A Comparitive Study and Analysis of some Supervised Learning Techniques used for Stock Forecasting

Aditya Mistry
Instrumentation and Control Engineering department
Institute of Technology, Nirma University
Ahmedabad, India

Manal Ghadawala
Instrumentation and Control Engineering department
Institute of Technology, Nirma University
Ahmedabad, India

Abstract— Stock trading is a very complex and dynamic phenomenon. People buy and sell stocks for the purpose of investments. Stock market trading can be beneficial if done in a wise and planned manner. With the help of technology, we can solve this cumbersome problem. The goal is to review different models and identify the best model that learns from the market datausing machine learning techniques and forecast future trends in stock price movement. In this paper, 6 different supervised learning methods are analyzed and observations are made, on the type of method that can be implemented, after analyzing the available data and conditions.

Keywords—Supervised Learning, forecasting, dataset, accuracy

INTRODUCTION

A stock market is a place where the trading of stocks takes place. This can be buying, selling, and the issuance of stock which belongs to publicly-held companies. There can be multiple venues for stock trading in a country. A stock that is listed on an exchange can be bought or sold. There is terminology associated with the stock market which will be very frequently encountered in this paper-

I) Initial Public Offering

The initial public offering is the process by which a private company can go public by selling its shares to the general public or retail investors.

II) Open(the opening price of the stock)

The starting price on a specific trading day at which the shares of a company is to be sold.

III) High

Highest price of the stock possible at a given instance of time.

iv) Low

The Lowest price of the stock, possible at a given instance of time.

v) Close(the closing price of the stock)

The final price on a specific trading day at which the shares of a company are sold.

vi)Volume

It shows the typical number of shares of stock that are traded during a specific period of time usually the daily trading volume. It also can convey the number of shares that you're allowed to get of a given stock.

vii)Volatility

It refers to the fluctuations within the price of an equity share. It has been observed that highly volatile stocks have severe ups and downs during trading sessions. These are highly risky bets which may bring a lot of profits for the skilled intra-day trader.

viii)Moving Average

It refers to the typical price per unit of an equity share with reference to a selected period of your time. Some popular time frames wont to study the moving average of stock include 50- and 200-day moving averages.

ix)Previous day's closing price

The previous day's close price mostly refers to the prior day's final price of a security when the market officially closes for the day.

x)P/E ratio

The price-earnings ratio, also called P/E ratio, P/E, or PER, is that the ratio of a company's share price to the company's earnings per share.

Stock Market trading is not an easy task. One needs to be very careful while trading in the stock market. So it becomes very essential to develop certain techniques, such that the people who trade in the stock market can get some assistance. The Stock Market has a unique characteristic which is very much highlighted which is that it is highly unpredictable in nature. Here we bring in the use of technology. Stock Market prediction can be used to determine the future values of the stock market in a given time frame. For the prediction of stocks, a fully automated model can be developed. There are various methods involved in developing this model-

- Feature extraction from the stock market dataset

- Feature selection for highest prediction accuracy
- Performing the dimensional reduction of the selected feature set

The robustness and accuracy of the prediction model developed[1].

This can be done using a variety of learning methods. In this paper, we have used supervised learning techniques and analysed their results. These Supervised Learning methods are- Support Vector Machine, Random Forest method, Linear Regression, Logistic regression, Naive Bayes, and KNN.We have analysed the works of people who have taken different sets of technical indicators. Technical indicators are parameters that are used for the purpose of forecasting. They are calculated from a time series stock data. In the case of stock price prediction, where the dataset is extremely random, one single regression model is not sufficient. Therefore an ensemble of different regression models is used to make predictions.

LITERATURE REVIEW

Stock Market prediction is a very wide domain. Several Machine learning and Deep learning techniques are used for this purpose. A dataset can be taken from any stock exchange, for a given interval of time. Using various attributes from that dataset, we can develop a model using any of the methods and can forecast the stock value. Support Vector machines are a very efficient and powerful method that can be used for stock prediction purposes. There are various challenges faced by traditional predictive regression techniques. (Challenges in out-of-sample predictability tests due to model uncertainty and parameter instability)[2].

For the stock-prediction purpose, we also use the classification trees, since the dependent variable is categorical. We continuously split the data on certain parameters. Here comes the Random Forest method which is a collection of decision trees in a randomized way. The Decision tree is constructed for every subset by calculating information gain and entropy. Another classification technique is the Naive Bayes method. It is much in use when we use analysis of the sentiments for the stock market prediction. This is because it is an efficient classification technique. It predicts the probability of each class such that the given data belongs to that class. The class that will have the maximum probability, will be the most suitable class.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (1\text{-}1)$$

The KNN method is a classification technique where we use a particular dataset, which is divided into test and train data. The test data and the training data is mapped on a graph as a set of vectors. The Euclidean distance between the test data and the individual train data (distance between the vector points) is calculated and the training datasets were figured out and sorted according to the Euclidean distance. The number of these datasets again depends on the value of "k". After this, the class with the majority vote is assigned as the predicted value, i.e. the output of the model. These were the few classification techniques. We also use regression techniques which are used for the prediction purpose. Linear Regression is a method of supervised learning, where we actually map the input variables from the available dataset to some constant function. We thus try to try to predict the results within a continuous output. Given a dataset, it is very easy to implement. Logistic Regression can be used to predict the closing stock price of the stock or identify the trend of the stock. It is a classification technique. It can be used to predict the categorical variables. The dependent variable is binary and the independent variable can be continuous or binary.

METHODOLOGY AND RESULTS

A.Support Vector Machine (SVM)

For the purpose of prediction of stock trends, the Support Vector Machine can be employed here as a classification technique. The support vectors are the extreme points and here as the name justifies, it is suitable for extreme cases. They are very important. The margin of the classifier is defined as the distance between the decision surface and the closest datapoint. For the prediction purpose, we aim to maximize the margin width. The maximal margin hyperplane should be selected. This will ensure the good accuracy of the prediction model [2][3]. There are various challenges faced by other traditional predictive regression techniques. (challenges in out-of-sample predictability tests due to model uncertainty and parameter instability). "Kernel" is a complex mathematical function that is used when we have the object which may/maynot be linearly separable and are used to separate the objects which are from different classes. The SVM dataset works well on a large dataset and also there is no problem of overfitting faced when this model is used.

B.*Random Forest*

A random forest is basically made by the collection of decision trees in a randomized way. It is a type of ensemble classifier. The method for implementing this model is- Firstly a bootstrap dataset is created by randomly taking samples from the dataset, and these samples are taken randomly. We create decision trees by taking the features from the bootstrap dataset. For deciding the nodes, the subset of the features is used at each step. And finally, voting is carried out. Based on this voting, the class that is supposed to be the output class is selected. Therefore as we observe, there is randomization in the process. The random forest method can be used to predict the closing price of a stock on a particular day. For such kinds of prediction, the input data attributes that we should have are- Open, high, low, and close data. Further, we can have more data[4]. We then need to identify a set of technical indicators and then calculate it [4].

D.K Nearest Neighbor

A prediction based model for the stock market was done based on the KNN method by Khalid Alkhatib et al, by extracting the data of 5 major listed companies. The sample was taken from the Jordanian stock exchange. With the k value which was equal to 5, the results, i.e. the predicted value of the stocks was accurate with a low error rate. The rational results also gave a very positive and impactful sign that when KNN could be used for data analysis the data mining techniques would be beneficial for the decision-makers at various levels. Further one additional point was raised by the authors regarding the lack of knowledge of financial econometrics in Jordan and explains various consequences for the same [5].

E. ***Naïve Bayes***

Naive Bayes is a classification technique based on Baye's theorem which involves conditional probability and taking some independent assumptions.
While Predicting the stock exchange volume of KSE (Karachi Stock Exchange) there were different matrices like- mean absolute error, accuracy, and root mean square error. The results showed us that the Naive Bayes model did not show good accuracy. Instead, Bayes Network (models relationship between the features) accuracy was found to be the third-best[6].
Also, we need to look upon the fact that The Naive Bayes method works well for a small dataset. It had the highest accuracy when compared with the accuracies of SVM, Random Forest, and KNN for the same dataset, attributes, and technical indicators. However, there are some other results contradicting this result [2].

F.Linear Regression

Accuracy plays a key role in the Linear Regression method. Various methods through which we can calculate accuracy are- RMSE values, the R2 value of the model, the Confusion matrix (for classification problems), etc.
For a model, the dataset was taken for Google by WIKI from www.quandl.com.
For a period of 14 years, and the closing price was to be predicted with the standard attributes which are taken for the purpose of predicting the closing price (open, high, low, close, Volume). The accuracy came out to be 97.67% which is a very good value. Also here we need to note that the features extracted here, were subject-specific [9].

Another study was performed by Mr Dinesh Bhuriya, et al. Calculated the Confidence value using different regression methods like linear regression, Polynomial regression, and RBF regression. Here they had taken the dataset from the TCS stock database (CSV file) and had converted these into pandas Dataframes that were indexed by date. The attributes were similar to the previous case with the close price being the dependent variable. The results showed us that Linear regression had the highest confidence value of 0.9774 followed by polynomial regression (0.468) and RBF regression (0.5652) [7].

G.Logistic Regression

For the purpose of stock prediction, we can select two different classes such that the stock value for the next month/year can be categorized as '0' or '1'. If we observe a downtrend, i.e. the average value of the stock for the next month/year is lesser comparatively, then it can be depicted by '0'. And if it is greater, then we can depict it as '1'. Based on this phenomenon, Mr. Jibing Gong and Mr.Shegtao Sun developed an equation based on Logistic Regression. It is given by [8]-

$$p = \frac{\exp(c_0 + c_1x_1 + c_2x_1 \ldots + c_kx_k)}{1 + \exp(c_0 + c_1x_1 + c_2x_1 \ldots + c_kx_k)} \quad (1\text{-}2)$$

OBSERVATIONS

The studies have shown that the SVM model gives us a good comparative accuracy even for different attributes [2][3].The SVM model was employed to see whether the market investment was "Good" or "poor". It was found that with 78 entries as a training set, we could get an accuracy of 96.15%. The error rate was found to be 30% for the test dataset. [2]The comparative study for the same dataset employing two different methods- SVM classification (Supervised Learning) and K-means clustering (Unsupervised Learning) shows us that the former has a greater amount of accuracy. Coming to the SVM software which employed the kernel method, Linear SVMs had the highest amount of accuracy and it was the most useful method despite the fact that the data was not related in any manner and it was not separable [2].The Random Forest method is the best suitable method when we are dealing with large datasets. We have observed that there is randomization in the process. It has the highest accuracy compared to SVM, KNN, Naive Bayes and softmax, for a large common dataset with attributes- open, high, close and values, and with similar 10 technical indicators namely-(Volatility, Stochastic oscillator, Rate of change (1), Rate of change (2), Volume price trend, Williams %R, Disparity Index, Commodity Channel Index and Moving Average (10), Moving Average (50)) [4].Seeing and analyzing various outcomes, we find that the error rate is inversely proportional to the number of trees in the forest.

While predicting whether the closing price will rise or fall after 30 days, there are 29 of the trees in the forest which predict the rise in price while a single tree predicts the fall in price. Hence this prediction matches the assigned actual label to the test sample. A point where such kind of method is not feasible for technical-minded people because the decision rules which are learned by the trees [9]. Linear regression was found to be the best method for calculating the confidence value and it gives good accuracy when used. [9] By implementing the formula (1-2) on a dataset taken by Mr. Jibing Gong and Mr. Shegtao Sun, it was observed that the process for selection of the best possible group of the coefficients improves the prediction accuracy of the model. 2 out of the 12 predictions were error some. That means the maximum accuracy that was obtained by this process was about 83.3%. Pros and cons of this project[8]-

TABLE I. PROS AND CONS OF THE LOGISTIC REGRESSION MODEL

Pros and Cons of the model	
Pros	***Cons***
Very easy to understand. Very easy user-interface. For the prediction of the stock price trend of the next month, the users could only consider the ongoing month's financial data, rather than going for the historical dataset.	Some of the features might fail, because of the negligible values of the parameter.

TABLE II. ADVANTAGES AND DISADVANTAGES DRAWN AFTER ANALYZING THE SUPERVISED LEARNING TECHNIQUES

Supervised Learning Techniques	**Advantages and Disadvantages concluded after analysis**	
	Advantages	***Disadvantages***
Support Vector Machine	SVM is simpler in high dimensional spaces and is comparatively memory efficient. It can solve the difficulty of overfitting.	These are training data sets during which the number of samples that fall in one among the classes far outnumbers people who are a member of the opposite class
	It can solve linear also as non-linear relationship problems.	The number of features for every datum exceeds the amount of coaching data samples, the SVM won't perform well.
Random Forest	The random forest technique reduces the overfitting (which is the problem in decision tree algorithm) since it is based on the ensemble modelling approach	Random Forest creates a lot of trees and combines their outputs which results in complexity.
	The random forest method can also handle big data with numerous variables running into more than thousands And can automatically handle missing values	Random Forest requires much more time to train as compared to decision trees as it generates a lot of trees and as a result, it makes the algorithm too slow and ineffective for real-time predictions
Linear Regression	It is one of the easiest and basic regression techniques that can be easily understood.	Assumption of a linear relationship between the input and the output variables.

	During the literature review, we have found that when Linear regression was implemented (different cases), it gave good accuracy.	This tendency can result in the oversimplification of the problem.
Naive Bayes	The Naive Bayes technique makes use of class conditional independence and as a result it is computationally faster.	Independent predictor assumption which results, in the decrease of accuracy
	It gives the best performance when we have a categorical input instead of a numerical input.	In the stock market scenario too, the dependencies exist among the variables.
K Nearest Neighbor	One of the main advantages of this method is that we have not taken any assumptions.	Here we saw that for the stock prediction purpose, we had a large dataset. It is a bit difficult to handle such a dataset.
	It is very easy to understand and implement.	As we saw in the studies that it requires very classified data.
Logistic Regression	It is feasible and is not complex.	Since, maximum likelihood calculations are less accurate at low sample sizes as compared to the standard least square.
	A linear relationship between the dependent and the independent variables is not necessary Logistic Regression shouldn't be used if the amount of observations is lesser than the number of features, otherwise, it's going to cause overfitting.	Logistic Regression shouldn't be used if the amount of observations is lesser than the amount of features, otherwise, it's going to cause overfitting.

CONCLUSION

Supervised Learning methods can be used for the purpose of Stock Forecasting. Here, we have used various regression as well as classification techniques that have been used for this purpose. For different features and datasets, we have seen different models. These models are then analyzed and conclusions are drawn based upon the analysis. It is pretty much evident, that apart from the dataset, there are many other factors that influence the predictive model. Thus a thorough understanding of the stock market and its various attributes becomes very necessary along with the knowledge of the Machine Learning technique which is to be used. For all the methods, we have figured out the advantages and disadvantages.

ACKNOWLEDGMENT

The foundation of research was supported by the Institute of Technology,Nirma University. We thank Dr. Ankit Sharma who provided insight and expertise that greatly assisted the research and for the comments that greatly improved the manuscript. We would also like to show our gratitude to Prof. DhavalPujara, Head of Department - Department of Electronics and Communication Engineering, Institute of Technology, Nirma University; for sharing their pearls of wisdom with us during the course of this research.

REFERENCES

[1] K. Pahwa and N. Agarwal, "Stock Market Analysis using Supervised Machine Learning," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, 2019, pp. 197-200, doi: 10.1109/COMITCon.2019.8862225.

[2] I. 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), Coimbatore, 2018, pp. 1003-1007, doi: 10.1109/ICICCT.2018.8473214.

[3] Z. Hu, J. Zhu and K. Tse, "Stocks market prediction using Support Vector Machine," *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, Xi'an, 2013, pp. 115-118, doi: 10.1109/ICIII.2013.6703096.

[4] Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*.

[5] Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (kNN) algorithm. *International Journal of Business, Humanities and Technology*, *3*(3), 32-44.

[6] A. Sharma, D. Bhuriya and U. Singh, "Survey of stock market prediction using machine learning approach," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 506-509, doi: 10.1109/ICECA.2017.8212715.

[7] J. Gong and S. Sun, "A New Approach of Stock Price Prediction Based on Logistic Regression Model," *2009 International Conference on New Trends in Information and Service Science*, Beijing, 2009, pp. 1366-1371, doi: 10.1109/NISS.2009.267...

[8] N. Powell, S. Y. Foo and M. Weatherspoon, "Supervised and Unsupervised Methods for Stock Trend Forecasting," *2008 40th Southeastern Symposium on System Theory (SSST)*, New Orleans, LA, 2008, pp. 203-205, doi: 10.1109/SSST.2008.4480220.