ENSEMBLE APPROACH TO STOCK MARKET PREDICTION

ABSTRACT

Stock market prediction is the method of determining the future value of a company stock or other financial instrument traded on a financial exchange. Investor's gain can be increased by a successful prediction. This paper proposes a machine learning model to predict stock market price. Researches have been done using many Machine Learning Algorithms like Support Vector Machine(SVM), to successfully predict the stock prices in the market . In this paper a scenario is proposed in which an ensemble approach is used with the best machine learning algorithms. The strategy will be considered profitable by judging its ability to identify stock prices accurately and consistently, proposing positive or negative returns and in the end it should use a learned model to produce the best results. The learned model is constructed as weighted support vector regression (SVR) classifier, Long-short term memory(LSTM) classifier, and Linear Regression classifier . The decision value would be chosen using a majority voting mechanism. Here, we are using simple ensemble technique – Max Voting. This ensemble approach has a lower error rate as compared to any other approach.

1.INTRODUCTION

A country’s capital market largely depends on the width and depth of the stock base.The financial growth of a country largely depends on the expansion and development of long term capital. The great renaissance noticed in the industrial world all over the world is due to shares and stocks. Interest in the purchase and sale of shares is not only shown by big companies and investors but small investors, individuals, salaried people, fixed income group. Shares are purchased when the prices are low in the market and sold when they are high. The margin is the profit to the investor. Even if there is minor improvement in the performance of stock market prediction , still great profit can be achieved.

Machine learning is the sub area of data mining where a model is developed in the computer by learning concept. This means that the model learns by training and testing over the given data. The model finally predicts new instances of data by making use of learning concept. In this paper, we have used Ensemble learning algorithm with classifier techniques namely Support Vector Regression, Linear Regression and Long short-term memory network(LSTM).

Ensemble model improves accuracy and robustness over single model methods.We have used ensemble in our research paper is because it has overcome the limitations of single hypothesis .The target function may not be implementable with individual classifiers, but may be approximated by model averaging. Thus we have used this method.

2.BACKGROUND STUDY

2.1 FINANCIAL MARKET PREDICTION

Financial Market Prediction in the past few decades has become a new hot research topic in the machine learning field. With the aid of powerful learning algorithms such as SVR (Support Vector Regression), Linear Regression and LSTM network, researchers overcame numerous difficulties and achieved considerable progress.

2.2 OVERVIEW OF THE ALGORITHM

Previously SVM and Linear Regression were used for many classification problem and there were many instances where reasonable accuracy was achieved. Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm. The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. Keeping that in mind our given problem statement was addressed to achieve as much accuracy as possible with tweaking algorithm. Multiple Regression is useful for finding relationship between two or more continuous variables. LSTM was used for many time series problems and the same approach is used to predict trend for stock by memorizing history data.

2.3 ENSEMBLE THEORY

An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles can be shown to have more flexibility in the functions they can represent. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to **decrease variance** (bagging), **bias** (boosting) or **improve predictions** (stacking).

2.4 ENSEMBLE LEARNING IN FINANCE DOMAIN

The accuracy of prediction of business failure is a very crucial issue in financial decision-making. Therefore, different ensemble classifiers are proposed to predict financial crises and distress. Also, in the train based manipulation problem, where traders attempt to manipulate stock price by buying and selling activities, ensemble classifiers are required to analyze the changes in the stock market data and detect suspicious symptom of stock price manipulation.

3.APROACH / PROPOSED WORK

Approach used in this paper, is Ensemble Learning.

Ensemble method is a machine learning technique that combines several base models in order to produce one optimal model. Here, unlike other single methods ensemble methods try to construct a set of hypotheses and combine them to use. Ensemble has various number of learners which called base learners. The ability of generalization of an ensemble is usually stronger than that of base learners. Actually, ensemble learning is appealing because that it is able to boost weak learners which are slightly better than strong learners which can make very accurate predictions. “Base learners” are also called as “weak learners”. In this paper base learners are generated from training data using base learning algorithms as Support Vector Regression (SVR), Long-short term memory(LSTM), and Linear Regression.

3.1 ENSEMBLER CONSTRUCTION

An ensemble is constructed in two steps. In the first step, all the base learners are produced, all these learners are generated in a parallel style where the generation of a base learner has influence on the generation of subsequent learners. In the next step, these base learners are combined to use, where among the most popular combination schemes that is majority voting for classification and weighted averaging for regression are used.

3.2 ALGORITHMS USED

Ensemble learning algorithm is formulated using three machine learning algorithms Support Vector Regression(SVR), Linear Regression and Long short-term memory network(LSTM).

3.2.1 MULTIPLE REGRESSION

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding the relationship between variables. Regression techniques mostly differ in two ways ,one the number of independent variables and second, type of relationship between the independent and dependent variables.

Regression is in which the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable is known as Linear Regression.

Multiple regression is an extension of simple **linear regression.**

Multiple regression is useful for finding relationship between two or more continuous variables. One is predictor or independent variable and others are responses or dependent variables. It looks for statistical relationship but not deterministic relationship. Relationship between variables is said to be deterministic if one variable can be accurately expressed by the others. Multiple regression can be expressed by following equation:

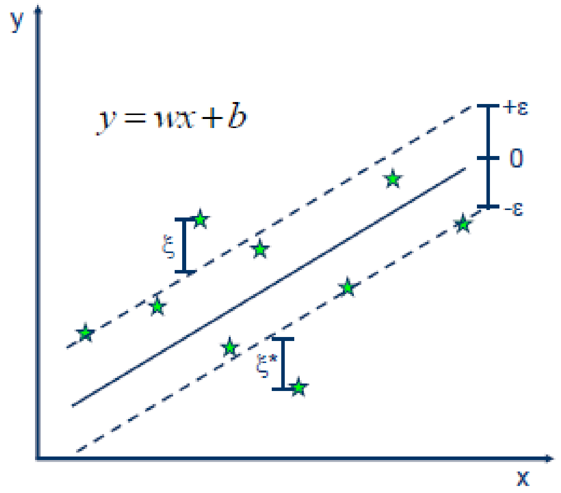
y = a0+(a1*s)+(a2*t)+(a3*u)+(a4*v)   : Multiple Regression

The motive of the linear regression algorithm is to find the best values for a0,a1,a2,a3 and a4.

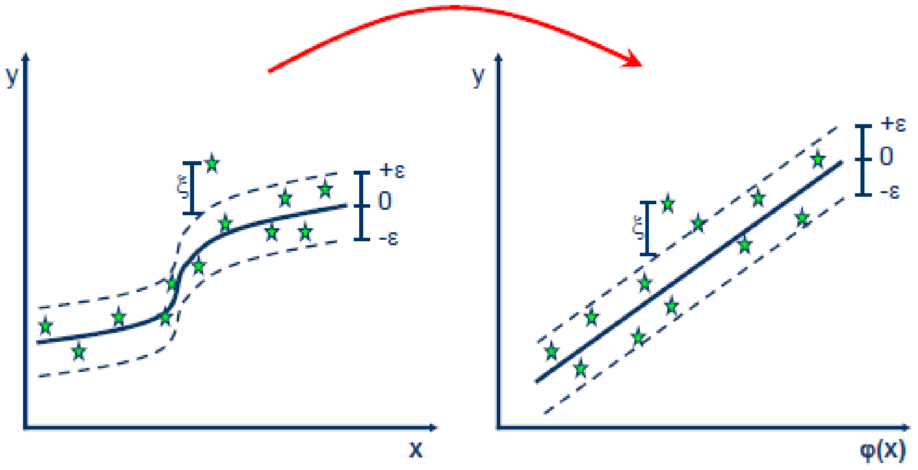
Here , input parameters are s ,t ,u and v. The number of times we train our model to find the best slope and bias for our model to fit the data is known as epochs. Finally, the speed of convergence, i.e how fast gradient descent finds the best parameters is known as the learning rate.

3.2.2. Support Vector Regression(SVR)

Support Vector Machine is also used as a regression method, keeping same all the main features that characterize the algorithm. The same principles as the SVM is used by Support Vector Regression (SVR) for classification, with some minor differences .Error occurs because output is a real number which makes it very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, epsilon (margin of tolerance) is set in approximation to the SVM which would have already requested from the problem. However ,main idea is: individualize the hyperplane which maximizes the margin ,to minimize error, , also keeping that part of the error is tolerated.



In this paper we have used NON LINEAR SVR which uses kernel functions that transform the data into a higher dimensional feature space to make it possible to perform the linear separation.



Kernel function that is used is:

Gaussian Radial Basis function

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\frac{\left\| \mathrm{x}_i - \mathrm{x}_j \right\|^2}{2\sigma^2} \right)$$

3.2.3 Long short-term memory network(LSTM)

Long Short Term Memory networks(LSTM)– are an special sort of RNN, equipped for learning long-term dependencies. LSTMs are explicitly designed to avoid the long-term dependency problem. A RNN is a sort of neural system where the yield of output of a block is fed as input to the next iteration, but it has various problems such as vanishing gradient that is overcome by LSTM.

In LSTM module called repeating module has four neural network layers interacting in a unique fashion as shown below in figure.The fundamental component of LSTMs is cell state, a line running from Memory from Previous Block ($C.t-1$) to Memory from Current Block ($C.t$). It allows the information to flow straight down the line.

Procedure:

• • Step 1: In the first step, we decide what information we are going to throw away and what information is to be used. It is controlled through first layer , called “Forget gate layer”

Eqn : $f.t = \sigma(W.f\ [h.t-1,x.t] +b.f)$

• • Step 2: In the second step, we decide what new information is to be stored.this is done in two steps:

• 1) First, in input gate layer we decide what information is to be updated, and we update that layer.

Eqn: $i.t= \sigma(W.i\ [h.t-1,x.t] +b.i)$

2)Second, in tanh layer we create a new vector of new candidates , that are to be added to the state.

Eqn: $c\sim.t=\tanh(W.c\ [h.t-1,x.t] +b.c)$

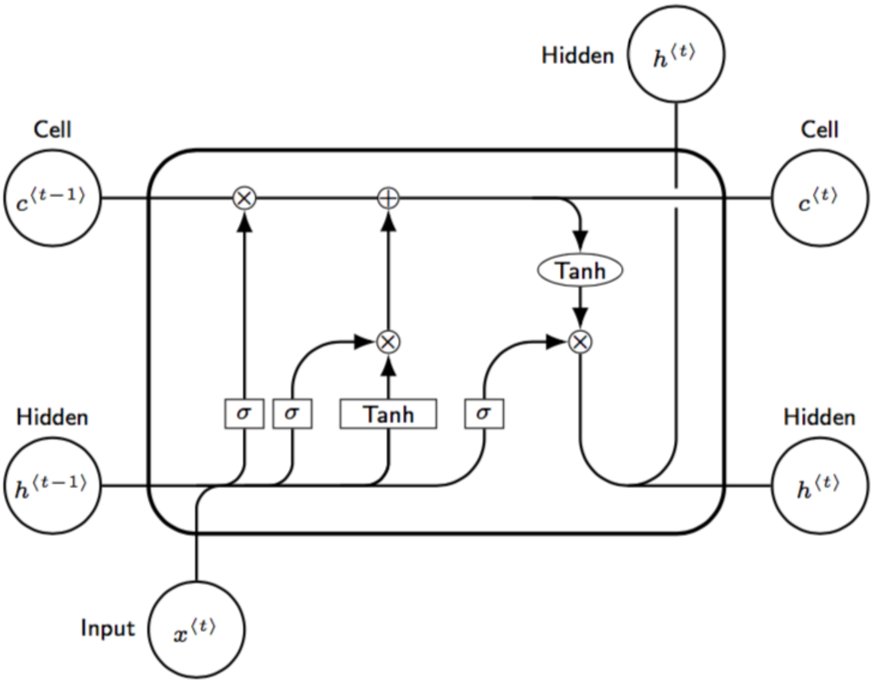
• • Step 3: In this step , we update the old cell state .

Eqn: $ct=f.t*c.t-1 + i.t * c\sim.t$

• • Step 4: Here, we compute the final output,using last sigmoid layer ,and then multiply it by new cell state.

Eqn: $o.t= \sigma(W.o[h.t-1,x.t] +b.o)$

Final output eqn : $h.t=o.t * \tanh[c.t]$



3.3 ENSEMBLE APPROACH

The ensemble method here used is WEIGHTING VOTING METHOD. In this method first we predict the prices using the base learners (Support Vector Regression classifier , LSTM classifier , Linear Regression classifier) , then accordingly assign the weights to the base learners on the basis of there accuracy results .

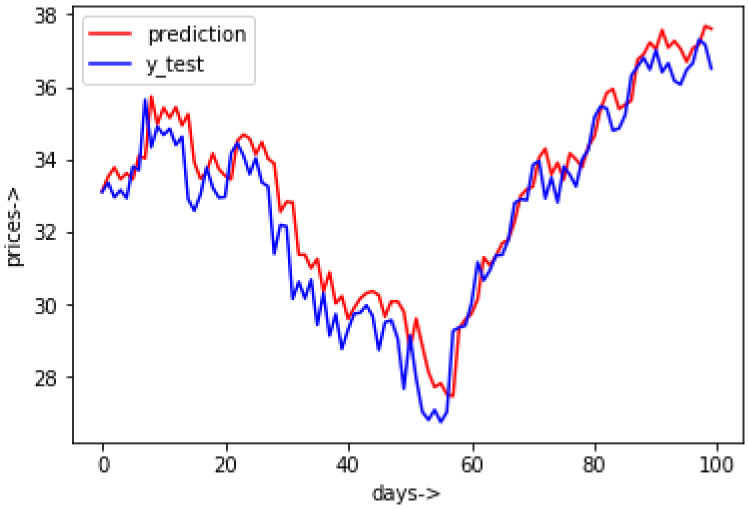
Then finally we will combined these base learners and apply the ensemble learning method.

4.EXPERIMENTAL & RESULTS

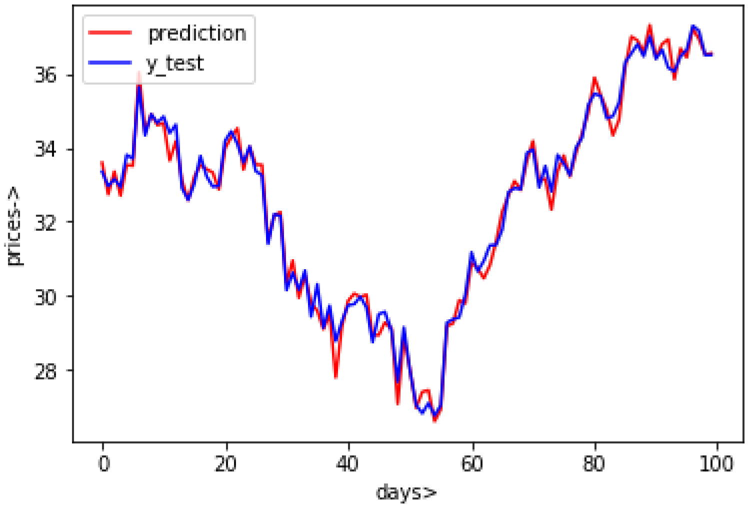
In this paper we have evaluated the performance of our various algorithms(*Support Vector Regression*, *Linear Regression and Long short-term memory network(LSTM)*) individually as well when they are combined using ensemble learning.

We have used Yahoo Stock () in evaluation of these algorithm, and below are the results given by these algorithms.

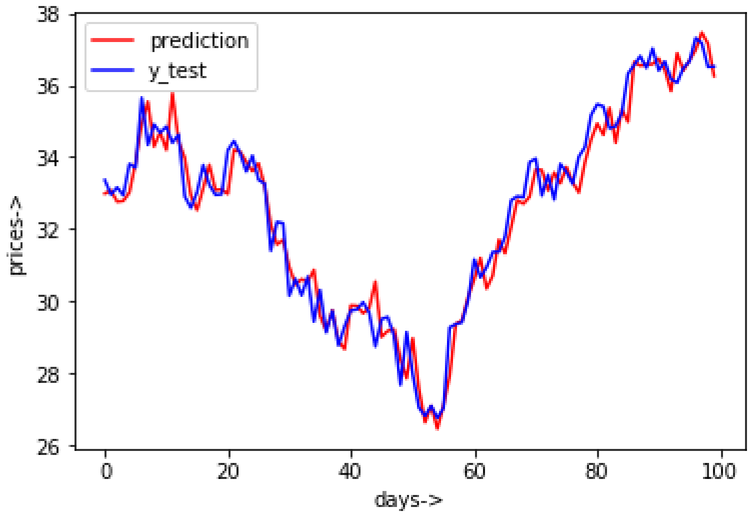
4.1 GRAPHICAL RESULTS



Long short-term memory network(LSTM)



Multiple Regression



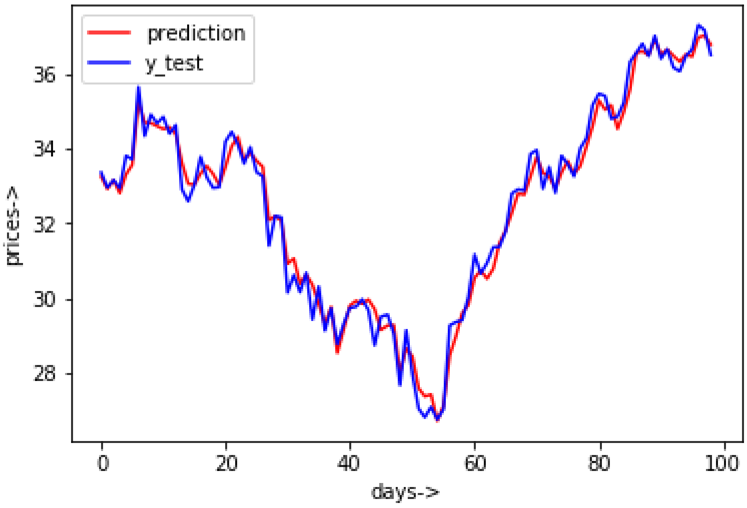
Support Vector Regression

4.2 MEAN ACCURACIES

ALGORITHM	MEAN ACCURACY
Support Vector Regression	98.56306691
Linear Regression	99.02800116
Long short-term memory network	97.63474000

These are the results when all these three algorithms are evaluated individually . Now, we will apply ensemble learning using these three machine learning algorithms.

RESULTS FOR ENSEMBLE LEARNING:



Ensemble model has a mean accuracy of **99.07562496.**

After observing the above results, we have found out that though there is not much increase in mean accuracy rates of the algorithm (because accuracy rate of linear regression is 99.0280116 and ensemble learning is 99.07562496),but from the graphs we get that the deviation between the actual and predicted price has significantly reduced in ensemble as compared to any other of the three algorithms for any particular day(in the ensemble model graph red and blue line almost completely overlap each other) .

Therefore, ensemble learning is more appropriate model than other single models.

5. REFERENCES

[1] Phayung Meesad and Risul Islam Rasel "Predicting Stock Market Price Using Support Vector Regression " in International Conference on Informatics, Electronics & Vision (ICIEV 2013), At Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh

[2] Chia-Hua Ho and Chih-Jen Lin "Large-scale Linear Support Vector Regression" in Journal of Machine Learning Research 13 (2012)

[3]Lucas Nunno "Stock Market Price Prediction Using Linear and Polynomial Regression Models" At Albuquerque, New Mexico, United States

[4]Thomas G Dietterich "Ensemble Methods in Machine Learning "At Oregon State University Corvallis Oregon USA

[5] David Opitz and Richard Maclin "Popular Ensemble Methods: An Empirical Study "At Journal of Arti?cial Intelligence Research 11 (1999)

[6]Klaus Greff And Rupesh K. Srivastava And Jan Koutn´ik And Bas R. Steunebrink And J¨urgen Schmidhuber "LSTM: A Search Space Odyssey" At TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

[7]Sepp Hochreiter And Jurgen schmidhuber " LONG SHORT TERM MEMORY " at Neural Computation 9 In 1997

[8] DANIEL SKANTZ and WILLIAM SKAGERSTRÖM “Stock forecasting using ensemble neural networks” at STOCKHOLM, SWEDEN 2018