INTERNATIONAL RESEARCH JOURNAL OF HUMANITIES AND INTERDISCIPLINARY STUDIES
(Peer-reviewed, Refereed, Indexed & Open Access Journal)
DOI : 03.2021-11278686 ISSN : 2582-8568 IMPACT FACTOR : 5.828 (SJIF 2022)

INDIAN STOCK PRICE PREDICTION USING MACHINE LEARNING ALGORITHM

Sushant Kumar Mohanty[1] Manish Chandra Roy[2] Tusharkanta Samal[3] Amiya Ranjan Kanungo[2]

[1]Dept of Information Technology, J.K.B.K. Govt. College, Cuttack (Odisha, India)
E-mail: Sushant.mty@gmail.com
[2]Dept. of Information Science and Telecommunication, Ravenshaw University, Cuttack (Odisha, India)
[3]Dept. of Computer Science & Engineering, C.V. Raman Global University, Bhubaneswar (Odisha, India)

DOI No. 03.2021-11278686 **DOI Link ::** https://doi-ds.org/doilink/04.2022-54388925/IRJHIS2204006

Abstract:

Prediction of financial market trends is an extremely important task and one of the biggest challenges for investors as forecasting future asset value of stock prices successfully may lead to an increase in the profit ratio. Stock prices, market influences news, and social media involvements have a great impact on stock market investment. An intelligent model helps investors and traders to increase their profit percentage. Several models of machine learning and artificial neural network have been used in recent years to predict stock price trends and these models help humans to take error-free decisions in a shorter time span. It is necessary to access the algo models through profitability metrics and validate the model's performance (because of the prediction of the price of the high-risk asset). The recently used algorithms are Linear Regression, Logistic Regression, Artificial Neural Network, Random Forest, Support Vector Machine, and K-nearest neighbor. However, many other factors can affect the market and change the price of the stocks.

Keywords: *Machine Learning, Random Forest, KNN, ANN, LSTM, Stock Market, SVM.*

I. Introduction:

Expert analysts and investors have placed a high value on improvements in stock price prediction. Due to intrinsically noisy surroundings and high volatility in relation to market trends, the stock market forecast for evaluating trends is difficult. The complexity of stock prices is influenced by a variety of factors such as quarterly earnings releases, market news, and social media platform comments [1]. It is vital to design a system that will operate with maximum precision and account

for all key factors that may influence the outcome. As a result of the increasing relevance of machine learning in numerous sectors in recent years, several traders have been encouraged to apply machine learning techniques to the industry, and some of them have produced promising outcomes. We will focus on Artificial Neural Networks and Recurrent Neural Networks in this article. Artificial intelligence is used in machine learning, allowing the system to learn and improve from previous experiences without having to be programmed repeatedly. Backward Propagation, often known as Backpropagation Errors, is a technique used in traditional machine learning prediction approaches [2]. Many researchers are increasingly employing ensemble learning approaches.

The stock market prediction tools can play a critical role in bringing new individuals and current investors together in one area. People's mindsets can be changed as a result of the more promising findings of prediction methods. Some of the methods used to predict the stock market include time series forecasting, technical analysis, machine learning modeling, and predicting the variable stock market. The stock market prediction model's datasets contain information such as the closing price, opening price, data, and a variety of other factors that are required to forecast the object variable, which is the price on a particular day. The goal is to create a model that uses machine learning methods to acquire insight into market data and forecast future stock value trends.

The Support Vector Machine may be used for both classification and regression (SVM). In the SVM approach, each data component is plotted as a point in n-dimensional space, with the value of a feature being the value of a certain coordinate, and classification is accomplished by identifying the hyperplane that explicitly distinguishes the two classes. For this, predictive approaches such as the Random Forest methodology are utilised. For classification and regression, the random forest method uses an ensemble learning technique. The random forest takes the average of the dataset's multiple subsamples, which improves prediction accuracy and decreases dataset over-fitting. Stock market movement may be predicted using logistic regression, which can be either a growing trend, an unchanged trend, or a declining trend.

II. RELATED WORK:

The stock market has attracted investors because of improved applications, such as forecasting, which can lead to effective market predictions. Stock trend forecasting is inextricably linked to stock data investing and trading. Several studies have been conducted on integrating machine learning algorithms for stock market forecasting. Based on the stock market prediction, this study suggests techniques such as the Bayesian model, Fuzzy classifier, Artificial Neural Networks (ANN), Support Vector Machine (SVM) classifier, Neural Network (NN), Machine Learning Methods [3]. For each algorithm, the accuracy of the aggregate result is verified. The goal of this taxonomical study is to reduce the mistake rate and improve the accuracy of a prediction model. If invested properly and sensibly, market investments are regarded as the fastest way to profit. Multiple

variables known as Market Risks, however, can influence the stock market and price. This necessitates the tracking and scoring of specific data or entities that reflect these criteria [4]. This focuses on using datasets to apply machine learning techniques including Random Forest, Support Vector Machine, KNN, and Logistic Regression. Performance measures like precision score and recall are used to assess the algorithms [5]. The emphasis of this systematic study is on Deep Learning models used for stock market forecasting utilizing technical analysis. Predictor methods, trading strategies, profitability measures, and risk management were the four primary areas of discussion. The LSTM method is commonly used in this circumstance, according to this study (73.5 percent) [6][13]. The deep learning models were similar, with the only variation being the deep learning approach employed, which was either FF, RNN, LSTM, or BP [7]. This seeks to develop a model for predicting future stock market values using Recurrent Neural Networks (RNN) and, in particular, the Long-Short Term Memory model (LSTM).

The major goal is to explore how accurate a Machine Learning algorithm can predict and how much the epochs can help our model [8]. To achieve the right predictions with respect to current affairs obtained from Twitter, the SVM approach is coupled with sentiment analysis methodology. The combination of these two approaches distinguishes this project from others since it focuses more on current events and has a better accuracy rate [9]. For stock prediction and analysis, linear regression and logistic regression are used, however this study suggests SVM for more accurate findings. A classification-friendly dataset is one that is utilised in prediction. Machine learning algorithms may be implemented using the following tools. All of the tools offer regression and classification methods, and users can use any of them according to their comfort and knowledge [10]. To forecast the direction of the closing price, a variety of machine learning techniques, including deep learning approaches, are used to stock data. Based on the learned findings, this framework may provide an appropriate machine learning prediction technique for each pattern. Ensemble machine learning techniques are used to create the investing strategy[12]. Empirical data from China's stock market from 2000 to 2017 indicate that our feature engineering has an effective predictive ability, with some trend patterns having a forecast accuracy of more than 60%. Data noise may be efficiently solved using a variety of methods, including big data, feature normalization, and the removal of aberrant data [11].

III. PROPOSED SYSTEM:

A. Data Collection:

Data gathering is a basic module and the project's first phase. It mostly concerns the gathering of the appropriate dataset. The dataset that will be used to make market predictions must be filtered in a variety of ways. Data gathering also contributes to the enhancement of the dataset by including more

external data. The dataset contains data from India's National Stock Exchange nifty 50 stocks from 2000 to 2021. Table 1 contains a description of the dataset.

Table -1 Description of dataset

Fields	Descriptions
Date	Details Capture Date
Symbol	Stock symbols used by data provider
Series	Category of the series(EQ, BE, BL, BT, GC, IL)
Prev Close	Previous day closing price
Open	Stock trading starting price
High	Highest price in the entire day trading
Low	Lowest price in the entire day trading
Last	Last Trade price in the entire day
Close	Final settlement price of the day
VWAP	volume-weighted average price of the day
Volume	Total traded volume
Turnover	Total number of shares traded during the day
Deliverable Volume	Net Deliverable Volume is the quantity of shares carry for holding

B. Data filtering:

The data is in an unprocessed state. The dataset must be transformed into an analysis-ready format. Data pre-processing is a subset of data mining that entails converting unstructured data into a more usable shape. Raw data is sometimes inconclusive or partial, and it frequently contains several mistakes. Checking for missing values, looking for categorical values, separating the data set into training and test sets, and lastly doing feature scaling to restrict the range of variables so that they may be compared on common environments are all part of the data pre-processing. It's conceivable that if data isn't normalised, the column with the highest values will be given greater weight in the forecast. To deal with this, we scale the data.

C. Training the Machine:

Feeding the data to the algorithm to clean up the test data is comparable to training the machine. The models are tuned and fitted using the training sets. The test sets haven't been altered because a model shouldn't be assessed based on data that hasn't been seen. Cross-validation is used

during model training to provide a well-grounded estimated performance of the model utilising the training data. The aim of training a model is to identify a set of weights and biases that have a low loss across all cases on average.

D. Classifiers:

Classification is a type of supervised learning in which a group of items is examined and classified based on a common characteristic. When given training data, classifiers build a model. After that, testing data is provided, and the model's accuracy is calculated. When more than one input is provided, classification will attempt to predict one or more outputs. Here are a handful of the classifiers that are utilised are Random Forest Classifier, SVM (Support Vector Machine), KNN (K-nearest neighbour), Logistic Regression, and Artificial Neural Networks.

(a) Random Forest Classifier-

The random forest classifier is a type of ensemble classifier that is supervised. It's a flexible method that can do both regression and classification. It essentially generates a series of decision trees that provide a certain conclusion. Only a subset of characteristics is taken into account in this method. Random Forest has the advantage of being highly successful on huge datasets. It increases the model's predictability, making it more accurate.

(b) Support Vector Machine-

It's a supervised machine learning approach that uses a separator to break instances into groups. The support machine algorithm's main goal is to find an N-dimensional space that categorises input points in a distinct way. N stands for a variety of characteristics. There may be numerous hyperplanes to choose from between two classes of data points. This algorithm's goal is to find a plane with the greatest margin. There must be a maximum margin on this plane. The distance between data points from both classes is referred to as maximising margin. The advantage of increasing the margin is that it gives some reinforcement, making subsequent data points easier to classify.

(c) K-nearest neighbor-

The KNN method is a basic supervised machine learning technique that may be used to tackle classification and regression issues. It's a classification method for comparable scenarios. It only provides results when you request them. Because it just needs to compute the value of k and the Euclidean distance, KNN has the benefit of being one of the simplest algorithms. Because of its slow learning characteristic, it is sometimes quicker than other algorithms. It's ideal for problems with several classes. Data normalisation is required for the KNN algorithm to produce the best results.

Algorithm for KNN:

1. Load the dataset
2. Set the value of k
3. Calculate the distance between test data (m) and each row of training data(n). Here we will

use Euclidean distance as our distance metric.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Where,

xi is the ith element of the occurrence x,

yi is the ith element of the occurrence y,

n is the total number of features in the data set.

4. Get the top k rows from the sorted array

5. Get the most common kind of these rows. (a) Random Forest Classifier

Classification is a type of supervised learning in which a group of items is examined and classified based on a common characteristic.

(d) Logistic Regression-

Logistic regression, on the other hand, predicts the likelihood of an event or class based on other variables. As a result, the result of logistic regression is always between 0 and 1. It is widely used for categorization because of this characteristic. The most accurate results come via logistic regression, but it involves selecting the appropriate feature to fit.

(e) ANN (Artificial Neural Networks)-

Nonlinear statistical models, or ANNs, are nonlinear statistical models that exhibit a complicated relationship between inputs and outputs in order to identify a new pattern. ANN may also produce an output result based on a sample of data rather than the full dataset. Because of their excellent prediction capabilities, ANNs may be used to improve existing data analysis approaches.

IV. RESULTS:

Using the padas data reader library the raw data load from yahoo finance and we are published our findings based on this data. There are mainly seven columns that describe the important daily movement of the stocks. The columns are date, High, Low, Open, Close, Volume, and Adj Close. High means daily high price, Low means daily low price, Open means the first traded price, and Close means the last traded price of the stocks. The volume column represents the total traded volume and the adjusted close price means the price calculated after adjustment for all applicable dividend and split distributions.

	Date	High	Low	Open	Close	Volume	Adj Close
1							
2	2011-01-03	528.242798	521.358032	527.499817	522.843933	4804792.0	478.54913
3	2011-01-04	534.879883	523.685974	525.023315	533.493042	10163093.0	488.29605
4	2011-01-05	539.882446	529.976379	534.929382	532.849121	11972802.0	487.7067
5	2011-01-06	540.575867	532.229980	533.938782	537.703125	10043704.0	492.14954
6	2011-01-07	538.768005	524.032715	535.573303	527.697937	8249359.0	482.99188

Fig. 1. head()

We will initially print the dataset's structure in this phase. To verify that there are no null values in the data frame, we check for them. The presence of null values in the dataset presents problems during training since they act as outliers, causing a broad range of results.

Fig. 2. Time Series Plot based on closing price

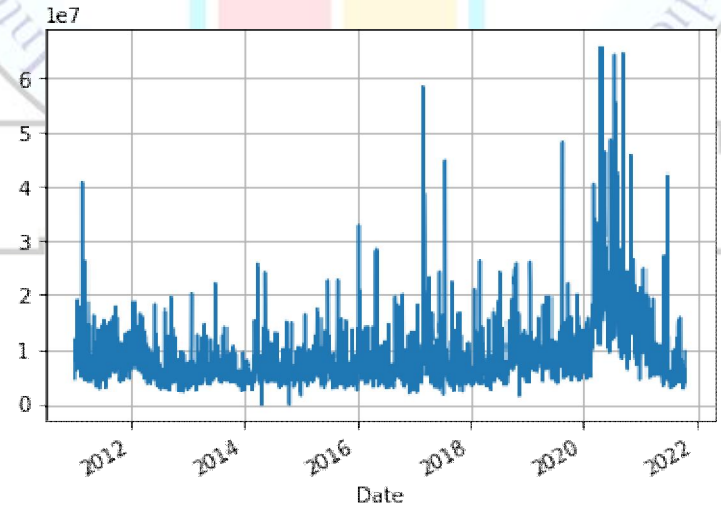

Fig. 3. Time Series Plot based on volume

Matplotlib help to generate time series plot of reliance based on closing price. This is showing the trend of the share for the last 10 years.

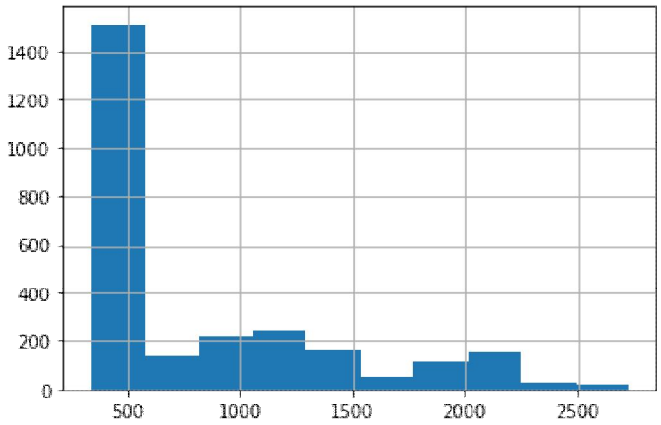

Fig. 4. Historogram Plot

The example above plots the observations in the Minimum Daily Closing price dataset as a histogram. The frequency or count of observations in each bin can reveal information about the data's underlying distribution in a histogram, which splits values into bins. The charting tool selects the size of the bins based on the dispersion of values in the data.

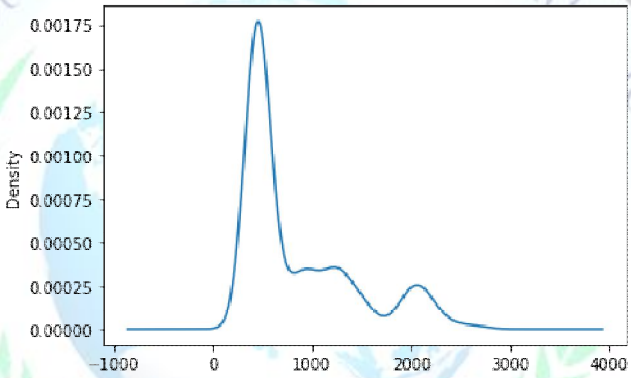

Fig. 5. Density map

A density map can help us understand the form of the distribution of observations. This is similar to the histogram, except instead of using a function to match the distribution of data, a lovely, smooth line is used to describe it.

The Random Forest has the greatest confidence accuracy rate among other machine learning algorithms as per the above dataset.

V.CONCLUSION:

Machine learning techniques for predicting stock market price have been successfully deployed on the dataset. On the dataset, we used feature selection and data preprocessing. On the dataset, we used Random Forest, SVM, KNN, Logistic Regression and ANN. The performance measures (accuracy, precision, recall, and f-score) were used to assess the differences between the algorithms. We observed that the random forest algorithm is the best suited algorithm for forecasting the market price of a stock based on numerous data points from historical data by assessing the accuracy of the different algorithms. Some benefits and limitations also associated with these algorithms. This paper's future scope would include the addition of new parameters that drive stock

market forecasting. Increasing the number of parameters will result in more accurate estimate. Investors will have a better grasp of the situation and be able to make more accurate predictions as a result of this.

VI. FUTURE WORK:

The greater the number of parameters considered, the higher the accuracy. Market news, social media news is also a vital parameter to add as a parameter to increase the accuracy. Various machine learning algorithm used to analysed the market sentimate.

VII. REFERENCE:

[1] Mehar Vijha, Deeksha Chandolab, Vinay Anand Tikkiwalb, Arun Kumar, "*Stock Closing Price Prediction using Machine Learning Techniques*", International Conference on Computational Intelligence and Data Science (ICCIDS), 2019.

[2] Shah D, Isah H, Zulkernine F. "*Stock Market Analysis: A Review and Taxonomy of Prediction Techniques*", Int. J. Financial Stud., 7(2), pp. 1-22, 2019.

[3] Gandhmal DP, Kumar K., "*Systematic analysis and review of stock market prediction techniques*", Computer Science Review, 2019 Nov 1;34:100190, 2019.

[4] Sarvadnya Navti, Dr. Nilesh Fal Dessai "*Quantitative Analysis of Different Algorithms for Stock Price Prediction*", International Research Journal of Engineering and Technology (IRJET), 2020.

[5] Sarvadnya Navti, Dr. Nilesh Fal Dessai "*Quantitative Analysis of Different Algorithms for Stock Price Prediction*", International Research Journal of Engineering and Technology (IRJET), 2021.

[6] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar," *Stock Closing Price Prediction using Machine Learning Techniques*",Procedia Computer Science,Volume 167, 2020

[7] A. W. Li and G. S. Bastos, "*Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review*" in IEEE Access, vol. 8, pp. 185232-185242, 2020.

[8] A. Menon, S. Singh and H. Parekh, "*A Review of Stock Market Prediction Using Neural Networks*" 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019.

[9] Hiransha M, Gopalakrishnan E.A., Vijay Krishna Menon, Soman K.P., "*NSE Stock Market Prediction Using Deep-Learning Models*", Procedia Computer Science, Volume 132, 2018.

[10] A M Pranav, Sujooda S, Jerin Babu, Amal Chandran, Anoop S, 2021, "*StockClue: Stock Prediction using Machine Learning*", INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCREIS, 2021

[11] Singh, Nirbhey & Khalfay, Neeha & Soni, Vidhi & Vora, Deepali. (2017). "*Stock Prediction using Machine Learning*" a Review Paper. International Journal of Computer Applications. 163. 36-43. 10.5120/ijca2017913453, 2017.

[12] Y. Lin, S. Liu, H. Yang and H. Wu, "*Stock Trend Prediction Using Candlestick Charting and Ensemble Machine Learning Techniques With a Novelty Feature Engineering Scheme*" in IEEE Access, vol. 9, pp. 101433-101446, 2021.

[13] Adil MOGHAR,Mhamed HAMICHE, "*Stock Market Prediction Using LSTM Recurrent Neural Network*",International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI), 2020.