Analysis of Bitcoin Price based on Historical Data

Yeshwanth Bharadwaj Mellachervu
Masters in Data Science
University of Michigan, Dearborn

Abstract

Bitcoin price analysis has been very important for bitcoin traders as one can buy or sell bitcoins based on the value it has in the market. In this paper, bitcoin price analysis is done based on various times of year, so that we can know the value of bitcoin at given point of time. Various models have also been built to analyze the pattern and predict the values from a given set of values.

I. INTRODUCTION

Bitcoin is popular know cryptocurrency and block chain technology was first developed for this. Cryptocurrency is digital form of money which has encryption techniques to verify the transfer of money and creation of monetary units. Bitcoin historical data consists of factors like Open, High, Low, Close, Volume.

Open is the opening price of bitcoin when the market starts. High is the highest prices sold in time between open and close time of market. Low is the lowest price sold, Close is the closing price of bitcoin and finally Volume is the number of bitcoins sold in the time when market is open and closed.

These factors are used in the analysis of bitcoin price prediction. The Close price of Bitcoin is important for a short-term trader to know based on which he can know what rate the market open and when to buy or sell bitcoins.

Developing a forecast model may not exactly predict the price of bitcoin but will be helpful in getting an estimated value by analyzing the historical data.

II. RESEARCH QUESTIONS

1. What can be done with this huge data?
2. What attributes are going to be useful in this analysis?
3. What model fits best to perform analysis?
4. What is successful output of this analysis?
5. How to handle the missing values in the dataset?

III. METHODOLOGY

The dataset is collection of 1-min intervals of OHLC data from the year 2012 to 2021 from Kaggle. The aim is to plot different visualizations which help in understanding the trends and seasonality of the data over time.

Preprocessing: The dataset contains 4857377 rows and 8 columns with min-min interval bitcoin OHLC (Open, High, Low, Close) data and there are also missing values/NaN. There are different ways how missing values can be handled. In first method the missing values are filled with aggregate functions like mean or median i.e., one column would be selected and the average or the median value for that column is calculated and then the same values are replaced in place of the missing values. But mean is in general sensitive to outlier so it is better to use median which is a bit less sensitive to outliers. While dealing with time series data the mean/ median is not the right way to replace the missing values as the data is not stationary.

So, in the second method the missing values are filled based on the nearest values either in forward direction or in the backward direction. This process which is called Interpolation has two methods linear and polynomial. This method in most cases gives better output but in the current scenario it wasn't useful.

Third method used here to handle missing values is to simply drop the rows wherever the missing values are observed. Now the dataset in reduced to 3613769 rows and 8 columns.

The data is now converted with index as Date Time and analysis using visualization and models are built on this data.

Regression Models:

Simple Linear Regression: The Simple Regression model is built for open price and close price. The model is built to understand how the opening price has a significant impact on closing price and the patterns associated to both. A linear equation is formed so that given an opening price we can predict the approximate the closing price. Also, a confidence interval has been built to tell how confident we are that the point estimate will fall in the given interval.

$$\text{CLOSE} = -0.076 + \text{OPEN}$$

Multiple Linear Regression: Multiple Linear Regression model is built for all the columns together to understand how they impact the closing price of the bitcoin. The multi linear equation of all the different predictors has been formulated so that given the values of predictors we can get a point estimate of nearly how much the closing price of bitcoin would be. Also, a confidence interval for the same has been built.

A Step towards Forecasting of Time series data:

Time series data is collection of data at regular intervals of time. This interval may be minutes, hours, days, months, years. In certain cases, it may be seconds also when it comes to sensor data. A time series data is analyzed to understand the nature of data so that a meaningful insights and accurate forecast can be done.

STL Decomposition:

The time series data generally is combination of level, trend, seasonality and noise/residual. Level is nothing but the average values in time series data. Trend tells if the values are going up / coming down. It is quantified as either upward or downward trend. Seasonality is the repeating cycle in a series data. Noise is the random series in the given data. The time series can be either additive or multiplicative model based on the series. The additive model is the sum of all the above factors i.e., level, trend, seasonality and noise whereas the multiplicative model is the product of all the factors.

The time series is decomposed using STL decomposition into separate plots to understand and analyze the trend, seasonality and if the model is additive or multiplicative. Once the analysis is done then it is checked if the time series is stationary or not.

The time series data is converted to stationary series as the forecast are more reliable. The test for stationary time series can be done using statistical test called Augmented Dickey Fuller test (ADH). Once this check is done the timeseries is detrended and de-seasonalized.

After that the missing values in data are handled and autocorrelation functions are calculated and forecasted with proper values. This paper only focuses on the analysis and not on forecasting.

IV. Results

Any model when built needs to be evaluated with some metrics. The below are some analyses with visualizations followed by model validations. The metrics used here are R-squared and Root Mean Square Error. R-squared tell us how much variation of target is explained by predictors and RMSE is the deviation of residuals and residuals are the measure how far data points are from the actual regression line.

Analysis using visualizations:

In the below Fig1 the plot shows the upward trend of the bitcoin High price over years. It can be observed that there is a sudden hike in the plot from end of year 2020 and starting of year 2021.

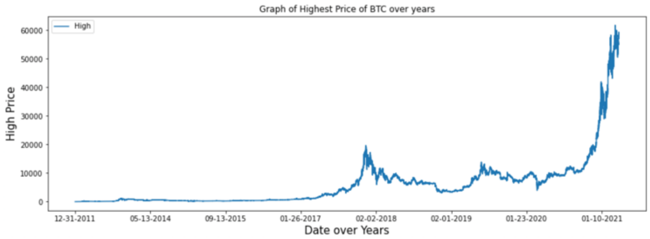

Fig1: Bitcoin High price over years

In the below Fig2 the plot shows the upward trend of the bitcoin Open price over years. It can be observed that there is a sudden hike in the plot from end of year 2020 and starting of year 2021.

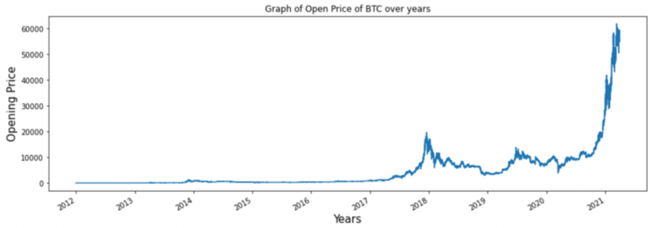

Fig2: Bitcoin Open price over years

In the below Fig3 the plot shows the upward trend of the bitcoin Close price over years. It can be observed that there is a sudden hike in the plot from end of year 2020 and starting of year 2021.

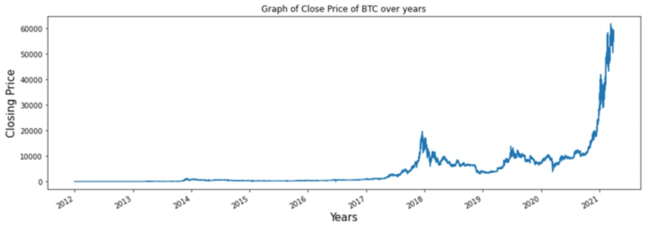

Fig 3: Bitcoin Close price over years

In the below Fig4 the plot is a closer look of plot in Fig3 i.e., from 2020 to 2021. The closing price has been gradually increasing over months in the given range.

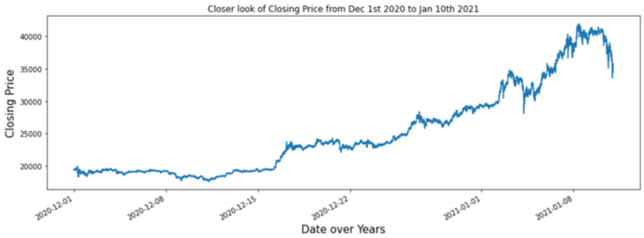

Fig 4: A Closer look at the Closing Price Dec 1st 2020 to Jan 10th 2021

In the below Fig5 the plot shows the upward trend of the bitcoin Low price over years. It can be observed that there is a sudden hike in the plot from end of year 2020 and starting of year 2021.

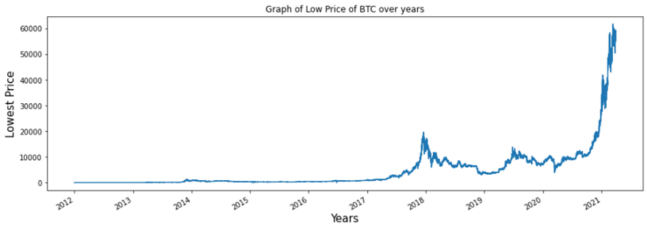

Fig 5: Bitcoin Low price over years

In the below Fig6 the plot shows the trend of the bitcoin Volume sold over years. It can be observed that there are a greater number of bitcoins sold around the year 2015.

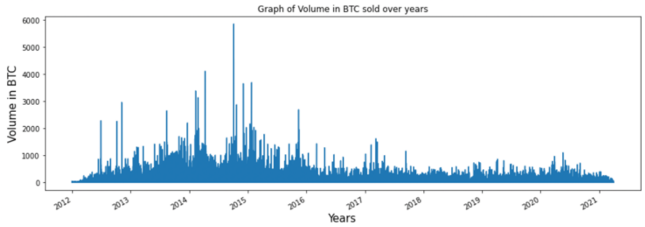

Fig 6: Bitcoin Volume sold over years

In the below Fig7 the plot shows the trend of the bitcoin Volume in currency sold over years. It can be observed that the price of bitcoin sold over years is high in the year 2021 though the number of bitcoins sold/bought is less but the volume in currency is high.

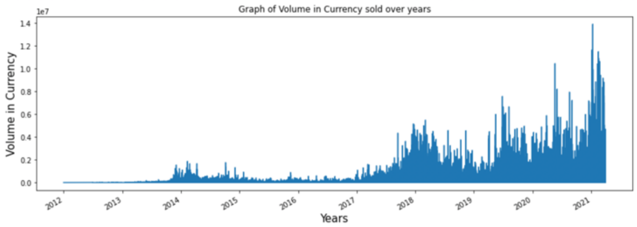

Fig7: Bitcoin Volume sold in currency over years

In the below Fig8 the plot shows the upward trend of the bitcoin Weighted price over years. It can be observed that there is a sudden hike in the plot from end of year 2020 and starting of year 2021.

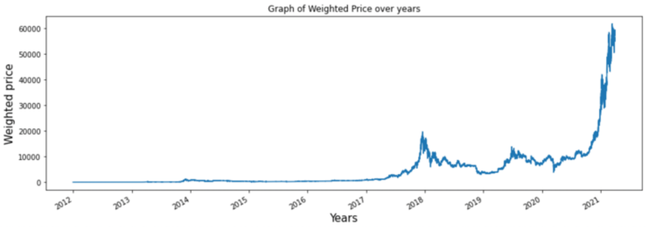

Fig8: Bitcoin Weighted price over years

In the below Fig9 the plot shows the 3 months moving average of volume of bitcoins sold vs the volume of bitcoins sold in currency over different years. It is observed that the price of bitcoins started to increase from 2018 and there has been a slight hike down in year 2019 but again it started gradual increase from 2021.

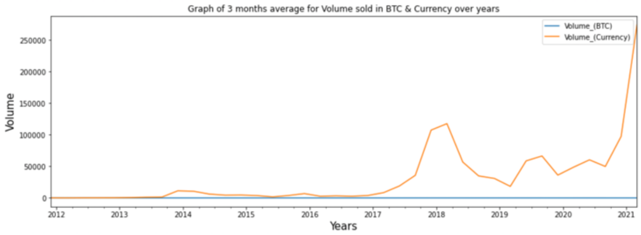

Fig 9: Volume in BTC vs Volume in Currency sold 3months average

In the below Fig10 the plot shows the 4months, 8 months and 12 months moving average of Closing price of bitcoins over years. It can be observed that average closing price is very big spike from 2014 and from then it is gradually increasing with little down hikes too.

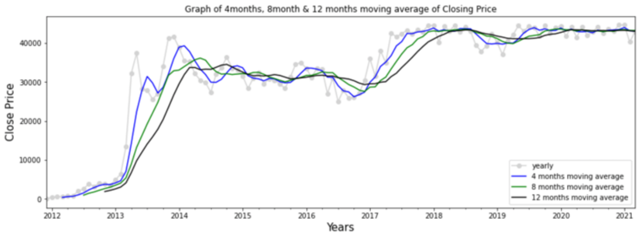

Fig 10: Closing Price moving average for 4,8 and 12 months

In the below Fig11 the graph is a line plot showing the Volume of BTC sold over different months in each year. The lines represent different years and how the volume of BTC is affected each month in a year is shown by the ups and down of the lines. We can observe that approximately in February in all the years there is a spike down in the volume of bitcoins sold.

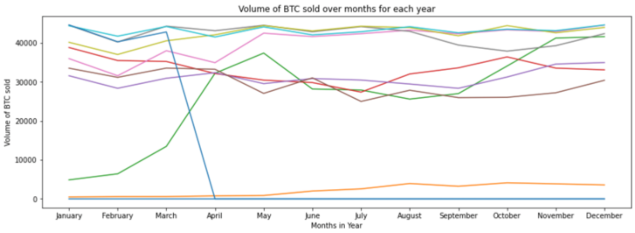

Fig 11: Volume of bitcoins sold over months for each year

In the below Fig12 the graph shows the strip plot with a boxplot on it visualizing the volume of bitcoins sold over years. From the plot it is can be understood that average volume being sold in gradually increasing i.e., the demand of bitcoins/ digital currency has been increasing over years. One other observation which can be seen is in the year 2021 the box plot and the points are in two extremes this is because of insufficient data of in year 2021 (the data collected is only till march and after that the values are took as 0s which affected the plot)

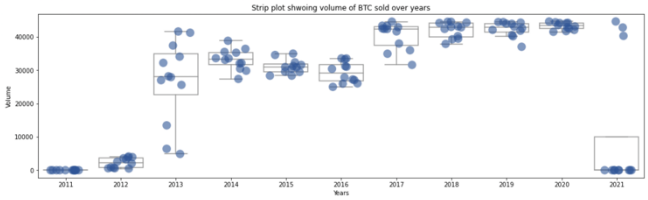

Fig 12: Strip plot showing volume of BTC sold over years

In the below Fig13 the plot shows seaborn heatmap explaining the correlation between each pair and it is observed that most of the pairs are highly correlated i.e., both the factors are connected and depend on one another.

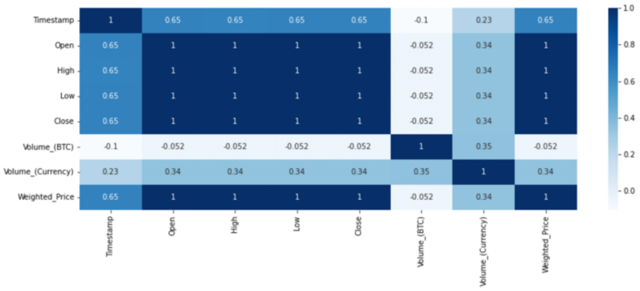

Fig 13: Heatmap showing correlation between each pair

Model Validation:

Simple Linear Regression:

In the below Fig14 the graph shows the scatter plot between open and close price with linear regressor line. It is observed that all the points in scatter plot are on the line with less residuals. So, the accuracy of the model with R-squared metric score close to 1.

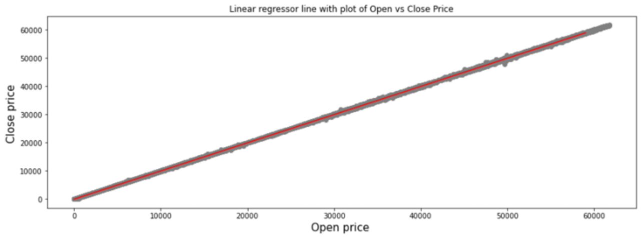

Fig 14: Linear regressor line plot between open and close price

The model has R-squared value of 0.999

Multiple Linear Regression:

Multiple linear regression has been performed using two different models once with train_test_split method and other is OLS method. The model has been built with Closing price as the target variable and all other factors being the predictors of the target class.

Model 1: train_test_split:

In this method the model has an accuracy score and RMSE score as below.
The model has accuracy score of 0.999
The RMSE value for the model is 7.228

Model 2: OLS Model

In this method the model has an R-squared value of 0.999.

Time Series Models:

STL Decomposition:

Additive Model:

In the below Fig15 the plot shows the decomposed time series with Level, trend, seasonal and noise in additive model.

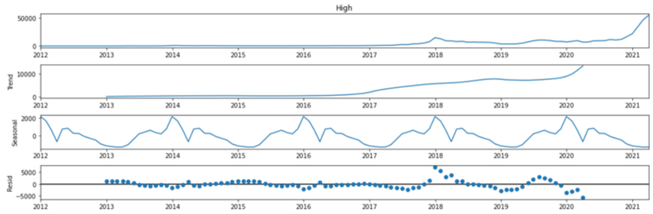

Fig 15: Decomposed trend, seasonal, noise data in additive model

Multiplicative Model:

In the below Fig16 the plot shows the decomposed time series with Level, trend, seasonal and noise in multiplicative model.

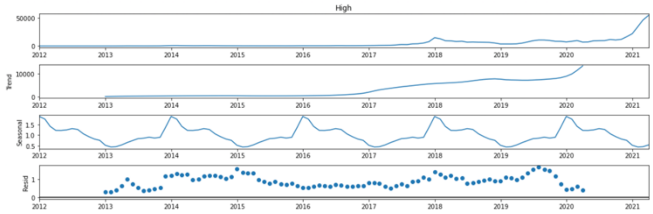

Fig 16: Decomposed trend, seasonal, noise data in multiplicative model

Limitations: Although Regression models like Linear regression and multiple linear regression can find the pattern, they are not very accurate as the data is not stationary.
Performing these models and evaluating them with different metrics is not much helpful. Because the columns in dataset are highly correlated and the metrics always shows that the model built is perfect. In reality it's not the case as the value of bitcoin may suddenly drop or increase at any point of time.

Future Work: ARIMA and Neural Networks models can be used to forecast the price values and increase the model performance.

V. Discussions

1. In Big Data analysis a lot of data is always helpful in training a model accurately. In this dataset a huge data has been given but most of the data has missing values, (so as already discussed in Section 3. Methodology) certain pre-processing techniques are used to filter the missing values or fill them based on various techniques. Once the data is pre-processed different plots are used as visualizations which help in understanding the relation between OHLC data and time. Then models are built in understanding the pattern of the data so that future values can be predicted based on model built.
2. From the models built i.e., simple linear regression and multiple linear regression it is evident that all the Open, High, Low, Close and Volume are important factors in predicting the future price.
3. The models built are simple linear regression, multiple linear regression and a step towards forecasting i.e., STL decomposition. Logistic regression model cannot be built as this is not a categorical data. Out of the models built both the simple linear and multiple linear regression models exhibits high metrics values of R-squared nearly 1 (0.999). This may be a case where the model is overfitting. The OHLC data when closely seen it is observed that all of them have nearly same values. This is one observation because of which the models exhibit r-squared metrics nearly 1. The best models that can be built to this dataset are ARIMA model / Neural networks models which gives better results that simple regression models.
4. The successful output is being able to build a model where it can analyze the pattern and predict the output for a given specific inputs. Here the outputs are successful but all the models generating successful outputs need not be best model. The model with good metrics should the best model. A step towards finding the best model is done which is STL decomposition where the trend, seasonality and

residual are separated from time series data and analyzed individually.

5. The missing values in the dataset can be handled in different ways, one way is aggregate function like mean, median and calculate the average/ median values for column and fill the same values in the missing values. But the time series data is not stationary so filling missing values based on that won’t be helpful. In the second strategy a method called interpolation is used where the missing values are filled in forward or backward direction based on the nearest values i.e., the values above or below them. Interpolation is a better way of filling the missing values rather using aggregate functions. This method of replacing the missing values didn’t help in the current dataset. So, we opt for a third strategy where the missing value rows are completely dropped and then the models are built on it.

VI. Conclusions

The objective of this paper was to perform an analysis on price of bitcoin and what factors affect the price and how the trader can get benefit by predicting the close price from open price. The models implemented and evaluated on sample data may not always be true as the regression models are performed on stationary data. So, better models like ARIMA (Auto Regressive Integrated Moving Average), Neural networks can be implemented which can predict the pattern more accurately and more data samples are added in the model along with feature engineering.

VII. References

1. Prabhakaran, Selva. “Time Series Analysis in Python - A Comprehensive Guide with Examples - ML+.” *Machine Learning Plus*, 19 Dec. 2021, https://www.machinelearningplus.com/time-series/time-series-analysis-python/
2. Brownlee, Jason. “How to Decompose Time Series Data into Trend and Seasonality.” *Machine Learning Mastery*, 9 Dec. 2020, https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/.
3. Uras, Nicola, et al. “Forecasting Bitcoin Closing Price Series Using Linear Regression and Neural Networks Models.” *PeerJ. Computer Science*, PeerJ Inc., 6 July 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7924725/.
4. Zielak. “Bitcoin Historical Data.” *Kaggle*, 11 Apr. 2021, https://www.kaggle.com/mczielinski/bitcoin-historical-data.