Empirical Study on Stock Market Prediction Using Machine Learning

Rachna Sable
Bennett University,
Department of Computer Science and Engineering,
Greater Noida, U.P. India
rachna.sable@gmail.com

Dr. Shivani Goel
Bennett University,
Department of Computer Science and Engineering,
Greater Noida, U.P. India
Shivani.goel@bennett.edu.in

Dr. Pradeep Chatterjee
Head Digital Transformation
Change Management & Customer Experience
GDC,Tata Motors,Pune
pchats2000@yahoo.com

***Abstract* — Stock market prediction is a crucial and challenging task due to its nonlinear, evolutionary, complex, and dynamic nature. Research on the stock market has been an important issue for researchers in recent years. Companies invest in trading the stock market. Predicting the stock market trend accurately will minimize the risk and bring a maximum amount of profit for all the stakeholders. During the last several years, a lot of studies have been done to predict stock market trends using Traditional, Machine learning and deep learning techniques. This survey will assist the readers & researchers in selecting algorithms that can be useful for a predicting the stock market. A survey of various algorithms and its parameters for stock market prediction is presented in this paper.**

Keywords— Stock market prediction, machine learning, SVM, ARIMA, DAN2 Naïve Bayes, KNN, RBF

I. 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[1]. The Stock market trend varies due to several factors such as political, economics, environment, society, etc. [2][3][18][19]. There are two types of stock analysis. One is a fundamental analysis, which requires study of the company's basics such as balance sheet, expenses and revenues, annual returns, company's profile, and position, etc. The other one is a technical analysis ,which deals with studying the statistics generated by market activities such as historical data, past price, and volumes [4]. There are two essential theories used in conventional approach for stock market prediction namely Efficient Market Hypothesis (EMH) introduced by Fama in 1964 which states that stock price future is unpredictable based on the historical data [5] and Random Walk Hypothesis (RWH) which states that stock's future price is independent of its history. Tomorrow's stock price has nothing to do with today's worth, but tomorrow's information [2][26]. The objective of this paper is to compare the traditional , Machine and Deep learning algorithms and assist the researchers for further analysis.

The rest of the paper is organized as follows. In part 2, similar work is discussed based on various Traditional, Machine learning and Deep learning algorithms used for stock prediction. In the second part of related work, details about widely used datasets, various evaluation metrics , and features used in the stock market are discussed. In part 3, the number of technical Indicators used for stock market prediction over different periods are highlighted. Conclusion ise presented in section 4.

I. RELATED WORK

This section discusses the similar work for stock market prediction, implemented by various authors using different algorithms. There are many datasets that can be used for data related to stocks. For example, S&P 500, NASDAQ, Karachi Stock Exchange (KSE), London Stock Exchange (LSE), NewYork Stock Exchange (NSE). Cao and Tay [1] used the S&P 500 dataset. Time series forecasting is an essential area in which variables past observations are collected and analyzed for development of the model, to describe the underlying relationships, so that the model can be used to predict the future [10]. ARIMA is one of the widely used time series models due to its statistical properties, and it can represent various types of time series like Pure Autoregressive (AR), Pure Moving Average (MA) and combined AR and MA (ARMA) series [17]. Univariate model ARIMA was parametric and based on the assumption that the nature of time series is linear and stationary. ARIMA failed to capture the nonlinear patterns because it is a linear model. To overcome this limitation , ANN was used due to its flexible nonlinear capabilities [17]. ANN's performance was found to be satisfactory because it could rely on more considerable information, technical indicators, fundamental indicators, etc. to make the predictions. But ANN had some critical issues like

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

overfitting , which lead to poor generalization and out of sample data. To resolve the mentioned issues the procedures such as Cross -Validation, Non-parametric Probability Density Estimation [6], Modifying Training Algorithms and Pruning parameters or hidden notes[7] were used.

To take advantage of the unique strength of both the models, Zhang proposed a hybrid model comprising of combination of both ARIMA and ANN[10]. In hybrid model, ARIMA was used to analyze the linear part of the problem, and ANN was used to model the residuals provided by ARIMA. By using a hybrid model , one could model linear and nonlinear patterns separately and then combine the forecast with improving the model's forecast and performance. Three datasets namely, Wolf's sunspot data, the Canadian lynx data, and British Pound/US dollar exchange rate data were used. To forecast the accuracy performance, MSE and MAD were selected. Based on the evaluation measures, it was concluded that the hybrid model outer performed each component model used separately. Another version of ANN, Dynamic ANN (DAN2), was used by Guresen et al.[12] using Multilayer perceptron (MLP) and Hybrid Neural Network- which used Generalized Auto-Regressive Conditional Heteroscedasticity (GARCH) to extract new features. The dataset used as NASDAQ, and the evaluation of the models was done by using two evaluation metrics namely, MSE and MAD. The paper concluded that the overall results showed that the ANN model MLP gave the best and reliable result in time series forecasting , whereas the hybrid model failed to improve the outcome.

Amin Hedayati, Moein[27]in their study investigated the ANNs ability to forecast the stock market. The daily NASDAQ Dataset was used for investigation. Two types of input dataset was used one was four prior days and another was nine prior days. The Paper discussed the various Training and Transfer functions and 16 different network architectures using the same were built and the two datasets were tested. The author concluded that there is no difference between the prediction ability of the four and nine prior working days as input parameters.

SVM improved the generalization property of neural networks with the introduction of insensitive loss function and cost penalty[8]. So SVM was modified to solve non-linear regression problems. The objective of the paper was to examine SVM's feasibility and functional characteristics in financial forecasting. The comparison of SVM with MLP trained by Back Propagation (BP) algorithms was done. The dataset S&P 500 daily index in the Chicago mercantile was used. Two groups of variables were taken as inputs to the neural network. The first group G1 was constructed from four lagged Relative Difference in Percentage (RDP) value based on five-day periods, and the transformed closing price was obtained by subtracting a fifteen-day exponential moving average to eliminate the trend in price. The second group G2 was obtained by adding another three technical indicators, the Moving Average Convergence Divergence (MACD), On Balance Volume (OBV) and Volatility. The test was conducted for both the groups of datasets, and the results were compared. The authors concluded that the forecast was better with SVM. but the numbers of free parameters were less in SVMs as compared to BP.

Osman and Mustafa [9] used the same dataset to predict the stock market price using Particle Swarm Optimization (PSO) and Least Square Support Vector Machines (LS-SVM). Traditional SVM was reformulated, and the LS-SVM algorithm was presented. LS-SVM comprised of least square function with equality constraints, leading to a linear system which meets the KKT conditions to get an optimal solution. The methodology to accurately select the LS-SVM must be robust against noise and will not need prior user knowledge about the free parameter influence. The proposed model was based on the study of historical data, technical indicators and optimizing LS-SVM with PSO algorithm to make daily stock price prediction. For comparison of LS-SVM and LS-SVM PSO model LM algorithm was used a benchmark. The architecture model consisted of six input vectors representing the historical data and derived technical indicators and one output representing the next price. The algorithm was tested for many companies that cover all stock sectors in the S&P 500 stock market. LS-SVM PSO model converged into a global minimum and was able to overcome the overfitting found in ANN. The parameters of LS-SVM PSO could be tuned quickly and, the performance of the model was better than LS-SVM and various algorithms compared. It also achieved the lowest error value.

SVM was used with Radial Basis Function(RBF) to predict the stock prices for large and small capitalizations, in three different markets, using price feature for predicting both daily and up-to-date minute frequencies[16]. SVM outputted the optimal hyperplane , which categorized new examples. Weka and YALE data mining environments were used for experimenting . The dataset used was IBM Inc , and the features used were price volatility, price momentum, sector volatility, and sector momentum. SVM was found to work well on massive datasets and did not give an overfitting problem and generated better results.

Zhu et al. [11] tried to investigate whether the trading volume features could improve the neural network performance under short, medium , and long-term forecasting scale. Component- based three-layer Feed Forward Neural Network (FFNN) was employed , and trading volume under different input selection schemes with the basic models was set up as three augmented network models to test if trading volume could significantly improve the performance. Three datasets namely, NASDAQ, DJIA , and STI, were used to investigate the impact of trading volume on daily, weekly and monthly predictions. The observations of the experiment were that trading volume could not fundamentally improve the forecasting performance for the short term, in contrast it modestly improved for the medium term, but the augmented model gave the best prediction for the long term.

Khan et al. [13][23][24][25] studied the effect of applying the Principal Component Analysis (PCA) to various machine learning classifiers and they checked the errors and accuracy of the classifiers at both times i.e., before and after applying PCA. The prediction of the future trends was performed on three different stock markets, namely the Karachi stock market, London stock market , and New York stock market. The three machine learning classifiers namely , SVM, K Nearest Neighbor (KNN), Naïve Bayes(NB), were employed. SVM increased the margin between the data and the hyperplane and tried to reduce the generalization error. It provided sparseness as it used a very less number of support vectors for finding class labels. KNN is the simple machine learning classifier in which the stock prediction problem is mapped into a classification based on similarity. Training and Testing data is mapped into vectors, which represent dimensions of features. The Euclidian distance was used to find the distance, and the clusters were made based on similarity and decisions were made. Naïve Bayes is a statistical classifier as it can make the prediction of class membership-based on probabilities. It is based on Bayes theorem and makes use of class conditional independence in which attributes are independent of each other given the class. After applying classifiers on all the given features, PCA was used for feature reduction , and again, all the classifiers were employed for testing. The evaluation measures used were RMSE and MAE. From the experiment, the conclusion drawn was that KNN had the highest accuracy and lowest MAE as compared to SVM and NB.

Tang and Chen [2][18][26] proposed a hybrid method of combining both historical prices and news as input to the model for predicting stock prices. RNN-LSTM for time series and CNN for high dimensional data were combined to make prediction. RNN is a form of neural network which has a looping/feedback structure, which gives it the ability to work with temporal data. RNN comprises of three layers: one input layer, the recurrent layer, and one output layer. The output of the first state is passed to the next state in combination with new input. The network was trained based on the comparison between the prediction and, target due to which RNN suffers from vanishing gradient problem and the training process took a very long time. To overcome this, LSTM was used where the hidden layers were replaced by special units/individual units called memory blocks also called memory cells. These were used to store the temporal state and multiplicative unit called gates. The input, output , and forget gate were used to control the information flow. In normal FFNN, the adjacent layers are fully connected, and if the input vector has a huge dimension, it could cause an over fitting problem. To overcome this problem, CNN was used. It had a special unit called filter or kernel to replace this part , and hence , each unit of output layer was connected to small portion of the input. Therefore, it required less memory and was faster to train, as it focused on a smaller region. The DJIA was used as a dataset ,which was divided into two parts Historical Data, and News every day’s world news headlines crawled from Reddit World News Channel[https://www.reddit.com/r/worldnews].

Top 25 headlines were considered for a single date. The evaluation metrics used were Mean Accuracy. The paper concluded that the performance of the hybrid model outer performed the other models that take historical data as the input feature only.

Hiransha et al.[14][21][22]implemented four deep learning architecture namely, MLP, RNN, LSTM, and CNN ,for predicting the company’s stock price based on historical prices available. Two datasets NSE and NYSE were used , and the day wise closing price was recorded. The models were trained on Tata Motors company’s data from NSE and for testing five companies from NSE, such as automobile, financial, IT sector were selected, and financial, petroleum sectors were chosen from NYSE for prediction. The Mean Absolute Percentage Error (MAPE) was used as an evaluation metric for the result obtained. The performance of ARIMA and Neural network for a specific time period was compared , and it was stated that neural network architecture performance was better than that of ARIMA.

Perwej et al. [15] used two models namely Particle Swarm Optimization (PSO) and Least Mean Square (LMS) and comparison of both the models was made using historical data. The dataset used was the Bombay stock exchange (BSE). The features of the dataset used were ending price, opening price, the lowest and highest price in the day , and total volume of stock traded each day. The model was used to predict the ending price on each day of the prediction period. The models were compared using the MAPE evaluation metric. The paper concluded that deep learning has excellent capability to extract features from a large set of raw data without trusting the prior knowledge of predictors , and this makes them convenient for stock market forecast. PSO algorithm works optimal during the higher number of days ahead prediction.

Omer, Murat [28] used Genetic algorithm and DNN to propose a stock trading system for creating buy-Sell points based on Optimized technical parameters. In GA the best RSI value for buy-sell points in downtrend & uptrend are found and passed to MLP and SMA is used to show the trend. Dow Jones 30 stocks are evaluated using financial evaluation. The author concluded that more technical parameters can be combined to get better results.

Tian Xia, Qibo Sun [29] focused on feature selection problem and proposed a feature selection algorithm by extending GA. Daily data of CSI 300 index is used. GA is used for feature selection and SVM is used for classification. With proper feature selection the classification accuracy can definitely be improved.

Xi Zhang, Siyu Qu [30], focused on the importance of using multiple data source which can help in improving the prediction of stock market. They proposed to combine the events, sentiments & quantitative data and state the importance of each data source their correlation and effect on prediction accuracy. Multi source Multiple Instance model was proposed for the prediction and concluded that the model not only integrates the heterogeneous data source but their effect on prediction of stock market is improved.

Xu Jiawei, Tomohiro [31], these author proposed the feature selection for selecting useful stock indexes. The model worked on Quantitative & Qualitative data by using Long Short Term Memory (LSTM).

A. Datasets, evaluation matrix and features

The evaluation metrics used for stock prediction are NMSE,MAE,DS,CP,CD, MSE and MAD. Stock prediction is made based on multiple features like closing price, days, relative frequency, trading volume, price volatility, stock momentum, index volatility, index momentum. Various algorithms, evaluation metrics , and features used by different datasets are summarized in table 1.

TABLE I. COMPARATIVE OF DATASET, ALGORITHM, FEATURES AND EVALUATION METRICS USED

Sr. No	Dataset Used	Algorithms used	Evaluation Metrics Used	Features
1	S&P 500	MLP-BP SVM	NMSE MAE DS CP CD	Days, Relative Frequency
2	S&P 500	NN-BP LS-SVM LS-SVM PSO	MSE	Days, Closing Price
3	The Wolf's sunspot data the Canadian lynx data British Pound/US dollar exchange rate data	ARIMA ANN Hybrid	MSE MAD	Days , Price
4	NASDAQ DJIA STI	Component- based three- layer feed forward neural network	One-step Sign prediction rate MSE	Trading Volume, Price
5	NASDAQ	MLP DAN2 GARCH	MSE MAD	Day, Price
6	Karachi, London and New York stock exchange	KNN SVM Naïve Bayes	MAE RMSE Accuracy	Date, Open, Low, High, Close, Volume, Trend, Sentiment & Future trend value
7	DJIA	RNN-LSTM CNN	Mean Accuracy	Price and Date , News
8	NSE,NYSE	MLP RNN LSTM CNN	MAPE	Closing Price , Days
9	BSE	PSO LSM	MAPE	Closing Values, Days
10	IBM Inc	SVM Radial Basis Function	Log2c Log2g	Price volatility, Stock Momentum, Index volatility, Index Momentum

It is clear from Fig 1., MSE is the most widely used metrics. Mean accuracy, MAPE and MAE are also used by most of the researchers for measuring the performance of stock prediction.

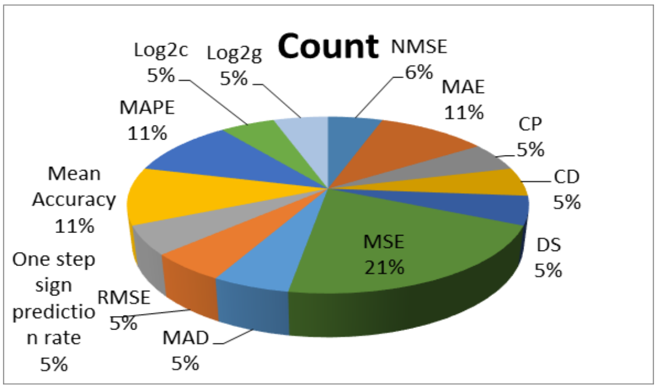

Fig 1. Percentage of Evaluation Metrics Usage from 2010 to 2018

B. Technical Indicators Used

TABLE 3: YEARWISE TECHNICAL INDICATORS USED

Year	Technical Indicators
2019	SMA,EMA,OBV,ADO
2018	MA,SMA,WMA,EMA,SD,ROC,MACD, RSI,ADO,TR,ATR
2017	R,RSI,SMA,SO
2016	MA,RSI,BB,OBV
2015	MA,EMA,RSI,BIAS,MACD,SO,BB,ROC,TR,OBV

CONCLUSION

This paper aims to study the stock market prediction using multiple Traditional, Machine learning, and Deep learning algorithms. Along with the algorithms, the survey has focused on various datasets used for stock market prediction, features of these datasets selected as input parameters and the evaluation metrics used for comparing the results of predictions.

REFERENCES

[1] Lijuan Cao and Francis E.H. Tay, “Financial Forecasting Using Support Vector Machines”, 2001 Springer-Verlag London Limited.

[2] Jinqi Tang, Xiong Chen, “Stock Market Prediction Based on Historic Prices and News Titles”, ACM ISBN 978-1-4503-6432-4/18/05.

[3] Billah, M., and Waheed, S. 2017. “Stock market prediction using an improved training algorithm of neural network”, International Conference on Electrical, Computer and Telecommunication Engineering. pp. 1-4. IEEE.

[4] Nischal Puri, Avinash Agarwal, Prakash Prasad, “A Survey on Machine Learning Approach for Stock Market Prediction”, Helix Vol. 8(5): 3705- 3709.

[5] Chang Sim Vui, Gan Kim Soon, Chin Kim On, and Rayner Alfred, “A Review of Stock Market Prediction with Artificial Neural Network (ANN)”, 2013 IEEE International Conference on Control System, Computing and Engineering, 29 Nov. - 1 Dec. 2013, Penang,Malaysia..

[6] Pan ZH, Wang XD. “Wavelet-based density estimator model for forecasting”, J Computational Intelligence in Finance 1998; 6:6–13.

[7] Dorsey R, Sexton R. “The use of parsimonious neural networks for forecasting financial time series”, J Computational Intelligence in Finance 1998; 6: 24–30.

[8] Vapnik VN. “The Nature of Statistical Learning Theory”, New York, Springer-Verlag, 1995.

[9] Osman Hegazy , Omar S. Soliman , and Mustafa Abdul Salam, “A Machine Learning Model for Stock Market Prediction”, International Journal of Computer Science and Telecommunications volume 4, Issue 12, December 2013.

[10] G. Peter Zhang, “Time series forecasting using a hybrid ARIMA and neural network model”,2002 Elsevier Science B.V. All rights reserved.

[11] Xiaotian Zhu, Hong Wang , Li Xu , Huaizu Li, “Predicting stock index increments by neural networks: The role of trading volume under different horizons”, 2007 Elsevier Ltd. All rights reserved.

[12] Erkam Guresen, Gulgun Kayakutlu ,Tugrul U. Daim , “Using artificial neural network models in stock market index prediction”, 2011 Elsevier Ltd. All rights reserved.

[13] Khan, W. Ghazanfar, M. A. Asam, M, Iqbal, A. ,Ahmad, S. ,Javed Ali Khan, “Predicting trend in stock market exchange using machine learning classifiers”, Sci.Int.(Lahore),28(2),1363-1367,2016.

[14] Hiransha M, Gopalakrishnan E.Ab, Vijay Krishna Menon, Soman K.P, “NSE Stock Market Prediction Using Deep-Learning Models”, 1877- 0509 © 2018 The Authors. Published by Elsevier Ltd. and Data Science (ICCIDS 2018).

[15] Asif Perwej, K. P. Yadav, Vishal Sood, Yusuf Perwej, “An Evolutionary Approach to Bombay Stock Exchange Prediction with Deep Learning Technique”, IOSR Journal of Business and Management(IOSR-JBM)

[16] V Kranthi Sai Reddy, “ Stock Market Prediction Using Machine Learning”, International Research Journal of Engineering and Technology, Volume: 05 Issue: 10 | Oct 2018

[17] G.E.P. Box, G. Jenkins, “Time Series Analysis, Forecasting and Control, Holden-Day”, San Francisco, CA, 1970.

[18] Chen. K. Zhou. Y., & Dai, F. 2015. A LSTM-based method for stock returns prediction: A case study of China stock market. IEEE International Conference on Big Data. 2823-2824. IEEE.

[19] Mahalakshmi, G., Sridevi, S,. & Rajaram, S. 2016. A Survey on forecasting of time series data. International Conference on Computing Technologies and Intelligent Data Engineering. 1-8. IEEE.

[20] S. SElvin, R. Vinayakumar, E.A. Gopalakrishnan, V.K. Menon and K.P. Soman. (2017) “Stock Price Prediction using LSTM, RNN and CNN- sliding window model.” International Conference on Advances in Computing, Communications and Informatics; 1643-1647.

[21] Rather A. M., Agarwal A., and Sastry V. N. (2015). :Recurrent neural network and a hybrid model for prediction of stock returns.” Expert System with Applications 42(6): 3234-3241.

[22] Zhang G., Patuwo B. E., and Hu M/ Y. (1998). “Forecasting with artificial neral networks: The State of Art.” International Journal of Forecasting 14(1): 35-62.

[23] Mittermayer, Marc-Andre. “Forecasting intraday stock price trends with text mining techniques.” System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on IEEE, 2004.

[24] Schumaker, Robert P., H. Chen. “Textual analysis of stock market prediction using breaking financial news: The AZFin Txt System.” ACM Transactions on Information Systems (TOIS) Volume 27, Issue 2, pp:12, 2009.

[25] Shen, Shunrong, H. Jiang, T. Zhang. “Stock Market Forecasting using machine learning algorithms.” , 2012.

[26] O, Bernal. S, Fok. R, Pidaparthi. 2012. Financial Market Time Series Prediction with Recurrent Neural Networks. Citeseer.

[27] Amin Hedayati Moghaddam, Moein Hedayati Moghaddam, Morteza Esfandyari ,” Stock Market Index Prediction Using Artificial Neural Network”,2077-1886© 2016 Universidad ESAN, Elsevier.

[28] Omer Berat Sezera,c, Murat Ozbayoglua1, Erdogan Dogdub ,” A Deep Neural-Network Based Stock Trading System Based on Evolutionary Optimized Technical Analysis Parameters “, Procedia Computer Science 114 (2017) 473–480.

[29] Tian Xia, Qibo Sun, Ao Zhou, Shangguang Wang, Shilong Xiong, Siyi Gao, Jinglin Li, Quan Yuan, «Improving the Performance of Stock Trend Prediction by Applying GA to Feature Selection", 2018 IEEE 8th International Symposium on Cloud and Service Computing(SC2).

[30] Xi Zhang 1, (Member, Ieee), Siyu Qu1, Jieyun Huang 1, Binxing Fang1, And Philip Yu2, “Stock Market Prediction via Multi-Source Multiple Instance Learning”, 2169-35362018 IEEE. Translations