Forecasting Bitcoin Price using Machine Learning Algorithms.

Smt.Abha Marathe

Vishwakarma Institute of Technology Pune, India

Vyankatesh Kulkarni
Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Kush Agarwal
Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Shrihari Mangle
Computer Engineering
Vishwakarma Institute of Technology
Pune, India

***Abstract*—Today's growing world has totally became the great example for the use of digital technology. Over the recent years, everything is going online and internet has become the keen provider for the required resources. Considering, these development in our mind the most important domain is the rapid and sudden development of crypto currency during these times is the most controversial development in the global economy. Adding to it, today's market is most critical that even good practitioner cannot predict the things since, the fluctuation and the volatile market of this Bitcoin currency led the confusion among the investors. So, this paper mainly focusses on the various approaches to predict and hence to forecast the fluctuating price of the Bitcoin. The various machine learning based approaches are discussed here with good accuracy so that there should not be any blind decision taken regarding investments particularly in the Bitcoin trading. The results of this paper verifies the applicability of model and give a direction to investors on how these machine learning techniques are used in decision making.**

Keywords—Bitcoin, Machine Learning, crypto currency, multiple linear regression, Random Forest, SVR.

I. INTRODUCTION

Transfer of direct money from a person to another is hectic and also there is a chance of loss of money in many ways so the digital currency which is named Crypto currency is introduced. Bitcoin is the worlds' most valuable cryptocurrency and is traded on over 40 exchanges worldwide accepting over 30 different currencies. It is one of the most popular cryptocurrencies around the world. However, it is not the first cryptocurrency that appears in the world. Satoshi Nakamoto first introduced Bitcoin and the concept of Blockchain in one of his papers in 2008[1].It has a current market capitalization of 9 billion USD according to https://www.blockchain.info/ and sees over 250,000 transactions taking place per day.[3] Digital currency is a method for trade which is web based and utilizes cryptographical capacities to perform money related exchanges. . The principle highlight of cryptographic money is that it isn't constrained by any fundamental power: the circulated pith of the blockchain makes digital forms of money hypothetically invulnerable to the old methods of government control and obstruction. Digital currencies support blockchain innovation to pick up decentralization, permanence, and straightforwardness. Bitcoin is an advanced installment that uses cryptographic money and distributed (P2P) mechanization to produce and oversee financial exchanges instead of a focal power. Time arrangement estimating or expectation is a notable issue [2]. With a market capitalization of around 170 billion US dollar (September 2020), bitcoin represents about 58% of the cryptocurrency market.[5]

As Bitcoin's price has very high volatility, many people are attracted to the research of predicting the price trend of Bitcoin because investors have a chance to gain high profit from the price change. As more and more researches focus on the topic, the methods used for prediction also expand from only time series models to machine learning models and deep learning models. The input feature sets also expand from historical price and exchange volume to text data from social media, news, internet search, and more.[1].

In this paper , we are going to predict the closing price and will be showing the price trends for the same. These predicted prices will be based on the previously observed and studied data form a dataset taken for building the project. The main input parameters which will be taken into consideration is that the opening price of the Bitcoin , how much it has gone high during the span of the day and the low price it has acquired. On this basis, the price at which it is going to close at the end of the day is the key prediction we have done while framing the various machine learning models. We have used the Multiple Linear Regression[MLR], Random Forest Algorithm and Support Vector Regression[SVR] for prediction purpose. Also we have plotted various graphs which will depict the actual and the predicted value trends for clear visualization purpose.

II. LITERATURE REVIEW

The studies which are carried out for the prediction of Bitcoin prices are quite good to study since they are very few by implementing through machine learning algorithm concepts. But there are also some relevant studies carried out which gives the total review of how should we carry out our implementation towards this domain of Bitcoin price prediction.

[3] performed the study on the dataset where they have used various deep learning algorithms and finally came to the conclusion that deep learning models such as the RNN and LSTM are evidently effective for Bitcoin prediction with the LSTM more capable for recognizing longer-term dependencies. Their results mainly says that, the LSTM achieved the highest classification accuracy of 52% and a RMSE of 8%. The popular ARIMA model for time series forecasting is implemented as a comparison to the deep learning models. As expected, the non-linear deep learning methods outperform the ARIMA forecast which performs poorly.

[2] mainly depicted the top deep learning approaches for the price prediction. As only predicting the price does not mean that we are done, so they have prepared the Python based GUI application (i.e API) for the real time calculation of the price based on the given input parameters. The most common deep learning approaches Long Short Term Memory (LSTM) , Gated Recurrent Unit (GRU), Gated Recurrent Unit (GRU) and Support Vector Machine are used for model building and evaluation. They tried to predict the value using deep learning and want to compare it with the machine learning model (SVM) but they got negative approach at that point. So it can be inferred that deep learning models take long time to train. Prediction of bitcoin price was a complicated task as it is based on large dataset.

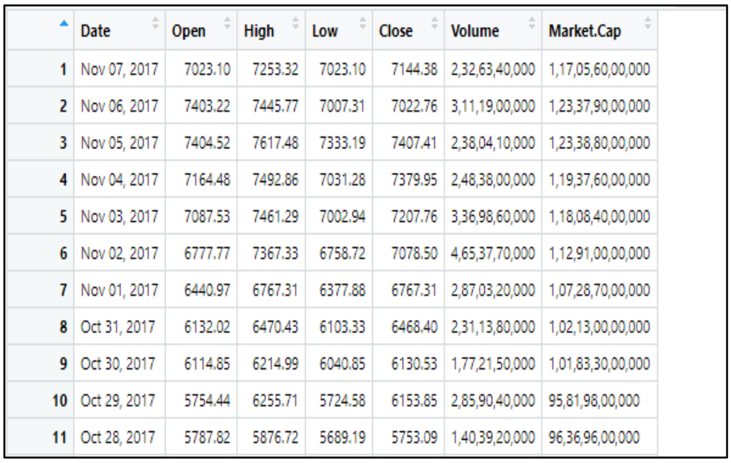
[1] In this research, they used historical Bitcoin transaction data, Twitter data, and COVID-19 data. They collected data from 1st January 2020 to 31st July 2020 from online open sources. Features of the transaction data includes Open, High, Low, Close prices, and Trading Volume. They formulated the data to four input feature sets, including: (1) Historical Bitcoin exchange data; (2) Historical Bitcoin exchange data + COVID-19 data (recovery, confirmed, death); (3) Historical Bitcoin exchange data + Twitter data; (4) Historical Bitcoin exchange data + COVID-19 data (recovery, confirmed, death) + Twitter data. And implemented four machine learning models including (1) Random Forest; (2) Decision tree; (3) AdaBoost; (4) Support vector machine. Finally, they concluded that (1) Twitter data improve the performance of models. (2) People consider the information within 5 days when they make decisions on investments. (3) Support vector machine does not perform well in predicting Bitcoin return or price trend. (4) COVID-19 data does not help improve the prediction.

[6] predicted Bitcoin price by using various machine learning algorithms including Linear Regression Model, Random Forest, K-nearest Neighbour Algorithm, Naïve Bayes, etc .They performed the analysis by mentioning the how they carried out the steps to build the models. Also they mentioned the various techniques to perform such predictions with the advantages and disadvantages to get proper idea and concluded that survey report will be just introducing modules of Bitcoin price prediction and machine algorithms. The Comparison table of ML algorithm model accuracy which tells that the Linear regression model will have most accuracy then the other algorithms. In that paper they concluded that the linear regression algorithm is more efficient then the other algorithms.

III. Methodology

A. Dataset Description

The btc train dataset from Kaggle has been used for the project. This dataset consists of 7 attributes and 1655 observations. In this dataset, the 7 attributes are Date, Open, High, Low, Close, Volume and Market Cap. In these 7 attributes, 6 attributes are decimal and 1 attribute is datetime. In this dataset, all the cryptocurrency values have been given from 28th April 2013 to 7th November 2017. Open attribute gives us the information about starting price of the bitcoin on that day. High attribute tells us about the highest bitcoin price reached on that day. Low attribute gives us the information about lowest price of bitcoin acquired on that day. Close attribute tells us about the closing price of bitcoin on that day. Volume attribute gives us the information about how much of the cryptocurrency was traded in the spam of 24 hours. Market cap gives us the information about total value of the cryptocurrency, and it is calculated by multiplying price of the cryptocurrency with number of coins in circulation.



	Date	Open	High	Low	Close	Volume	Market.Cap
1	Nov 07, 2017	7023.10	7253.32	7023.10	7144.38	2,32,63,40,000	1,17,05,60,00,000
2	Nov 06, 2017	7403.22	7445.77	7007.31	7022.76	3,11,19,00,000	1,23,37,90,00,000
3	Nov 05, 2017	7404.52	7617.48	7333.19	7407.41	2,38,04,10,000	1,23,38,80,00,000
4	Nov 04, 2017	7164.48	7492.86	7031.28	7379.95	2,48,38,00,000	1,19,37,60,00,000
5	Nov 03, 2017	7087.53	7461.29	7002.94	7207.76	3,36,98,60,000	1,18,08,40,00,000
6	Nov 02, 2017	6777.77	7367.33	6758.72	7078.50	4,65,37,70,000	1,12,91,00,00,000
7	Nov 01, 2017	6440.97	6767.31	6377.88	6767.31	2,87,03,20,000	1,07,28,70,00,000
8	Oct 31, 2017	6132.02	6470.43	6103.33	6468.40	2,31,13,80,000	1,02,13,00,00,000
9	Oct 30, 2017	6114.85	6214.99	6040.85	6130.53	1,77,21,50,000	1,01,83,30,00,000
10	Oct 29, 2017	5754.44	6255.71	5724.58	6153.85	2,85,90,40,000	95,81,98,00,000
11	Oct 28, 2017	5787.82	5876.72	5689.19	5753.09	1,40,39,20,000	96,36,96,00,000

Fig 1. First 11 observations of btc train dataset

B. Prediction and ensemble algorithms

Prediction is a supervised procedure used for predicting continuous values or quantitative values. This paper proposes an approach for bitcoin price prediction using prediction algorithms.

Step 1- Data Preprocessing

First, we have performed some datatype conversion. Market cap is converted from character to numeric datatype. Date is converted from character to date datatype. Also, we have replaced commas with blank spaces for Market cap and Volume. Also, we have arranged all the observations in ascending order of dates.

Step 2- Splitting the dataset into training and testing.

We have split the dataset into training and testing. 70% is used for training and 30% is used for testing. We have built the models using training datasets. Efficiency of prediction is tested using testing dataset.

Step 3- Building a model

The working of the individual algorithms is as follows.

B.1. Multiple Linear Regression

If there is one dependent variable and then more than one independent variable and the relation between dependent and independent variable is linear, then multiple linear regression is used. Equation of multiple linear regression is if there are n independent variables

y = b0 + b1 * x1 + b2 * x2 + b3 * x3 + …… bn * xn

The values of b0,b1,b2, …,bn is calculated in such a way that sum of squares of errors will be minimum. Where error is the difference between actual value and predicted value. The basic goal of multiple linear regression is to fit the best fit hyperplane into n+1 dimensional space such that it captures maximum number of points. After calculating all the values of b0,b1,b2,…,bn , these values are put into the equation and the predicted value is calculated for given attributes.

B2. Random Forest

Random forest algorithms are often used for both classification and regression. For classification problem, the ultimate output is taken into consideration by using majority voting classifier. Within the case of regression problems, final output is that the mean of all the outputs. this can be called as Aggregation. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the utilization of multiple decision trees and a method called Bootstrap and Aggregation, commonly called bagging. The fundamental idea behind this is often to mix multiple decision trees in determining the ultimate output instead of counting on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and have sampling from the dataset forming sample datasets for each model. This part is termed Bootstrap.

B3. Support Vector Regression

The Support Vector Regression (SVR) uses the identical ideas because the SVM for classification, with some small differences. A margin of tolerance (epsilon) is supplied within the case of regression as an approximate estimate to the SVM that the problem would have already requested. except that, there's a tougher reason: the algorithm is more complex; thus, it must be considered. However, the fundamental idea remains the same: to minimize error by customizing the hyperplane to maximize the margin while keeping in mind.

In a supervised-learning environment, the support vectors are the foremost influential cases that determine the shape of the tube, and also the training and test data are assumed to be independent and identically distributed (iid), obtained from the identical fixed but unknown probability data distribution function. SVR is defined as an optimization problem by first constructing a convex-insensitive loss function to be reduced and so determining the flattest tube that has the bulk of the training cases. As a result, the loss function and also the geometrical parameters of the tube are combined to create a multiobjective function.

Step 4- Prediction on testing

The model which we built using algorithms is used to predict the results on testing data. Also, we have plotted the plots of actual and predicted using all the algorithms implemented.

Step 5- Evaluating performance metrics

For regression, we have used 4 performance metrics.

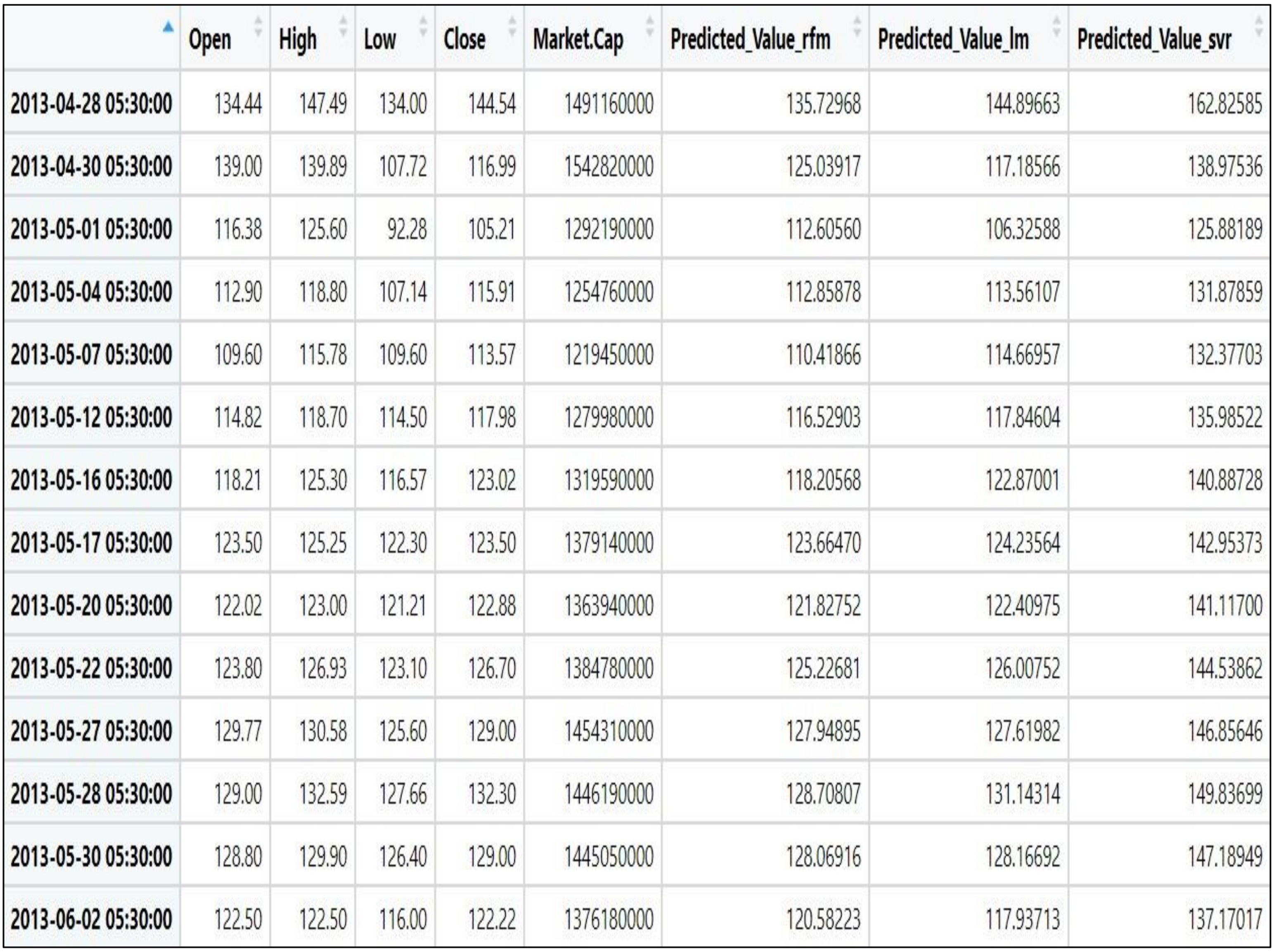
1. **R-squared(R^2)** – R-squared is the ratio of sum of squares due to regression and sum of squared deviations of y from its mean. SSR is basically sum of squares of difference between predicted value and mean value. SST is sum of squares of differences between actual value and mean value. R-squared is also known as measure of goodness of fit. More the value of R-squared, the better will be the model. Its value ranges from 0 to 1. If its value is 0, the model is worst.

2. **Mean Absolute Error(MAE)**- In this, error is calculated using formula actual-predicted and its modulus is taken. Then the mean of all the modulus of errors is calculated by using formula sum of absolute errors divided by number of observations. Less the value of mean absolute error, the better will be the model.

3. **Mean squared error(MSE)**- In this, square of error is calculated. Then, addition of squares of errors is done. Then, divide the addition of squares of errors by number of observations. Less the value of MSE, the better will be the model.

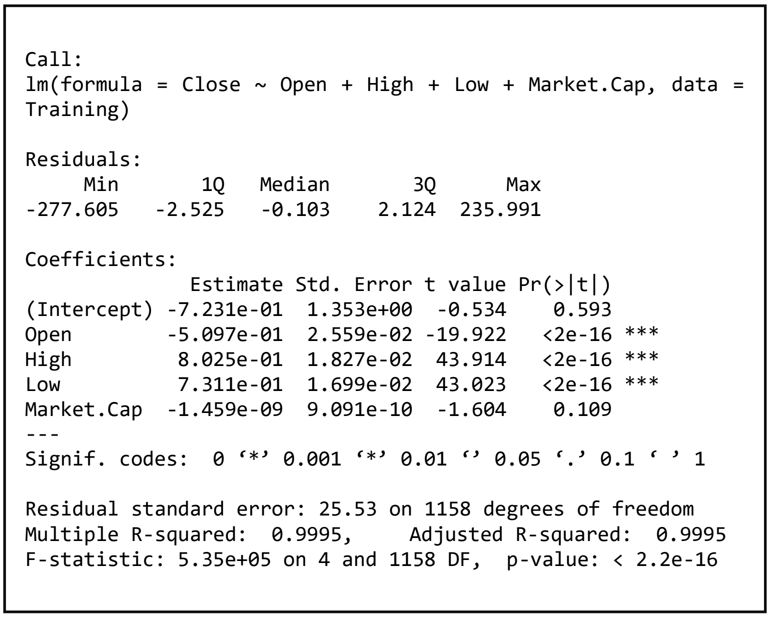
4. **Root mean squared error(RMSE)**- It is square root of mean squared error. Less the value of RMSE, the better will be the model.

IV .RESULTS AND DISCUSSIONS



	Open	High	Low	Close	Market.Cap	Predicted_Value_rfm	Predicted_Value_lm	Predicted_Value_svr
2013-04-28 05:30:00	134.44	147.49	134.00	144.54	1491160000	135.72968	144.89663	162.82585
2013-04-30 05:30:00	139.00	139.89	107.72	116.99	1542820000	125.03917	117.18566	138.97536
2013-05-01 05:30:00	116.38	125.60	92.28	105.21	1292190000	112.60560	106.32588	125.88189
2013-05-04 05:30:00	112.90	118.80	107.14	115.91	1254760000	112.85878	113.56107	131.87859
2013-05-07 05:30:00	109.60	115.78	109.60	113.57	1219450000	110.41866	114.66957	132.37703
2013-05-12 05:30:00	114.82	118.70	114.50	117.98	1279980000	116.52903	117.84604	135.98522
2013-05-16 05:30:00	118.21	125.30	116.57	123.02	1319590000	118.20568	122.87001	140.88728
2013-05-17 05:30:00	123.50	125.25	122.30	123.50	1379140000	123.66470	124.23564	142.95373
2013-05-20 05:30:00	122.02	123.00	121.21	122.88	1363940000	121.82752	122.40975	141.11700
2013-05-22 05:30:00	123.80	126.93	123.10	126.70	1384780000	125.22681	126.00752	144.53862
2013-05-27 05:30:00	129.77	130.58	125.60	129.00	1454310000	127.94895	127.61982	146.85646
2013-05-28 05:30:00	129.00	132.59	127.66	132.30	1446190000	128.70807	131.14314	149.83699
2013-05-30 05:30:00	128.80	129.90	126.40	129.00	1445050000	128.06916	128.16692	147.18949
2013-06-02 05:30:00	122.50	122.50	116.00	122.22	1376180000	120.58223	117.93713	137.17017

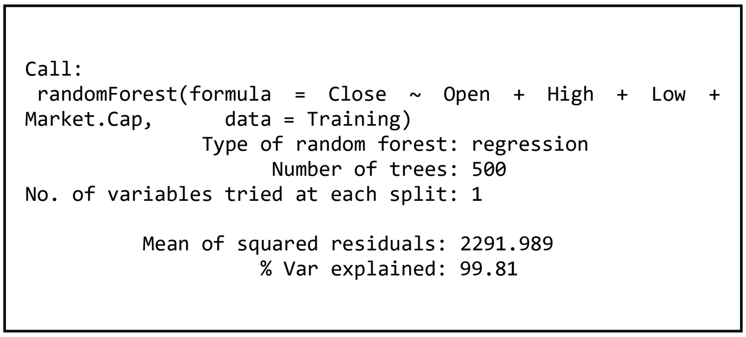
Fig 2. Predicted Values Sample Using Different Algorithms



```
Call:
lm(formula = Close ~ Open + High + Low + Market.Cap, data =
Training)

Residuals:
     Min       1Q   Median       3Q      Max
-277.605   -2.525   -0.103    2.124  235.991

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.231e-01  1.353e+00  -0.534    0.593
Open        -5.097e-01  2.559e-02 -19.922   <2e-16 ***
High         8.025e-01  1.827e-02  43.914   <2e-16 ***
Low          7.311e-01  1.699e-02  43.023   <2e-16 ***
Market.Cap  -1.459e-09  9.091e-10  -1.604    0.109
---
Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.53 on 1158 degrees of freedom
Multiple R-squared:  0.9995,     Adjusted R-squared:  0.9995
F-statistic: 5.35e+05 on 4 and 1158 DF,  p-value: < 2.2e-16
```

Fig 3. Summary of MLR model

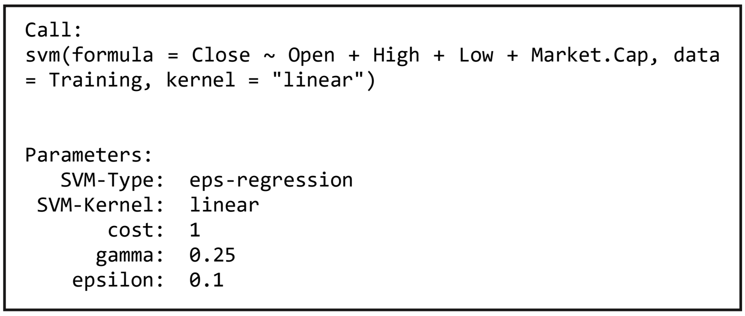
By using fig3, we can say that the attributes open, high and low are important for predicting close value. Also, R-squared value of this model is 0.9995 and residual standard error is 25.53.



```
Call:
 randomForest(formula  =  Close  ~  Open  +  High  +  Low  +
Market.Cap,      data = Training)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 2291.989
                    % Var explained: 99.81
```

Fig 4. Random Forest Model

In fig 4, we can see that % of var explained is observed to be 99.81 and mean of squared residuals is 2291.989.



```
Call:
svm(formula = Close ~ Open + High + Low + Market.Cap, data
= Training, kernel = "linear")


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.25
    epsilon:  0.1
```

Fig 5. SVR Model

In fig 5, as we can see that kernel is linear and gamma value is 0.25. Number of support vectors are 15. Epsilon value is 0.1.

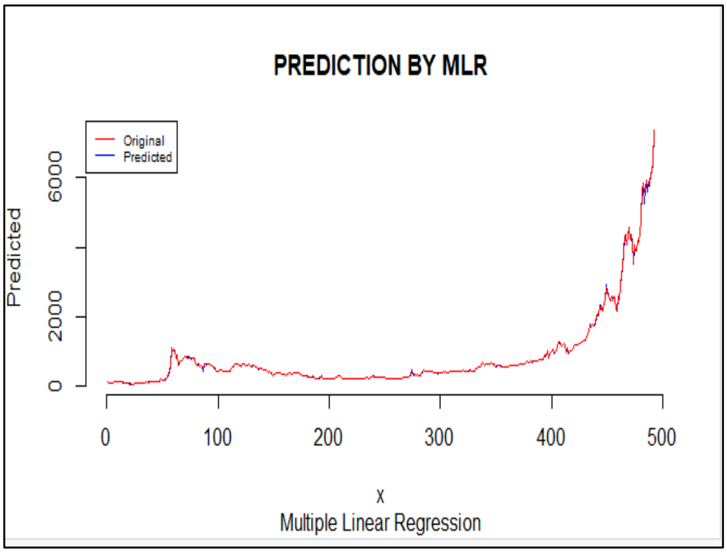


Fig 6. MLR prediction plot

By using figure 6, we can see that red line denotes actual values and blue line denotes predicted values.

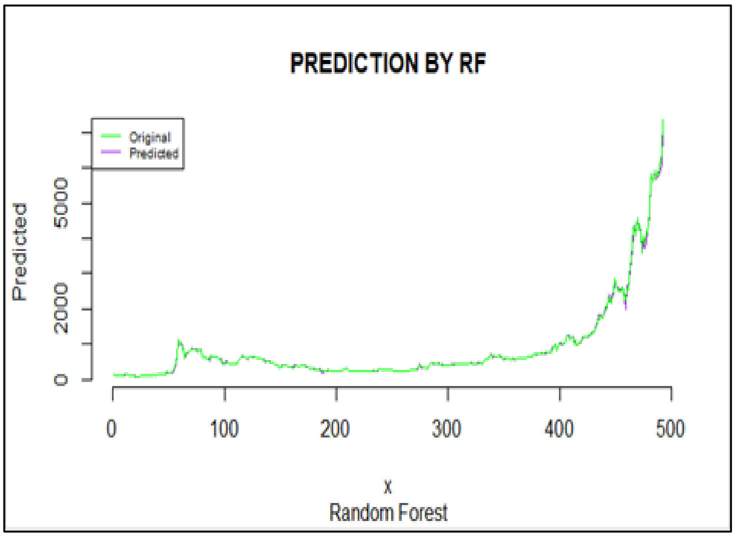


Fig 7. Random Forest prediction plot

In figure 7, we can see that green line denotes actual values and purple line denotes predicted values.

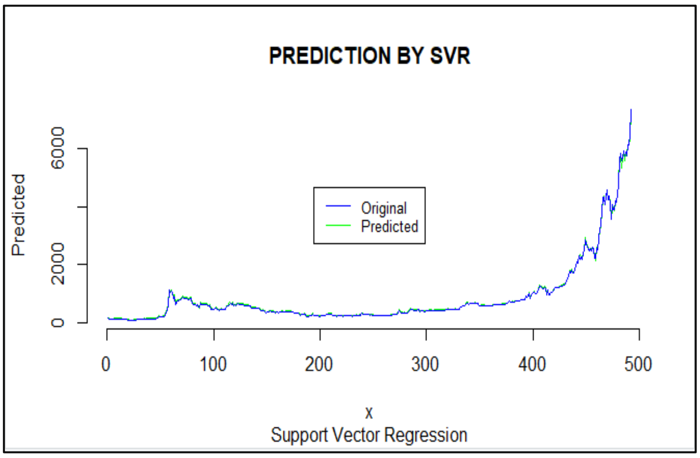
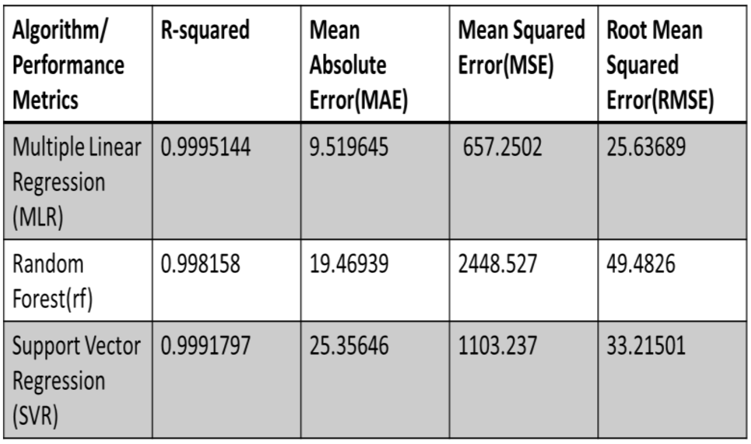


Fig 8. SVR Prediction Plot

In fig 8, blue line is used for original points and green line is used for predicted values.



Algorithm/ Performance Metrics	R-squared	Mean Absolute Error(MAE)	Mean Squared Error(MSE)	Root Mean Squared Error(RMSE)
Multiple Linear Regression (MLR)	0.9995144	9.519645	657.2502	25.63689
Random Forest(rf)	0.998158	19.46939	2448.527	49.4826
Support Vector Regression (SVR)	0.9991797	25.35646	1103.237	33.21501

Fig 9. Performance Metrics Table

*V.*CONCLUSION

Prediction of anything paves a way for the easy access of the future actions. Hence the term forecasting which is our topic helps us to act for the future plans. Here, we predicted the Bitcoin price using various machine learning algorithms of which the standard parameters to judge the model is mentioned in the table of Fig 9 above. Considering all these algorithms (Multiple Linear Regression[MLR], Random Forest Algorithm and Support Vector Regression[SVR]), the most accurate and effective over the dataset which has given almost close accuracy is by Multiple Linear Regression[MLR] then ranks Random Forest and finally the Support Vector Regression. Fig 9 precisely depicts the various types of terms regarding error in which the Multiple Linear Regression[MLR] has the least values as compared to the other two. Similarly R^2 value also can be observed which can be termed as accuracy is highest for the Multiple Linear Regression[MLR].So, these are the conclusions which we came across while building of framing our predictive model.

*VI.*REFERENCES

[1] Jiayun Luo1 “Bitcoin price prediction in the time of COVID-19”1Department of Statistics, University of California-Los Angeles, Los Angeles, 90024, USA.

[2] E.Mahendra, H.Madan, S.Gupta, S.V.Singh “Bitcoin Price Prediction Using Deep Learning and Real Time Deployment” 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)

[3] Sean McNally∗ Jason Roche† Simon Caton∗, “Predicting the Price of Bitcoin Using Machine Learning” 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing

[4] Mohammed Mudassir1 • Shada Bennbaia1 • Devrim Unal2 • Mohammad HammoudehR. Nicole, “Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach” Received: 15 April 2020 / Accepted: 16 June 2020.

[5] Patrick Jaquart*, David Dann, Christof Weinhardt Institute of Information Systems and Marketing, Karlsruhe Institute of Technology, Germany “Short-term bitcoin market prediction via machine learning” revised 3 March 2021; accepted 4 March 2021

[6] Lekkala Sreekanth Reddy, Dr.P. Sriramya “A Research On Bitcoin Price Prediction Using Machine Learning Algorithms” INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 04, APRIL 2020

[7] Alvin Ho1 , Ramesh Vatambeti1∗, Sathish Kumar Ravichandran1 1 Computer Science and Engineering, CHRIST (Deemed to be University), Bengaluru, India “Bitcoin Price Prediction Using Machine Learning and Artificial Neural Network Model”. Received: 20.05.2021 Accepted: 15.07.2021 Published: 11.08.2021.

[8] S M Rajua,∗ , Ali Mohammad Tarif “Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis” Computer Science, International Islamic University Malaysia, Gombak, Malaysia.

[9] Mahboubeh Faghih Mohammadi Jalali and Hanif Heidari, “Predicting changes in Bitcoin price using grey system theory”.

[10] Mr. Shivam Pandey1, Mr.Anil Chavan2, Miss. Dhanashree Paraskar3, Prof. Sareen Deore4, “Bitcoin Price Prediction using Machine Learning” Volume: 08 Issue: 05 | May 2021.