ENSEMBLE APPROACH TO STOCK PREDICTION

MINOR PROJECT I

Submitted by:

ANKITA SHARMA (9916103082)
PRIYANKA RANA (9916103155)
SHIVAM SINGH (9916103193)
MAYANK MATHUR (9916103119)

Under the supervision of:
Dr. SHIKHA MEHTA

Department of CSE/IT
Jaypee Institute of Information Technology University, Noida

NOVEMBER 2018

ABSTRACT

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable.

Here, trading strategy is formulated using machine learning algorithms. The strategy will be considered profitable by judging its ability to identify stock indices accurately and consistently, proposing positive or negative returns.

The learned model is constructed using ensemble learning in which following algorithm are used: Support vector regression (SVR), Multiple Regression, and LSTM. The weights are assigned according to the accuracies of the respective algorithms taken. The ensemble method is augmented by a weighted average method.

We have used these algorithms as previously SVR and Multiple Regression were used for many regression problem and there were many instances where reasonable accuracy was achieved. Keeping that in mind our given problem statement was addressed to achieve as much accuracy as possible with tweaking algorithm. LSTM was used for many time series problems and the same approach is used to predict trend for stock by memorizing history data.

ACKNOWLEDGEMENT

We would like to place on record our deep sense of gratitude to Dr. Shikha Mehta , Dept. of Computer Science and Engineering , Jaypee Institute of Information Technology, India ,for her generous guidance, continuous encouragement and supervision throughout the course of present work.

We also wish to extend our thanks to our seniors and classmates for their insightful comments and constructive suggestions to improve the quality of this project work.

Signature(s) of Students

Ankita Sharma (9916103082)
Priyanka Rana (9916103155)
Shivam Singh (9916103193)
Mayank Mathur (9916103119)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LSTM	Long Short Term Memory
SVM	Support Vector Machine
SVR	Support Vector Regression
RNN	Recurrent Neural Network

Chapter 1: INTRODUCTION

A country's capital market largely depends on the depth and width 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. Not only big companies and investors are widening but small investors, individuals, salaried people, fixed income group are also nowadays interested in the purchase and sale of shares. 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. A minor improvement in the performance of stock market prediction can give a great profit.

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. Here we have used Ensemble method with regression techniques namely Support Vector Regression(SVR) , Multiple Regression , Long short-term memory network(LSTM). We are using ensemble method to avoid the loss of capital in stock prediction.

Chapter 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), Multiple Regression and LSTM network, researchers overcame numerous difficulties and achieved considerable progress.

2.2 OVERVIEW OF THE ALGORITHM

Previously SVM and Multiple 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.

Chapter 3: REQUIREMENT ANALYSIS

3.1 Software

a. Jupyter Notebook

b. Spyder 3.1

3.2 Hardware

a. A computer system

b. RAM – 4GB

3.3 Functional requirements

a. Retrieval of Datasets

b. Data Pre-processing

c. Feature extraction

d. Classification and segregation on the basis of accuracies

3.4 Non-Functional requirements

a. Secure access of data

b. Scalability and performance

c. Reliability

d. Time response

e. Normalization

Chapter 4: DETAILED DESIGN

4.1 DATASET CREATION :

For data collection we used Yahoo finance. We collected Yahoo Stock data from Yahoo Finance. We collected data from April 1996 to april 2016. The stock market took a toll during 2007-2008 financial crisis. This time around, companies went in loss and stock data of companies completely unpredictable. Training our model using this data would cause our system to be less accurate because of a lack of trend during the crisis period. So,we avoid data that can result into uncertain behaviour.

For our problem we kept technology and software services as a sector. The stock data obtained contains the following parameters:

- Date
- Open
- High
- Low
- Close
- Adj Close
- Volume

We took the daily closing values of each of the stock as the stock value for a day.

4.2 ALGORITHMS :

LSTM:

LSTM stands for Long Short Term memory. It is building block of a neural network . LSTM blocks are used to build a recurrent neural network. An RNN is a type of neural network where the output of a block is fed as input to the next iteration. An LSTM block is composed of four main components: a cell, an input gate, an output gate and a forget gate. The cell is responsible for "remembering" values over arbitrary time intervals; hence the word "memory" in LSTM. Each of the three gates can be thought of as a "conventional" artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum.

SVR:

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. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated. We have used the non linear SVR(Radial Basis Function) .

Gaussian Radial Basis function

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

Multiple 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.

4.3 METHODS USED:

We have used ensemble method.

Ensemble method helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. In stock prediction it also helps to reduce the difference between predicted and actual stock on a particular day.

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).

Ensemble Techniques

1) Max Voting:

The max voting method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a 'vote'. The predictions which we get from the majority of the models are used as the final prediction.

2) Averaging:

In this method, we take an average of predictions from all the models and use it to make the final prediction. Averaging can be used for making predictions in regression problems or while calculating probabilities for classification problems.

3) Weighted Average:

This is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction. In our model we have used weighted average.

4.4 TOOLS:

Language Used: Python

The implementation is carried out in python. This language supports both object oriented and functional programming and is compatible with machine learning methods. It has many well documented and easy to use libraries that can help accomplish lot of different programming tasks. We have used the following libraries:

- **Numpy**
- **Scikit-learn**
- **Pandas**
- **Keras**

Chapter 5: IMPLEMENTATION

Multiple Regression:

We are using the linear model from scikit learn library. We are using Open , High, low, Volume as our dependent variable in multiple regression for the prediction of closing price of stocks of the particular day. We are predicting the closing stock price of latest 100 days of the dataset and we are using the 90% of the remaining stock details for the training set and 10% for testing set. After, the prediction of the stocks and for the purpose of comparing accuracy with the actual closing price we are plotting graphs between them using the matplotlib library and are calculating the mean accuracy.

LSTM:

We are implementing LSTM using keras deep learning library . We are using open and high features for the prediction of closing stock prices of latest 100 days . We have used our dataset in sets of 5 and then predicted every 5^{th} value of that set . We are using the 90% of the remaining stock details for the training set and 10% for testing set. For training purpose we have added 128 classes in the hidden layer than by adding another dense layer with 64 classes we have narrowed it down. We have further narrowed it down to 16 classes then to the final output, we have used Rectified Linear Unit as our activation function. We are then computing the mean accuracy and are plotting graph between predicted and actual stock prices.

SVR:

We are implementing SVR using Scikit learn library. We have used open feature as a base for prediction of closing price for our regression

technique .We have used Radial Basis Function for the prediction of closing stock.

Ensemble Method:

We have multiplied the outcomes of the SVR, LSTM, Multiple Regression with the respective weights that were assigned to them on the basis of their accuracy the output is then divided by the sum of all weights assigned to the algorithm.

Chapter 6: EXPERIMENTAL RESULTS AND ANALYSIS

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

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

6.1 GRAPHICAL RESULTS

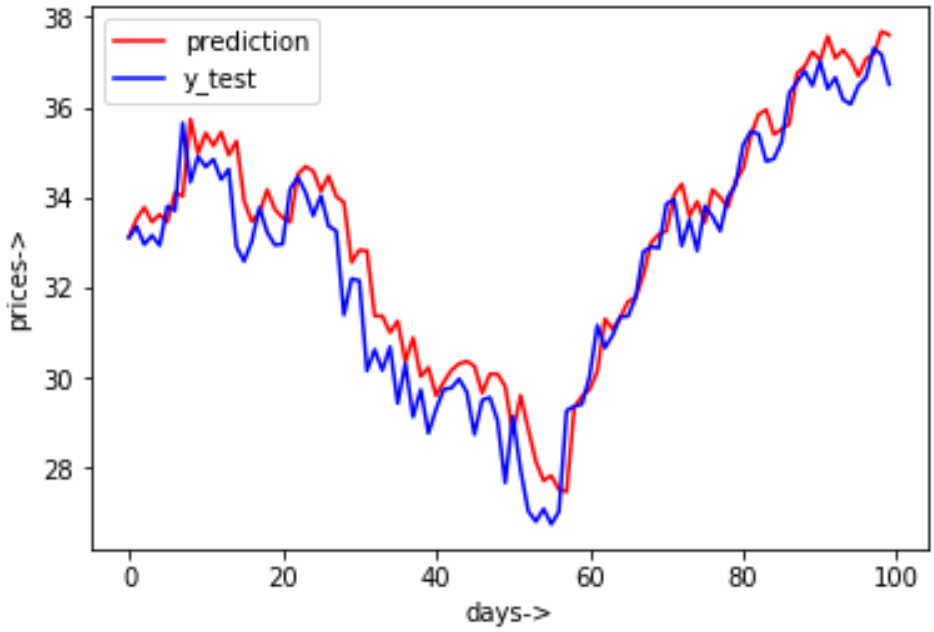


Fig 6.1 Long short-term memory network (LSTM)

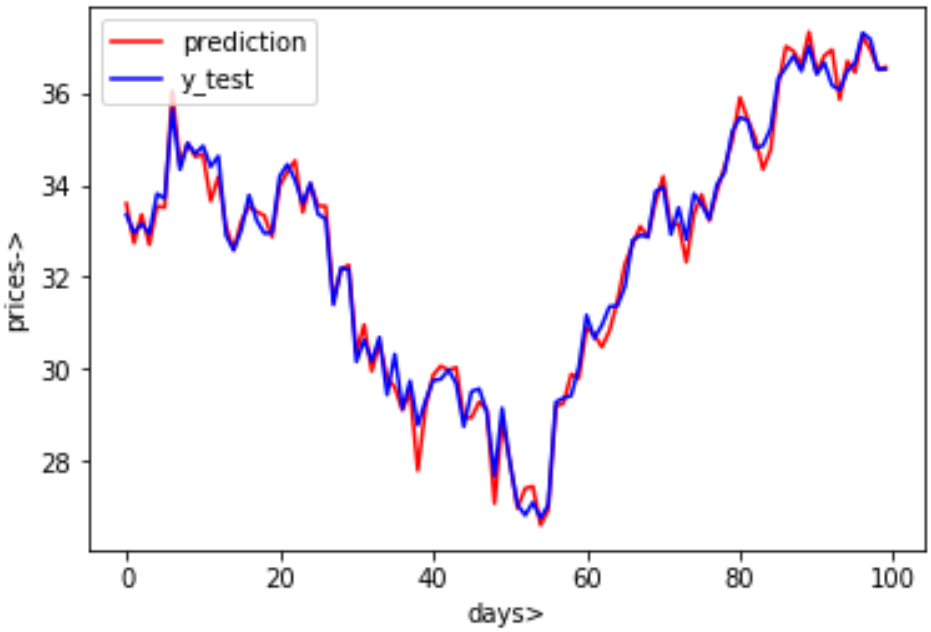


Fig 6.2 Multiple Regression

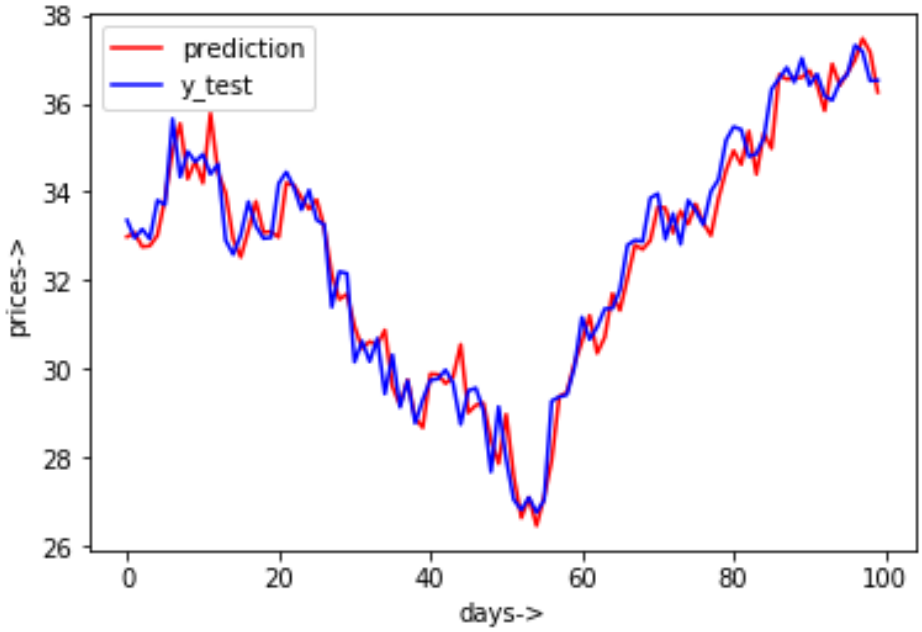


Fig 6.3 Support Vector Regression

6.2 MEAN ACCURACIES

ALGORITHM	MEAN ACCURACY
Support Vector Regression	**98.56306691**
Multiple Regression	**99.02800116**
Long short term memory network	**97.63474000**

Table 6.1 Mean Accuracy

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

6.3 RESULTS FOR ENSEMBLE LEARNING

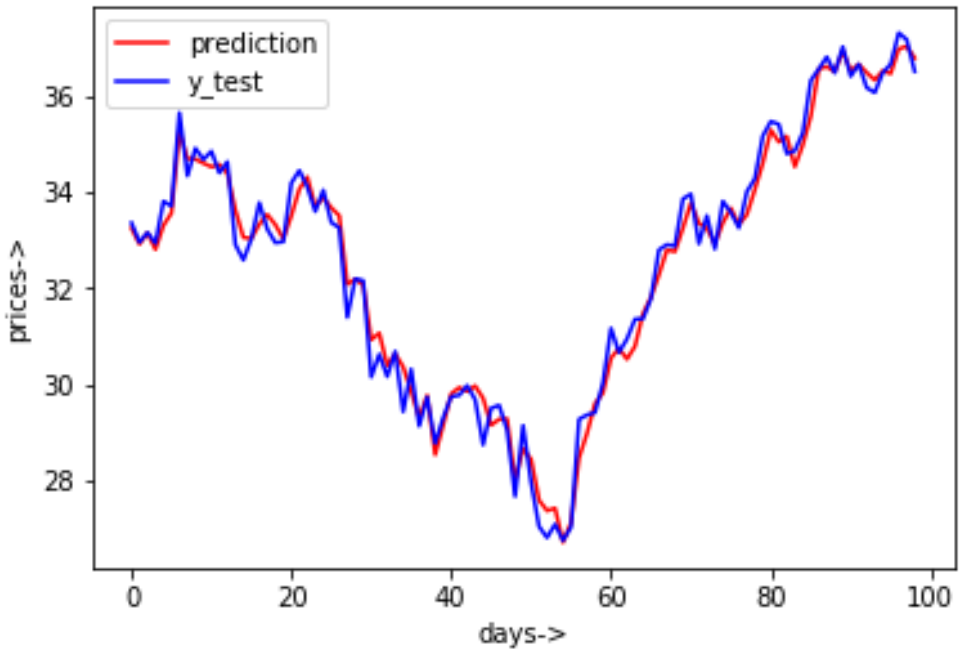


Fig 6.4: Ensemble Model

Mean Accuracy: 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.

Chapter 7: CONCLUSION AND FUTURE SCOPE

In this project, we have demonstrated an ensemble approach to predict stock market trend using different machine learning algorithms. Result shows how we can increase the accuracy and efficiency of stock predictions using ensemble learning .For this implementation, we would like to conclude that if we incorporate all the factors that affect stock performance and feed them to our algorithms with proper data preprocessing and filtering, after training the network we will be able to have a model which can predict stock prices very accurately and this can result into better stock forecasting and profit for financial firms.

REFERENCES IN IEEE FORMAT

[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

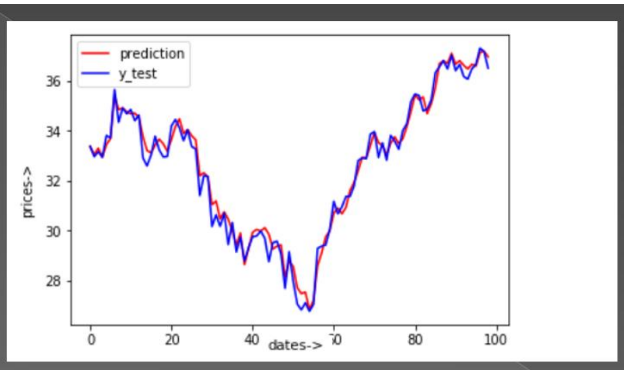
[4] Thomas G Dietterich "Ensemble Methods in Machine Learning "At Oregon State University Corvallis Oregon USA

[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



ENSEMBLE APPROACH TO STOCK PREDICTION

Problem statement

In this project, we have demonstrated an ensemble approach to predict stock market trend using different machine learning algorithms. Result shows how we can increase the accuracy and efficiency of stock predictions using ensemble learning.

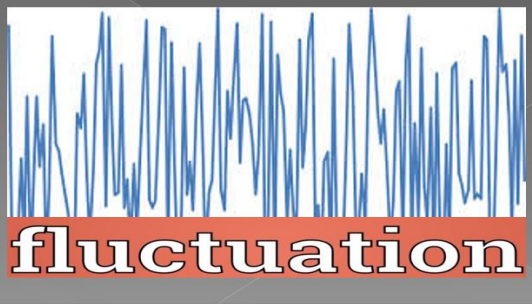
State-of-the-art and their limitations

1) State-of-the-art :-

- Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one.
- In our project we will be focussing on the best of the algorithms we have studied so far from different research papers and then based upon their accuracies, we will assign them preferences and finally applying them in ensemble learning to predict the future stocks for a particular company.

2) Limitations:-

- Presence of Nan values in the dataset.
- Limitation of parameters.
- Complexity of enviornment .



Objectives

In our project we our going to implement a novel way of predicting future stock prices by combining all these algorithms into a single unit (Ensemble learning) which will give better accuracy .

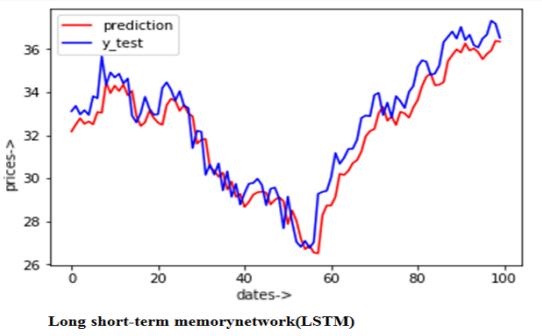
	LSTM	+ MULTIPLE REGRESSION	+ SVR
WEIGHTS	(1)	(3)	(2)

Ensemble Learning

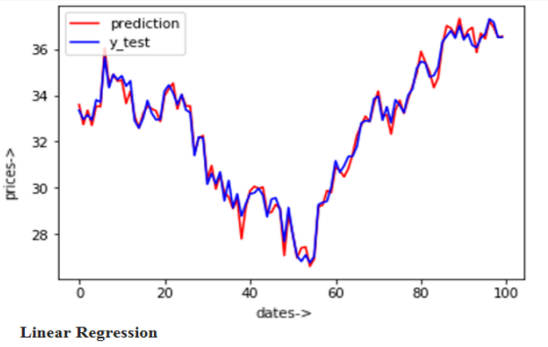
IMPLEMENTATION

- We have multiplied the outcomes of the SVR , LSTM , Multiple Regression with the respective weights that were assigned to them on the basis of their accuracy.
- The output is then divided by the sum of all weights assigned to the algorithm.

GRAPHICAL RESULTS

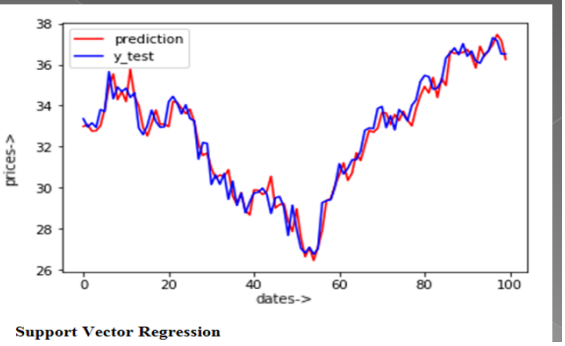


Long short-term memorynetwork(LSTM)



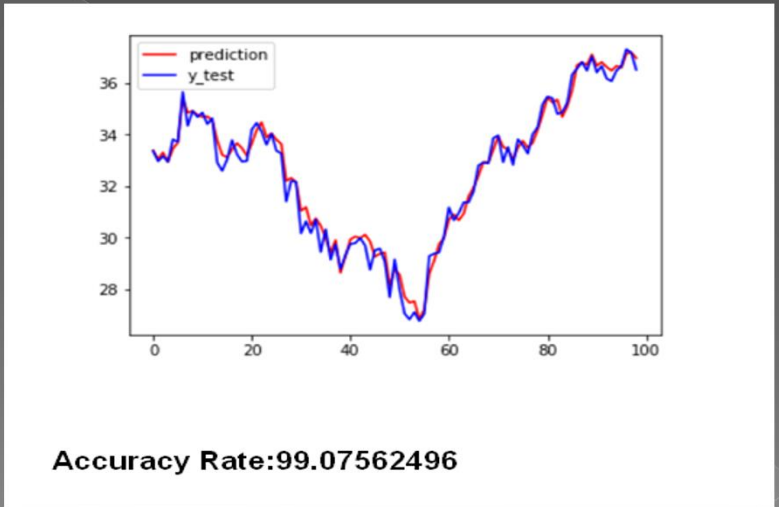
Linear Regression

ALGORITHM	ACCURACY RATE
Support Vector Regression	98.56306691
Linear Regression	99.02800116
Long short term memory	97.63474000



Support Vector Regression

RESULTS FOR ENSEMBLE LEARNING



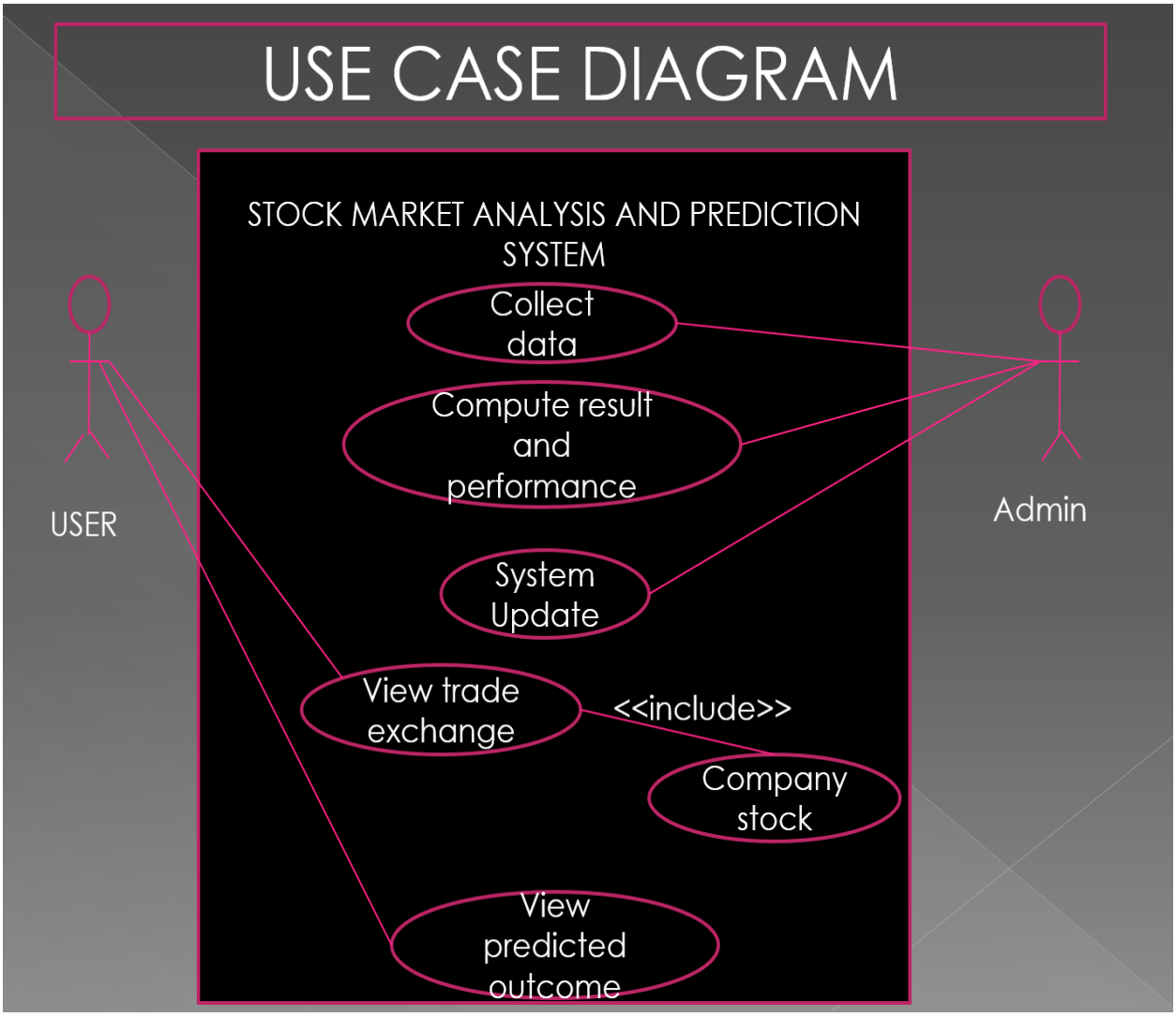
Conclusion

- In this project, we have demonstrated an ensemble approach to predict stock market trend using different machine learning algorithms.
- We predict the prices using the base learners (Support Vector Regression , LSTM , Multiple Regression) , then accordingly assign the weights to the base learners on the basis of there accuracy results .
- Result shows how we can increase the accuracy and efficiency of stock predictions by taking best part of every algorithm

Future scope

We hope that our project can predict stock prices in real time by taking more parameters with a great accuracy .

We dream to make a bot which predict stock prices by using ensemble learning for users to get some profit without any loss.


USE CASE DIAGRAM
STOCK MARKET ANALYSIS AND PREDICTION SYSTEM
Collect data
Compute result and performance
System Update
View trade exchange
<<include>>
Company stock
View predicted outcome
USER
Admin