models. Many of the other studies used a single stage approach. News is not included in this study. News can be categorized into various categories and their statistical weights can have an influence on the prediction.

The approach shown in figure 5 is again utilized by the paper presented in [10]. The work develops and studies three hybrid models ARIMA-SVM, ARIMA-ANN and ARIMA-RF. Thus, the linear and non-linear models are combined to capture the relationship in the time series data. The results prove that the hybrid ARIMA-SVM outperforms other models in terms of profit, volatility and risk adjusted performance. The residuals of ARIMA became the input to the SVM. The daily closing values of NIFTY was obtained from the NSE website. The study also concluded that hybrid model works well...
and out-performs the independent models. The proposed hybrid model can also be used by policy makers in forecasting financial and economic data.

The hybrid framework suggested by the study [3] is similar to the multi-layer approach shown in figure 5. The proposed research has considered the SVM as first layer and LSTM at second layer. The results of the hybrid model are compared with SVM, LSTM and Back Propagation. The results clearly indicate that the proposed model outperforms other algorithms both in terms of accuracy and time. The dataset of TATA Global from NSE from 2012 to 2018 was considered. The prediction was made for both one day ahead and one week ahead. Social media data was not considered in this study.

The research [5] carried out not only forecasts one day ahead stock price, but also recommends a trade or no trade for a stock. The study reveals a multi-layered approach as shown in figure 5. First the technical indicators were optimized using GA technique and then used in decision tree. Moreover, two more systems one using Decision Tree – SVM and the other using Decision Tree – SVM with feature selection on the optimized technical parameters were studied. The GA optimized technical indicators decision tree SVM with feature selection gave the best results. Cipla and HUL from Bombay Stock Exchange as well as Glaxo Smithkline and Royal Bank of Scotland from London Stock Exchange were taken up for study.

A hybrid framework combining Support Vector Machine (SVM) with K-Nearest Neighbour (K-NN) is proposed in paper [14]. Initially SVM uses different kernel functions to predict the profit or loss. This output becomes the input for K-NN and computes the nearest neighbour from the training set to predict the stock index in the horizon of 1 month, 1 week and 1 day. The results also guide the user to analyse the closing price, volatility and momentum of the stock market.

The formulation of an appropriate stock selection strategy enables the development of stock market portfolio and is covered in [17]. Generally, any portfolio tries to maximize the profit while minimizing the risk. In the first step of this study, k-means clustering is applied to a heterogeneous sample of 53 stocks listed on the NSE. The monthly closing return is considered as a basis to classify the stocks in three categories (such as low stock price, medium stock price, and high stock price). The Generalized Reduced Gradient (GRG) method under the dynamic environment is used next for distributing weights to the stocks belonging to the portfolio. The study bridges the effects of unsupervised learning to supervised learning and is unique.

Stocks are classified into different clusters using data mining to optimize the stock portfolio as presented in [18]. A portfolio is built by selecting stocks from each cluster ensuring diversification and few other investment criteria. Results prove that K-Means cluster builds the most compact cluster as compared to SOM and Fuzzy C-means. The research uses timely stock returns and valuation ratios to measure the performance of the stock.

One more hybrid model [2] predicts the returns of stocks from the NSE. The model consists of one non-linear model i.e. RNN and two linear models – Exponential Smoothing and Autoregressive Moving Average. By merging the predictions from both linear and non-linear model the authors have come up with a unique Hybrid Prediction Method (HPM). In this research prediction from RNN, ES and ARMA are merged to form the HPM. Here the dataset is taken from NSE and consists of 6 stocks from different sectors, for a period of 16 weeks. Training and Test dataset surprisingly were in equal proportion i.e. 50% each. One needs to note that for training only RNN is used whereas for testing the HPM was used. The performance of the predictive models is measured using MSE and MAE. The correlation coefficient between the actual returns and predicted returns is calculated and is significantly positive. The main drawback of this research is that it does not give similar results on a different dataset. Moreover, the optimal regression order for the linear model needs to be chosen after trial and error. The computation time taken by this method is also significantly high.

A very unique and different kind of study is covered in [12] where increase or decrease in stock market performance is predicted. Normally basic indicators like (open, close, low, high etc.) or technical indicators like (MA, RSI, MACD, etc.) are used for any kind of stock prediction using machine learning. Here some of the most common macroeconomic factors such as commodity price, market history, Foreign exchange rates are used to predict BSE. In this study the four machine
Learning algorithms considered are SVM, RF, Gradient Boosting and Adaptive Boosting. Study includes market history data of 2 years from Jun-2015 to Dec-2016. The commodity and foreign exchange rates for 9 months are used from investing.com. The dataset is processed using Python and different proportions of training (70% & 90%) are used. The AdaBoost algorithm resulted in best accuracy. The results revealed that the performance of BSE is best correlated to only Gold price. Moreover, the research revealed a negative correlation of BSE with other factors considered and is equally useful.

The research [7] is unique in the sense it is carried out keeping in mind the fact that as the dimension of data increases, the complexity of the model also increases. Hence it tries to implement minimal variability OWA (Order Weighted Averaging) operator. The operator ensures that high dimensional data is aggregated to a single dimension. The research proves that when subtractive clustering is combined with hybrid fuzzy based network inference system based on the proposed model, the results are better compared to other models. The experiment is carried out on the Bombay Stock Exchange Index (BSE30).

4. Discussion
This survey focuses on the different types of researches carried out on Indian stock market. The survey identifies that there are few major approaches used mainly while predicting the stock price trend, stock price or the stock index trend. Almost all ML algorithms are implemented on the Indian stock market data. It is observed that the same algorithm say for example neural network may be used by multiple researchers where the input parameters are different and hence the output. Moreover, the performances of various researches are all measured using different methods. Hence it becomes extremely difficult to compare any two researches.

Some of the researches focus only on a small subset of the stocks, making it difficult to imagine as to how it will perform when a different set of stock data is used. The stock market is extremely volatile; hence the performance of one particular research may vary from time to time. However, in general it is observed that fine-tuned SVM and ANN seem to give better results compared to other algorithms. In order to improve the accuracy most of the current researches employ a hybrid or multi-layered approach as shown in figure 4 or figure 5.

5. Conclusion
Many of the researches are single dimensional i.e. only one major parameter is used in the prediction process. There are many hybrid or multilayer researches, but again they fail to cover all the parameters that affect the stock price / stock price index. Most of the work have not considered the fundamental parameters and macro-economic variables that significantly affect the stocks. One factor that cannot be ignored at present is the impact of social media data on stock market. Clearly the research that tries to cover all the parameters is missing. Moreover, when all parameters are considered, it needs to be checked whether a multilayer or hybrid approach is better. If possible based on historical evidence weights can be given to parameters and it needs to be evaluated from time to time. If possible, the research should focus on customized set of parameters and weights for each different stock to improve the accuracy.

References
[1] Rout A K, Biswal B, and Dash P K 2014 A hybrid FLANN and adaptive differential evolution model for forecasting of stock market indices International Journal of Knowledge-Based and Intelligent Engineering Systems 18(1) 23–41
[2] Rather A M, Agarwal A, and Sastry V N 2015 Recurrent neural network and a hybrid model for prediction of stock returns Expert Systems with Applications 42(6) 3234–3241
[3] Mundra A, Mundra S, Verma V K, and Srivastava J S 2020 A deep learning based hybrid framework for stock price prediction Journal of Intelligent and Fuzzy Systems 38(5) 5949–5956
[4] Nayak A, Pai M M M, and Pai R M 2016 Prediction Models for Indian Stock Market Procedia Computer Science 89 441–449
[5] Nair B B, and Mohandas V P 2015 An intelligent recommender system for stock trading Intelligent Decision Technologies 9(3) 243–269
[6] Vanipriya C H, and Thammi Reddy K 2014 Indian Stock Market Predictor System Advances in Intelligent Systems and Computing 249 17–26
[7] Kaur G, Dhar J, and Guha R K 2014 Stock market forecasting using ANFIS with OWA operator Int. J. Artif. Intell. 12(2) 102-114
[8] Patel J, Shah S, Thakkar P, and Kotecha K 2015 Predicting stock market index using fusion of machine learning techniques Expert Systems with Applications 42(4) 2162–2172
[9] Srinivas K, Srinivasa Rao V, and Sreemalli M 2018 Binarized genetic algorithm with neural network for stock market prediction ACM International Conference Proceeding Series 8–11
[10] Kumar M, and Thenmozhi M 2014 Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN and ARIMA-random forest hybrid models International Journal of Banking, Accounting and Finance 5(3) 284–308
[11] Abe N, Institute of Electrical and Electronics Engineers, & IEEE Computer Society. (n.d.). 2018 IEEE International Conference on Big Data : proceedings : Dec 10 - Dec 13, 2018 Seattle WA USA
[12] Kohli P P S, Zargar S, Arora S, and Gupta P 2019 Stock prediction using machine learning algorithms In Advances in Intelligent Systems and Computing 698 405–414
[13] Charkha P R 2008 Stock price prediction and trend prediction using neural networks Proceedings - 1st International Conference on Emerging Trends in Engineering and Technology ICETET 2008 592–594
[14] Nayak R K, Mishra D and Rath A K 2015 A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices Applied Soft Computing Journal 35 670–680
[15] Suganya R, and Sasi Kumar A 2020 Stock price prediction using tech news based soft computing approach International Journal of Advanced Trends in Computer Science and Engineering 9(2) 2049–2054
[16] Nivetha R Y, and Dhaya C 2017 Developing a Prediction Model for Stock Analysis Proceedings - 2017 International Conference on Technical Advancements in Computers and Communication ICTACC 2017-2017-October 1–3
[17] Gupta S, Bandypadhyay G, Biswas S and Upadhyay A 2019 A Hybrid Machine Learning and Dynamic Nonlinear Framework for Determination of Optimum Portfolio Structure In Lecture Notes in Networks and Systems 74 437–448
[18] Nanda S R, Mahanty B, and Tiwari M K 2010 Clustering indian stock market data for portfolio management Expert Systems with Applications 37(12) 8793–8798
[19] Nadu T, Nadu T, and Nadu T 2015 Cuckoo Optimized SVM for Stock Market Prediction
[20] Upadhyay V P, Panwar S, Pancariya R and Merugu R 2016 Forecasting stock market movements using various kernel functions in support vector machine ACM International Conference Proceeding Series 12-13-August-2016