Analysis of various machine learning algorithm and hybrid model for stock market prediction using python

Sahil vazirani, Abhishek sharma and Pavika sharma
Dept. of ECE, Amity School of Engineering & Technology, Amity University Uttar Pradesh
abhiiisharmaaa003@gmail.com , sahilvazirani1998@gmail.com and psharma12@amity.edu

Abstract—With the up-gradation of technology and exploration of new machine learning models, the stock market data analysis has gained attention as these models provide a platform for businessman and traders to choose more profitable stocks. As these data are in large volumes and highly complex so a need of more efficient machine learning model for daily predictions is always looked upon. Therefore, in this work an extensive comparative analysis of already implemented models for stock trade market has been performed. On the basis of results obtained, new linear regression models are proposed which provide much significant error reduction. It was observed that appending two linear regression models where output of first block is fed to the input of second linear regression model gives the most efficient prediction model.

Keywords: Stock Market Prediction, Machine Learning, KNN, SVM, decision tree, MAE, MSE, RMSE

INTRODUCTION

The stock market prediction attempts to determine the stock values and provides the idea to public to know about the product in the market and also about its stock prices. It also symbolizes country's overall economy. Stock market produces large amount data in short time and it changes every moment, this makes it unpredictable. Stock price depends on companies buying and selling shares of publicly-held companies and data is released regularly in terms of opening value, closing value, high value and other details of stock. This data is public and protected by the regulation for transparency. Stock price prediction is highly important for everyone, especially for traders and businessman, who make large investments in these stocks. Due to the stock market arbitrary nature people lose and gain money randomly most of the times. After investing in the stocks, investor expects the capital in which he invested will give him high returns and when he sells the same profit is earned. Because of this high risk associated with the high returns, large people are investing and so in turn the increased demand of shares that will give a profitable outcome is there. The only solution to this problem is to analyze previous stock market data of the company and understand the pattern depending upon different factor that were associated for that specified company. Since the data is in large volume and it can't be processed instantly so machine learning algorithms are used for analyses and predicting such values.

Supervised Learning: At the point where a calculation gains from model information, related objective reactions which will comprise of Numerical qualities, for example, Features and labels, so as later foresee right reaction which is presented with new models goes in the classification of the supervised learning technique. The methodology for sure like that the person is learning under the guidance of the instructor and instructor gives genuine guides to the understudy to remember, and the understudy at that point gets general standards from these particular models.[1]

Unsupervised Learning: in this a known dataset has the set of observations with the features but the response is not known. The predictive model uses features to identify how to classify and represent the data points of new or unseen data Reinforcement Learning: At the point we also need names when we present our calculation with model it is like learning solo. With the positive or negative criticism, we can go with the model as per the arrangement the calculation proposes goes under the class of Reinforcement realizing, which is associated with applications for which the calculation must decide, and the choice bear the outcomes. In the human world, it is much the same as learning by experimentation. Blunders assist you with learning since they have a punishment included (cost, loss of time, lament, torment, etc.), instructing you that a specific strategy is less inclined to prevail than others. An intriguing case of support learning happens when PCs figure out how to play computer games without anyone

CONCEPTUAL FRAMEWORK FOR BASIC MACHINE LEARNING FOR STOCK DATA ANALYSIS

1. KNN algorithm

KNN stands for k nearest neighbour. It is very good algorithm for nonlinear classified datapoints. It can used for both classification and regression problem. it is mostly used in cases where the plot is nonlinear and used in search application.

Working of KNN model

Suppose there is a dataset with two attributes/classes and after fitting the training dataset we will get our new datapoint somewhere in between the two attributes. Then it will choose the nearest neighbour and the number of nearest neighbours

depends upon value of k (k is the parameter to include the majority of voting process) [3][4]

Euclidian distance is the method used to calculate distance between the new datapoint and neighbours

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

Whereas Manhattan distance can be used instead

$$\sum_{i=0}^{m} |x_i - y_i| \quad (2)$$

The new datapoint will choose the class which has more datapoints at that time. The value of k should be odd because if we choose k as even value then situation can occur where we have equal number of neighbours from both the classes and due to which it is difficult to choose a particular class for a datapoint

2. Decision tree

Decision Tree comes under the category of Supervised Learning Where we have the labeled dataset and the desired outcome is known to us. We have the input variable and the output variables. It is used for both the classification and regression problems and for continuous dataset we use Decision Tree Regressor.[5]

Working - Suppose we have the labeled dataset. the first step is to choose the target attribute (it is the dependent attribute) and other attributes is an independent one. We have to calculate the information about Gain of the target attribute and the equation is used to calculate the information gain where P is the element of target variable and P+N is total elements of target variables

$$I_g = \frac{-P}{P+N}[\log\left(\frac{P}{P+N}\right) - \frac{N}{P+N} log \frac{N}{P+N}] \quad (3)$$

After calculating the information gain, we need to calculate the Entropy of all the other attributes by using the formula of Entropy Finally, the Gain of each of the attribute

$$E(A) = \sum_{i=1}^{V} \frac{P_i+N_i}{P+N} I(P_i N_i) \quad (4)$$

Finally, the Gain of each of the attribute (the formula is given below):

$$\text{Gain} = \text{I}_\text{G} - (\text{A}) \quad (5)$$

3. Random forest

It is another supervised learning algorithm which uses ensemble learning technique in which the final output value does not depend on single classifier, it is dependent on multiple classifier that is why the output using random forest is more accurate. It can be used for classifier and regressor. Ensemble learning is divided into two parts that are Bagging and Boosting and random forest comes under Bagging technique in which we divide the dataset in multiple training dataset, and each classifier have its own dataset. Although we can provide single training dataset to every classifier but different dataset can provide more accurate results. After going through the result of each classifier it decides the final result.[6]

Working – It uses decision tree algorithm, its working involves Randomization where we choose every attribute, feature or sample randomly. Suppose we have a dataset first it will convert the original dataset to a Bootstrap dataset (it will pick the samples randomly and make another dataset). Then there will be the target attribute. Random forest uses multiple Decision Trees and it uses Random fashion to build the decision tree. Now it randomly nominates the attribute for the root node and out of the nominated attributes the one which splits the data better will be chosen as the root node, similarly second one will choose, till it forms the decision tree. And all the decision trees are made with the help of bootstrap data.. Now before choosing the value for the target tupple it will go through the result of each decision tree and then decides the final result. And with help of multiple decision trees the result will be more accurate.[7][8]

4. Support vector machine

It is a supervised learning technique used for classification and regression analysis. In this it has classified and labeled dataset

Working - It is known that it contains classified and labelled dataset. In this SVM uses decision boundary or hyper plane to separate the classes of the data. There is another line which is called Boundary line. It is the line which touches the data points which are much closer to the data points of another class, these data points are also known as support vectors, the boundary lines are parallel to the hyper plane and also helps in creating the margins. Margin is the distance between two boundary lines. There can be number of hyper plane but only that hyper plane will be chosen whose margin or width is higher because it separates the data more clearly and gives more accuracy.[9]

5. Linear regression

It is a linear model having a linear relationship between the input variable and output variable (single output). The model in which there is single input and single output is called simple linear regression [1].

Working – Linear regression try to make predictions on Linear scale. All these predictions are done on the basis of previous values or previous data, we need to train our data by using the previous dataset by using the fit method and then applying Linear regression to predict values.[1][2]

Where y is dependent variable is independent variable and m is the slope and c are the intercept

$$y = mx + c$$

Cross validation

In the train test split, it is known that when we re-run the program the values will get shuffle, the accuracy and results may change after re-run. Also, a random state cannot help in this situation because different random states will have different pattern of divided data set. To avoid this problem and to get proper accuracy there is cross-validation.

KFOLD - This cross-validation technique divides the data set into k number of subsets which are called folds. Out of which one subset will be taken as the testing data and other subsets are testing data, similarly one by one each subset will become testing data and the other becomes the training data. The iteration will take place k times. Spearman's correlation: As we know covariance is very important for data preprocessing and data analysis and also covariance will help us to get the relation between features. It tells that the two features vary or not, if the covariance is positive then it means features vary, and in case of negative it does not vary. The spearman's correlation is shown in

between features. It tells that the two features vary or not, if the covariance is positive then it means features vary, and in case of negative it does not vary. The spearman's correlation is shown in equation (6) where is ρ spearman rank correlation coefficient, d is total mean and n is number of elements

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (6)$$

But covariance doesn't give the result how much it varies so the concept of correlation arrives. We can use Pearson correlation shown in equation (7) where $Cov\,(x, y)$ is covariance of x and y and $\sigma_x\sigma_y$ is standard deviation of x and y respectively. But it does not focus on outliers and does not give as much accurate result which is given by spearman's correlation because of its focus on outliers also. In spearman, we take the rank of the parameters used in the formula.

$$\delta_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y} \quad (7)$$

Model Accuracy can be checked by calculating the error between the predicted values and observed values. Mean absolute error, mean square error and root mean square error should are used. the more output is closer to zero, the better and accurate model is predicting the results. Equation (8), (9) and (10) are used to calculate mean absolute error, mean square error and root mean square error respectively

$$MAE = \frac{1}{n}\sum|y - \hat{y}| \quad (8)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y - \hat{y})^2 \quad (9)$$

$$RMSE = \sqrt{MSE} \quad (10)$$

Proposed hybrid model

In hybrid model two algorithms are appended one after another. Further, the dataset is imported using pandas library hence the dataset is divided in 70:30 ratio for training and testing. by using linear regression, we predict the values of testing dataset which will act as input for next algorithm and the output of first algorithm is further divided into 70:30 ratio for training and testing of dataset and hence result is obtained from applying two algorithms in a hybrid form model. We can take combinations of any two algorithms but results from the model testing done, depict that linear regression is providing much accurate results as compared to other algorithms so it will be beneficial if first algorithm used is linear regression and in combination with this any other algorithm can be used, as a result of this hybrid method much accurate result are expected .

SIMULATION RESULTS

Results using single algorithm

Red is for predicted plot and blue is observed plot

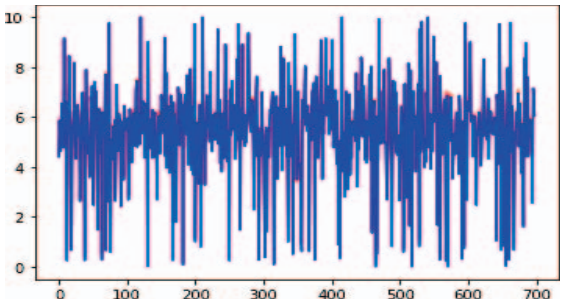

Fig 1. Plot between predicted and observed values for Linear regression

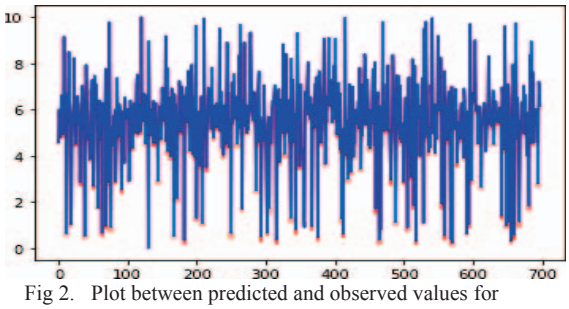

Fig 2. Plot between predicted and observed values for KNN

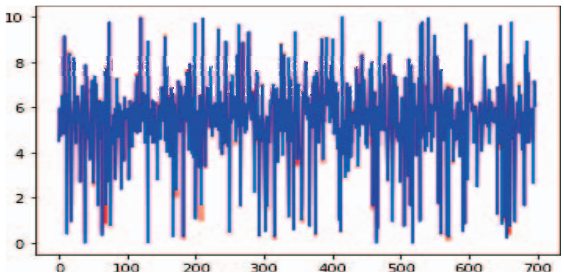

Fig 3. Plot between predicted and observed values for decision tree

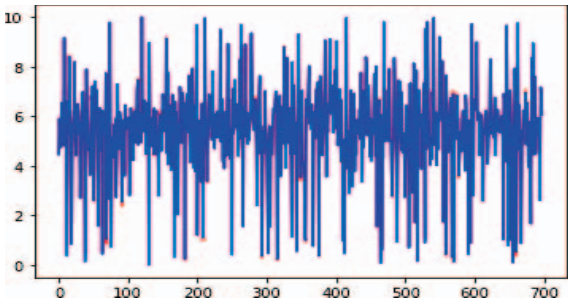

Fig 4. Plot between predicted and observed values for random forest

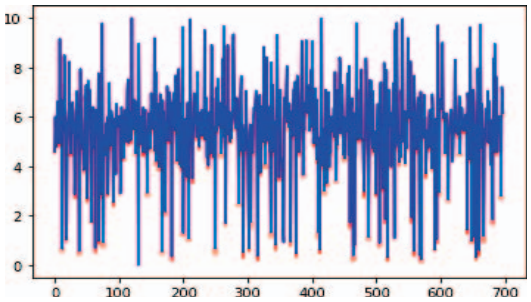

Fig. 6. Plot between predicted and observed values for support vector machine

Model result using Hybrid model

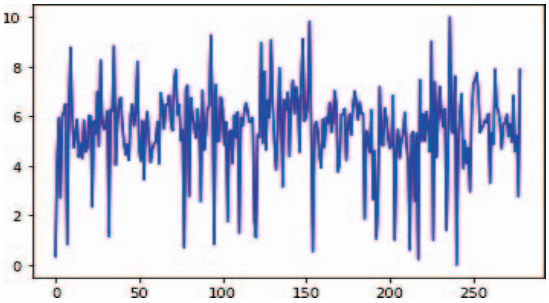

Fig. 7. Plot between predicted values and observed values for Linear regression with Linear regression

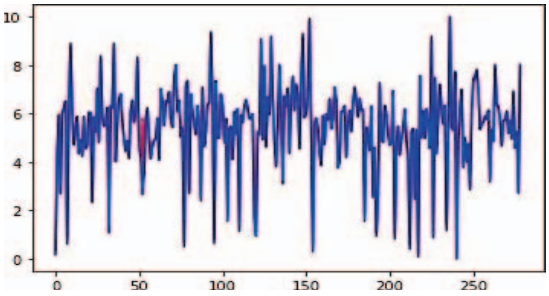

Fig. 8. Plot between predicted values and observed values for Linear regression with KNN

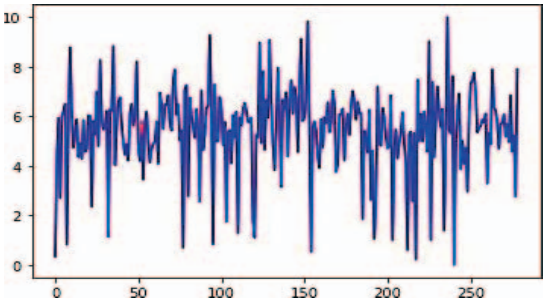

Fig. 9. Plot between predicted values and observed values for Linear regression with Support vector machine

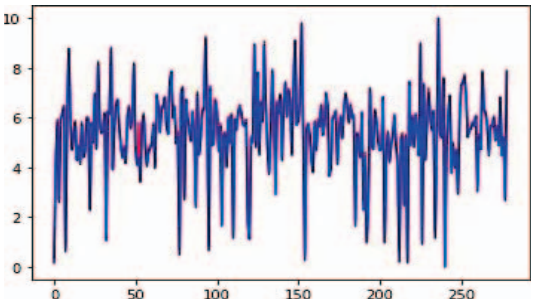

Fig. 10. Plot between predicted values and observed values for Linear regression with Decision tree

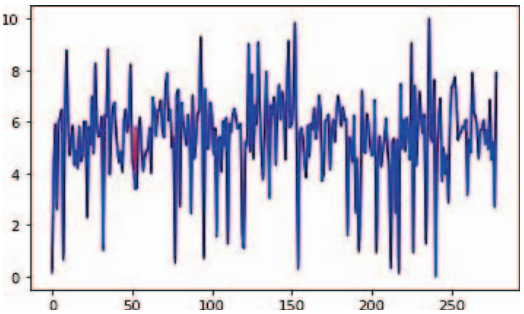

Fig. 11. Plot between predicted values and observed values for Linear regression with Random forest

Comparison table

	MAE	MSE	RMSE	SCORE MEAN
LINEAR REGRESSION	0.02469680074102	0.00128933686784	0.1571616131863	0.998110
KNN	0.0254100463	0.0017226097	0.1594052896	0.815525
SVM	0.056233657	0.003849879	0.237136694	0.994762
DECISION TREE	0.028295466	0.002116064	0.168212563	0.839632
RANDOM FOREST	0.232856129	0.001541146	0.152596241	0.832489

Model accuracy using single algorithm

	MAE	MSE	RMSE
LR + LR	$3.0899217871739\ e^{-16}$	$3.12975351019934\ e^{-31}$	$1.75781733612272\ e^{-8}$
LR + KNN	0.0168824060	0.000717925	0.12993231
LR + SVM	0.0217266217	0.000875966	0.14739953
LR + DT	0.0268039095	0.0016172521	0.16371899
LR + RF	0.0147869564	0.0004964863	0.121601630

Model accuracy using Hybrid model

CONCLUSION

In this paper we use two methods to predict the price which were firstly using single algorithm and other a hybrid model. using MAE, MSE, RMSE we calculated the error between predicted price and real price. The more value was nearer to zero the more accurate model would be and so linear regression gave most accurate results among KNN, support vector machine, decision tree and random forest followed by support vector machine. The graph plotted by linear regression and SVM was quite accurately predicting the real time price with score mean of 0.9981 and 0.9982 respectively

When hybrid model was applied with base algorithm as linear regression, although the computation complexity increased but it was observed that when linear regression was used as second algorithm then MAE was $6.0103e^{-15}$, MSE was $5.6911e^{-29}$ and RMSE was $7.7526e^{-8}$ which are very small thus highly accurate graph was obtained. Also, SVM gave second best results where MAE was 0.072338, MSE was 0.030724and RMSE was 0.268957 where the hybrid model with second algorithm as linear regression and SVM gave better results than linear regression and SVM alone.

Thus, it can be concluded that hybrid model with linear regression appended with another linear regression achieves efficient, accurate and better results when compared to KNN, SVM, decision tree and random forest alone since the error was minimal in these which makes it highly optimal to predict the real stock price and minimize the uncertainty of future value.

REFERENCES

[1] H. I. Bulbul and Ö. Unsal, "Comparison of Classification Techniques used in Machine Learning as Applied on Vocational Guidance Data," *2011 10th International Conference on Machine Learning and Applications and Workshops*, Honolulu, HI, 2011, pp. 298-301, doi: 10.1109/ICMLA.2011.49.

[2] H. Hirose, Y. Soejima and K. Hirose, "NNRMLR: A Combined Method of Nearest Neighbor Regression and Multiple Linear Regression," *2012 IIAI International Conference on Advanced Applied Informatics*, Fukuoka, 2012, pp. 351-356, doi: 10.1109/IIAI-AAI.2012.76.

[3] D. Wang, Y. Gao and Z. Tian, "One-Variable Linear Regression Mathematical Model of Color Reading and Material Concentration Identification," *2017 International Conference on Smart City and Systems Engineering (ICSCSE)*, Changsha, 2017, pp. 119-122, doi: 10.1109/ICSCSE.2017.37.

[4] Okfalisa, I. Gazalba, Mustakim and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," *20172nd International conferences on Information Technology, Information Systems and Electrical Engineering(ICITISEE)*, Yogyakarta, 2017, pp. 294-298

[5] H. b. Jaafar, N. b. Mukahar and D. A. Binti Ramli, "A methodology of nearest neighbor: Design and comparison of biometric image database," *2016 IEEE Student Conference on Research and Development (SCOReD)*, Kuala Lumpur, 2016, pp. 1-6, doi: 10.1109/SCORED.2016.7810073.

[6] S. Patil and U. Kulkarni, "Accuracy Prediction for Distributed Decision Tree using Machine Learning approach," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019, pp. 1365-1371,

[7] H. Elaidi, Y. Elhaddar, Z. Benabbou and H. Abbar, "An idea of a clustering algorithm using support vector machines based on binary decision tree," *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, Fez, 2018, pp. 1-5, doi: 10.1109/ISACV.2018.8354024

[8] A. Behnamian *et al*., "Dimensionality Reduction in The Presence of Highly Correlated Variables for Random Forests: Wetland Case Study," *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 9839-9842, doi: 10.1109/IGARSS.2019.8898308

[9] Y. Guo, Y. Zhou, X. Hu and W. Cheng, "Research on Recommendation of Insurance Products Based on Random Forest," *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*

[10] Learning, Big Data and Business Intelligence (MLBDBI), 2019, pp. 308-311, doi: 10.1109/MLBDBI8998.2019.00069