Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)
IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

A comparative study of supervised machine learning algorithms for stock market trend prediction

Indu Kumar
Computer Engineering Department
NIT Kurukshetra
kumarindu22@gmail.com

Kiran Dogra
Computer Engineering Department
NIT Kurukshetra
kirandogra1@gmail.com

Chetna Utreja
Computer Engineering Department
NIT Kurukshetra
chetna4395@gmail.com

Premlata Yadav
Computer Engineering Department
NIT Kurukshetra
yadav.premlata15@gmail.com

***Abstract*- Impact of many factors on the stock prices makes the stock prediction a difficult and highly complicated task. In this paper, machine learning techniques have been applied for the stock price prediction in order to overcome such difficulties. In the implemented work, five models have been developed and their performances are compared in predicting the stock market trends. These models are based on five supervised learning techniques i.e., Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor (KNN), Naive Bayes, and Softmax. The experimental results show that Random Forest algorithm performs the best for large datasets and Naive Bayesian Classifier is the best for small datasets. The results also reveal that reduction in the number of technical indicators reduces the accuracies of each algorithm.**

Keywords- machine learning, classifier, Random Forest, SVM, KNN, Naïve Bayes, Softmax

I. INTRODUCTION

Stock market plays a very important role in fast economic growth of the developing country like India. So our country and other developing nation's growth may depend on performance of stock market. If stock market rises, then countries economic growth would be high. If stock market falls, then countries economic growth would be down. In other words, we can say that stock market and country growth is tightly bounded with the performance of stock market. In any country, only 10% of the people engaging themselves with the stock market investment because of the dynamic nature of the stock market. There is a misconception about the stock market i.e., buying or selling of shares is an act of gambling.

Hence, this misconception can be changed by bringing the awareness across the people for this. The prediction techniques in stock market can play a crucial role in bringing more people and existing investors at one place. Among the popular methods that have been employed, Machine Learning techniques are very popular due to the capacity of identifying stock trends from massive amounts of data that capture the underlying stock price dynamics. In this paper, we applied supervised learning methods for stock price trend forecasting.

The details of the structure of paper are as follows. In the next section, related work in this field has been mentioned. In section 3, the paper discusses research data details. In section 4 proposed work has been presented. Finally, in section 5 obtained results are discussed and section 6 concludes the proposal.

II. RELATED WORK

Correct Prediction of stock market trends is of great importance for the investors as it helps in determining whether the investment would pay off or not. Many methods have been deployed for the same. Artificial Neural Network based method is the first

technique to be used for the stock market trend prediction [1]. Machine learning has been used for prediction of movement sign of stock market index. Kim [2] in last two decades applied SVM for the first time for predicting stock market price. Random Forest is another machine learning model used for predicting trend direction of stocks [3]. Five-days and ten-days ahead models have been used. In this research paper, the comparative study of the supervised machine learning algorithms using the time window of size 1 to 90 has been proposed. The algorithms have been compared based upon the parameters: Size of the dataset and Number of technical indicators used. Accuracy and F-measure values have been computed for each algorithm. Long term model has been used to compute the accuracy and F-measure.

III. RESEARCH DATA

The data used in this research paper has been collected from data sources like Yahoo Finance, Quandl, NSE-India, and YCharts. The data available has the following attributes: Date Open, High, Close, and Volume. Twelve technical indicators have been used for the model prediction. First technical indicator used is Moving Average (MA10 and MA50). It is responsible for smoothening the stock price signal and making the identification of trends easier. Moving averages for 10 and 50 days have been used in this paper. Next technical indicator is Relative Strength Index (RSI) which detects whether the stock is overbought or oversold or not.

Next indicator is Rate of Change (RoC) which simply measures the rate of change of price from one period to another. RoC1 and RoC2 have been used in the paper. Next indicator which has been used is Volatility which gives the measure of the dispersion of returns for a given security. Volatility for a period of 10 days has been calculated.

Disparity Index (DI) that measures the relative position of selected moving average to the most recent closing price. Disparity Index for 10 days has been calculated. Next indicator is Stochastic Oscillator which depicts the location of the closing price relative to the high-low range. Williams%R is momentum indicator which shows the level of the closing price relative to the highest high. Next indicator is Volume Price trend which relates the volume and the price. Commodity Channel Index (CCI) calculates the current price level relative to an average price level over a given period of time

IV. PROPOSED METHODOLOGY

The proposed architecture for the implemented work mainly consist of four steps: feature extraction from the given dataset, supervised classification of the training dataset, supervised classification of the test dataset, and result evaluation. Flow chart for the proposed methodology is described in Figure1.

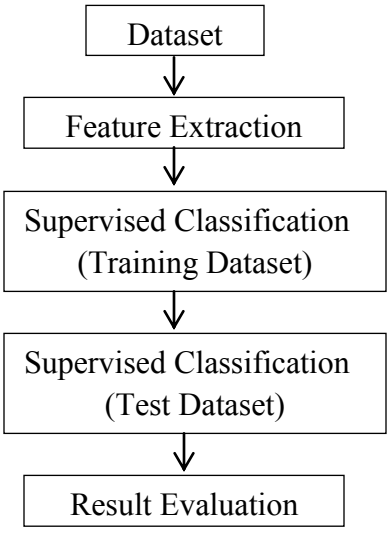

Figure 1. Flow chart of proposed methodology

A. Dataset

The dataset has been collected from data sources Yahoo Finance, Quandl, NSE-India, YChart and features have been extracted. The dataset for sites Amazon, Cipla, Eicher, Bata and Bosch have been collected for a period of five to ten years. Amazon, Bata and Bosch are the large dataset which has approximately 4500 entries. Cipla and Eicher are the small dataset which contains approximately 1800 entries.

B. Features Extraction

For feature extraction, twelve technical indicators have been calculated from the downloaded dataset for each site. The indicators calculated are Moving average (MA10, MA50), Rate of Change (RoC1, RoC2), Relative Strength Index (RSI), Volatility10, Williams%R, Stochastic Oscillator, Channel Commodity Index (CCI), Disparity (Disparity5, Disparity10) and Price Volume trend.

C. Supervised Classification (Training Dataset)

The data has been divided into two parts i.e., training and testing data in the 70:30 ratios. Learning algorithms have been applied on the training data and based on the learning, predictions are made on the test data set.

D. Supervised Classification (Test Dataset)

The test dataset is 30% of the total data. Supervised learning algorithms have been applied on the test data and the output obtained is compared with the actual output.

E. Result Evaluation

Results have been evaluated where accuracies and F-measure values for each learning algorithm have been calculated. Time window that has been used is of size 90 i.e. models from day 1 to 90 are evaluated. Total accuracy has been calculated by taking the average of accuracies of 1 to 90 day models. The comparison of accuracies of the algorithm has been done on the basis of the following parameters: Size of the dataset and Number of technical indicators.

V. EXPERIMENTAL RESULTS

For comparative study of the supervised learning algorithms for stock market prediction, accuracy and F-measure are used in this paper. Accuracy is mathematically expressed using equation (1) and F-measure is mathematically expressed by equation (2), where TP is true positive, TN is true negative, FP is false positive and FN is false negative

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$F-measure = 2 * \frac{\frac{TP^2}{(TP+FP)*(TP+FN)}}{\left(\frac{TP}{TP+FP}\right)+\left(\frac{TP}{TP+FN}\right)} (2)$$

The results evaluated for twelve indicators are shown in Table 1 and for six indicators are shown in Table 2.

Table 1. Long term model for twelve indicators

Algorithms	Large Dataset (Accuracy/F-measure in %)			Small Dataset (Accuracy/F-measure in %)	
	Amazon	Bosch	Bata	Cipla	Eicher
SVM	67.16/ 75.98	64.56/ 73.85	62.35/ 75.20	58.51/ 65.84	58.98/ 65.80
Random Forest	72.36/ 80.55	64.51/ 73.30	66.28/ 75.88	55.71/ 64.28	55.80/ 63.95
KNN	65.56/ 77.00	55.06/ 69.22	60.89/ 74.20	45.94/ 57.26	45.81/ 57.06
Naive Bayes	70.80/ 60.42	63.36/ 50.04	50.93/ 50.24	63.84/ 62.32	64.03/ 50.14
Softmax	57.80/ 64.74	53.18/ 66.13	60.00/ 70.62	45.90/ 48.11	46.93/ 46.94

Table 2. Long term model for six indicators

Algor	Large Dataset	Small Dataset

ithms	(Accuracy/F-measure in %)			(Accuracy/F-measure in %)	
	Amazon	Bosch	Bata	Cipla	Eicher
SVM	66.14/ 78.36	59.80/ 71.74	66.14/ 78.36	55.28/ 61.04	54.42/ 60.02
Random Forest	69.73/ 79.87	60.49/ 71.12	69.58/ 79.69	52.40/ 61.92	54.66/ 59.89
KNN	65.56/ 76.99	55.08/ 69.23	65.60/ 77.04	45.91/ 57.25	43.45/ 55.89
Naive Bayes	67.61/ 50.72	60.78/ 59.58	67.61/ 50.72	60.07/ 43.32	60.09/ 48.33
Softmax	57.22/ 63.16	52.94/ 65.48	56.61/ 61.21	45.7/48.72	45.87/ 44.32

Comparison of different algorithms for small dataset has been shown in Figure 2. It verifies our result that Naïve Bayes performs best for small dataset. Comparison for large dataset is shown in Figure 3. It verifies our result that Random Forest performs best for large dataset. Technical indicators are removed to see the effect on accuracy of the algorithms when features to train the algorithm are removed. It is observed that accuracy of algorithms decreases when technical indicators are reduced.

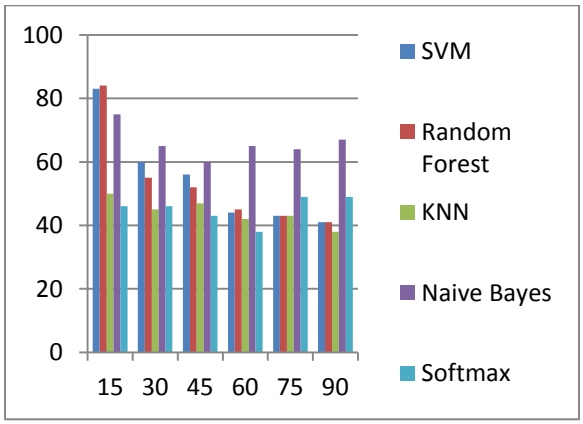

Figure 2. Comparison of algorithms for small dataset (Naive Bayes performs best).

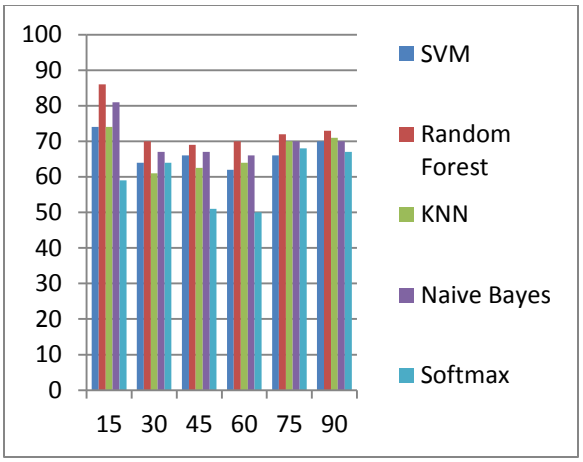

Figure 3. Comparison of algorithms for large dataset (Random forest performs best).

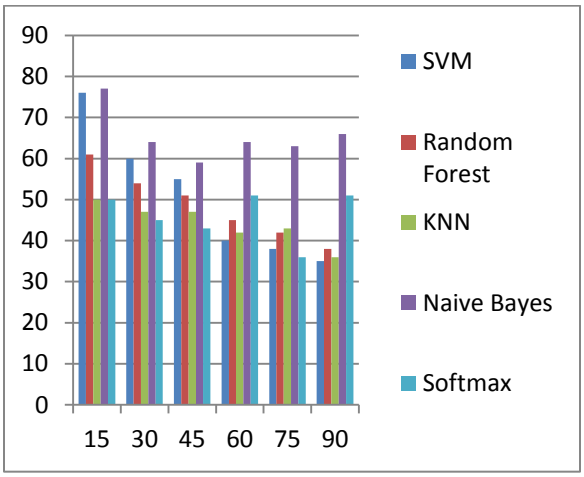

Figure 4. Comparison of algorithms for less number of technical indicators for small dataset (Compare with Figure 2).

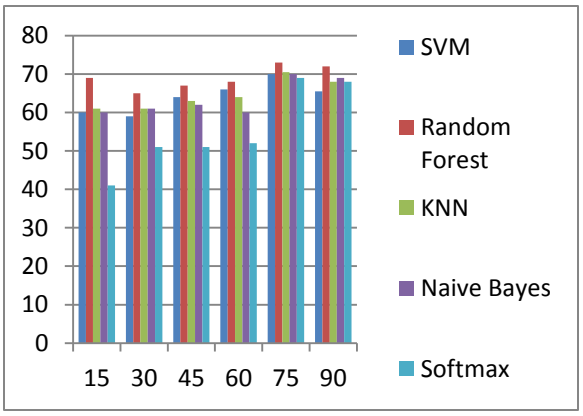

Figure 5. Comparison of algorithms for less number of technical indicators for large dataset (Compare with Figure 3)

Six technical indicators have been removed i.e RSI, Williams%R, MA 50, Disparity 10, RoC 2, CCI. Results after removal of technical indicators are shown in Figure 4 for small dataset and Figure 5 for large dataset.

VI. CONCLUSION

In this paper, Supervised machine learning algorithms SVM, Random Forest, KNN, Naive Bayes Algorithm, and Softmax Algorithm have been applied for the stock price prediction. The results reveal that for large dataset, Random Forest Algorithm outperforms all the other algorithms in terms of accuracy and when the size of the dataset is reduced to almost half of the original, then Naïve Bayes Algorithm shows the best results in terms of accuracy. Also, reduction in the number of technical indicators reduces the accuracy of each algorithm in predicting the stock market trends.

REFERRENCES

[1] G. Zhang, B.E. Patuwo, M.Y. Hu, Forecasting with arti-cial neural networks: the state of the art, Int. J. Forecasting 14 (1998) 35–62.
[2] K. Kim, Financial time series forecasting using support vector machines. Neurocomputing, 55, pp. 307–319, 2003.
[3] T. Manojlović* and I. Štajduhar*, Predicting Stock Market Trends Using Random Forest: A Sample of the Zagreb Stock Exchange, IEEE International Convention, pp. 1189-1193, 2015.
[4] Yuqing Dai, Yuning Zhang, Machine Learning in Stock Price Trend Forecasting, 2013.
[5] Steven B. Achelis, "Technical Analysis from A to Z", 2nd ed., McGraw-Hill Education, 2000.
[6] Koosha Golmohammadi, Osmar R. Zaiane and David Díaz, Detecting Stock Market Manipulation using Supervised Learning Algorithms, IEEE International Conference on Data Science and Advanced Analytics, pp. 435-441, 2014.
[7] P. Hajek, Forecasting Stock Market Trend using Prototype Generation Classifiers, WSEAS Transactions on Systems, Vol.11, No. 12, pp. 671-80, 2012.