INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 6 - V7I6-1138)
Available online at: https://www.ijariit.com

Stock Market prediction made easy with Machine Learning algorithms

Vaibhav Kumar
thejiffy09@gmail.com
Indian Maritime University, Kolkata, West Bengal

Rishabh Raj
rishabhraj.dhn@gmail.com
Aryabhatta College, University of Delhi, Delhi

Abstract: ***The main objective of this paper is to find the best model for predicting the stock market movement. We have tested various models based on machine learning that were previously implemented and during the process we found out that the Random Forest and Support Vector Machine algorithms were not exploited well. In this paper we are going to find out a more feasible method to predict the stock market with a higher accuracy. We have taken dataset of stock market prices from previous years and pre-processed the data for real analysis. So, our paper will also be focusing on pre-processing of the raw dataset. After pre-processing, we will be reviewing the use of random forest and support vector machine on the datasets and the outcome it generates. The paper also examines the feasibility of the prediction system in real-world settings and issues associated with the accuracy of predicting the market. If this model achieves higher accuracy than previously implemented machine learning algorithms then it can prove to be a great asset for the stock brokers, institutions and individual investors.***

Keywords— *Stock Market, Machine Learning, Support Vector Machine*

1. INTRODUCTION

Share Market is basically a microcosm of the demand-supply mechanism which affects economy around the globe. A stock is a small portion of a business which an investor owns. Investors and traders rely on predicting the future of stock markets' movement, either upside or downside, to make profits. The attempt to determine the future movement of stock prices is fundamental to share market. The prediction must be accurate in order to consistently benefit from the stock market. The system must take into account various datasets, variables and real-world events which affect a stock's price movement for accurate prediction. There are various procedures for implementing the prediction system like Technical Analysis, Machine Learning, Market Mimicry, and Time series aspect structuring. With the advent of technology, various advanced methods of prediction have come up lately. The most reliable technique involves the use of Artificial Neural Networks, Recurrent Neural Networks, which is basically the implementation of machine learning. Machine learning involves artificial intelligence which enables the system to learn from earlier available datasets and improve its functioning without being programmed time and again. Traditional methods of prediction in machine learning use algorithms like Backpropagation errors. Many researchers are using ensemble learning techniques aimed at improving prediction of stock price movement. It would use low price and time lags to predict future highs while another network would use lagged highs to predict future highs. While stock market prediction over a short time span may seem a random process, but over a longer duration it usually develops a patterned movement. Investors tend to buy those stocks whose prices are expected to rise in near future. However, many refrain from investing owing to the uncertainty in the stock market. Hence, there is a need to develop a system for more accurate prediction which can be used in a real-life scenario. The methods used to predict the stock market includes a time series forecasting along with technical analysis, machine learning modeling and predicting the variable stock market. The datasets required for prediction are opening price, closing price, average trading price, etc. The aim is to design a system which reads information from the market using machine learning strategies and predict the future patterns in a stock's price movement.

2. IDENTIFYING THE PROBLEM

Predicting the stock price movement is not easy and will remain a difficult task as long as a robust and accurate prediction algorithm is not developed. The movement in stock market depends on the sentiments of millions of investors. Therefore, stock market prediction must take into consideration the effect of various recent events on investors' sentiments. These events can be political in nature such as a political statement by a political leader, riots, scams; economic events such as bankruptcy of bigwigs, recession or any other micro or macro event which can affect the movement in stock prices. All these events affect businesses and

their earnings, which in turn affects investors' sentiments. These entire factors make predicting movement in stock prices correctly and consistently a next to impossible job. However, if we have the right set of data available, it can then be used to train the machine to predict better.

3.LITERATURE SURVEY

During a literature survey, we collected some of the information about Stock market prediction mechanisms currently being used.

i. Survey of Stock Market Prediction Using Machine Learning Approach

Stock Market prediction has increasingly become important as it can create enormous wealth for investors. One such method is technical analysis, but it does not yield accurate results. Stock Markets generate enormous amount of data at any given point of time and hence such model fails to bring accurate results. Such huge amount of data needs to undergo proper analysis and testing before it can be employed for predicting the movement in prices. The technique that was used in this case was a regression. Each of the techniques which come under regression has its own advantages and barriers over its peers. One of the most used techniques is linear regression. The linear regression models are often fitted using the least squares approach; however, they may alternatively be fitted in other ways too, such as by diminishing the "lack of fit" in some other norm, or by diminishing a handicapped version of the least square's loss function. Contrariwise, the least squares approach can be utilised to fit nonlinear models.

ii. Impact of Financial Ratios and Technical Analysis on Stock Price Prediction Using Random Forests

Using machine learning and artificial intelligence to predict stock market movement is an increasing trend. More and more researchers are investing their time to come up with a model which can predict the market movement more accurately. Today, there are 'n' numbers of options available to predict the price movement of stocks, but all methods don't work the same way even when presented with same datasets. In this paper, we are using Random Forest Algorithm to predict the price movement using financial ratios from previous quarters. This is just one way to predict the market by analysing historical data. However, there are several factors which influence the stock market, such as sentiments of investor about a company, weather, latest news, etc. By using financial ratios along with a model which factors in the above-mentioned parameters, the accuracy of prediction can be enhanced.

iii. Stock Market Prediction via Multi-Source Multiple Instance Learning

Predicting stock market movement is quite a challenging task, however, the modern web has proved to be useful in this feat. The interconnectedness of data has made it possible to extract various variables at a given time, thus, helping to establish relationships between variables and chart out a pattern of investment. Investment patterns from various techniques show some similarities, and to predict a stock's future movement these consistencies between the datasets are exploited. The stock market can be predicted successfully by using more than just historical technical data, such as sentiment analyser which can derive connections between investors' emotion and how they are influenced by investment in specific stocks. One more significant part of prediction is extraction of important events from news available on the web.

iv. Stock Market Prediction: Using Historical Data Analysis

Predicting stock market is filled with uncertainty and influenced by multiple factors. Prediction plays an important role in business and finance. In this method, the technical and fundamental analysis is done by sentimental analysis process. Social media data is collected to gauge the sentiments of individuals. Due to high usage of social media, it can be helpful in predicting trend of stock markets. Technical analysis is done by applying machine learning algorithm on historical data of stock prices. The method gathers social media data and news to extract sentiments expressed by individuals. Other data like previous years stock prices are also considered. The relationship between these datasets is then established and predictions are made out of it.

v. A Survey on Stock Market Prediction Using SVM

Many studies have concluded that most of the predictive regression models are inefficient; the reason being parameter instability and model uncertainty. Support Vector Machine, abbreviated as SVM, provides with the kernel, decision function, and sparsity of the solution. SVM is a training algorithm for classification and regression. It works well with larger dataset. There are many algorithms available in the market, but SVM provides better efficiency and accuracy.

vi. Predicting Stock Price Direction Using Support Vector Machines

Financial organisations have made several exclusive algorithms and attempted to beat the market, but only occasionally anybody could accomplish more-than-average profitability. Nevertheless, stock price forecasting becomes interesting owing to the fact that even a slight improvement in prediction accuracy can create huge amount of wealth for these organisations.

vii.A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)

The prediction model, which is based on SVM and independent analysis, combined called SVM-ICA, is proposed for stock market prediction. The SVM solves regression problems in non-linear classification and time series analysis. And ICA provides a mechanism for deconstructing a given signal into statistically independent components.

viii. Machine Learning Approach in Stock Market Prediction

A vast majority of stockbrokers utilized specialized, fundamental or the time series analysis for stock prediction, however, these techniques do not work in every case and therefore cannot be trusted completely. So, there emerged the need for a strong strategy for stock market prediction. Machine learning and AI along with a supervised classifier was implemented for better accuracy.

Results were tried on the binary classification utilizing SVM classifier with an alternate set of a feature list. The greater part of the Machine Learning approach for taking care of business issues had their benefit over factual techniques that did exclude AI, even though there was an ideal procedure for specific issues. Swarm Intelligence optimization method named Cuckoo search was most easy to accommodate the parameters of SVM. The proposed hybrid CS-SVM strategy exhibited the performance to create increasingly exact outcomes in contrast with ANN. Likewise, the CS-SVM display performed better in the forecasting of the stock value prediction. Prediction stock cost utilized parse records to compute the predicted, send it to the user, and autonomously perform tasks like buying and selling shares utilizing automation concept. Naïve Bayes Algorithm was utilized.

ix. Movements: Insights from Data Mining

This paper finds that corporate communication can have a very significant effect on a company's performance. This paper proposed a technique to reveal the performance of a company. The technique finds the relationship between the frequencies of email exchange of the key employees and the performance of the company reflected in stock values. This paper proposed to use a data mining algorithm on a publicly available dataset of Enron Corporation. The Enron Corporation was an American energy, commodities, and services company founded by Kenneth Lay in 1985.

4. DISADVANTAGES OF EXISTING SYSTEM

- The existing system turns out to be a failure where there are limited outcomes or predictors.
- When the traditional classifier is used, the results indicate that the stock price is unpredictable.
- When there is a change in operating environment the existing system does not work well.
- The existing system does not factor in external events in the environment, such as news events or social media.
- It uses only one data source, thus making it highly biased and inadequate.
- It fails to evaluate the inconsistencies and incompleteness in the data collected.

5. PROPOSED SYSTEM

In the system proposed here, we have focused on predicting stock values using machine learning algorithm like Random Forest and Support Vector Machines. In this system we trained the machine using various data points from the past to make a forecast. The data of stocks was taken from previous years to train the model. And, two machine learning libraries were employed to solve the problem. First, Numpy was used to clean and manipulate the data and getting it into a form ready for analysis. Second, Scikit was used for real analysis and prediction. The datasets used here are from publicly available database online from previous years.80 per cent of the data used were to test the machine and the rest 20 per cent data for testing the data. The basic approach is to learn the relationship between the data from the training set and then reproduce them for the test data. In this paper, Python Pandas library has been used for data processing which tunes different datasets into a data frame. The combined data frame prepares the data for feature extraction. The data frame features date and the closing price for a particular day. All these features were used to train the machine on Random Forest model and predicted the future price on a given day. The accuracy was determined using the predictions for the test set and the actual values. The proposed system involves different areas of research including data pre-processing, random forest, and so on.

6.METHODOLOGIES

i. Classification

It is the arrangement of dataset in taxonomic groups according to their observed similarities. Classification helps to draw some conclusion from the values or the data available. Some of the classifiers used here for stock market prediction are the Random Forest classifier and SVM classifier.

Random Forest classifier– It is a supervised machine learning algorithm, which combines the output of multiple decision trees to reach a single result. The basic approach of this classifier is to take the aggregate of random subset decision trees and come up with a final class or result based on the votes of the random subset of decision trees.

Parameters- The parameters included in the random forest classifier are n_estimators which is total number of decision trees, and other hyper parameters like oob- score to determine the generalization accuracy of the random forest, max_features which includes the number of features for best-split. min_weight_fraction_leaf is the minimum weighted fraction of the sum total of weights of all the input samples required to be at a leaf node. Samples have equal weight when sample weight is not provided.

SVM classifier – SVM is used for working out classification or regression problem. The SVM solves regression problems in non-linear classification and time series analysis. It provides with the kernel, decision function, and sparsity of the solution

Parameters- The tuning parameters of SVM classifier are kernel parameter, gamma parameter and regularisation parameter.

- Kernels can be categorised as linear and polynomial kernels, which calculates the prediction line. In linear kernels prediction for a new input is calculated by the dot product between the input and the support vector.
- C parameter is known as the regularisation parameter. It determines whether the accuracy of model is increasing or decreasing. The default value of c is 10. Lower regularisation value leads to misclassification.
- Gamma parameter measures the influence of a single training on the model. Low values signify far from the plausible margin and high values signify closeness from the plausible margin.

ii. Random Forest algorithm

Random Forest algorithm has been termed as one of the easiest to use and flexible machine learning algorithm. It is used in stock market prediction because it gives good accuracy in prediction. The algorithm is usually used in classification tasks. The task of

predicting stock price movement is quite challenging owing to the high volatality in the market. We are using this algorithm for prediction since it has the same parameters as of a decision tree. The decision tool has a model similar to that of a tree. It takes the decision based on possible consequences, which includes variables like event outcome, resource cost, and utility. The random forest algorithm randomly selects diiferent observations and features to build several decision tree and then takes the aggregate of the outcomes of all the decision trees. The data is sorted into partitions based on the questions on a label or an attribute.

iii. Support Vector Machine algorithm

The main task of the SVM algorithm is to identify an N-dimensional space t categorises the data points in a distinguished manner. Here, N stands for the number of features. Between two classes of data points, there can be multiple possible hyperplanes that can be chosen. The objective of this algorithm is to find a plane that has maximum margin. Maximising margin refers to the distance between data points of both the classes. The benefit of maximising the margin lies in the fact that it provides some reinforcement so that future data points can be more easily classified. Decision boundaries that help classify data points are called hyperplanes. Based on the position of the data points relative to the hyperplane they are attributed to different classes. The dimension of the hyperplane relies on the number of attributes, if the number of attributes is two then the hyperplane is just a line, if the number of attributes is three then the hyperplane is two dimensional.

7. SYSTEM ARCHITECTURE

Kaggle is huge repository of community published data and code. The community contains datasets from various fields contributed by data miners. The contributors there compete to create best models for predicting and depicting the information. The dataset used in this project has been taken from Kaggle. The dataset that we have used here is in raw format and is a collection of stock market information about a few companies.

Firstly, the raw data needs to be converted into processed data; this is done using feature extraction. In the raw data collected there are various information which are not very useful for the purpose of prediction. The first step feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset. Feature extraction basically involves diminishing of raw variables to progressively reasonable features for better management of data.

Soon after feature extraction, the next step involves classification process. Here the processed data that was obtained after feature extraction is split into two different and distinct parts - training and testing. The data is split in a way that training data maintain a higher proportion than the test data. The training data set is used to train the model whereas the test data is used to predict the accuracy of the model.

To analyse the data, Random Forest algorithm utilises a collection of random decision trees to analyse the data. In data splitting a cluster of the decision trees look for specific attributes in the data from the total number of decision trees in the forest. In this case, the end goal of our propose project is to predict the future price of the stock by analysing its historical data.

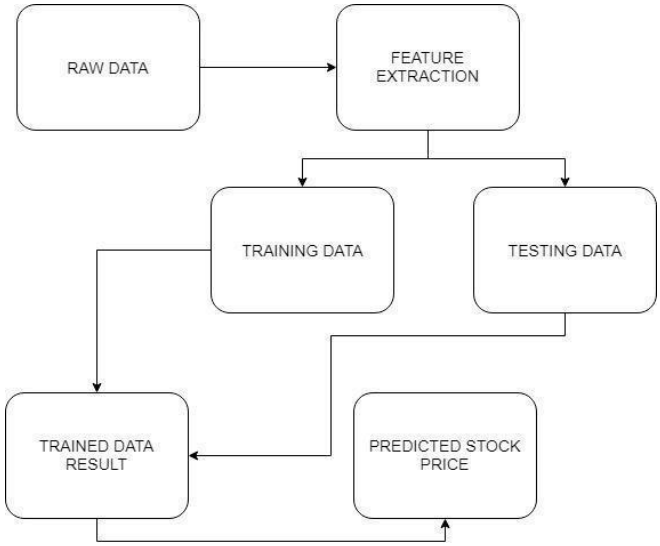

Fig 1: System Architecture

8. MODULE IDENTIFICATION

i. Data collection- Data collection involves the process of collecting the right dataset for stock market prediction. Here we will be analysing the Kaggle dataset initially, and according to the accuracy, we will be using the model with the data to analyse the prediction accurately.

ii. Pre-processing- Data pre-processing involves transforming raw data into a more comprehensible format. Raw data usually contains many errors and much information is missing. The data pre-processing involves checking for the missing values, looking for categorical values and splitting the dataset into training and test set. And finally perform a feature scaling to limit the range of variables so that they can be compared on common environs.

iii. Training the machine- Training the machine is like feeding the data to the algorithm to refurbish the test data. The training of the model involves cross-validation. Here we get a well-grounded approximate performance of the model using the training data.

Tuning models are meant to specifically tune the hyperparameters like the number of trees in a random forest. We perform the entire cross-validation loop on each set of hyperparameter values. At the end, we will calculate a cross-validated score for individual sets of hyperparameters. Then we select the apt hyperparameters. The basic approach behind training the model is to get some initial values with the dataset and then optimize the parameters which we want to in the model. This is repeated time and again until we get the optimal values. Thus, from the trained model on the inputs from the test dataset we take the predictions. Hence, it is divided in the ratio of 80:20 where 80% is for the training set and the rest 20% for a testing set of the data.

Data scoring - Data scoring is the process of generating values based on a trained machine learning model, given some new input data. Random forest algorithm involves an ensemble method, which is usually used, for classification and as well as regression. Based on the learning models, we achieve interesting results. The last module thus describes how the result of the model can help to predict the probability of a stock to rise and sink based on certain parameters. The algorithm also shows the vulnerabilities of a particular stock. To make sure that only authorized entities have the access to the results, the user authentication system control is implemented.

EXPERIMENTAL RESULTS

There are total eleven attributes which determines the rise and fall in stock prices. We assigned the following variables to these attributes:

52 week high price of a stock - HIGH
52 week low price of a stock - LOW
Opening price of a stock on a given day - OPENP
Closing price of a stock on a given day - CLOSEP
Last Traded Price at any given time - LTP
Number of trades executed in a given stock in a given day - VOLUME

Similarly, there are other attributes such as YCP, TRADE, VALUE, DATE and TRADING CODE.

DATE	TRADING CODE	LTP	HIGH	LOW	OPENP	CLOSEP	YCP	TRADE	VALUE (mr	VOLUM
28-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.4	6.5	79	1.888	2,94,7
27-12-2017	1JANATAMF	6.5	6.5	6.4	6.5	6.5	6.5	73	1.295	2,00,0
26-12-2017	1JANATAMF	6.5	6.6	6.4	6.5	6.5	6.5	103	4.119	6,30,5
24-12-2017	1JANATAMF	6.6	6.6	6.4	6.5	6.5	6.5	46	0.654	1,01,1
21-12-2017	1JANATAMF	6.6	6.6	6.4	6.4	6.5	6.4	24	0.241	37,0
20-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.4	6.4	37	0.296	45,8
19-12-2017	1JANATAMF	6.4	6.6	6.4	6.5	6.4	6.5	55	1.387	2,16,5
18-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.5	6.4	36	0.141	21,8
17-12-2017	1JANATAMF	6.5	6.5	6.4	6.5	6.4	6.6	118	2.904	4,52,1
14-12-2017	1JANATAMF	6.5	6.6	6.5	6.6	6.6	6.6	36	0.596	90,5

Fig 2: Raw Data

	DATE	TRADING CODE	LTP	HIGH	LOW	OPENP	CLOSEP	YCP	TRADE	VALUE (mn)	VOLUME
0	2018-08-16	1JANATAMF	6.2	6.3	6.1	6.2	6.2	6.2	56	0.757	122741
1	2018-08-16	1STPRIMFMF	11.2	11.2	10.9	11.0	11.1	10.9	145	2.640	238810
2	2018-08-16	AAMRANET	80.1	80.4	78.5	78.5	79.7	78.3	545	15.488	195035
3	2018-08-16	AAMRATECH	30.8	31.6	30.7	31.0	30.9	31.0	195	5.100	164899
4	2018-08-16	ABB1STMF	6.1	6.1	5.9	6.0	6.1	6.0	109	11.214	1857588

Fig 3: head()

This is the result of using the head(). Since pandas library is used to analyse the data, it returns the first five rows. Here, five is the default value of the number of rows it returns unless given command otherwise. There are some attributes from the processed dataset which are not relevant, such as the TRADING CODE. So, we use the strip() to remove it and replace them with a value "GP".

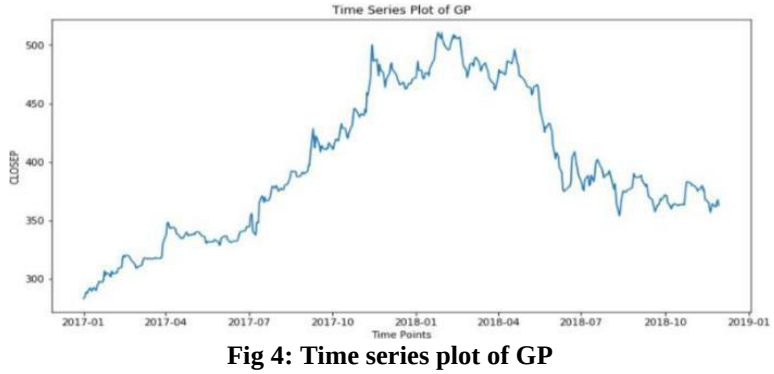

Fig 4: Time series plot of GP

This is a time series plot generated from using the “matplotlib.pyplot” library. This plot shows the relation between the attributes “CLOSEP” vs “DATE”. This shows the trend of closing price of stock over a span of two years. The figure shown below depicts the candle stick plot, which was generated using the library “mpl_finance”. The candle stick plot was generated using the attributes 'DATE', 'OPENP', 'HIGH', 'LOW','CLOSEP'.

Fig 5: Candlestick plot

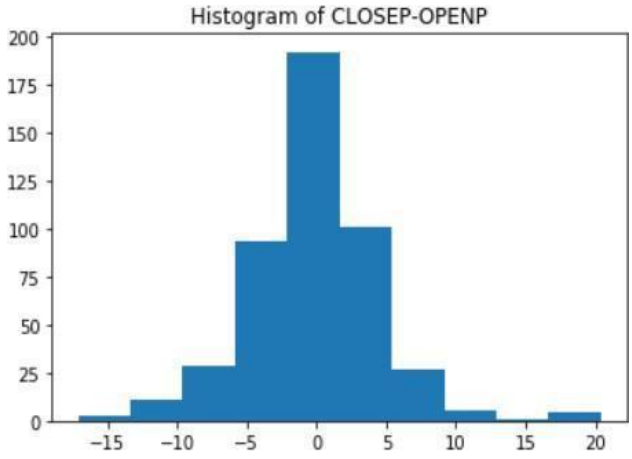

Fig 6: Histogram of CLOSEP-OPENP

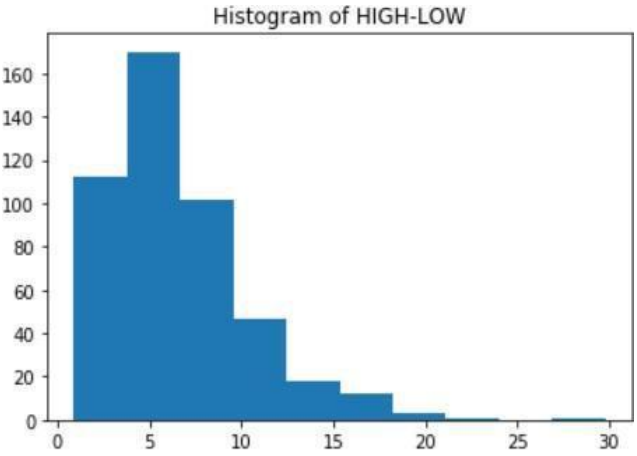

Fig 7: Histogram of HIGH-LOW

The above two figures is a depiction of the plotting between "CLOSEP" and "OPENP" and "HIGH" and "LOW". This is done because we believe that today's closing price and opening price along with the highest and lowest price of the stock during last year will have an affect over a stock's future values. Based on such reasoning we devised a logic "iftoday's CLOSEP is greater than yesterday's CLOSEP then we assign the value 1 to DEX or else we assign the value -1 to DEX. The whole dataset is then processed and upon using head() we get a glimpse of the data obtained so far.

The next step involves the setting of feature and target variable. Along with it, the setting of train size is to be done too. We import SVC classifier and fit it with the training data using the sklearn libraries. So, after we train the model with the data and run the test data through the trained model, we get a confusion matrix, which is shown below.

	precision	recall	f1-score	suppor
-1.0	0.76	0.93	0.84	2
1.0	0.85	0.58	0.69	1
micro avg	0.79	0.79	0.79	4
macro avg	0.81	0.75	0.76	4
weighted avg	0.80	0.79	0.78	4

Fig 9: Confusion Matrix

We have used the same dataset to train another model. This model puts the Random Forest Classifier belonging to the ensemble technique to work. We will take this as v.2; here the decision trees have the default values so that makes the "n_estimator" value to be 10. However, in v.22 the value will change to 100. When we feed the data in the model and run it against predicted data, the accuracy turns out to be 0.808. To put in a nutshell, the accuracy of the SVC model in test set is 0.787 while the accuracy of the Random Forest classifier is 0.808.

9. CONCLUSION

After calculating the accuracy of different models, we arrived at the conclusion that the most appropriate algorithm for predicting the price movement of a stock based on historical data is the Random Forest algorithm.

This algorithm will give an upper hand to the stock brokers and investors for investing money in the stock market. This project demonstrates the supremacy of Random Forest algorithm over the previously implemented machine learning models to predict the stock market.

10. FUTURE ENHANCEMENT

The accuracy of this model can be enhanced by adding more parameters and factors like the financial ratios, multiple instances, etc. In future the algorithms can also be applied to gauge the public sentiments regarding any particular stock by analysing the contents of public comments by senior management of a company and thus determining patterns and relationships between the public sentiments and stock price movements.

REFERENCES

[1] Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2017.
[2] Loke. K.S. “Impact of Financial Ratios and Technical Analysis on Stock Price Prediction Using Random Forests”, IEEE,2017.
[3] Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, “Stock Market Prediction via Multi-Source Multiple Instance Learning.” IEEE2018.
[4] Vivek Kanade, Bhausaheb Devikar, Sayali Phadatare, Pranali Munde, Shubhangi Sonone. “Stock Market Prediction: Using Historical Data Analysis”, IJARCSSE 2017.
[5] Sachin Sampat Patil, Prof. Kailash Patidar, Asst. Prof. Megha Jain, “A Survey on Stock Market Prediction Using SVM”, IJCTET2016.
[6] https://www.cs.princeton.edu/sites/default/files/uploads/Saahil_magde.pdf
[7] Hakob GRIGORYAN, “A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)”, DSJ 2016.
[8] Raut Sushrut Deepak, Shinde Isha Uday, Dr. D. Malathi, “Machine Learning Approach In Stock Market Prediction”, IJPAM2017.
[9] Pei-Yuan Zhou, Keith C.C. Chan, Member, IEEE, and Carol Xiaojuan Ou, “Corporate Communication Network and Stock Price Movements: Insights from Data Mining”, IEEE2018.

