2023

Bitcoin Closing Price Prediction



**CS4551- MACHINE LEARNING**

**Project**

**Table of Contents**

[1. Background 2](#_Toc135295081)

[2. Problem Definition 2](#_Toc135295082)

[3. Dataset 3](#_Toc135295083)

[4. Solution Statement 3](#_Toc135295084)

[5. Benchmark Model 3](#_Toc135295085)

[6. Evaluation Metrics 3](#_Toc135295086)

[6.1. Mean Squared Error 3](#_Toc135295087)

[6.2. Root Mean Squared Error 4](#_Toc135295088)

[6.3. Mean Absolute Error 4](#_Toc135295089)

[6.4. Coefficient of Determination 4](#_Toc135295090)

[7. Project Design 5](#_Toc135295091)

[7.1. Import Necessary Libraries 5](#_Toc135295092)

[7.2. Load Data 5](#_Toc135295093)

[7.3. Exploratory Data Analysis 6](#_Toc135295094)

[7.4. Data Cleaning 8](#_Toc135295095)

[7.5. Building the Gradient Boosting Regression Model 9](#_Toc135295096)

[7.5.1. Hyperparameters Tunning 10](#_Toc135295097)

[7.5.2. Final Model 12](#_Toc135295098)

[7.6. Linear Regression Model 13](#_Toc135295099)

[8. Models Comparison 14](#_Toc135295100)

[9. References 15](#_Toc135295101)

# Background

Bitcoin is a decentralized digital currency that allows for peer-to-peer transactions without the need for a trusted intermediary such as a bank or government. It was created in 2009 by an unknown person using the pseudonym Satoshi Nakamoto and has since become the largest cryptocurrency by market capitalization. Bitcoin's underlying technology, blockchain, provides a secure and transparent ledger of all transactions and allows for the creation of new tokens through a process called mining. Bitcoin can be used for online payments, remittances, e-commerce, and other applications that require fast, secure, and low-cost transactions [1].

Bitcoin is not only a currency, but also a technology and a social movement. It represents a paradigm shift in the way money and value are created, exchanged, and stored. It challenges the existing power structures and institutions that control the global financial system. It empowers individuals to have more control over their own money and data. It inspires people to imagine new possibilities for the future of money and society. While initially met with skepticism, Bitcoin has gained widespread adoption and attention from both individuals and institutions, leading to a volatile market and the rise of numerous other cryptocurrencies. However, predicting the price of Bitcoin remains a challenging task due to its complex and highly speculative nature [1].

# Problem Definition

The closing price of a stock is the final price at which it is traded on a particular trading day. It is the last price at which a trade occurred before the market closed for the day. The closing price is an important metric for investors, as it is used to calculate gains or losses on investments made in a particular stock. The closing price can also be used to track trends and patterns in the market, and to inform decisions about future investments or trades. It is one of the most commonly used metrics in financial analysis and is widely observed by investors around the world [2].

Consequently, predicting the closing price of bitcoin is important for investors looking to make informed decisions about buying and selling this volatile cryptocurrency. As the price of bitcoin is highly unpredictable and sensitive to various market factors, accurately forecasting its closing price is a challenging task.

# Dataset

The dataset used is downloaded from Kaggle [3]. The dataset has a size of 97 kB. In addition, it consists of 5 columns and 1885 records. Each record in the dataset represents the information of the bitcoin in a specific day. Specifically, the dataset consists of bitcoin data from the 1st of February 2018 till the 31st of March 2023. In other words, the 1885 records represent the 1885 days in the range specified.

# Solution Statement

In this project, a predictive model is developed to utilize the historical price data of bitcoin to predict its closing price. The outcome of the project will inform potential investors and traders about the future price of bitcoin and help them make informed decisions. In particular, Gradient Boosting regression model will be used.

# Benchmark Model

The designed model will be compared with simple linear regression model.

# Evaluation Metrics

The problem addressed by this project is of a regression type. Therefore, to evaluate the performance of the model, the following evaluation metrics are used.

## Mean Squared Error

The mean squared error (MSE) is a metric used to evaluate a regression model by measuring the average of the squares of the residuals , which are the differences between the predicted and actual values. It represents the average squared difference between the predicted and actual values and is calculated as the sum of the squared residuals divided by the number of samples [4].

## Root Mean Squared Error

The root mean squared error (RMSE) is a variant of the mean squared error , which takes the square root of the mean squared error to obtain the square root of the average squared difference between the predicted and actual values. It is a useful metric for evaluating the accuracy of a model, as it is expressed in the same units as the target variable [4].

## Mean Absolute Error

The mean absolute error (MAE) is a metric used to evaluate a regression model by measuring the average of the absolute differences between the predicted and actual values. It represents the average difference between the predicted and actual values and is calculated as the sum of the absolute residuals divided by the number of samples.

## Coefficient of Determination

The coefficient of determination, also called R-squared score (R2), is a metric that measures the proportion of the variance in the target variable that is explained by the regression model. It takes values between 0 and 1, where 0 indicates that the model does not explain any variance and 1 indicates that the model explains all of the variance.

# Project Design

The steps of performing the project are as follows:

## Import Necessary Libraries

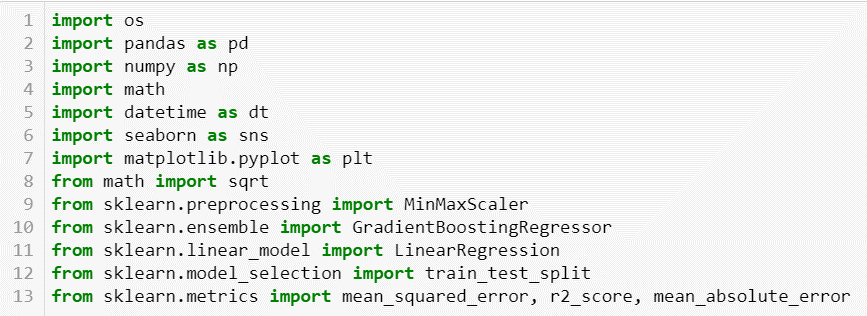


Figure Importing libraries.

## Load Data

The data is loaded, and a summary of the columns are printed as is shown in Figure 2.

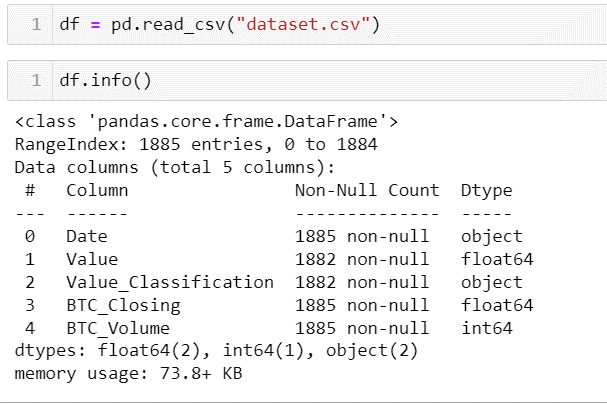


Figure Loading Data.

## Exploratory Data Analysis

First, a summary description of the data is printed to get a general view of the dataset as is shown in Figure 3. From the description, it is noted that there is high variation within the numerical features such as the volume feature. Therefore, the data needs to be normalized.

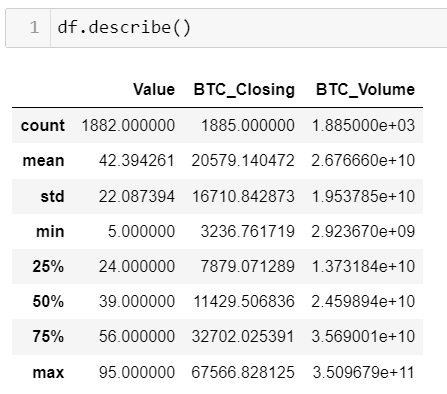


Figure Data description.

Following that, the data is checked for any missing values as is shown in Figure 4. It was found that there are three records with missing Value and Value\_classification. In addition, the dataset was checked for duplicated for which none were found.

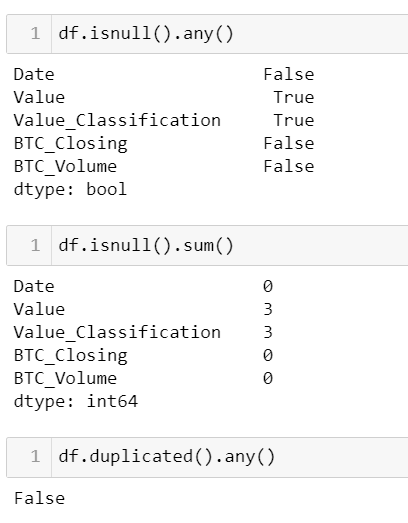


Figure Checking for missing and duplicate values.

The value classification feature is then plotted in Figure 5. It is clear that Fear is the prominent class.

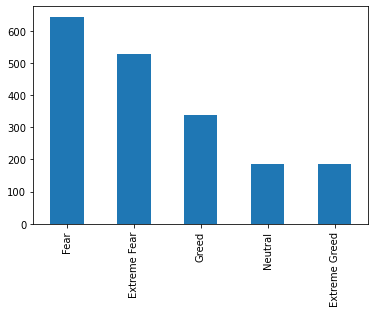


Figure Value\_classification feature.

Following that, the closing price, BTC\_volue, and value are plotted against date as is shown in Figures 6 to 8. From the figures, it is clear that the closing values were high in 2021 especially toward the end of the year. On the other hand, it is clear from Figure 7 that there is an outlier in the BTC\_volume that needs to be removed from the data.

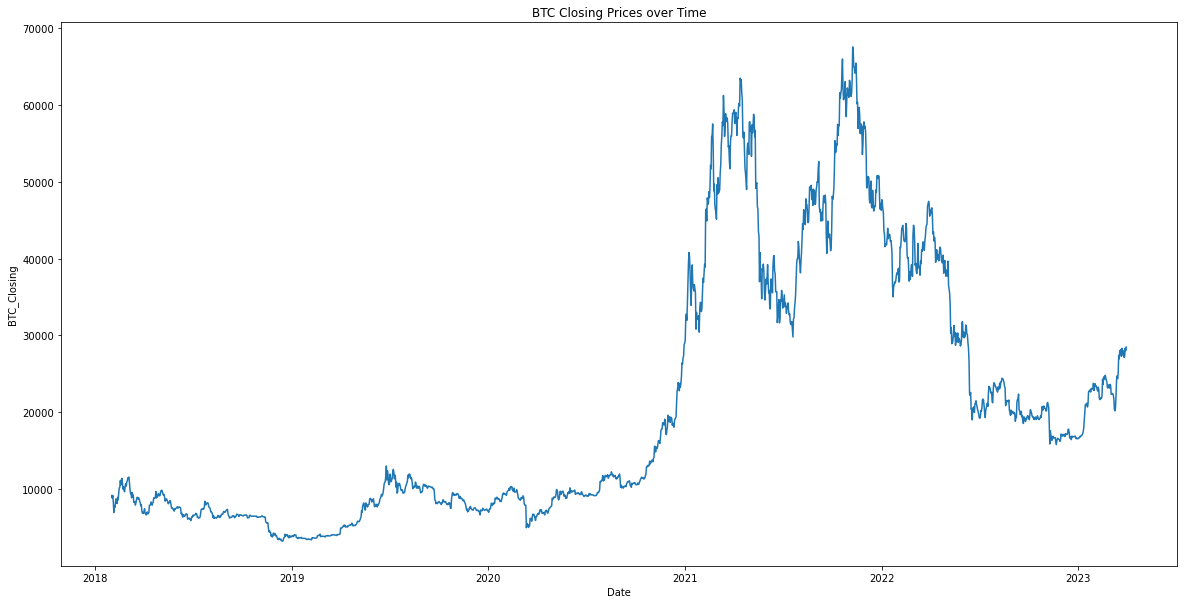


Figure Closing price over time.

A blue line graph on a white background

Description automatically generated with low confidence

Figure BTC volume over time.

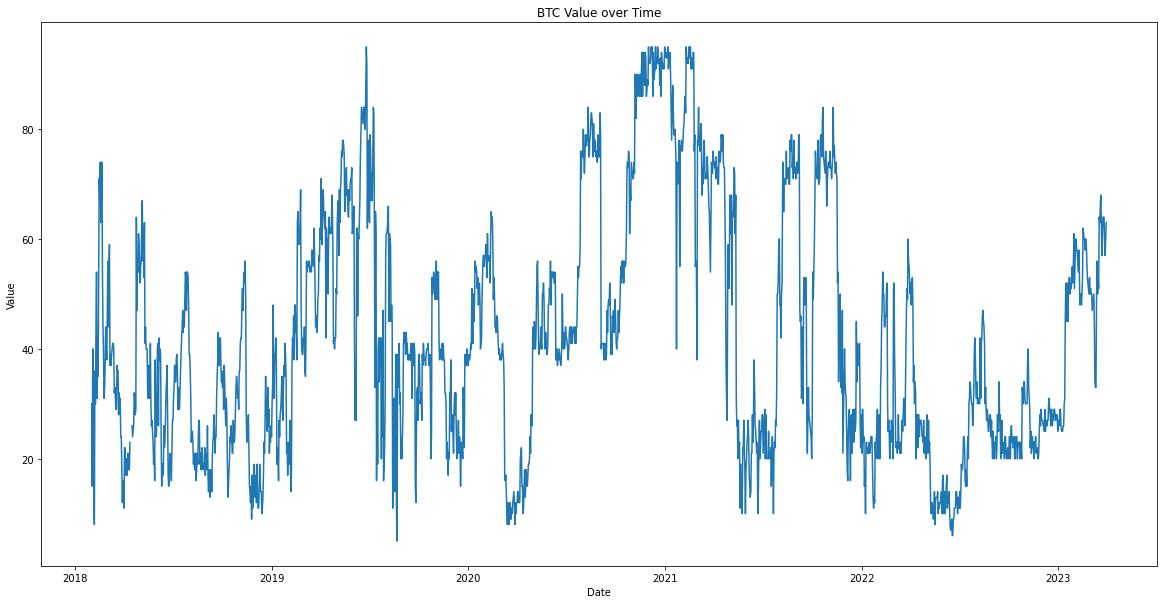


Figure BTC value over time.

## Data Cleaning

First, the outlier in the BTC\_volume and the rows with missing value and value classification are removed. Following that, the date is split into month and year column as the full date is of no use in building the model. Then the Date column is dropped from the dataset. These processes are shown in Figure 9.

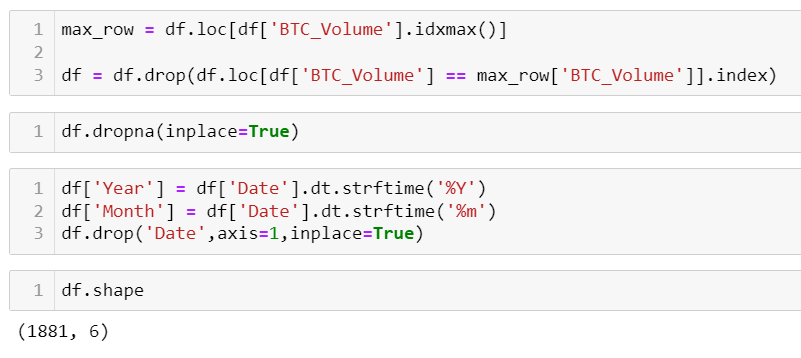


Figure Removing outliers/missing values and splitting date column.

Following that, the value\_classification column is one-hot encoded as is shown in Figure 10.

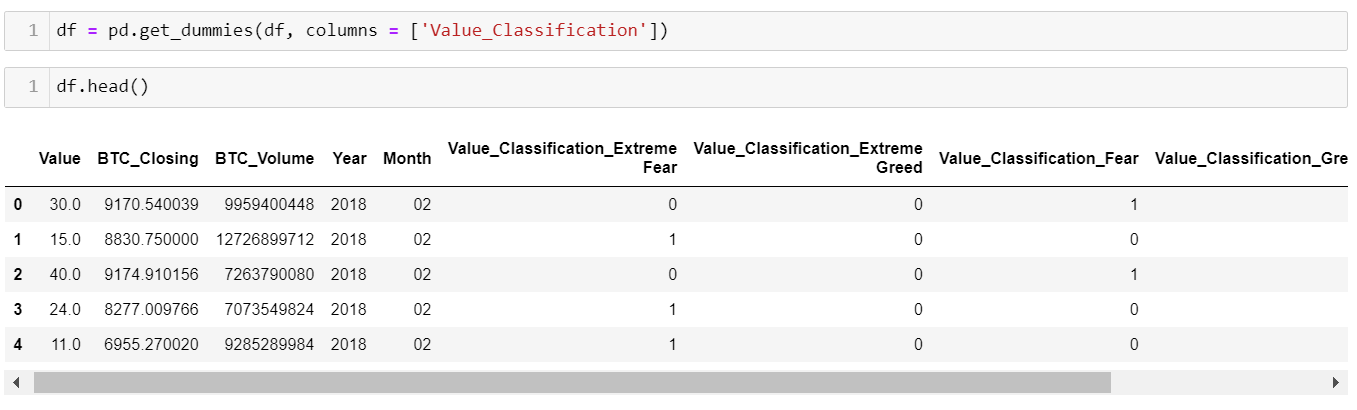


Figure Value\_classification One-hot encoding.

## Building the Gradient Boosting Regression Model

First, the data is split to the input features (9 features) and the output (i.e., BTC\_closing). After that, the data is split into training and testing where the testing data is 20% of the total dataset. Finally, before building the model, the input data is scaled using the MinMaxScaler. These processes are shown in Figure 11.

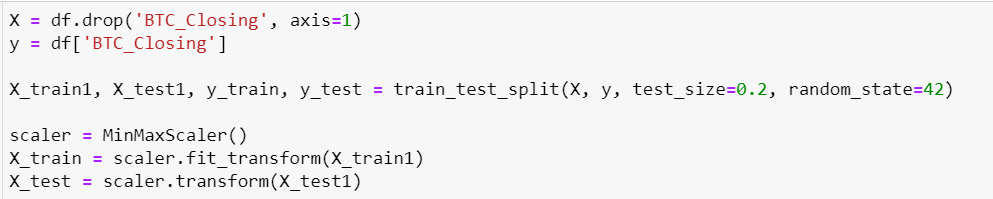


Figure Data splitting.

### Hyperparameters Tunning

#### Depth

A loop is created to train the gradient boosting model using different max\_depth values ranging from 1 to 29 as is shown in Figure 12. In each iteration, the four-evaluation metrics i.e., MSE, RMSE, MAE, and R2, are calculated.

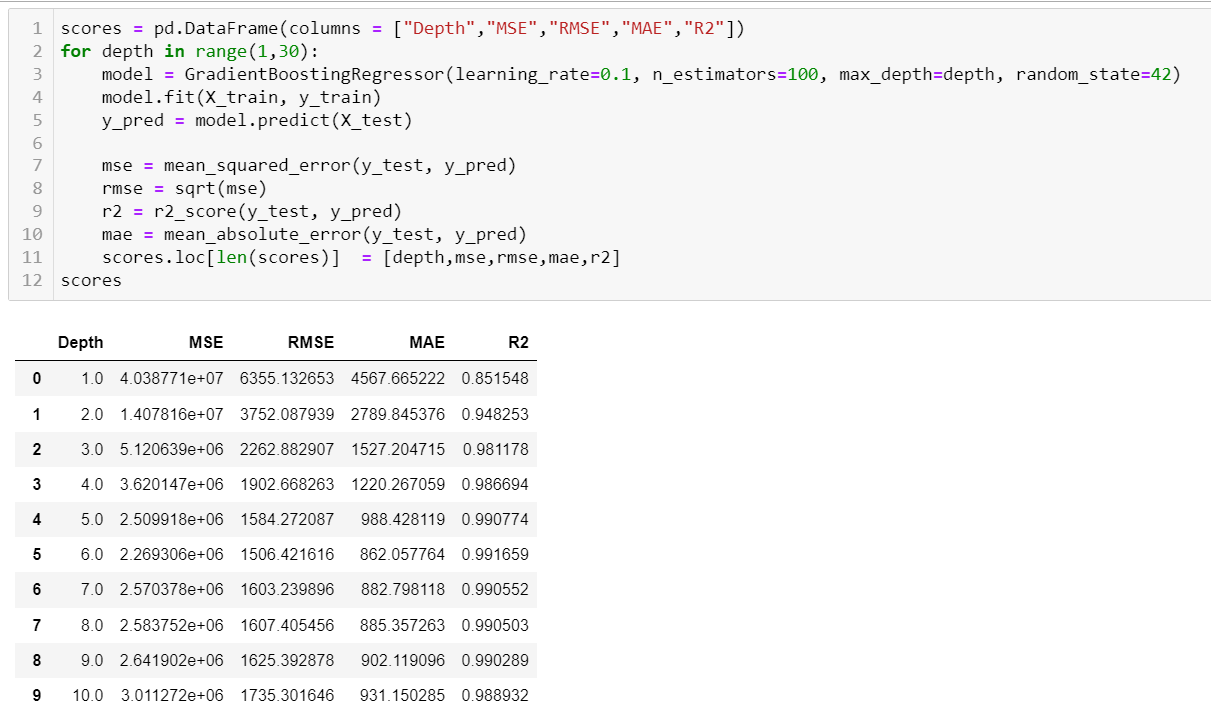


Figure Depth hyperparameter tunning.

The graphs showing the RMSE, MAE, and R2 against depth are shown in Figure 13. It is clear that the errors (RMSE and MAE) values decreases and the R2 score increases in the beginning. However, after the depth 4, the error starts to increase slightly and the R2 score decreases. Therefore, the optimal depth is 4.

A picture containing screenshot, line, rectangle, parallel

Description automatically generated

Figure RMSE, MAE, and R2 against depth.

#### Learning Rate

Similarly, a loop is created to train the model using different learning rates starting from 0.01 till 0.1 with a step of 0.01. In addition, the MSE, RMSE, MAE, and R2 are calculated for each iteration. This process is shown in Figure 14.

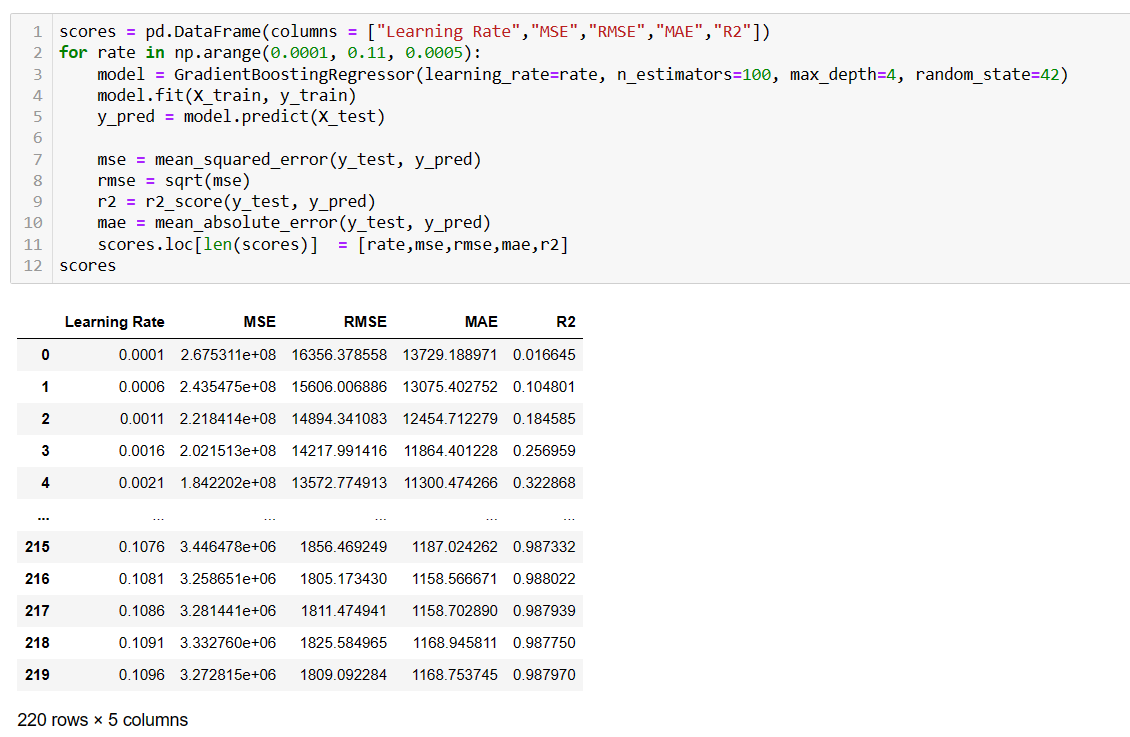


Figure Learning rate hyperparameter tunning.

The graphs showing the RMSE, MAE, and R2 against depth are shown in Figure 15. It is clear that the errors (RMSE and MAE) values decreases and the R2 score increases in the beginning. It is clear that optimal learning rate is 0.08.

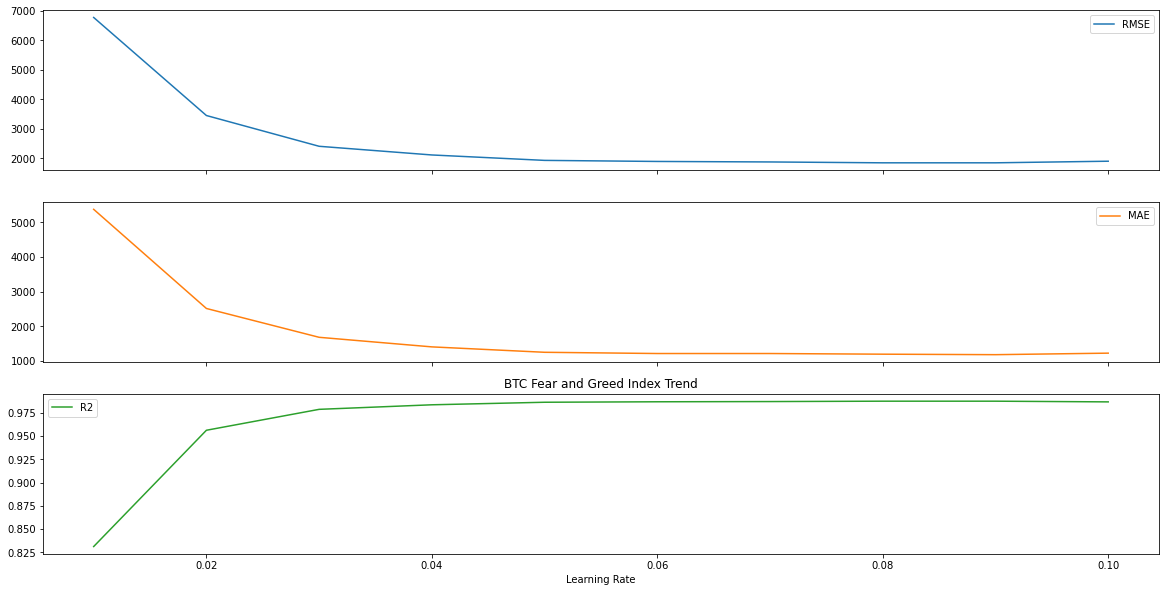


Figure RMSE, MAE, and R2 against learning rate.

### Final Model

The final model will have the learning rate as 0.08 and the max\_depth as 4. The training and evaluation of the final model is shown in Figure 16. It is clear that model performs well as the R2 score is very close to 1. In addition, the predicted and true values are plotted in Figure 17.

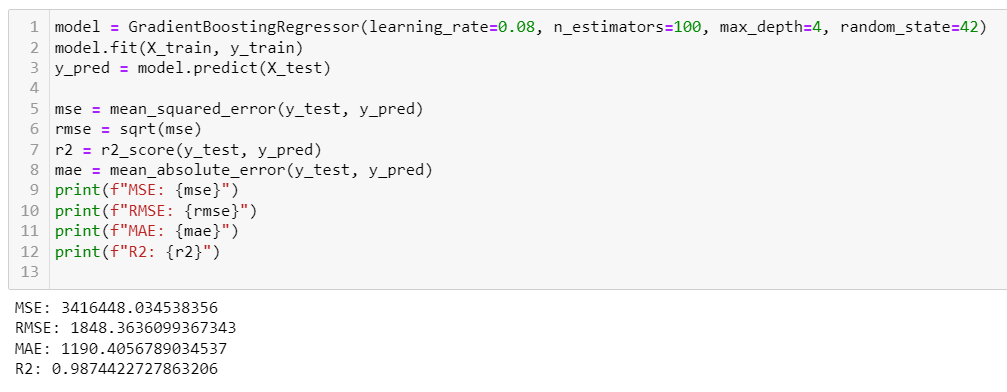


Figure Final model training and testing.

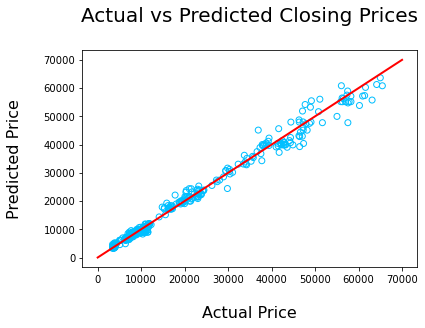


Figure Actual vs predicted closing prices.

## Linear Regression Model

A simple linear regression model is used to train and test the data as is shown in Figure 18. Similarly, the predicted and true values of the closing price are plotted in Figure 19. It is clear that this model performs poorly on the dataset.

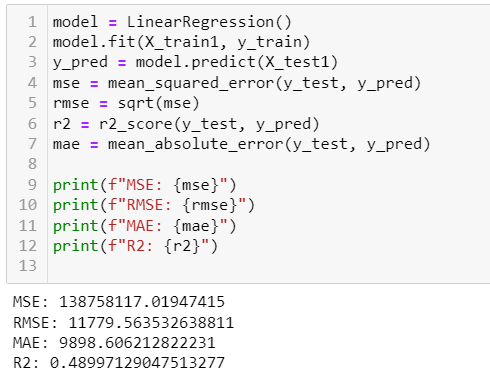


Figure Linear regression model.

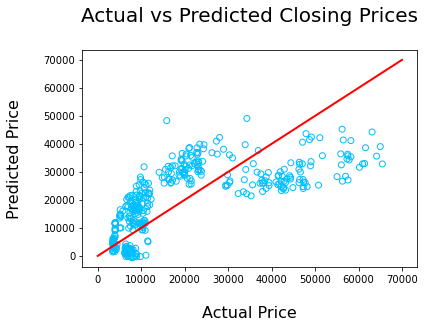


Figure Actual vs predicted closing prices.

# Models Comparison

The evaluation metrics of the gradient boosting regression model and the linear regression model are compared in Table 1. It can be concluded that Gradient Boosting Regression outperformed Simple Regression in terms of overall performance.

The Gradient Boosting Regression model showed a higher R-squared score, which indicates that it explained a higher percentage of the variance in the Bitcoin closing price data. Additionally, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were lower for the Gradient Boosting model, meaning that it had a lower average squared difference and a lower average absolute difference between the predicted and actual Bitcoin closing prices. Overall, the results suggest that Gradient Boosting Regression is a more effective method for predicting Bitcoin's closing price compared to Simple Regression.

Table Models comparison.

|  |  |  |
| --- | --- | --- |
| **Test Split ratio** | **Gradient Boosting Regression** | **Linear Regression** |
| MSE | 3416448 | 138758117 |
| RMSE | 1848 | 11779 |
| MAE | 1190 | 9898 |
| R2 | 0.987 | 0.489 |

# References

1. S. Nakamoto, “Bitcoin: a Peer-to-Peer Electronic Cash System,” Oct. 2008. Available: <https://bitcoin.org/bitcoin.pdf>
2. A. Hayes, “What Is a Closing Price?,” Investopedia, Oct. 18, 2021. <https://www.investopedia.com/terms/c/closingprice.asp>
3. “Bitcoin & Fear and Greed,” www.kaggle.com. <https://www.kaggle.com/datasets/adilbhatti/bitcoin-and-fear-and-greed>
4. A. Chugh, “MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?,” Medium, Dec. 08, 2020. <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>