# Bitcoin Price Prediction Using Machine Learning

**Zuber Khan** **Aurko Mitra** **Debarshi Dasgupta**
**IIIT Delhi** **IIIT Delhi** **IIIT Delhi**

## Abstract

The objective of this paper is to predict Bitcoin price accurately with help of various machine learning algorithms taking into consideration various features that affect the price value. The dataset consists of the values of the parameters and target variable from July 2010-Sept 2021. At first the dataset is preprocessed by dealing with missing values and then certain visualizations are performed of daily returns and also moving averages are calculated which are useful in identifying the trends in stock movement. Then multiple regression models are created and analyzed which use different machine learning algorithm for predicting the closing price and a comparison is made between the various models taking into account the visualizations and model performance metrics in each case.

## 1 Introduction

Bitcoin is a digital currency which is not operated by any government or bank. It was invented in 2009 by a person having a false name called Satoshi Nakamoto whose main goal was to create an electronic cash system with no central authority. It creates a public ledger where each transaction block is added after it is verified by miners who secure the network by checking the validity of a transaction by solving a computational problem.

The problem statement for this paper is predicting the bitcoin price accurately. Predicting bitcoin prices accurately can lead to more timely investments and thus increased profits. As the price prediction is a complex and challenging task we have used different machine learning algorithms and compared them for better price prediction.

## 2 Literature Review

Temesgen Awoke et al. [1] used deep learning based approaches in order to predict the price of Bitcoin and analyze the volatility in Bitcoin prices. They utilized Bitcoin daily price data set from 1st January 2014 to 28th February 2018 from the Kaggle repository with train-test split ratio to be 80:20. Long short term memory(LSTM) and gated recurrent unit (GRU) are employed for the purpose of prediction and the accuracy is evaluated using the metrics such as root mean square error and mean absolute percentage error. It is found that the GRU model works better for most of the cases. Also, the compilation time of GRU (5 msec) is much less than that of LSTM (53 msec). However, there is no information about the R-square metric in this research work. Also, there is no hypertuning involved and the predictions are made on default models.

Ashutosh et al. [2] implemented machine learning as well as deep learning models for the same purpose. The data set for this research which includes 1-day trading data is taken from the crypto-compare website. The data involves 1501 bitcoin prices (in US dollars) during the period 6 August 2016 to 14th September 2020. Among the machine learning models, they used linear regression, theil-sen regression and huber regression and each model predicted with an accuracy of around 99.85% with the execution time of linear regression being the least. On the other hand, in deep learning algorithms, they used GRU and LSTM with R-square score of prediction being greater than 0.99 for both.

Alvin et al. [3] utilized linear regression and long short term memory model for predicting the prices of Bitcoins. The data set used involves day transactions from 29th August 2017 to 9th August 2020. It is found that the linear regression model is 99.87% accurate while the error rate for LSTM model is around 0.08%. However, the number of features in the data set is limited in this study and hence the results will become unreliable in case more features are taken into consideration.

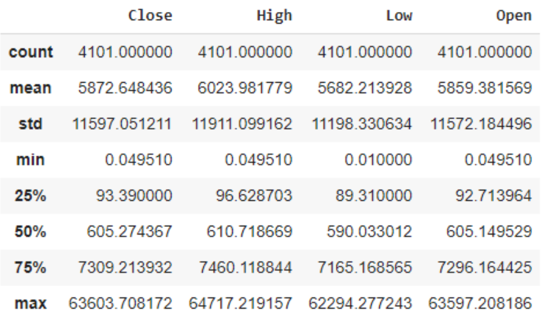
Sean McNally et al. [4] used Bayesian recurrent neural network, long short term memory (LSTM) model and ARIMA model in order to predict the direction of Bitcoin price trend. The data set included the data regarding Bitcoin from 19th August 2013 to 19th July 2016 and is taken from CoinDesk website as well as blockchain.info. The LSTM model attained an accuracy of around 52% while the optimized Bayesian RNN as well as ARIMA model had accuracy of around 50%. However, the root mean square error of the ARIMA model is around 50% while that of RNN and LSTM is 5.5% and 6.8% respectively. The major limitation of this study reveals that the algorithms employed perform only above average including ARIMA being the worst performer.

Alex Greaves et al. [5] used Bitcoin transaction graph in order to predict the prices of Bitcoins. The transaction data set is collected from CS224W website that involved all the Bitcoin transactions before 7th April, 2013. The data set is reduced over 3 million unique users and then represented using a directed graph such that each node denoted a user while each edge denoted a transaction. This representation is used to obtain the features for price prediction while the prices of Bitcoin were acquired from api.bitcoincharts.com. Linear regression and support vector machine algorithms were used to predict 1 hour ahead Bitcoin prices with mean square error of 1.94 and 1.98 respectively. Also classification algorithms such as Logistic regression, SVM and neural network were employed to classify whether the Bitcoin prices would increase or decrease in the upcoming hour with percentage accuracy of 54.3%, 53.7% and 55.1% respectively. However, this research work does not take into account the exchange behavior in the transactions used for prediction purpose and hence the accuracy is limited to 55% only.

S.Velankar et.al [6] proposed two approaches for bitcoin price prediction namely Bayesian Regression and GLM/Random forest. In the former method, they divide the dataset into three parts and the first third of data is further divided into 180 s, 360 s and 720 s interval sizes. K-Means clustering is applied to obtain clusters for each interval and then sample entropy is used to get twenty best clusters. Then the weights are calculated for features using second third of data while the last set of data is used to evaluate the algorithm. In case of GLM/Random Forest approach, three-time series data sets for 30, 60, and 120 minutes is constructed. Then Random Forest/GLM is run on each of the two sets and two linear models are obtained which are used to predict the price change. This paper majorly intends to predict the direction of shift (positive/negative) in the daily bitcoin prices but do not throw much light on the actual prices

## 3 Data Collection and Analysis

The dataset that is being used to predict the bitcoin price consists of daily values of open, high, close, low prices. OHLC price data with daily resolution has been sourced from glassnode. The missing values are dropped and the date for a respective day is set as an index for the rows.
The description of the dataset is shown in Fig. 1



| | Close | High | Low | Open |
|---|---|---|---|---|
| count | 4101.000000 | 4101.000000 | 4101.000000 | 4101.000000 |
| mean | 5872.648436 | 6023.981779 | 5682.213928 | 5859.381569 |
| std | 11597.051211 | 11911.099162 | 11198.330634 | 11572.184496 |
| min | 0.049510 | 0.049510 | 0.010000 | 0.049510 |
| 25% | 93.390000 | 96.628703 | 89.310000 | 92.713964 |
| 50% | 605.274367 | 610.718669 | 590.033012 | 605.149529 |
| 75% | 7309.213932 | 7460.118844 | 7165.168565 | 7296.164425 |
| max | 63603.708172 | 64717.219157 | 62294.277243 | 63597.208186 |

Fig. 1 Dataset Description

The plot for closing prices for the entire range of July 2010 to Sept 2021 is shown in Fig. 2

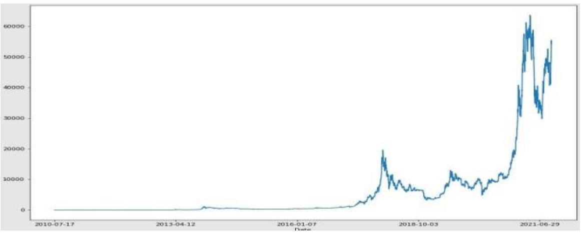


Fig. 2 Closing Price vs Date

The datatype for all the entries in the columns are checked and found to be of float64.

After that three columns are added to the dataset which are MA20, MA50 and MA100 which are the moving averages which help to capture the trend in stock movement.

The daily return which is a comparative measure of the closing price on a particular day with the previous day is plotted and the Gaussian plot obtained is shown in Fig. 3.

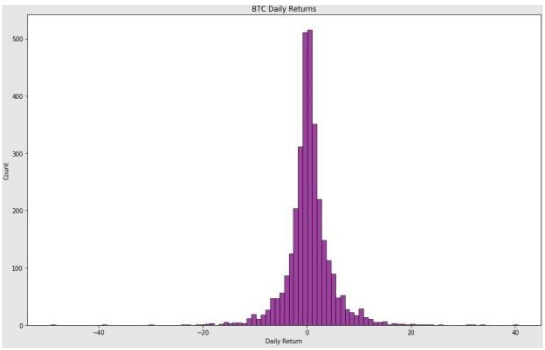
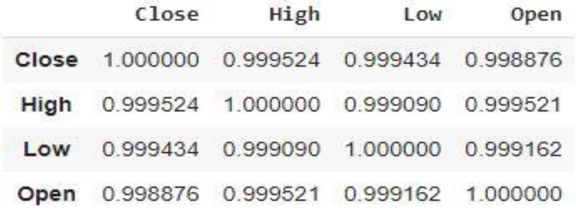


Fig. 3 Density of Daily Returns

The correlation between the target variable and the features is determined and shown in Fig. 4



| | Close | High | Low | Open |
|---|---|---|---|---|
| **Close** | 1.000000 | 0.999524 | 0.999434 | 0.998876 |
| **High** | 0.999524 | 1.000000 | 0.999090 | 0.999521 |
| **Low** | 0.999434 | 0.999090 | 1.000000 | 0.999162 |
| **Open** | 0.998876 | 0.999521 | 0.999162 | 1.000000 |

Fig. 4 Correlation Table

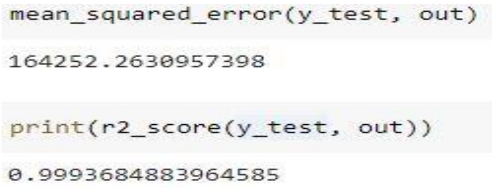
# 4 Baseline Models

## 4.1 Linear Regression

Linear Regression is an algorithm in which the task of the model is to capture a linear relationship between the input variables and the output variable. If there are multiple input variables then it is also known as multiple linear regression.

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p$$

First the dataset is being split in such a way that the training data constitutes 70 percent of it and the test data 30 percent. Then using the training data the model is trained with the help of Linear Regression from sklearn library and the following results and visualizations were obtained.



```
mean_squared_error(y_test, out)

164252.2630957398

print(r2_score(y_test, out))

0.9993684883964585
```

Mean Square error gives an average of squared errors between the predicted values and the actual data points. As it is convex in nature it has only one global minimum.

$R^2$ is a statistical measurement that indicates the percentage of variance of dependent variables that is explained by the independent variables. A $R^2$ value greater than 0.7 is usually a high level of correlation so our model performs fairly well.

The graph of residues (i.e. difference between actual and predicted values) is plotted and a normal distribution is obtained thus proving the assumption for linear regression that the errors are sampled from a normal distribution as shown in Fig. 5

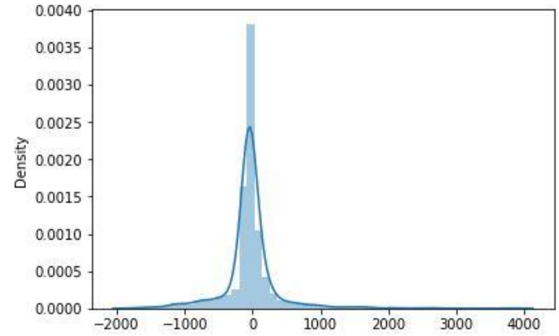


Fig. 5 Density of Residues

The actual values of the target variable, the predicted values of the target variable and the error difference between them are plotted and as shown in Fig. 6.

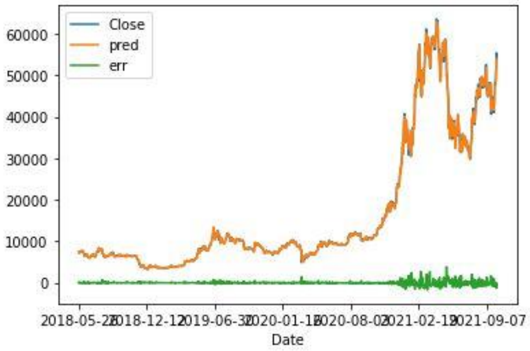


Fig. 6 Bitcoin price vs Date

It is observed that the curve of predicted values almost traces the one containing actual values.

Thus the accuracy of the linear regression model is very high which can be further improved by using different deep learning models.

### 4.2 Decision Tree Regressor

Decision Tree is a machine learning algorithm which has several applications and it can be used to solve both classification and regression problems. It has a tree like structure where the topmost node from which the other nodes originate is called the root node. The splitting of root node/any node occurs due to certain decisions which favor information gain. This continues until we reach a leaf node where no more splitting is possible. Here we are using decision tree regression as the variable (Bitcoin price) that we are planning to predict is continuous in nature.

The dataset is split into training and testing data in a ratio of 70:30 and then using the training data the model is trained with the help of Decision Tree Regressor from sklearn library and the following results and visualizations were obtained.

The $R^2$ score using the Decision Tree Regressor is 0.3549

The actual values of the target variable, the predicted values of the target variable are plotted in the same graph and as shown in Fig. 7.

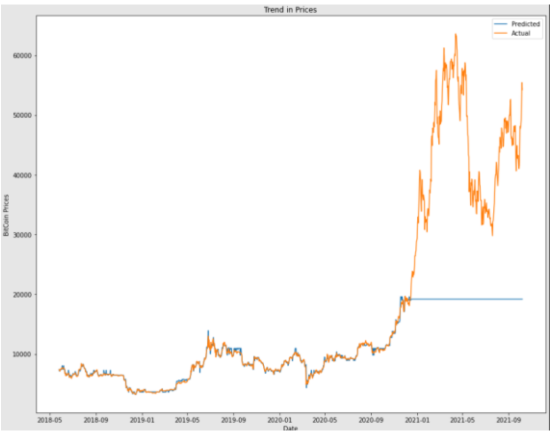


Fig. 7 Bitcoin price vs Date

## 5 Final Models

Here, we have done both long term and short term forecasting where long term implies three years ahead forecasting (2018-2021) while short term implies 6 months ahead forecasting.

### 5.1 Random Forest Regressor

Random forest is constituted of a large number of decision trees where each such tree arrives at a decision independent of other trees in the forest. The final prediction quantity is the average of all the predicted values. GridSearch CV is employed in order to tune the number of estimators and maximum depth of random forest. The best results arrive on max depth=50 and no. of estimators=100.

### 5.2 Multi-Layer Perceptron Regressor

It is a feed-forward neural network with multiple layers connecting the set of inputs with outputs forming an artificial neural network. GridSearch CV is employed in order to tune the maximum iterations and alpha hyper-parameter of MLP Regressor. The best results arrive on maximum iterations =1000 and alpha=0.07 for long term forecasting while alpha = 0.0001 with same max_iterations for short term forecasting.

### 5.3 K Nearest Regressor

KNN regression is an algorithm which uses mean method for prediction of new data points and approximates the relationship between independent variables and the outcome by averaging the k nearest observations. Grid Search was applied for tuning the hyper parameter n_neighbours which is found out to be 5.

### 5.4 Long Short Term Memory (LSTM)

Long short-term memory extends RNN, basically extending the memory. LSTMs enable recurrent neural networks to remember values over a longer time period as they contain information in a memory. The sigmoid activation function has been chosen in LSTM and the optimizer used is Adam. The number of epochs while model-fitting is 200 and the batch size is kept to be 32.

## 6 Results and Analysis

The actual values of the target variable and the predicted values of the target variable are plotted in the same graph for both short term and long term forecasting and as shown below for all the models (except Linear Regression and Decision Tree long term forecasting which is already shown in Fig.6 and Fig.7)

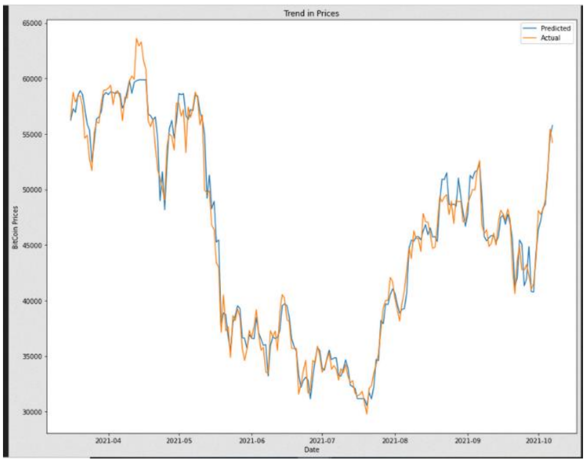


Fig. 8 Bitcoin price vs Date (Random Forest) (Short Term)

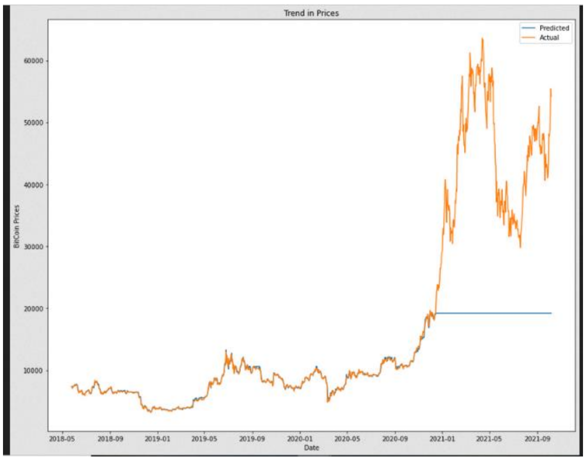


Fig. 9 Bitcoin price vs Date (Random Forest) (Long Term)

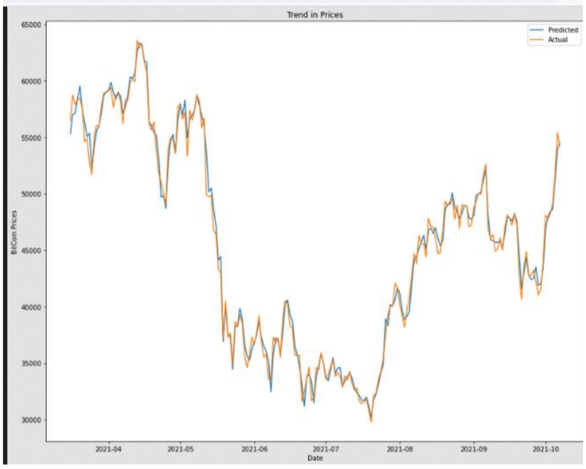


Fig. 10 Bitcoin price vs Date (MLP) (Short Term)

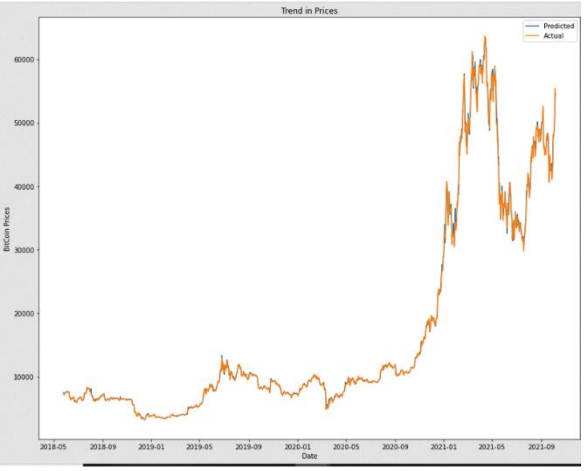


Fig. 11 Bitcoin price vs Date (MLP) (Long Term)

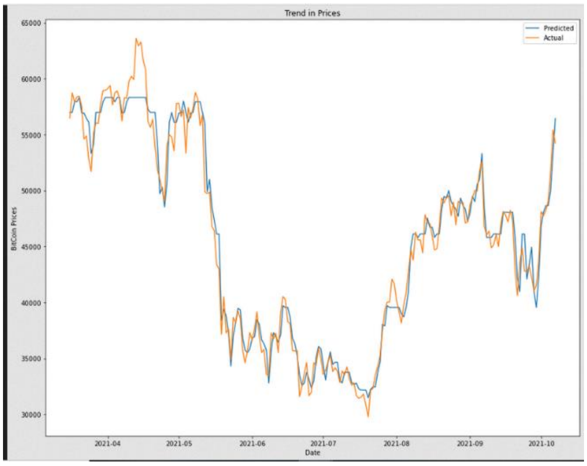


Fig. 12 Bitcoin price vs Date (KNN) (Short Term)

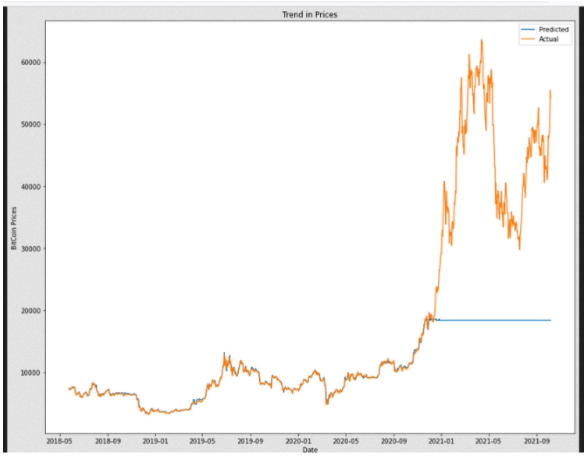


Fig. 13 Bitcoin price vs Date (KNN) (Long Term)

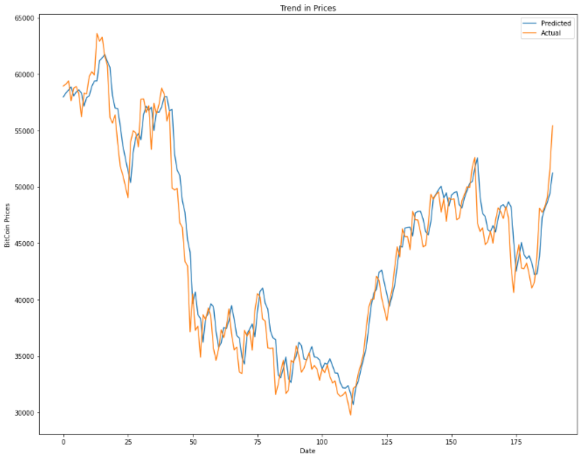


Fig. 14 Bitcoin price vs Date (LSTM) (Short Term)

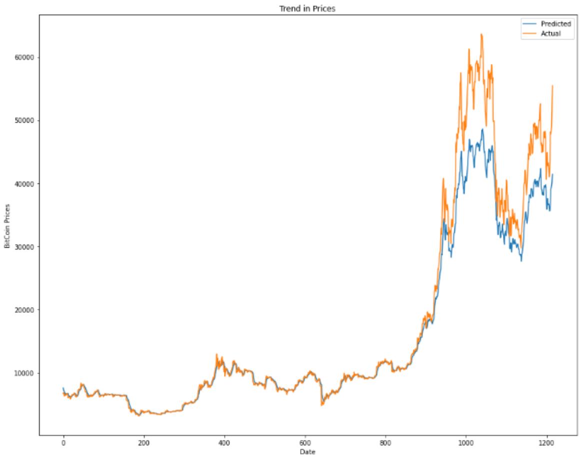


Fig. 15 Bitcoin price vs Date (LSTM) (Long Term)

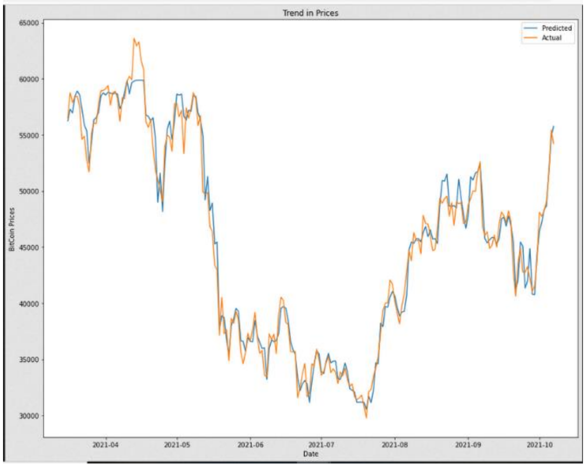


Fig. 16 Bitcoin price vs Date (Linear Regression) (Short Term)

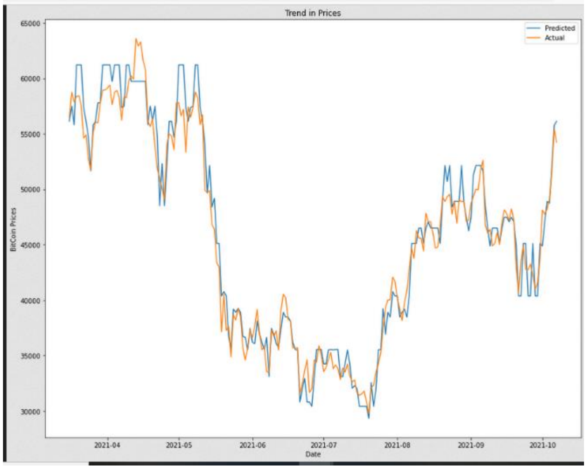


Fig. 17 Bitcoin price vs Date (Decision Tree) (Short Term)

The $R^2$ score for all the models is summarized in table 1 as follows:

Table 1: $R^2$ score of all Models

| **Model** | **Time Frame** | **$R^2$ score** |
|---|---|---|
| Decision Tree Regressor | Long Term<br>Short Term | 0.3549<br>0.9647 |
| Random Forest Regressor | Long Term<br>Short Term | 0.3568<br>0.9812 |
| KNN Regressor | Long Term<br>Short Term | 0.3189<br>0.9740 |
| Linear Regression | Long Term<br>Short Term | 0.9993<br>0.9931 |
| MLP Regressor | Long Term<br>Short Term | 0.9993<br>0.9922 |
| LSTM | Long Term<br>Short Term | 0.9402<br>0.9510 |

## 7 Conclusion

For long term price trends, linear regression and MLP proved to be the best with a R2 score of 0.9993. For short term price trends, linear regression was the best being marginally better than MLP with a R2 score of 0.9931 and 0.9922 respectively.

### Contribution of Each Group member

Zuber Khan-Literature Review, Decision Tree, Random Forest Regressor, K-Nearest Regressor, MLP, Report
Aurko Mitra-Dataset Collection, Processing, Presentation, LSTM
Debarshi Dasgupta-Linear Regression, LSTM, Report

## References:

1. Muniye, Temesgen & Rout, Minakhi & Mohanty, Lipika & Satapathy, Suresh. (2020). Bitcoin Price Prediction and Analysis Using Deep Learning Models. 10.1007/978-981-15-5397-4_63.
2. Shankhdhar, A., Singh, A. K., Naugraiya, S., and Saini, P. K., “Bitcoin Price Alert and Prediction System using various Models”, in *Materials Science and Engineering Conference Series*, 2021, vol. 1131, no. 1, p. 012009. doi:10.1088/1757-899X/1131/1/012009.
3. Ho A, Vatambeti R, Ravichandran SK. (2021) Bitcoin Price Prediction Using Machine Learning and Artificial Neural Network Model. *Indian Journal of Science and Technology*. 14(27): 2300-2308. https://doi.org/10.17485/IJST/v14i27.878
4. S. McNally, J. Roche and S. Caton, "Predicting the Price of Bitcoin Using Machine Learning," *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, 2018, pp. 339-343, doi: 10.1109/PDP2018.2018.00060.
5. Greaves, A., & Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin. *No Data*.
6. S. Velankar, S. Valecha and S. Maji, "Bitcoin price prediction using machine learning," 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018, pp. 144-147, doi: 10.23919/ICACT.2018.8323676.