CAPSTONE PROJECT

Predicting the price of Bitcoin

By: Rojesh Dhakal

15/04/2023

Table of Contents

1.0. Problem of Statement

Over the years, Bitcoin has gained significant recognition and attention all around the world as a digital currency which can also be the alternative to traditional financial systems. Predicting the price of Bitcoin has been challenged due to its volatile nature and potential for rapid fluctuations in value. Therefore, it is a major problem for investors and traders. The purpose of this project is to develop a machine learning model through which prediction of price of Bitcoin can be done by looking at its historical data.

The project will focus on downloading data from y-finance and analyzing data of historical Bitcoin prices. The data obtained will be preprocessed and then divided into training and testing sets and then fitted into two different machine learning models and evaluated.

Two machine leaning algorithms, ARIMA and LSTM model, will be trained and evaluated on the dataset to identify the best performing model. The model's performance will be evaluated using a range of metrics, including mean squared error, mean absolute percentage error.

The goal of this project is to develop a machine learning model that can accurately predict the future price of Bitcoin based on historical data. The insights gained from this paper can be used by investors to make informed decisions about buying and selling Bitcoin and mitigate risk in the rapidly evolving world of cryptocurrency trading.

1.1. Business Context

Bitcoin is a digital currency which was created in 2009 and operates in decentralized platforms known as blockchain. The total supply for bitcoin is 21 million which ensures that Bitcoin remains scarce and valuable over time. It can be used as a form of payment for buying goods and services without a central authority or bank as an intermediary. Similarly transfer of funds can be done internationally, more effectively and securely with the help of Bitcoin without financial institution. Bitcoin has attracted attention from the media as well as investors over the years because of its innovative features, such as its simplicity, decentralization, and traceability (Fry and Cheah, 2016). Over the years, the price of Bitcoin has appreciated significantly, therefore, it has become an increasingly popular investment for individuals or investors.

According to Dwyer (2015), the average monthly volatility for other gold or a set of foreign currencies is comparatively lower as compared to Bitcoin. Bitcoin is highly volatile and has large fluctuations due to various factors such as supply and demand, market sentiment and regulatory changes. Therefore, predicting the price of bitcoin is a must to make informed decisions for individuals or investors which help them to make a profit and minimize the risks.

Machine learning has become increasingly popular to the price of Bitcoin, because it allows to analysis large amounts of data which can help to identify different trends and patterns and helps to make predictions about future price movements.

1.2. Stakeholders

The key stakeholder for the project is individual investors who are interested in buying and selling the bitcoins. The main motive for this project is to predict the price of bitcoin accurately using machine learning models. This will give stakeholders more confidence in investing in bitcoin especially during this hard time.

1.3. Business Question

People always invest in assets like Bitcoin by thinking that the price will go up and generate wealth from it. The business question would be how we can use machine learning techniques to predict the price of Bitcoin and how can we identify patterns to minimize losses with accuracy?

1.4. Data Question

The data question would be what the historical price trend of Bitcoin would be. The main objective of this question is to collect and analyze the historical data on Bitcoin prices over time, to identify various patterns and trends which might be predictive of future price movements.

1.5. Purpose

The purpose of this project is to predict the price of Bitcoin with high accuracy using machine learning models. Bitcoin is considered more volatile than other currency such as USD, therefore it provides more potential and motivation to predict the price of Bitcoin. It helps the investors for the opportunity to make profit and minimize the risks. Similarly, the other motive is to look at the moving average of the Bitcoin to identify the trend so that we can identify when to buy and sell Bitcoins.

1.6. Data

The data was downloaded using API from y-finance. Data ranges from 2014 to 2023. The data is clean, and no imputations are required. Initially, the datasets had 3110 rows and 10 columns.

2.0. Data Science Process

2.1. Data Analysis

For further analysis, Year, Month & Days columns were combined as one column using pandas datetime format, which was saved as a new column named, date. Date was then set as an index. The initial year, month and days column were later dropped. Similarly, Dividends and Stock Splits columns were dropped as well, as it contains 0 as a value throughout.

Only the column named Close was used for further analysis because it helped in predicting the closing price of Bitcoin. Other columns were dropped because of multi collinearity.

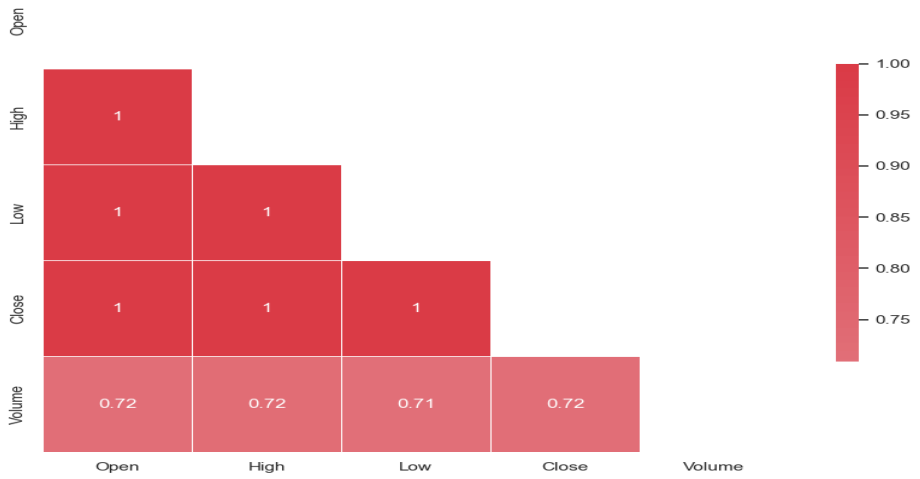


Fig: Correlation Matrix

2.1.1. Closing price Analysis

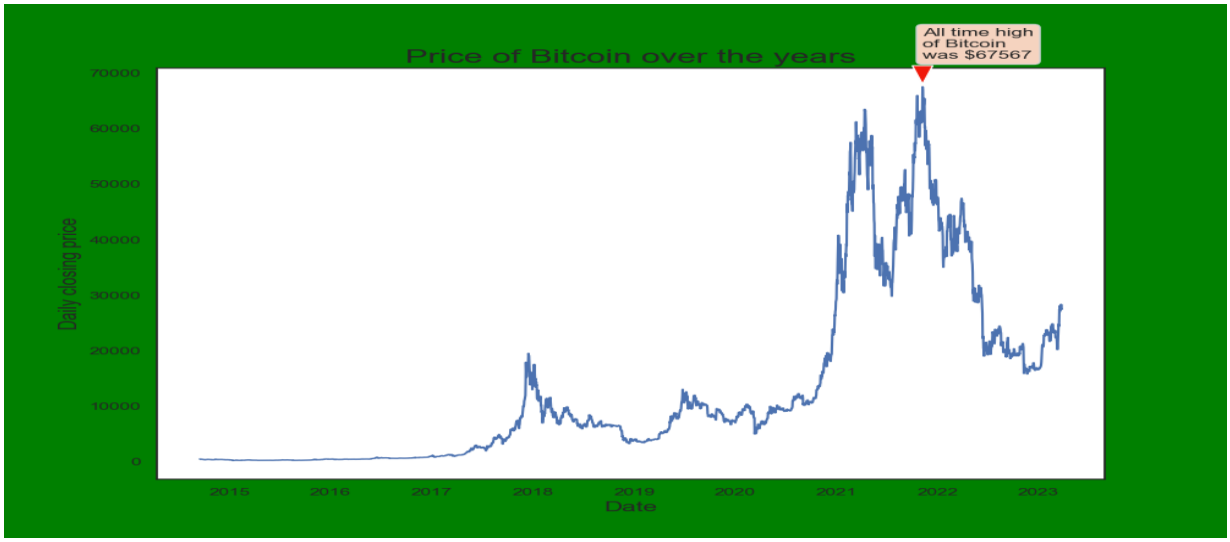


Fig: Closing Price of Bitcoin over the years

As we can see from the figure, the price in 2014 was considerably low but within 8 years the price has risen significantly making an all-time high in 2022 of $67567.

2.1.2. Investment Analysis

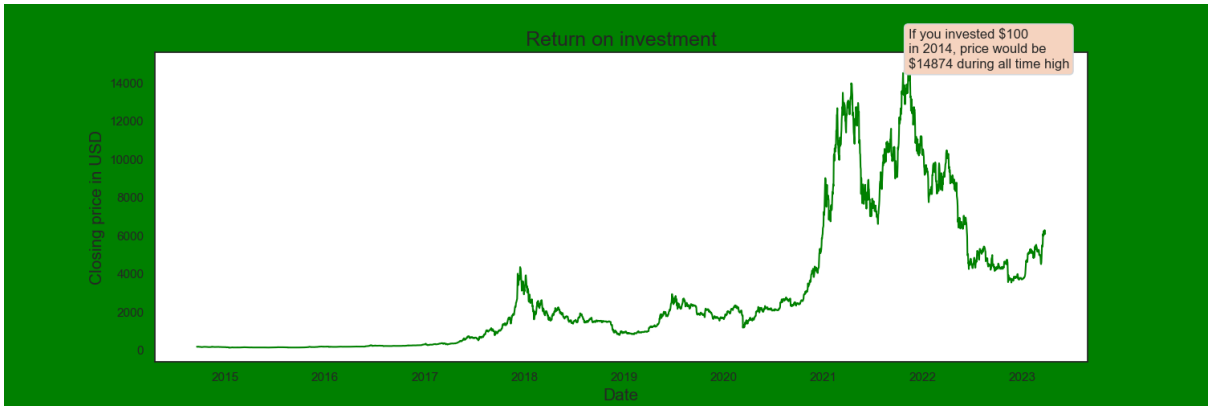


Fig: Investment Analysis

The figure above shows that if you invested $100 in 2014, then at one point your investment would be worth $14874 which is a whopping return of 14774%.

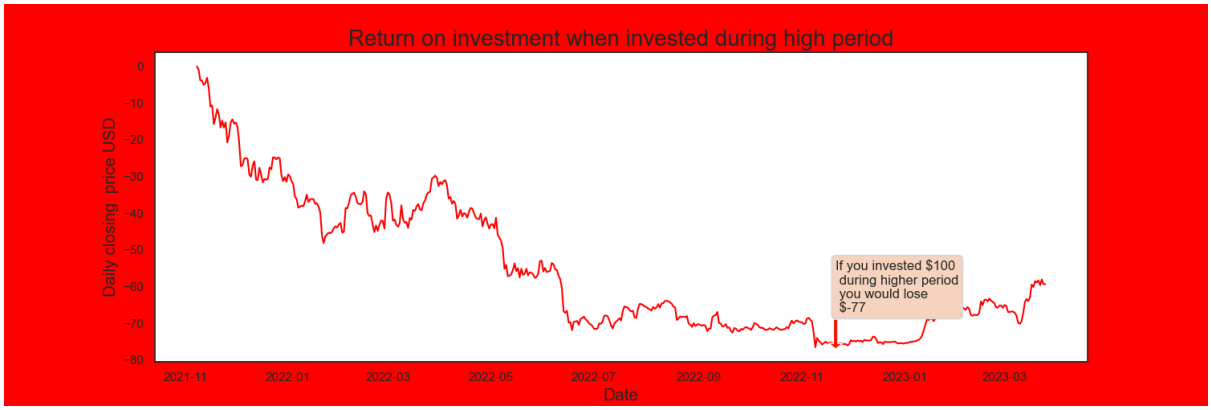


Fig: Investment Analysis

But if you have invested $100 when the price of bitcoin was all time high, you would lose -77 which is a loss of -177%.

2.1.3. Moving Average

Moving Average is a technique that helps to identify the trends and patterns of a Bitcoin price data over a certain period by calculating the average of a set of data points within specified time and is updated with each new data point. When the price of Bitcoin is above the moving average it might be a signal to buy, whereas if the price goes below the moving average it might indicate to sell.

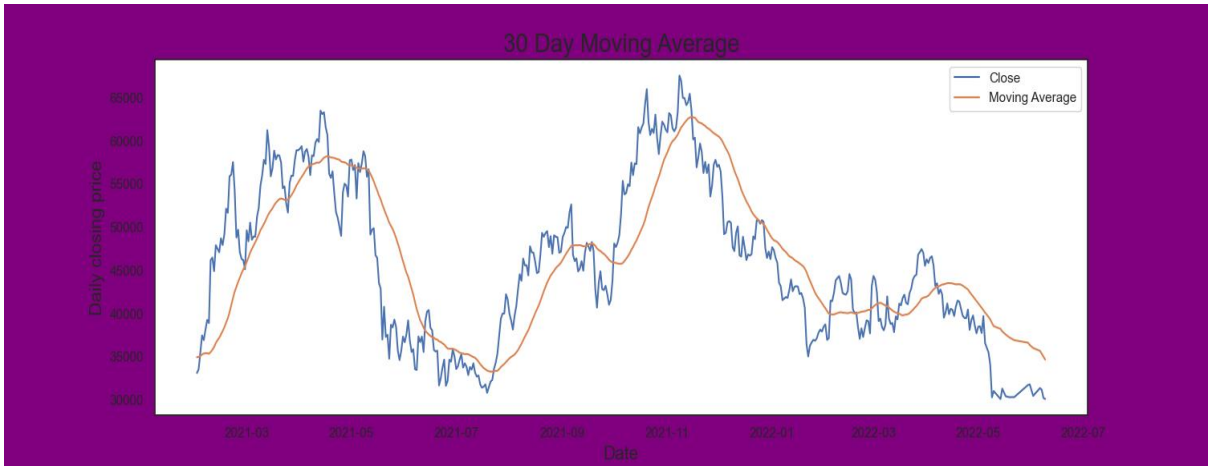


Fig: Moving average

2.2. Modelling

2.3. ARIMA model

ARIMA is short for Auto Regressive Integrated Moving Average. This model is best used when we have time series. It is based on its own past values, that is, own lags and the lagged forecast errors, so it can be used to forecast future values. It has three components:

AR- Autoregressive dependent relationship between observation and some lagged observations(p)

I- Integrated used for differencing raw observation subtracting one observation from another from previous timestamps to make time series stationary(d)

MA- Moving Average dependency between an observation and residual errors from a moving average model applied to lagged observations It takes account in past values to predict future values(q)

Lagging a time series means to shift its values forward one or more steps at a time, or equivalently, to shift the times in its index backward one or more steps.

The main assumption for ARIMA model is that time series should be stationary. We can check the stationarity by Augmented Dickey Fuller (ADF) test.

Stationary time series has statistical properties or moments (e.g., mean and variance) that do not vary in time.

2.3.1. Stationarity Test

We can see how time series can be stationary by ADF test. ADF tries to prove null hypothesis wrong.

Null hypothesis: Time series is not stationary (p-value>0.05)

Alternate hypothesis: Time series is stationary.

Results:

ADF Statistic: -1.549358057144826

p-value: 0.5089688390430713

Since p-value is more than 0.05, we can infer that it's not stationarity therefore we need to find the order of differencing.

Differencing

The purpose of differencing is to make the time series stationary. The order of differencing can be identified by visual inspection of time series plot, the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function). The PACF plot shows a sharp cut off after certain lag and ACF plot shows a slow decay.

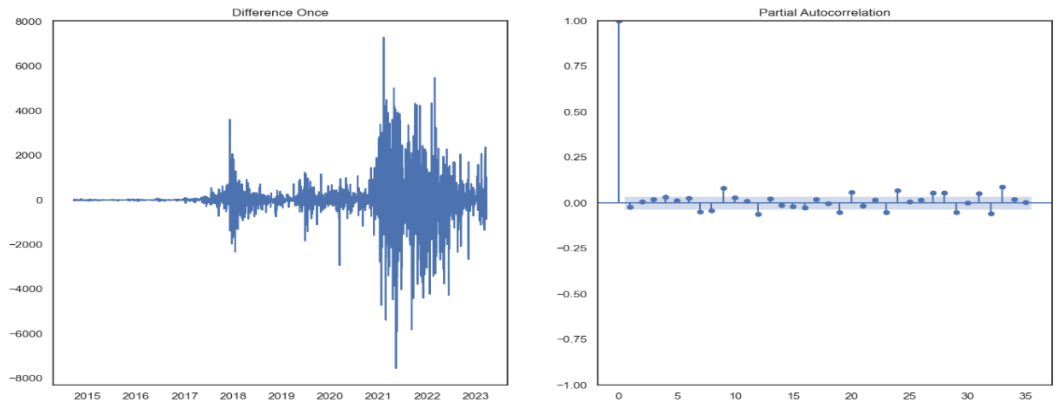


Fig: Differencing once

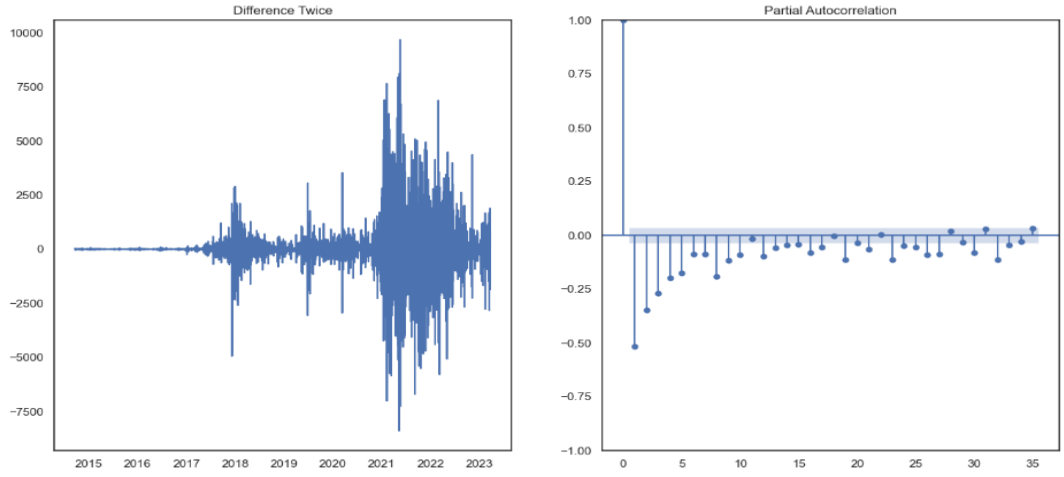


Fig: Differencing Twice

As we can see when we difference twice, the lags move to the extreme negative which suggests that it has been over differenced. The first figure shows the value for d might be 1 or 4.

Alternatively, the value for d can be found using stats models.

2.3.2. AR term(p)

AR term can be found by inspecting PACF plot which shows the corelation between series and its lag.

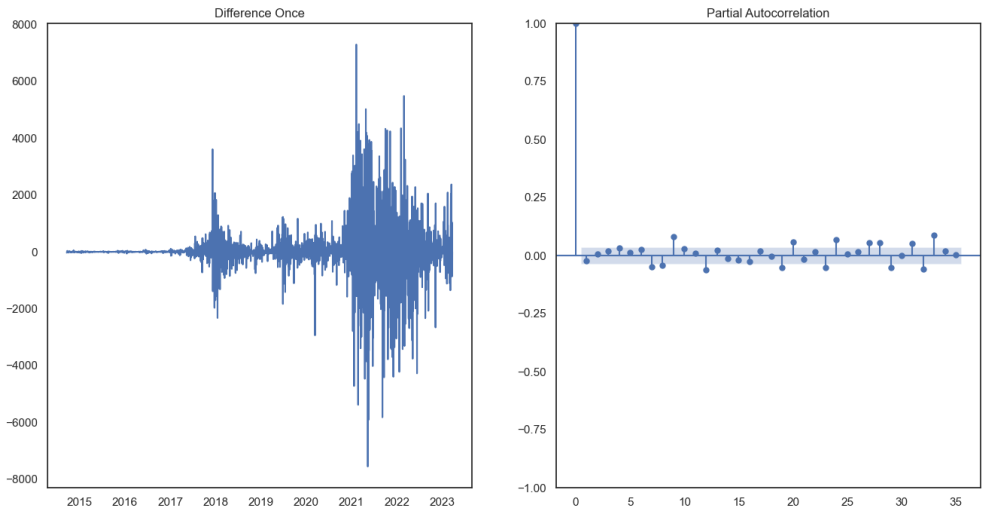


Fig: Partial Autocorrelation

The value for p might be 2 or 4 by inspecting the partial autocorrelation plot.

2.3.3. MA term(q)

MA term can be found by ACF plot which looked after the error of the lagged forecast.

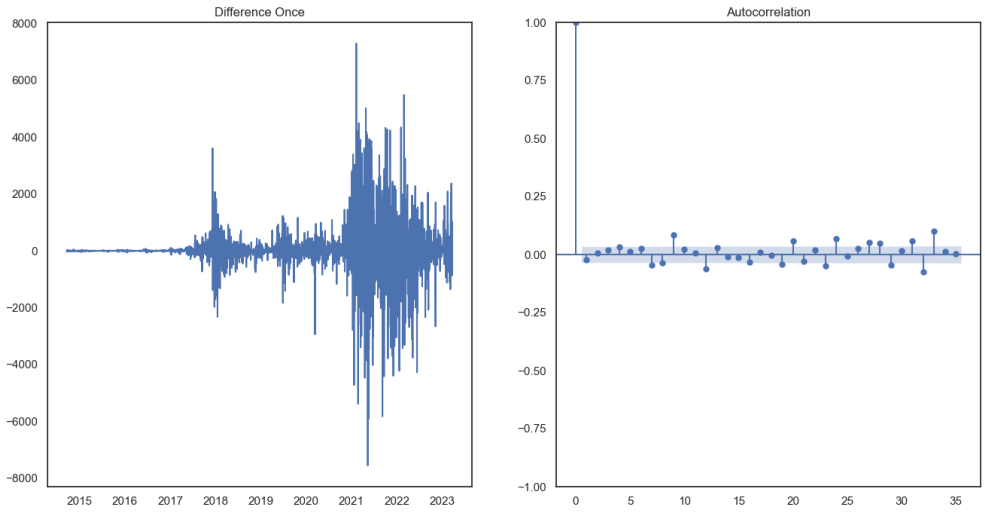


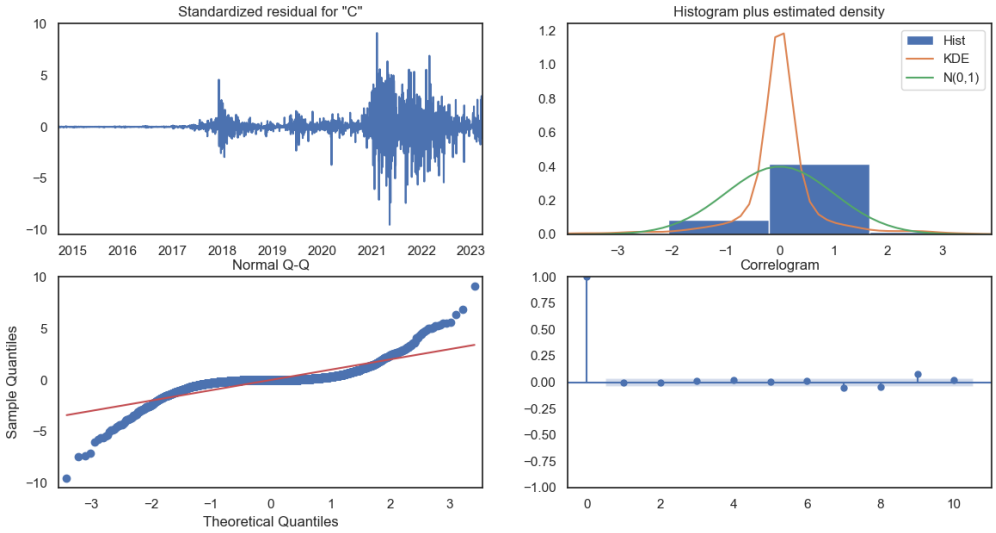
Fig: Autocorrelation

The value of q might be 2 or 3 by inspecting the autocorrelation plot.

2.3.4. Building the model

When p, d, and q are found the model is fitted with value of them as 2,1,2. Different values for p, d and q were evaluated, but the optimal results were not obtained.

Residuals



Top left: The residual errors seem to fluctuate around a mean of zero and have a uniform variance.

Top Right: The density plot suggests normal distribution with mean zero.

Bottom left: All the dots should fall perfectly in line with the red line. Any significant deviations would imply the distribution is skewed.

Bottom Right: The Correlogram, aka, ACF plot shows the residual errors are not autocorrelated. Any autocorrelation would imply that there is some pattern in the residual errors which are not explained in the model.

Overall, it seems to be a good fit. Let's predict.

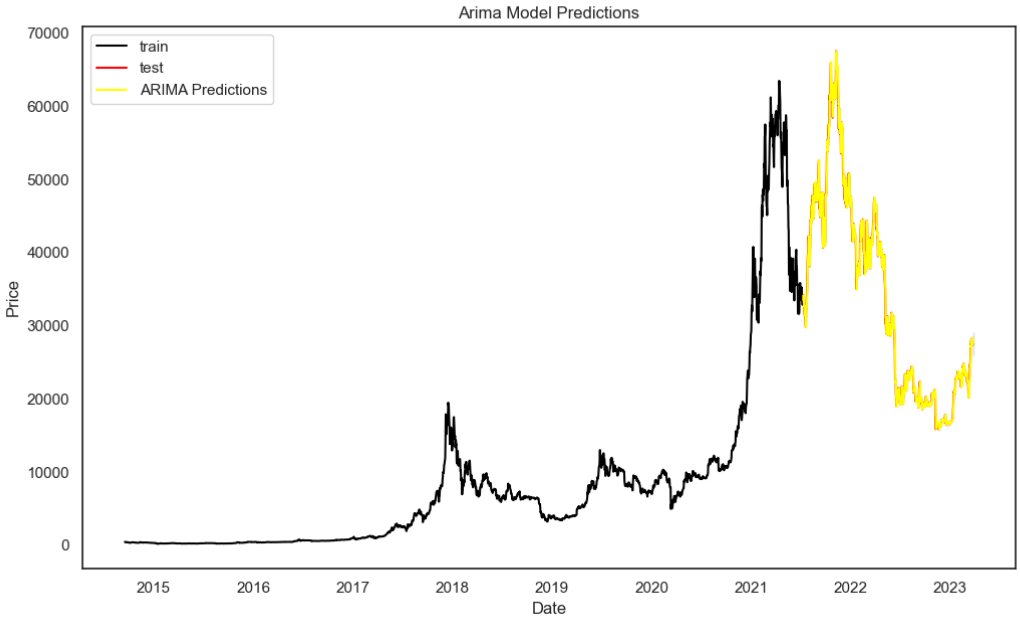


Fig: ARIMA predictions

The root mean square (RMSE) is 1214.39, which implies that average difference between the actual and predicted is 1214.39 which is not bad considering the upper range closing price of Bitcoin. Also, the mean absolute percentage error (MAPE) is 2.34, which implies that the predicted price is off by 2.34% which is great.

2.3.5. Forecast

The figure below shows the next 60 days forecasted price of bitcoin.

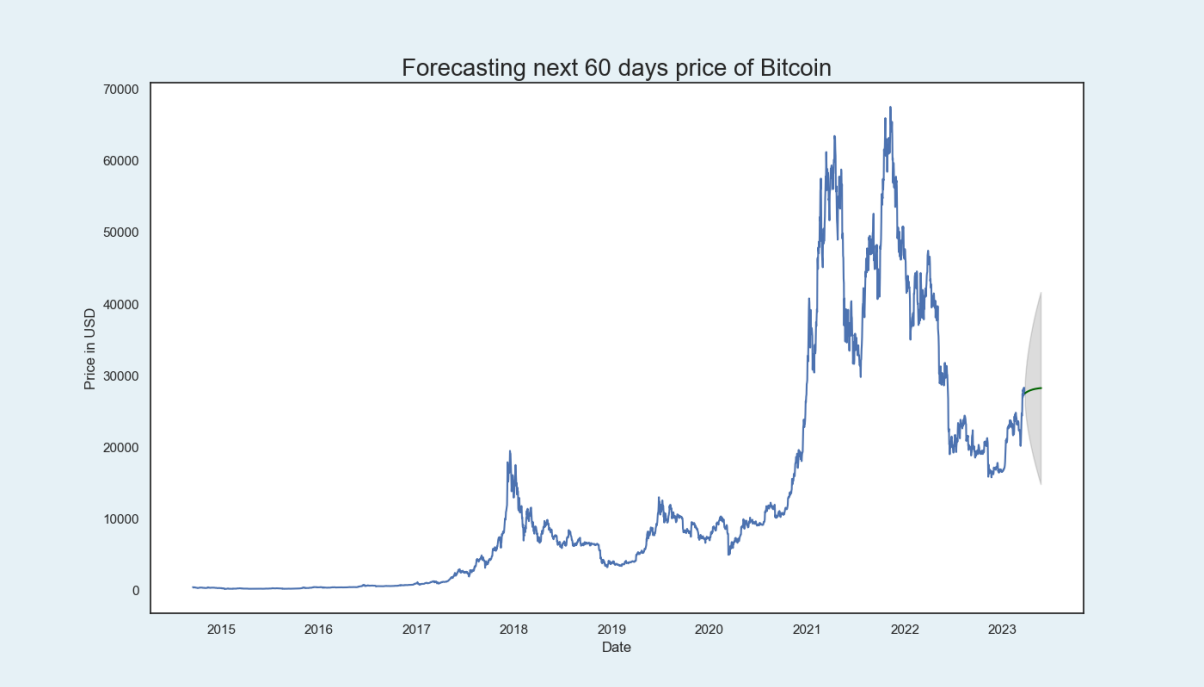


Fig: Forecasting 60 days ahead

The above figure shows that, in the next 60 days, the price of bitcoin can reach as high as USD 41,000 or go as low as USD 14,000, with 5% confidence interval.

2.4. LSTM model

LSTM, short for Long Short-Term Memory, is a type of recurrent neural network (RNN) which can learn long-term dependencies in a sequential data as timeseries.

The model has been built with different layers with dropout layer and dense output layer for predicting the price of bitcoin. LSTM model has been trained with training data which is 80% of the whole data, and the performance has been evaluated with the remaining 20% data also known as test data. Different metrics such as Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) have been used to evaluate the performance of the model.

When the model has been trained and evaluated by LSTM model, predictions have been done and it can be deployed for real-time predictions.

2.4.1. Predictions

Fig: LSTM predictions

The root mean square (RMSE) is 437.93, which implies that average price difference between the actual and predicted is 437.93 which is quite good, considering the upper range closing price of Bitcoin. Also, the mean absolute percentage error (MAPE) is 3.52, which implies that the predicted price is off by 3.52% which is good.

2.4.2. Actual Vs Predicted

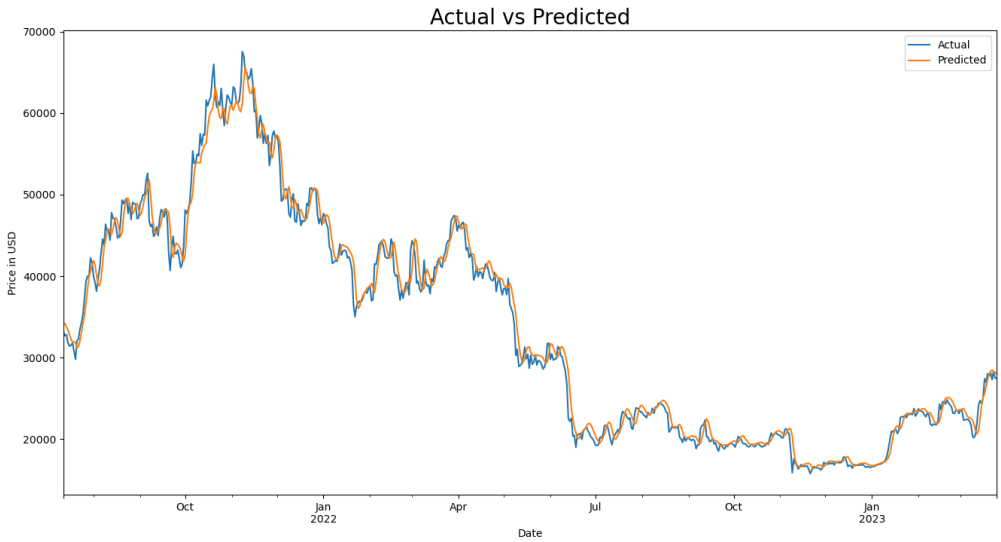


Fig: Actual vs Predicted

The figure above shows the actual vs predicted price of Bitcoin from LSTM model. It shows that the model is predicting accurately, and it suggests that model is a good fit for the data. The predicted values closely align with the observed values which indicates that the model is capturing the underlying patterns and trends in the data.

3.0. Outcomes

EDA and correlation process helped to identify highly correlated features and then later removed them because of multicollinearity. LSTM model gave better results than ARIMA model. The process confirms the possibility to predict the price of bitcoin with high accuracy.

4.0. Implementation

The model predicted the price of Bitcoin with high accuracy, and it can be implemented by the investors if they want to predict the price. However, more features like news can be added to predict because the price can be hugely affected by different news around the world which can be analyzed by sentiment analysis.

5.0. Data answer

After conducting the analysis on historical data of Bitcoin prices, the trends and patterns of price could be identified as a buying or selling opportunities based on historical price patterns.

6.0. Business answer

The business question was answered satisfactorily by developing predictive models which forecasted the price of Bitcoin with high accuracy.

7.0. Conclusion and Future Work

Bitcoin, the most popular cryptocurrency, has a huge impact on the economy and avoids financial institutions. The main motive for this project is to forecast the price of Bitcoin with accuracy using various models and minimize risks so that investors can take advantage from this. Two models were implemented for this project ARIMA and LSTM. Among those two LSTM performed well to predict the price of Bitcoin. The Root Mean Squared Error (RMSE) for LSTM is 437.93, which is comparatively low as compared to ARIMA model. This indicates the LSTM model's prediction is closer to actual values. However, the MAPE for ARIMA model is 2.34 which is lower than LSTM 3.52, which suggests that LSTM models' percentage error is slightly higher. However, the difference in MAPE values is not significant, and lower RMSE for LSTM models suggests that it is performing better overall in terms of accuracy.

However, the model cannot be fully implemented at this stage because prediction done by looking at only closing price might not give us better results. The price of Bitcoin is hugely affected by supply and demand, market sentiment, regulatory changes etc. Therefore, for the next stage analyzing the market sentiment will help us to predict the price more accurately because positive news such as regulations clarity, institutional adoptions etc. can push the price up whereas negative sentiment such as hacks or regulatory crackdowns will push the price down.

8.0. References

1. Fry, J., & Cheah, E.-T. (2016). *Negative bubbles and shocks in cryptocurrency markets. International Review of Financial Analysis*,47,343-352.
2. Dwyer, G.P. (2015). *The economics of Bitcoin and similar private digital currencies. Journal of Financial Stability*, 17, 81-91.
3. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
4. https://www.kaggle.com/code/meetnagadia/bitcoin-price-prediction-using-lstm