Stock market analysis using ensemble learning

Madhavi Katamaneni[1] **and Laith Abualigah**[2,3,4,5]

[1]Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India.
[2]Computer Science Department, Prince Hussein Bin Abdullah Faculty for Information Technology, Al al-Bayt University, Mafraq 25113, Jordan.
[3]Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan.
[4]MEU Research Unit, Middle East University, Amman, Jordan.

[5]aligah.2020@gmail.com

Abstract. A stock market or share market is the combination of shoppers and sellers of shares. Prediction of the stock market is a method for calculating the future value of a company's stock. Stock market can be regarded as a specific records of data mining as well as machine learning problem. The daily changes within the stock depends on the profits and losses and many people think that stock market is irregular and uncertain. Based on daily changes we can predict some movements in the stock. In the previous years, researchers have used many machine learning models to know the development of the stock market to enhance the accuracy. However researchers have implemented these machine learning models independently and compared the results of models. It is observed that the group of models produce quite much less noisy compared to independent models. Our vision is to study investment strategies to predict and analyses the stocks using machine learning models such as Random Forest, K Nearest Neighbor, Gaussian Naive bayes and Decision Tree. And these models are ensemble by using different techniques such as Voting Classifier, Adaboost classifier and Bagging Classifier to estimate their accuracies. From the results it is observed that ensemble approach gives maximum accuracy compared to individual machine learning models.

Keywords: ensemble approach, random forest, K Nearest Neighbor, Gaussian Naive bayes, Voting Classifier, Adaboost classifier, Bagging classifier.

1. Introduction

Stock market is also a share market in which buyers and sellers interact with each other to sell their shares. For understanding of the stock market we must have detail knowledge on the enterprise reports and economic status. By knowing the true stock price of an enterprise, helps us to invest or buy the stocks of the company when there is increase in price. Estimating the tendancy in stock market is very complex for people who are investing in stocks because of various ups and downs involved in the values of the stock market. The analysis of the historical data independently and manually is very difficult in real world. There are many aspects that will impact the behavior of the stock prices. It includes the aspects like impact of the government, impact of the political issues, economic factors and various international trade. As a result stock values keep varying and this changes provides a space for various activities on increasing the complexity of the stock market. Different analysis has been used by the trading people to estimate the feature behavior of the stock values. These analysis of the stock market includes the collecting and integration of information about the stock values to make a proper decision for the investment in the company having less risk with greater benefits. Thus it not only help the investors to earn the profits for themselves it also helps for increase in the national economy.

In this paper we implemented machine learning algorithms along with ensemble learning models for the analysis of the stock market value. However many researchers have uses the machine learning models independently and they compared the results with the ensemble models it produce less noise compared to independent machine learning models. Machine learning techniques such as Random Forest, K-Nearest Neighbor, Decision tree, Guassian Naive Bayes are implemented and accuracy is estimated. Ensemble techniques such as Voting classifier, Adaboost classifier and Bagging classifier are implemented and accuracy is estimated. It is observed that integration of the machine learning with ensemble classifier produces the better accuracy comparing to independent models.

2. Literature survey

In 2020, Naadun Sirimevan, I.G. U. H. Mamalgaha, Chandira Jayasekara, Y.S.Mayuran[1], and Chandimal Jayawardena presented a paper. They used data from the Dow 30 firms from Yfinance, as well as old Twitter data from 2016 to July 2019 and Google trend data. Techniques used are long-short term memory neural network and also Recurrent neural network. It states that for time series forecasting, long short term memory outperforms endure techniques and recurrent neural networks, and that the effect of twitter data and web news sources of data is reduced for long term predictions, and that twitter is useful for short term forecasting. Beta values for stock prices should be examined for feature inputs in future research.

In 2019, Shika Mehta, Priyanka Rana, Ankita Sharma and Parul Agarwal presented a paper in which they used Yahoo Stock data from Yfinance. It contains a dataset which is of 10 years that is from April 1996 to April 2016. There are 8 features and 5039 instances in this set. In this they used techniques like Support vector machine, Multiple Regression and also long short-term memory. It states that the techniques like support vector machine, long short-term memory and multiple regression are ensembled in-order to acquire high accuracy which respectively obtain 98.56,97.63 and 99.02 accuracies. from the obtained accuracies it states that proposed ensemble technique gives superior results compared to the normal existing learners. Restriction is that accuracy rate of the multiple regression and ensemble technique are not increased much.

In 2019, Sukhman Singh, Tarun Kumar Madan, Jitendra Kumar and Ashutosh Kumar Singh presented a paper. In this they used historical data that is obtained from Yahoo Finance and also from NSE india and BSE india and many more. Support Vector Machine, Multi-Source Multiple Instance Learning Artificial Neural Networks, Random Forest and Boosted Decision Tree, Convolutional Neural Network and Long Short Term Memory techniques are utilized in the study of the above-mentioned data. It shows that Multi-Source Multiple Instance Learning has the accuracy of 60.1% and Random Forest has 80.8% accuracy and support vector machine gets the accuracy of 76.65%. One of the most significant challenges in stock market forecasting is that historical data alone is insufficient to anticipate the stock market. Another issue is that certain strategies rely on assumed values for a variety of themes and moods, which are not properly managed.

3. Methodology

In this paper stock market analysis is done using four different machine learning algorithms namely Random Forest, Decision Tree, Guassian Naive Bayes, K-Nearest Neighbor and these four models are grouped by using three different ensemble models Voting Classifier, Adaboost Classifier and Boosting Classifier.

In the above feature we are considering the close column and the volume column for the evaluation of stocks. The graph is plotted on close column it is represented as shown:

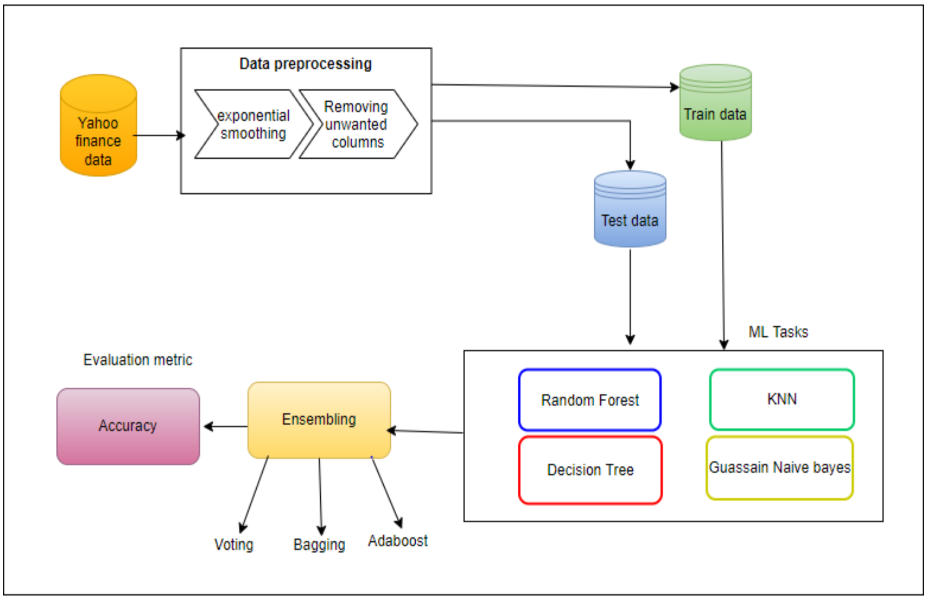


Figure 1. Architecture diagram.

As represented above the end output will be the accuracy of individual accuracies of the models and ensemble model accuracy.

Step-1:

At first we import historical yahoo stock dataset from the yahoo finance for the evaluation of the stocks. The data used in this is from the S&P 500, which has 8 characteristics and 6893 instances. The Yahoo set of data has a number of unique features [1]:

Step-2:

The head of the dataset is represented as shown i.e., the first six columns:It is observed that the above data is rough as it is time series

data it contains lot of spikes. Since the data is not smooth it is difficult to extract the trends of the stock. For this exponential smooth is done on the data. The graph obtained is:

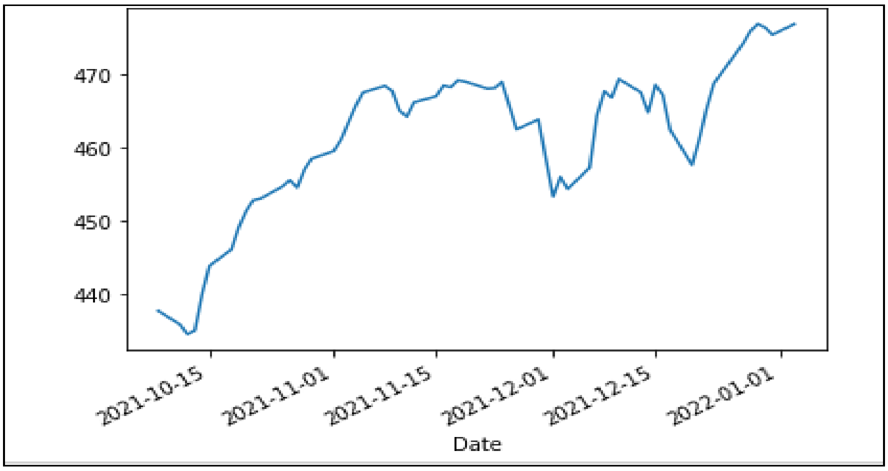


Figure 2. Close column after smoothing.

Now the data is much smoothed. The final library from the python eval function is used to compute the technical indicators. Computation of exponential moving averages at different lengths to a normalized volume. Columns like Open, High, Low, Adj close are removed. The ema's are calculated as:

Computation of ema's at different lengths: 1.data(ema50) =data(close)/data(close)ewm (50). mean ()

data(ema21) = data(close) / data(close)ewm(21). mean ()

data(ema15) = data(close) / data(close)ewm (15). mean ()

data(ema5) = data(close) / data(close)ewm (5). mean ()

The volume also normalized because it has correlation with price fluctuations. Rather than using the original volume column as:

data(normVol) = data(volume)/data(volume).ewm (5). mean () The dataset after calculating all the technical indicator values:

Step-3:

The data after preprocessing and computing the technical indicators are divided into train data and the test data.

Step-4: Algorithms used

Algorithms such as Random forest, K-Nearest Neighbor, Guassian Naive Bayes and Decision Tree are used.

3.1. Random forest

Random Forest is a supervised learning system that uses machine learning. In machine learning, random forest can be used for various classification and regression tasks. It is model that has a variety of decision tree for the given data set and this model helps in improving the accuracy of the dataset. It takes less time for the training comparing to the other machine learning algorithms [2]. Even in the case of large datasets its evaluates the output with high accuracy and works efficiently. In this work the computed technical indicators along with close column are used for the training of the model. Usually the stock market has the high noise data because of the huge size of the datasets, so based on the training columns it is easy to analyze the stocks [3].

3.2. K-nearest neighbor

K-nearest neighbor (KNN) is a machine learning algorithm that comes under supervised learning technique. It can be used in machine learning for both classification and regression analysis, however it is most commonly employed for classification. KNN is also called as Lazy Learner algorithm [4]. In KNN the test and historical stock data are mapped to set of vectors. Euclidean distance is taken as similarity metric for the prediction of decision. As it is lazy learning algorithm it does not have a model or the function which was previously built it uses the closest k-records of training data for testing. The prediction of stocks on close column is computed as by determining the number of k-neighbors and distance is computed between the records and the training values. The major portion of k-records are then used for stock market prediction after sorting all of the training data [5].

3.3. Guassian naive bayes

Guassian Naive Bayes is the extension of Naive Bayes which is a machine learning algorithm that comes under supervised learning technique. It is classification technique with high functionality. Guassian naive bayes is the simple algorithm that we can work beacuase we need to estimate only the mean and the standard deviation. It can be useful in the case of predictions by implementing all the predictions into Guassian Probability Density function [6].

p(x, mean, s) = (1 / (sqrt(2 * PI) * s)) * exp(-((x- mean^2)/(2*s^2)))

where p(x) is the guassian probability density function, mean and s are the mean and standard deviation, sqrt () is the function to compute the square root, PI is numerical value, exp () is the eulers number which is raised to power and x is the given input value.

3.4. Decision tree

Decision tree is one of the machine learning model that comes under supervised learning technique. Decision tree can be used for both classification and regression machine learning tasks. Decision tree is a graphical representation of all the solutions to a problem based on the given condition. It represents a tree like structure which starts with a root node and consists of branch nodes so it is called a decision tree. Decision tree works on the principle of the yes/no based on this its splits the tree into sub trees. In this paper the prediction of the stock market is made on the close column so it contains on decision node with different branch nodes containing the gini index values and sample values. The gini index for the obtained decision tree is calculated as [7]:

Gini Index= $1- \sum P^2$

Step-5: Ensemble learning

Ensembling is machine learning technique that combines various models and produces an optimized model. It is mainly used to increase the prediction, classification and many more.

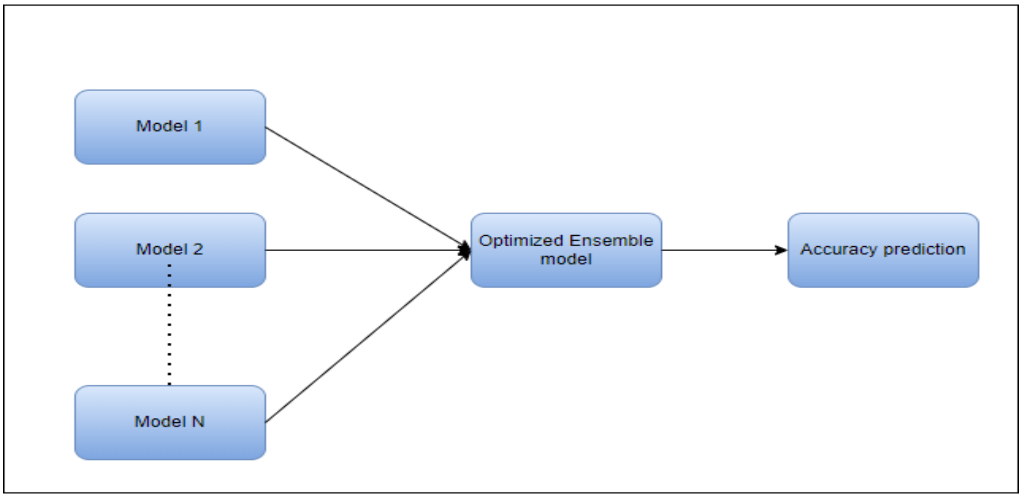


Figure 3. Ensemble flow diagram.

In this paper we have used three ensembling techniques namely Voting classifier, Adaboost Classifier and Bagging classifier.

3.5. Voting classifier

Voting classifier is an ensemble learning model. It trains the various models and predict them and then these models are aggregated together. Here each model is considered as a vote. The aggregation of each output is considered and the voting is majority classified into two types.

Hard voting

Soft voting

In case of hard voting each class output is considered and the class with the highest probability is taken as final one. Suppose if a class has both 1’s and 0’s if we get 1’s as the maximum then 1 will be the final output or viceversa.

In case of soft voting the output is taken as the average of all the models. Suppose if we have class A and class B and class A average is 0.345 and class B accuracy is 0.234 therefore class A will be considered as the final output. In this paper soft voting is implemented [8].

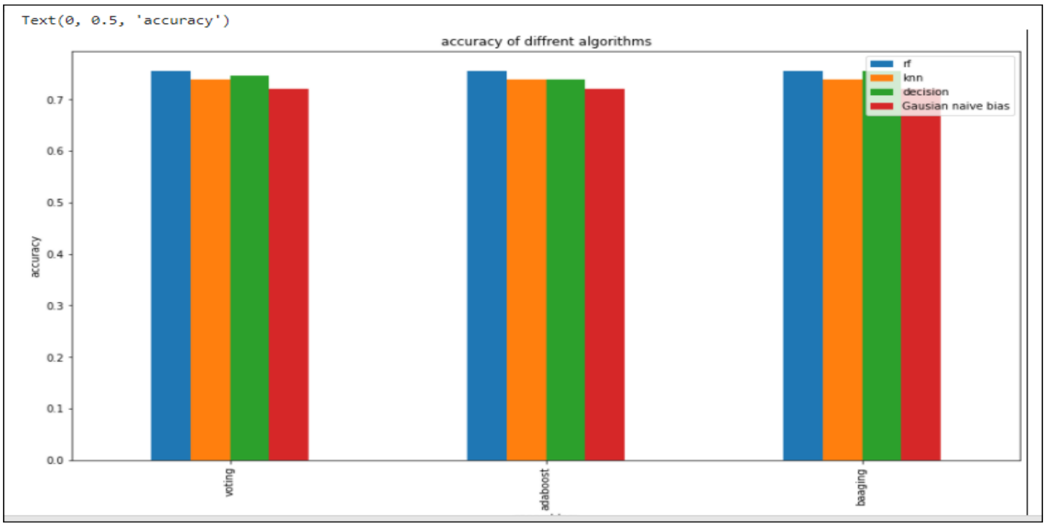
4. Results and observation

For the prediction of stocks we have use yahoo stock data which can be imported from yfinance api. Close column is considered for the analysis of stocks and it smoothen using exponential smoothing. The precision and recall values obtained from the close column are:

Accuracies of the models are calculated as:

1. Random forest accuracy: sum(rf_results)/len(rf_results)
2. KNN accuracy: sum(knn_results)/len(knn_results)
3. Decision tree accuracy: sum(dt_results)/len(dt_results)
4. Guassian Naive Bayes accuracy: sum(gnb_results)/ len(gnb_results)
5. Ensemble accuracy: sum(ensemble_results)/ len(ensemble_results)

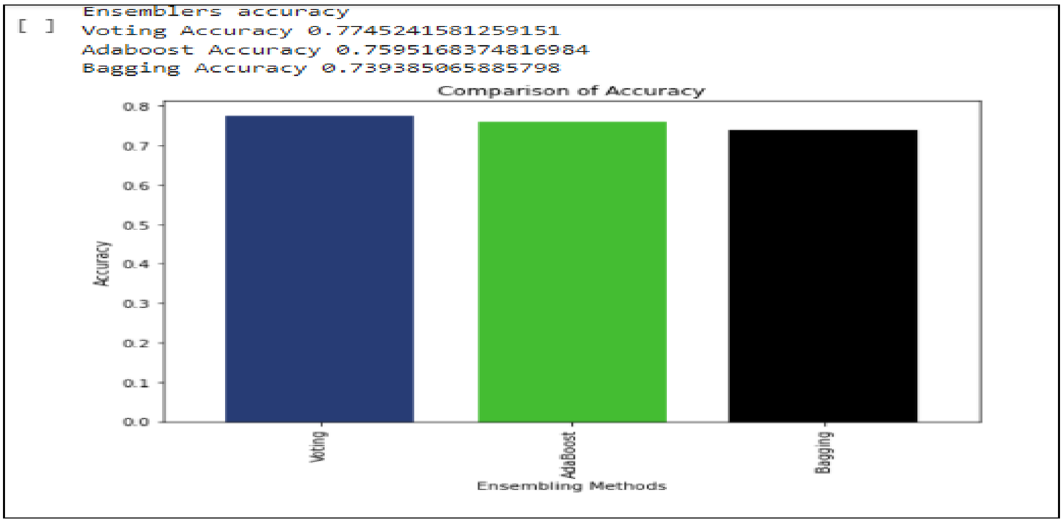
All the stated four machine learning models are incorporated with voting ensemble method to predict the accuracy. All the stated four machine learning models are incorporated with adaboost ensemble method to predict the accuracy. All the stated four machine-learning models are incorporated with bagging ensemble method to predict the accuracy. Accuracy comparisions of different machine learning algorithms such as random forest, knn, decision tree and guassian naive bayes.



```
Text(0, 0.5, 'accuracy')
```

Figure 4. Comparision of ML model accuracies.

Accuracy comparisions of ensemble models such as voting classifier, adaboost classifier and bagging classifier.



```
    Ensemblers accuracy
[ ] Voting Accuracy 0.7745241581259151
    Adaboost Accuracy 0.7595168374816984
    Bagging Accuracy 0.739385065885798
```

Figure 5. Comparision of ensemble model accuracies.

5. Conclusion

Stock market is viewed as a data mining and machine learning task. In this paper an ensemble model based on a variety of machine learning methods i.e. Random forest, K Nearest Neighbor, Decision Tree and Guassian Naive bayes are used. The model gives the accuracies obtained from Yahoo stock data. The ensemble models used in this project are voting classifier, adaboost classifier, bagging classfier. By comparing the accuracies from the ensemble models, voting is slightly higher than boosting and bagging.In future, more work will be done to find a specialized classification and regression algorithms which works best in our historical data set to produce high accuracy. Also to find the best fit ensemble algorithm which works best for the above algorithms to produce more accurate results while checking with cross validation in the field of Stock market.

References

[1] Yahoo finace:https://finance.yahoo.com/quote/SPY/history/

[2] Random forest: https://www.javatpoint.com/machine- learning-random-forest-algorithm

[3] Mehar vijh, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar, “Stock Closing Price Prediction using Machine Learning Techniques,” in *Internal Conference on Computational Intelligince and Datasceince. ICCIDS 2019.*

[4] KNN: https://www.javatpoint.com/k-nearest-neighbor- algorithm-for-machine-learning

[5] Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi, and Mohammed k. ali Shatnawi, “ Stock Price Prediction Using K-Nearest Neghbor (KNN) algorithm,” in *International journal of Business, Humanities and technology* vol.3

[6] Guassian Naïve Bayes: https://machinelearningmastery.com/naive-bayes-for- machine-learning/

[7] Decision Tree: https://www.javatpoint.com/machine- learning-decision-tree-classification-algorithm

[8] Voting Classifier: https://towardsdatascience.com/use- voting-classifier-to-improve-the-performance-of-your-ml- model-805345f9de0e

[9] Adaboost classifier: https://www.analyticsvidhya.com/blog/2021/09/adaboost- algorithm-a-complete-guide-for-beginners/