IEEE-SEM, Volume 9, Issue 11, November-2021
ISSN 2320-9151

Comparitive Study of Machine Learning Algorithms for Stock Market Prediction

S.S. Sambare[a], Utkarsh Pathak[b], Venu Sonavane[c], Mitali Gadiya[d], Suyash Musale[e]
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India

Email: [a]santosh.sambare@pccoepune.org, [b]pathak_utkarsh@outlook.com, [c]sonavanevenu@gmail.com, [d]mitaligadiya07@gmail.com, [e]suyashdmusale@gmail.com

***Abstract*— An attempt to determine the future of a particular stock is known as Stock Market Prediction. The ever-increasing number of variations in stock market parameters has made prediction complicated. So, to overcome this issue various machine learning algorithms are used to create relations in the not so obvious trends found in stock market for predicting future of stock market. For stock market prediction, the paper explores four machine learning techniques: Support Vector Machine (SVM), Random Forest (RF), Long Short-Term Memory (LSTM), and Artificial Neural Network (ANN). The Random Forest Algorithm has higher accuracy than the other algorithms.**

Keywords—Machine Learning, Unsupervised Machine Learning, Supervised Machine Learning, Reinforced Machine Learning, Random Forest, data sparseness.

I. INTRODUCTION

Initially, stock trading was just a method to share risks and profits between people and other companies by financing huge projects such as trading voyages, huge construction projects in a group. But now, even though the main motive behind stock trading remains the same the scale at which it performs is massive. The ease with which an individual can contribute with even a small sum of money has changed the dynamics of stock trading from a luxury to a necessity.

The complexity of the stock market makes it intimidating for people to dive into it. But it is observed that the major difference in financial traders, non-trading bank employees and non-experts is the higher risk-taking ability in financial traders than non-trading bank employees than non-experts [1]. By helping people with an unbiased opinion can help them in their decision-making process. This is possible with the use of machine learning. The results of machine learning algorithms in stock markets prediction are based upon interpreting patterns based upon past information.

There are various machine learning classification and regression algorithms such as Linear Regression, Logical Regression, Decision Trees, Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), etc., which are used to prediction stock market trends with the help of data collected by fundamental and technical analysis.

Fundamental Analysis has a high emphasis on the institution's earnings, assets, liabilities, sales, revenue, and various other information required to evaluate a specific institution. Technical Analysis, on the other hand, is the study of the market through the perspective of investors. "Fundamental analysis presumes the prospects of a security are best analysed through a proper assessment of the intrinsic value of the underlying company while technical analysis believes that the markets are efficient at discounting even future developments, price moves through trends, investors are both logical and emotional creatures, and past behaviours tend to repeat themselves more so when enough time has elapsed that the behaviours have been forgotten"- Titus SUCIU. [2]

II. LITERATURE REVIEW

Investor behaviour is a major topic which indirectly affects stock market prediction . According to a study conducted by Dr. Mohammad Shafi "some of the most common factors that have a noteworthy impact on the investor's behavior are herding, over-reaction, cognitive bias, irrational thinking, confidence (over or under), gender, age, income, education, risk factor, dividends, influence of people's opinion (friends or family) past performance of the company, accounting information, ownership structure, bonus payments, expected corporate earnings"[3]. Gurjar M., Naik P., Mujumdar G., & Tejaswita Vaidya according to their research suggest that Artificial Neural network is a technique that is used for prediction because "they are able to run nonlinear mappings between input and outputs. It is possible that ANN outperforms traditional analysis like Linear Regression"[4]. The study conducted by Kunal Pahwa and Neha Agarwal found that, the accuracy of the algorithm is directly proportional to the amount of data provided and the attributes selected for training [5]. This idea is backed up by the findings of the study conducted by Indu, Kiran, Chetna, Premlata, "reduction in the number of technical indicators reduces the accuracy of each algorithm in predicting the stock market trends."[6]. The accuracy can be improved by removing the data sparseness and increasing the amount of data used to train the machine learning algorithm [7]. Furthermore, the use of people's sentiments can be used to increase the accuracy of the prediction model and if for a particular day the dataset for sentiment analysis which can be news or tweets on Twitter is less, then Principle Component Analysis with numerous factors to solve the problem. [8]. Text mining with the help of Random Forest algorithm can be used to extract critical indicators, and classification of related news articles. By using unigram

features, the Random Forest classifier resulted in 98.34% classification accuracy. [9].

A. *Machine Learning*

Machine Learning is a group of techniques applied on certain devices that uses computer science fundamentals to support these devices to educate themselves from the data provided. It creates an effective model by testing various solutions against the data available to find the perfect fit for the given problem.

The "application of machine learning techniques is used to model the data and identify the hidden patterns in which the data is behaving (regression and classification results)".[10]

Depending on the different expected results, various algorithms are developed which come under the three subsections supervised machine learning, unsupervised machine learning, and reinforcement machine learning are three types of machine learning.

B. *Supervised Machine Learning*

Supervised machine learning is a method that takes a group of data with known characteristics (Labelled Dataset) and creates a model for future prediction. The data provided for training the model is analysed and the model is adjusted to fit the trends appropriately to get a final model which is neither overfitting nor underfitting for the data. Overfitting is a state in which the model satisfies the even training data perfectly. This results in extremely high accuracy during the training but test accuracy reduces. Underfitting is when the error is high during the training period itself. "A classifier with an accuracy less than 95% is practically useless." [5]- Kunal, Neha.

Supervised Machine learning includes different methods like Linear regression, Naïve Bayes, Logistical Regression, Support Vector Machine, Random Forest.

Random Forest Algorithm

Random forest is an example of a supervised machine learning algorithm that is used to solve regression and classification problems and is based on decision trees. It was first implemented in 1995 by Tin Kam Ho. Forest is a fast and easy-to-understand machine learning method that employs ensemble tactics and random sampling to make accurate predictions for a variety of datasets. [9]. The input data is fit into random subsets using multiple Classification and Regression Tree (CART) models in a random forest, and the forest's aggregated result is used for prediction. Random forests also take into account the weights of each independent variable when modelling the dependent variable. [11]. This helps to increase the accuracy and helps eliminate overfitting conditions which is a major drawback of the decision tree. Thus, Random Forest helps us achieve low bias and low variance.

Consider an example of a fruit bowl consisting of apples, oranges, and grapes, bananas etc. that needs to be sorted. A classical decision tree may accomplish it by differentiating them by shape and size to get the final result. But the same problem can be solved by using multiple decision trees and averaging their results. The different decision trees can look for different parameters like shape, size, texture, growth season etc. and different combinations of these parameters to increase the accuracy of the results.

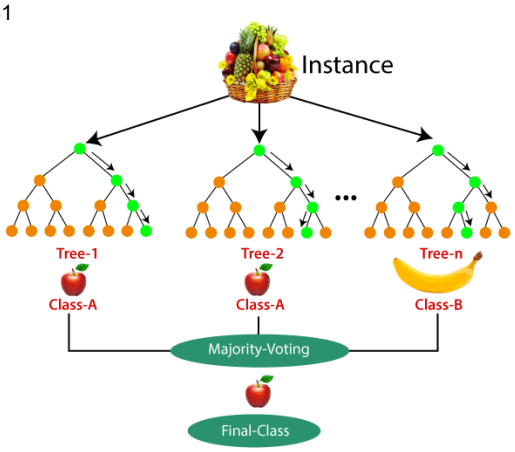

Fig. 1. Example of a Random Forest Classifier (source: JavaPoint)

Long Short-term Memory Algorithm

Long Short-term Memory abbrivated as LSTM is a deep learning artificial neural network algorithm. It was invented in 1997 by Hochreiter and Schmidhuber mainly for purpose of processing, classifying, and making predictions on data. "Long Short Term Memory network (LSTM) is an updated form of Recurrent Neural Network. Apart from the traditional network, there are connections between the layers of neural network." [12] Speech recognition, music composition, time series prediction, and human activity recognition are just a few of its significant applications. A remembering cell, input gate, output gate, and forget gate make up the LSTM. The value is stored in the cell for long-term propagation, and the gates control it. As we know as per fundamental analysis for investing in any stock, we need to consider two factors- the intrinsic values of the company and the prehistoric data of particular stock.

LSTM considers all the prehistoric data of stock before making any prediction, which makes the algorithm more accurate and efficient. The major purpose to use LSTM is that it dissolves the issue of vanishing gradient and it also has strong adaptability to data with different stability[14].The effectivess of the algorithm can be increased by stacking it under different layers. The stacked based algorithm basically increases the efficiency of both training and testing.The major disadvantage of LSTM is that any fluctuations in the prehistoric data can affect the efficiency of the algorithm . So a stable dataset with accurate time series is a must for Long short term memory algoithms to generate optimum results

Methodology of LSTM:

Step 1: Collection of raw data: prehistoric data is collected of a particular stock for estimating future prices.

Step 2: Processing and Cleaning of data:

This step involves:

- Data discretization: Here, reduction of data but with particular priority is carried out , mostly for numeric data.

- Data transformation: Normalization of data.
- Data cleaning: Here all the missing values are filled in.
- Data integration: In this, we integrate all the data files.

After processing data, it is transformed into a clear dataset, then it is splitted into training set and testing set for evaluation. Testing dataset is kept as 10-15 percent of the total dataset.

Step 3: Extraction of features:

The values to be predicated are extracted in this step. Majorly the 6 key values of the price bar are extracted namely: OPEN, HIGH, CLOSE, LOW, VOLUME, RANGE

Step 4: Training the Algorithm

In this step, data is fed into the recurrent neural network algorithm and tested on random iterations. All the layers of the algorithm are tested on basis of the data and results are analysed

Step 5: Output Generation:

The results of the tested data are analysed. the error factors are minimized and the data is again tested to increase accuracy and efficiency.

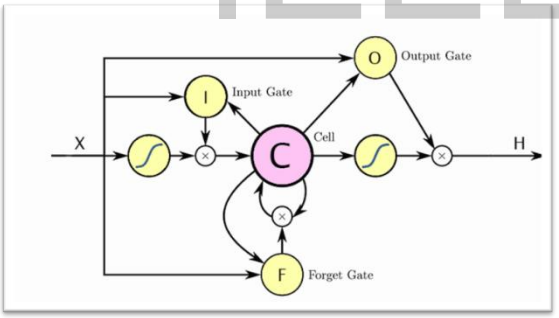

Fig. 2. LSTM Cell Structure (*Source: [8]*)

Support Vector Machine (SVM)

The support vector machine has been shown to be very accurate while using less computing power. A separating hyperplane defines SVM, which is a discriminative classifier. The goal of this technique is to find a hyperplane that categorises the data points in x-dimensional space. Hyperplanes are decision boundaries that help categorise data points on either side of the hyperplane and are assigned to different classes.

SVM is a method generally used to solve non-linear classification and regression problems in time series analysis. As high generalization and testing accuracy is found using the algorithm. The core idea underlying the SVM model is to represent provided samples of data as points in a high-dimensional feature space, with feature vectors linearly separated by a maximum margin hyperplane. [13]

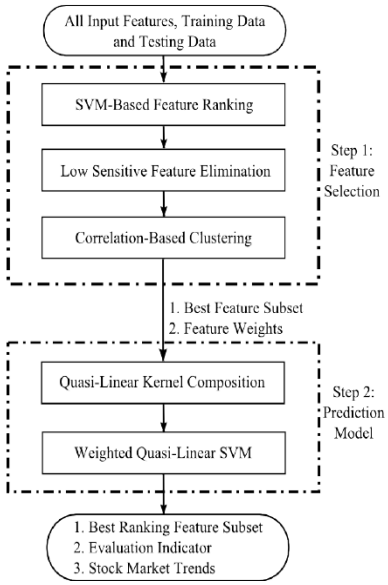

Fig. 3. Flow chart of SVM based stock market prediction

C. Unsupervised machine learning

Unsupervised Machine Learning uses different algorithms to compute and cluster unlabelled data. It tries to find the hidden patterns which may be similarities or differences in the test data to predict future patterns. This makes it a good choice for image recognition, pattern recognition, exploratory data analysis. It can also be used to reduce the number of features of a data model. Examples of algorithms that are under unsupervised machine learning are k- nearest neighbour (KNN), Neural Networks, k- means clustering, etc.

Artificial Neural Network (ANN)

Warren McCulloh and Walter Pitts in 1943 created a computational model that used artificial neural networks (ANN). As the name suggests “artificial neural network” applies biological phenomena in the field of artificial intelligence. “Artificial neural networks (ANN) mimic the human brain in processing input signals and transform them into output signals” [15].

ANNs are computational models based on biological neural networks. In this network, the nodes are grouped into certain layers beginning with an input layer and ending with an output layer. A group of neurons is connected in link to transfer a signal. The nodes learn from the past examples to minimize the errors that occurred in the prediction.

The ANN method uses a technique called “backward propagation”. This technique is used to avoid errors and is a two-phase method and is regularly used by the neurons to acquire near perfection.

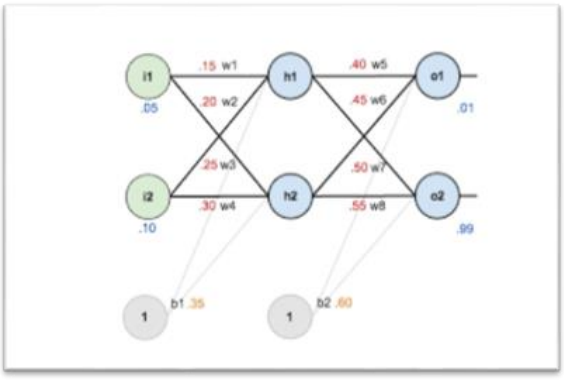

Fig. 4. Working of ANN (*Source: [14]*)

D. Reinforcement Learning

Reinforcement machine learning uses trial and error to train the algorithm to find a solution to the problem. The algorithm receives positive feedback or reward after taking correct action, and negative feedback or penalty on performing bad action. The objective of the algorithm is to receive the most possible reward.

III. ANALYSIS

Accuracy:

The accuracy ratio is the number of correct predictions divided by the total number of forecasts made.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions\ made} \quad (1)$$

Percentage Accuracy:

Percentage accuracy is defined as the ratio of correct predictions to the total number of forecasts made multiplied by 100 to yield a percent value.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions\ made} \times 100 \quad (2)$$

IV. RESULTS

The following table is a summation of evaluation parameters of the individual test results carried out for the SVM [17], Random Forest (RF) [17], LSTM [18], ANN [19] regressor.

TABLE I.

Comparison of Results		
Algorithm	***Accuracy***	***Percentage Accuracy***
SVM	0.846	84.6
RF	0.862	86.2
LSTM	0.781	78.1

Comparison of Results		
Algorithm	***Accuracy***	***Percentage Accuracy***
ANN	0.7451	74.51

Fig. 5. Results

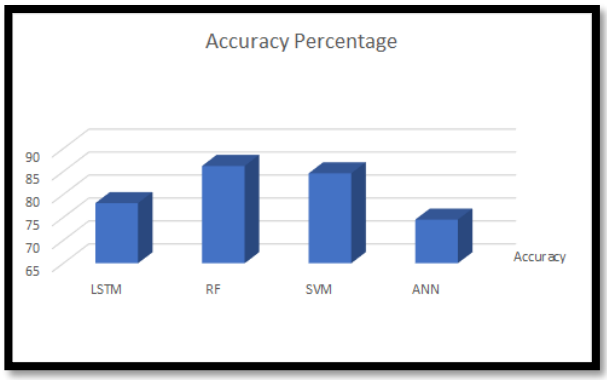

Fig. 6. Graph of Results

V. CONCLUSION

After comparing the accuracy of the four proposed machine learning algorithms namely Long-Short Term Memory (LSTM), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN), it is found that RF outperforms the other algorithms for stock market prediction when provided with large data set. In future, the results can be further improved by comparing the accuracy of the algorithms on a common input data set for training and testing.

REFERENCES

[1] Thoma V, White E, Panigrahi A, Strowger V, Anderson I (2015) Good Thinking or Gut Feeling? Cognitive Reflection and Intuition in Traders, Bankers and Financial Non-Experts. PLoS ONE 10(4): e0123202. https://doi.org/10.1371/journal.pone.0123202

[2] Titus SUCIU, Elements of Stock Market Analysis. Bulletin of the TransilvaniaUniversity of Braşov, Series V:Economic Sciences, Vol. 6 (55), No. 2 -2013.

[3] D.M.S. (2014a). DETERMINANTS INFLUENCING INDIVIDUAL INVESTOR BEHAVIOR IN STOCK MARKET: A CROSS COUNTRY RESEARCH SURVEY. Arabian Journal of Business and Management Review (Nigerian Chapter) Vol. 2, No. 1, 2014.

[4] Gurjar, M., Naik, P., Mujumdar, G., & Tejaswita Vaidya, P. (2018). STOCK MARKET PREDICTION USING ANN. International Research Journal of Engineering and Technology (IRJET), 05(3).

[5] KunalPahwa, Neha Agarwal. Stock Market Analysis using Supervised Machine Learning, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019.

[6] Kumar, K. Dogra, C. Utreja and P. Yadav, "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction,"2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1003-1007, doi: 10.1109/ICICCT.2018.8473214.

[7] Asad Masood Khattak, Ammara Habib, Habib Ullah, Muhammad Zubair Asghar, Hassan Ali Khalid, Fazal Masud Kundi. Stock Market Trend Prediction using Supervised Learning, SoICT 2019, December 2019, Hanoi Ha Long Bay, Vietnam. https://doi.org/10.1145/3368926.3369680

[8] Nayak, Aparna & M M , Manohara & Pai, Radhika. (2016). Prediction Models for Indian Stock Market. Procedia Computer Science. 89. 441 449. 10.1016/j.procs.2016.06.096.

[9] M. N. Elagamy, C. Stanier and B. Sharp, "Stock market random forest-text mining system mining critical indicators of stock market movements," 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), 2018, pp. 1-8, doi: 10.1109/ICNLSP.2018.8374370.

[10] Mondal, Saikat & Dutta, Abhishek & Chatterjee, Piyali. (2020). Application of Deep Learning Techniques for Precise Market Prediction. National Conference on Machine Learning and Artificial Intelligence (NCMLAI Delhi 2020).

[11] N. Sharma and A. Juneja, "Combining of random forest estimates using LSboost for stock market index prediction," 2017 2nd International Conference for Convergence in Technology (I2CT), 2017, pp. 1199 1202, doi:10.1109/I2CT.2017.8226316.

[12] Fei Qian, Xianfu ChenSchool of Information Science and Technology University of Science and Technology of China Hefei, China978-1-7281-1410-1/19/$31.00 ©2019 IEEE

[13] Smola, A. J., & Schölkopf, B.(2004). A tutorial on support vector regression.Statistics and computing,14(3), 199-222.

[14] David M. Q. Nelson, Adriano C. M. Pereira, Renato A. de OliveiraStock Market's Price Movement Prediction With LSTM Neural Networks978-1-5090-6182-2/17/$31.00 ©2017 IEEE

[15] Wesolowski M, Suchacz B. Artificial neural networks: theoretical background and pharmaceutical applications: a review. J AOAC Int 2012;95:652-68. 10.5740/jaoacint.SGE_Wesolowski_ANN.

[16] Gurjar, M., Naik, P., Mujumdar, G., & Vaidya, T. (2018). STOCK MARKET PREDICTION USING ANN. International Research Journal of Engineering and Technology (IRJET), 05(03). https://www.irjet.net/archives/V5/i3/IRJET-V5I3634

[17] S. S. Maini and K. Govinda, "Stock market prediction using data mining techniques," 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 654-661, doi: 10.1109/ISS1.2017.8389253.

[18] LSTM-1.Siyuan Liu1,2 , Guangzhong Liao1,2 ,Yifan Ding1,Stock Transaction Prediction Modeling and Analysis Based on LSTM978-1-5386-3758-6/18/$31.00 c 2018 IEEE

[19] ANN-Chang Sim Vui, Gan Kim Soon, Chin Kim On, and Rayner Alfred A Review of Stock Market Prediction with Artificial Neural Network (ANN) 978-1-4799-1508-8/13/$31.00 ©2013 IEEE