Analyzing Bitcoin Tweets and predicting its price movements

Project Milestone: Project Report

Group 24
Abhishek Hemantkumar Taware
Veer Pradyumna Karadia

857-245-8202
857-4379-678

taware.ab@northeastern.edu
karadia.v@northeastern.edu

Percentage of efforts contributed by Student 1: 50%
Percentage of efforts contributed by Student 2: 50%

Signature of Student 1: Abhishek Hemantkumar Taware
Signature of Student 2: Veer Pradyumna Karadia

Submission Date: 04/25/2022

Problem Setting

In today's world, a topic of discussion that has quickly grasped everyone;s attention over the past few years is Cryptocurrencies. Of all the cryptocurrencies out there, Bitcoin is the largest cryptocurrency based on its Market Capitalization due to its limited supply and even limited bitcoins mined. One of the platforms that has grown with increase in Topic Discussions is Twitter; wherein you can post microblogs also known as Tweets for the world to see your opinion and interact with you. Twitter has become a popular site for Bitcoin Aficionados to post their opinions about the Cryptocurrency. Using the datasets at hand, we plan on performing different sets of analysis on the datasets to try and draw some relations between the price movements of Bitcoin and Tweets that surface on Twitter.

Problem Definition

Knowing that predicting the price of cryptocurrency is not an easy task due to the sheer volatility in its price movements, we intend to check if there's any correlation between the price of bitcoin and number of tweets occurring at any given point of time within the available dataset. Along with this, using different prediction models, we intend to check which model gives the price movements which almost replicates the actual price of bitcoin. We also intend on checking which are the keywords which are regularly used in these tweets to understand how these keywords relate to some other content / cryptocurrencies.

Data Sources

Bitcoin Tweets: Kaggle
Includes the Twitter database containing Tweets which referenced the hashtags btc and bitcoin in them starting February 2021.
Bitcoin Price: Yahoo Finance API / Yahoo Finance
Includes the Opening Price, Day's High, Day's Low, Closing Price and Volume of Bitcoins Traded since the availability of Data.

Data Description

The tweets dataset consists of over 2 Million Rows with columns such as number of followers a user has, number of accounts a user is following, the date of the tweet, its content, the hashtags used, platforms on which the tweet was made etc.
The price dataset contains of 5 main columns which are date and time of the change in price, Opening Price of the Bitcoin for that time period, its closing price, how high did it go for the time period and what was the lowest price at which it traded and number of transactions that were made for that day.

Description of Variables (Default)

Table 1: Bitcoin Tweets

User Name	Represents the Name of the User Account
User Location	Represents the approximate location from which the Tweet was made
User Description	Contains the Bio of the User
User Created	Corresponds to the Date on which the account was created
User Followers	Corresponds to the Number of Followers a User has
User Friends	Corresponds to the Number of Accounts the User is Following
User Favorites	The Number of Favorites a Account currently has
User Verified	Checks if the User is Verified or not
Date	Date on which the Tweet was made
Text	Content of the Tweet
Hashtags	Hashtags used in the Tweet
Source	Platform / Application from which the Tweet was made
Is a Retweet	Whether a Tweet is a retweet or not.

Table 2: Bitcoin Price

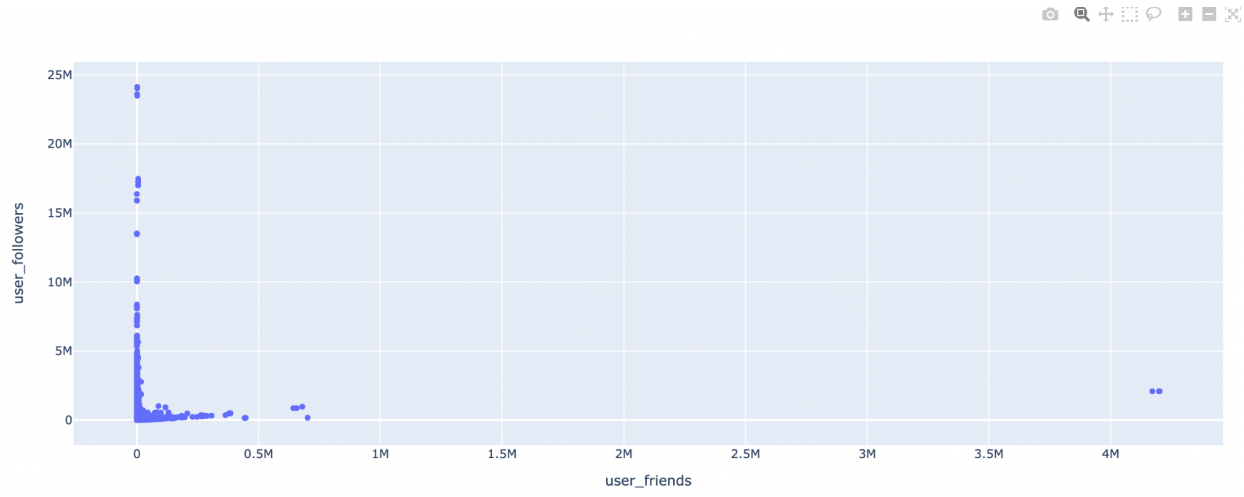
Open	Price at which Bitcoin Opened on that day
High	Highest Price which Bitcoin achieved on that day
Low	Bottom made by Bitcoin on that day
Close	Price at which Bitcoin closed on that day
Date	Date on which the data was captured
Volume	Number of Bitcoins traded / Trades made on that day
Adjusted Volume	Corresponds to corrections made in the Volume to account for losses if any

Data Exploration

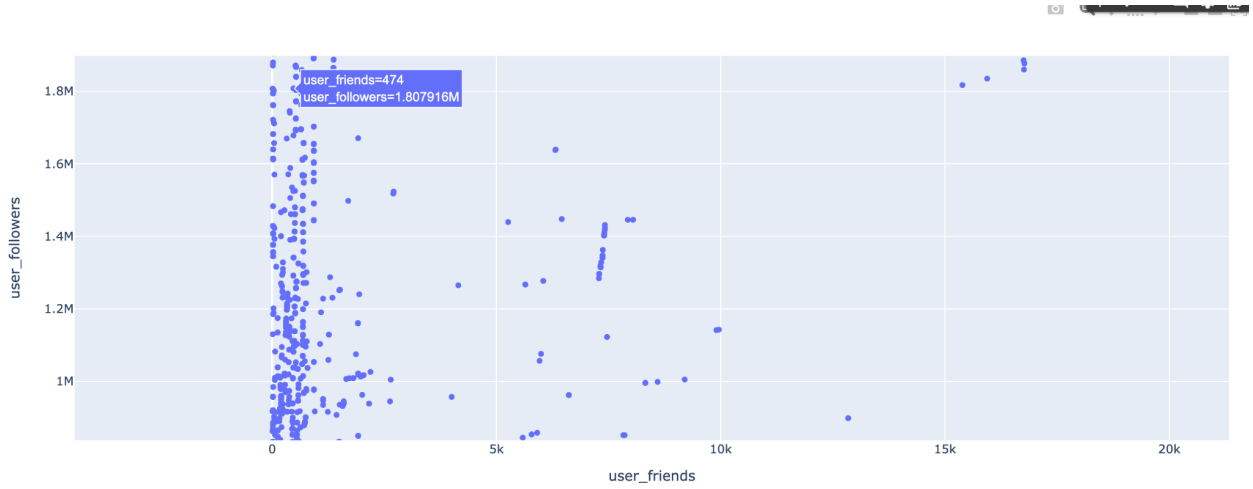
Since we have 2 different databases which correspond to different types of values, we have analyzed both of them separately. As the data within both the datasets was too big to be visually clear, we used the Plotly package which offers the functionality of Zooming into the data and having better visual clarity along with the hover function.

Price Trend of Bitcoin over the Years (since inception of data)

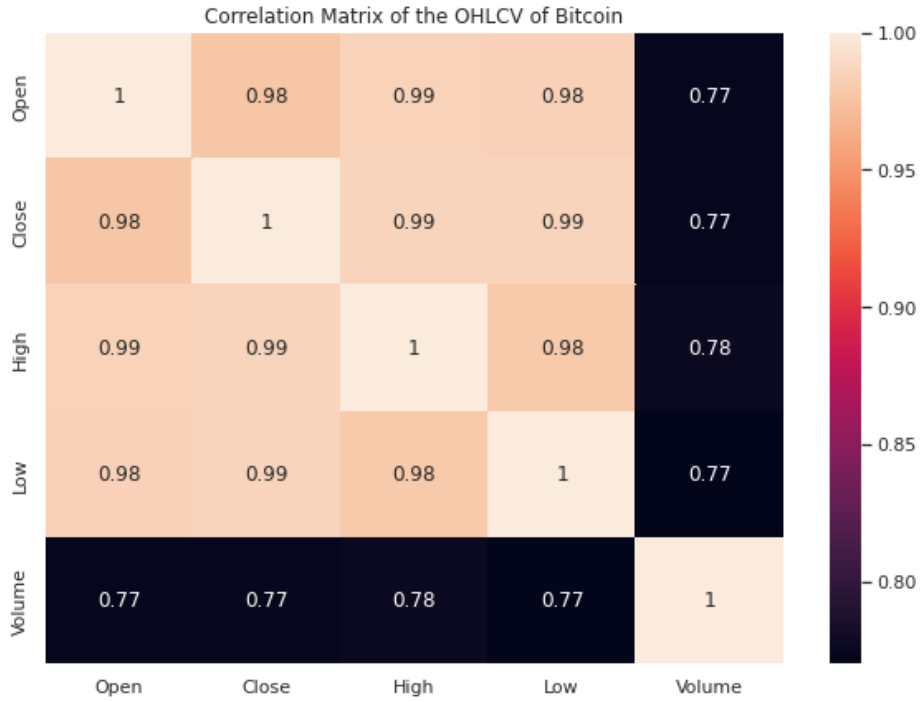
Scatter Plot of Followers and Friends



As we can see in the above graph, there are a few outliers towards the bottom right of the graph which show that they are following over 4 following accounts and have a follower count of less than 3 Million accounts which might indicate that these might be spam accounts. Whereas the left hand side of the graph indicates a better visibility of the Scatter Plot of Users as shown in the image below.



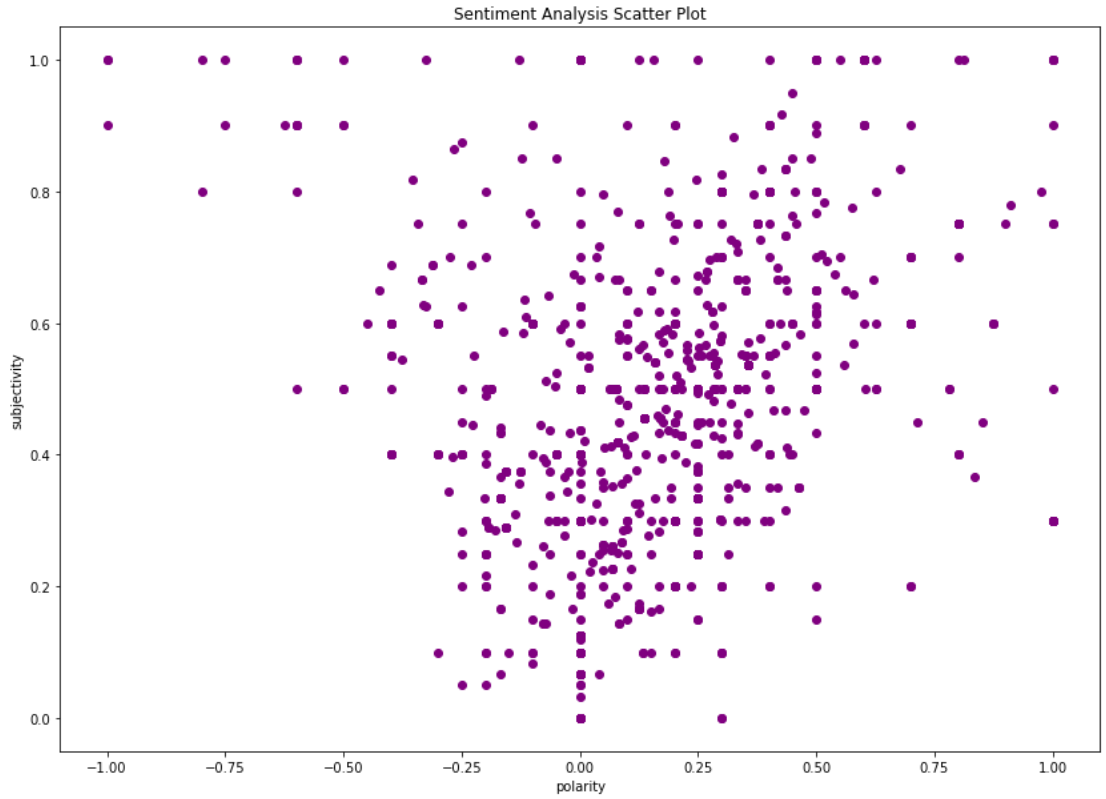
Correlation Matrix of OHLCV of Bitcoin



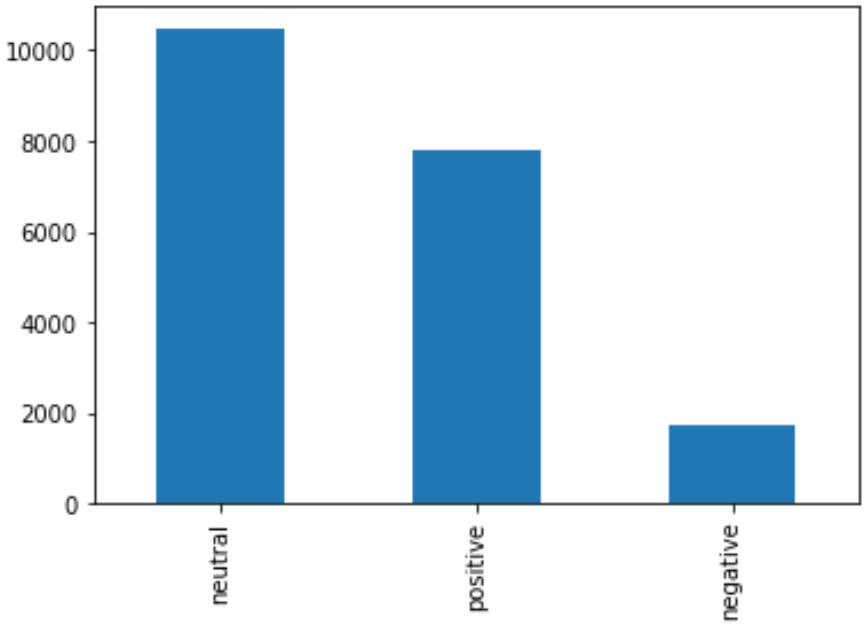
We wanted to find out if there's any strong correlation between any of the following parameters and most of them have a strong correlation with each other except for volume. In the stock market, where Volume is considered to be a crucial driver in the movement of stock prices, the inverse can be seen in the case of Bitcoins wherein Volume has a slightly weak correlation to the price of Bitcoin.

Sentiment Analysis (Sample Size: 20000)

Subjectivity & Polarity Correlation Scatter Plot

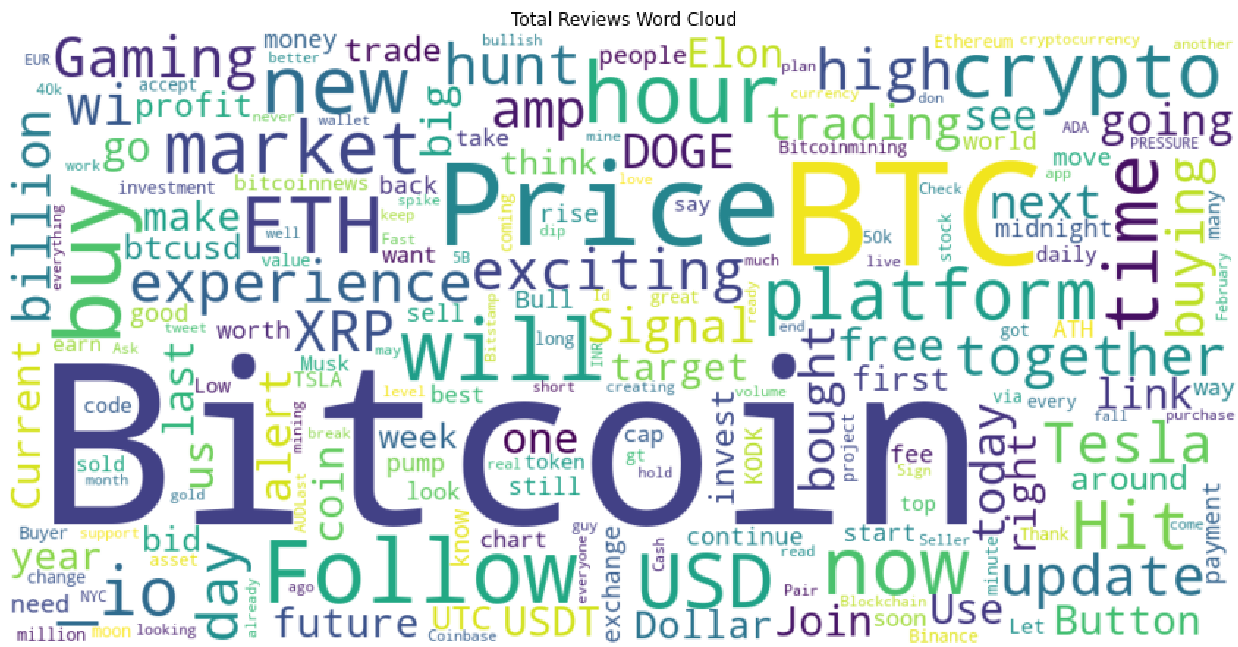


Sentiment of Tweets



Based on the Chart above, we can see that more than half of the Tweets from the Sample Set have a Neutral Sentimental Content in their Tweets followed by around 7500 Tweets being Positive and the Rest being Negative.

<u>Word Cloud of Most Frequently Used Words</u>



Data Mining Tasks

As we had a very large dataset at hand, we decided to divide the data frames into sub data frames to include only the necessary columns as and when needed for each task be it visualization or implementing a model.

First of all going on with the Tweets Database, the cleaning and processing part involved mutating the dates to Month & Year, finding out the NA Values and replacing / removing them, dropping columns which contained NA Values in the User_Name column. Also, with regards to cleaning the Tweets for the Word Cloud, we had to remove the Stop Words.

Whereas for the Pricing Database, since the data was 100% clean as it was from an official source, we only had to drop the column Adjusted Volume as it would be the same as Volume (only specific to cryptocurrencies) and mutate the date column to Month and Year for better clarity and use.

Cleaning of the Tweets Dataframe

```
Tweets.isna().sum()
```

```
user_name               31
user_location       1133881
user_description     280966
user_created              0
user_followers            0
user_friends              0
user_favourites           0
user_verified             0
date                      0
text                      0
hashtags              17009
source                 3566
is_retweet              252
Year                      0
Month                     0
```

Replacing all the nan values with Not Available, No Hashtags etc for better clarity

```
[ ]  Tweets["user_location"].fillna("Not Available", inplace = True)
     Tweets["hashtags"].fillna("No Hashtags", inplace = True)
     Tweets["source"].fillna("Not Available", inplace = True)
     Tweets["user_description"].fillna("Not Available", inplace = True)
     Tweets["user_followers"].fillna(Tweets['user_followers'].mean, inplace = True)
     Tweets["user_friends"].fillna(Tweets['user_friends'].mean, inplace = True)
     Tweets["user_favourites"].fillna(Tweets['user_favourites'].mean, inplace = True)

     #Here we are not cleaning the data for user_created,date as they are Datetime fields
     #Same goes for is_retweet as we will be using the column during visualization
```

Drop rows with nan values based on user_name column

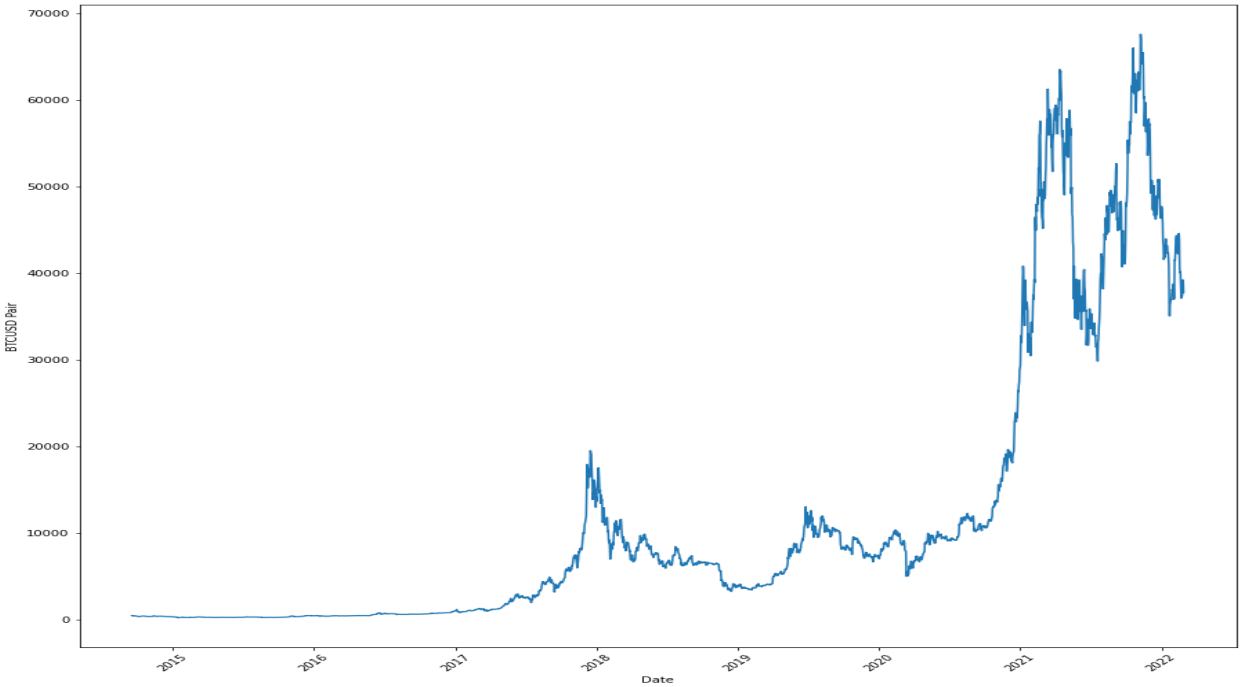
```
[ ]  Tweets = Tweets.dropna(subset=['user_name'])
```

Data Mining Models

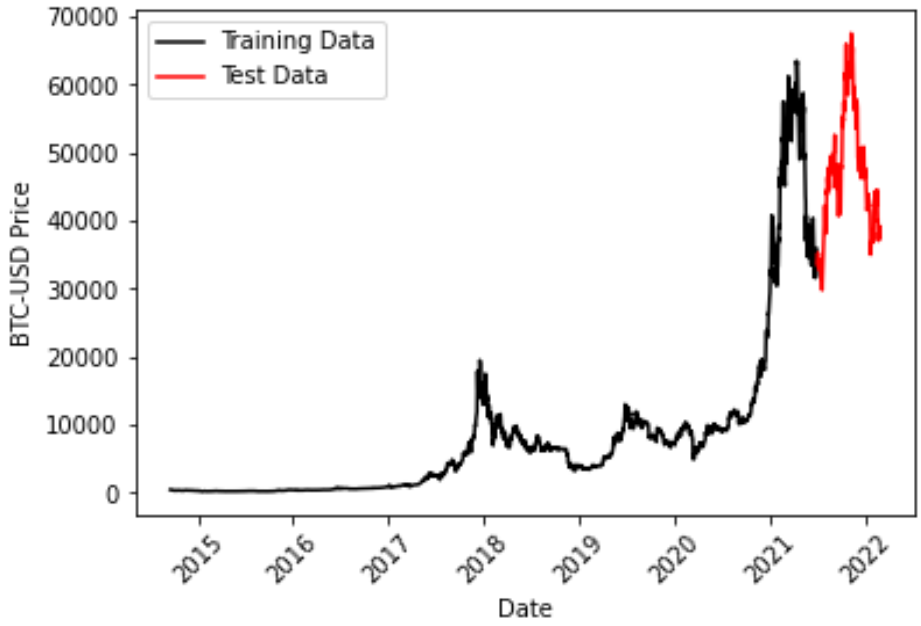
For Data Mining, we stuck to the Bitcoin price dataset which contained the Date and OHLCV Data of Bitcoin's Daily Price Movements since 2014. For implementation purposes, we decided to go with 2 established models which are used frequently when it comes to times series data: ARIMA & LSTM; and we decided to go with 1 comparatively newer model i.e Prophet Model by Facebook. All of these models are Time Series specific and are used frequently when it comes to Time Series data. For our purpose, we split the data into the ratio of 65:35 for Training and Testing Respectively.

ARIMA Range

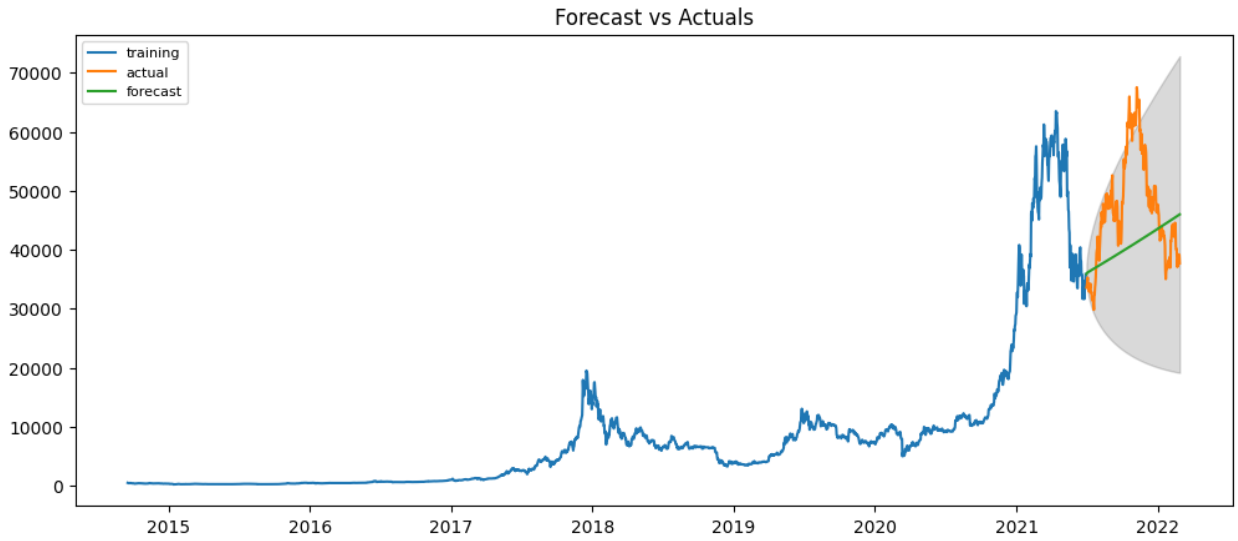
Actual Price Movements of Bitcoin



Training & Testing Range



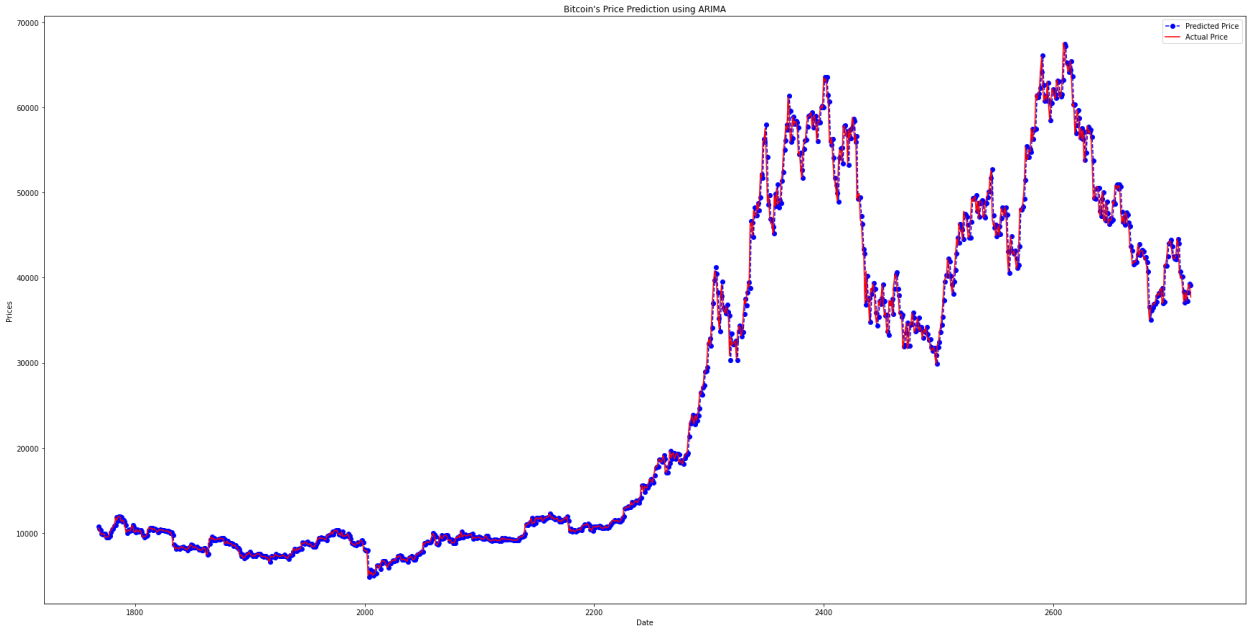
Range Predicted at Confidence Level 99



While showcasing the movement range of Bitcoin in both the directions, it can be seen that at 99% Confidence Interval, the Price tends to stay inside the trading zone for most of the part except for a minor outburst. To test the model a bit differently, we used 75% linear training data and 25% linear test data.

ARIMA Fit

High Resolution Image



The above ARIMA plot takes into consideration the Autoregressive Integrated Moving Average wherein we provided the parameters for Number of Lags (p),Degree of differencing (d), Order of Moving Average (q) as 4, 1, 0 which gave almost a perfect fit for the data at hand. The data was again split in the same ratio i.e 65% for Training and 35% for Testing.

Trendlines with 6 & 15 Day Moving Averages

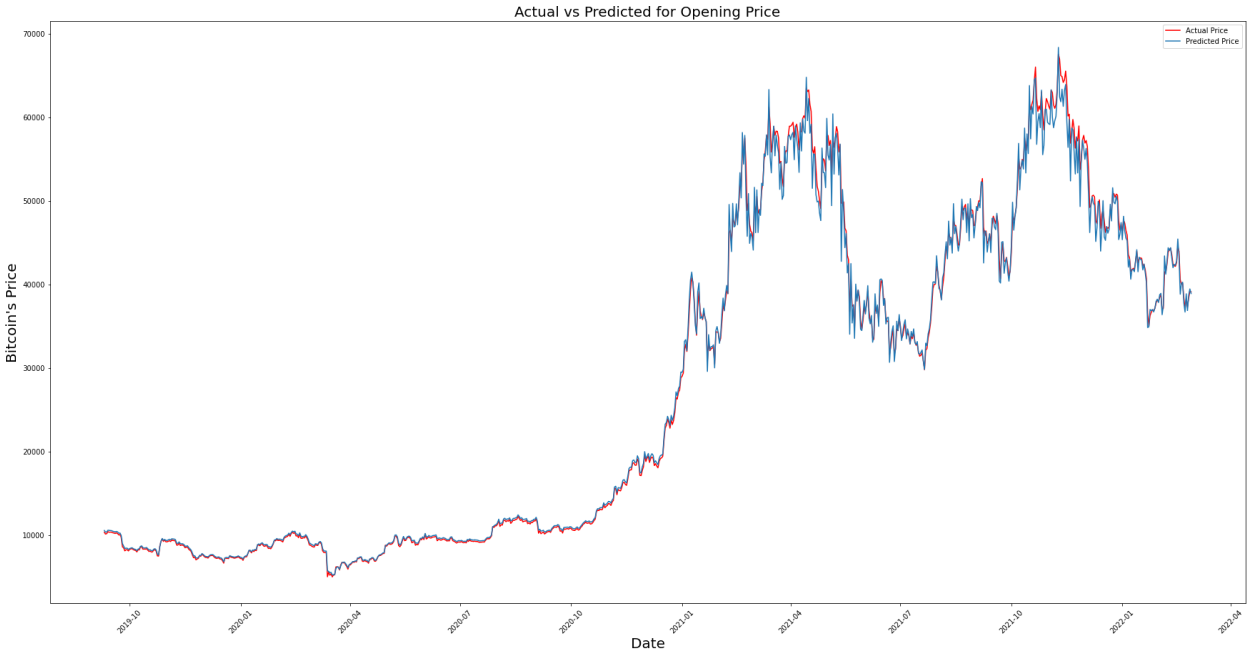
High Resolution Image

In Technical Analysis, Moving Averages are considered as Indicators based on which trends of different durations can be predicted. Normally these Moving Averages are used in conjunction with other indicators / with multiple Moving Averages. 2 of the most common pairs used are 6, 15 (for Short Term) and 21, 63 (For Medium Term). Without proving anything conclusively, we can say that at times, whenever the price of Bitcoin closes above the 2 Moving Averages, the price of bitcoin can be seen to go higher. The case is vice versa when the price closes below the 2 Moving Averages.

LSTM Model

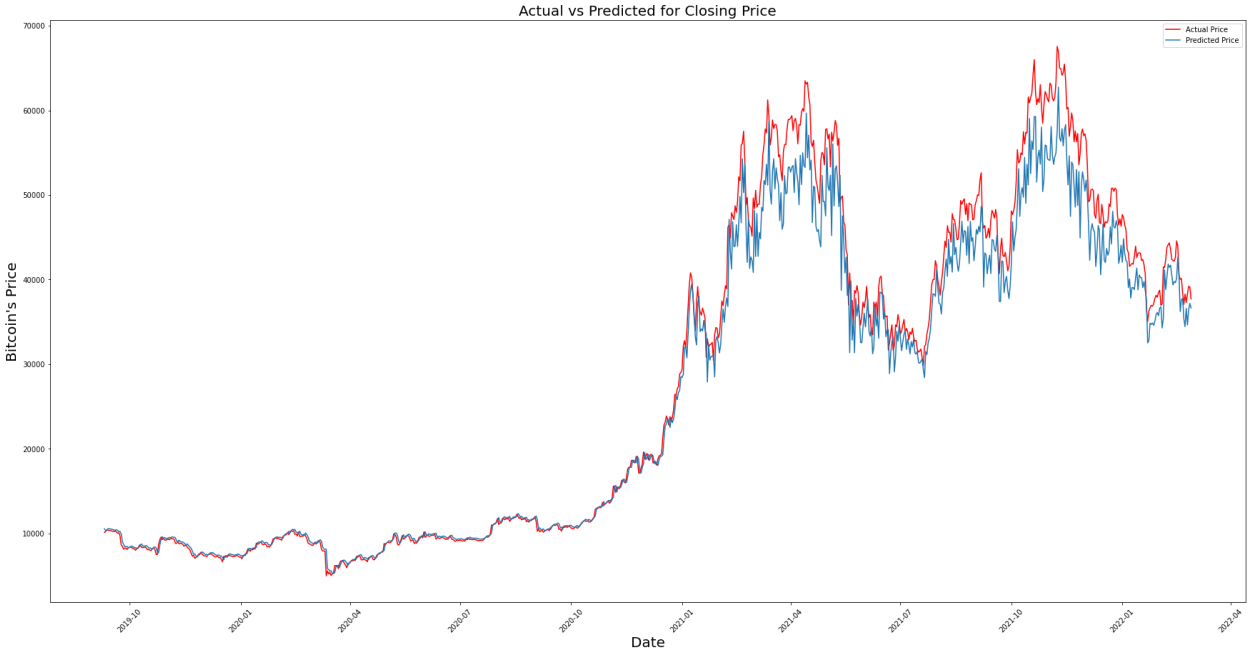
Implementation on Opening Price

High Resolution Image



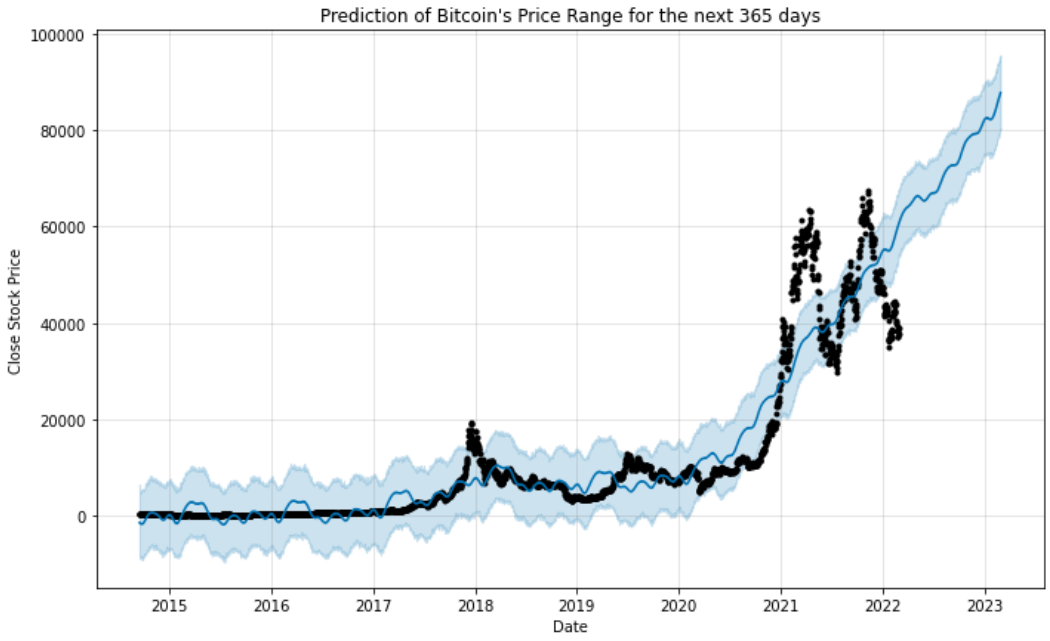
Implementation on Closing Price

High Resolution Image



LSTM is considered as one of the most popular Machine Learning Models when it comes to Time Series Data. For this use case, we split the data as 65% for the training data and 35% for the testing / validation data. Based on the charts above, we can conclusively see that the Model was able to perform up until the start of 2021 but since then, the predicted price was almost always slightly below the actual price which was conclusively seen in the Closing Price Chart.

Prophet Model



Prophet Model is developed by Facebook which acts as a procedure to forecast time series data based on additive models where non-linear trends are fit on a daily basis. It is new as compared to other Machine Learning Models. The black dots indicate the training data and the blue translucent area is the confidence interval. As per this model, the price of Bitcoin should reach around 88000. It seems unlikely since the model considers outburst in data as well, it is like a wait and watch to see if the prediction is accurate or not.

Performance Evaluation

Since the dataset that we are using consists of Time Series Data, creating Confusion Matrix and plotting the ROC Curve etc, were not possible. Even if we try to calculate the accuracy, it will come out to be very low as it would require us to match the exact values of Predicted vs Actual which in result would most probably be 0%. Due to this, we used the Standard Performance Metrics Parameters for the 3 Models that we have implemented.

Long Short-term Memory Model

Parameter	Opening Price	Closing Price
Forecast Error	137	1684
Bias	137	1684
Mean Absolute Error	578	1937
Mean Squared Error	1025314	10002410
Root Mean Squared Error	1012	3162
R2 Score	0.997	0.973

As we can see over here, we have implemented LSTM on the Opening Price as well as the Closing Price. Bitcoin being traded 24/7, technically there are no opening and closing prices as such. The Opening Price is treated as 12:00:00 AM and Closing Price is treated as 11:59:59 PM. Even though logically there shouldn't have been any difference between the Two Prices, we can see that since the day changes, there is a massive difference in the Performance Metrics of the two prices. With a Forecast Error of just 137 as compared to 1684, we can see that the Opening Price Index is a better tracker for predicting the price of Bitcoin when it comes to LSTM. Of all the 3 Models used, this would be ranked as the 2nd best model to track / predict the prices.

ARIMA Model

Parameter	Closing Price
Forecast Error	16
Bias	16
Mean Absolute Error	735
Mean Squared Error	538471
Root Mean Squared Error	1262
R2 Score	0.995

ARIMA being the preferred model for Time Series data, it was expected to perform well but the results were better than expected. Even though the R2 Score between the Opening Price (in LSTM) and Closing Price (in ARIMA) is almost the same, 1 parameter wherein there is a massive difference is the forecast error. With a forecast error of just $16 as compared to $137 in LSTM's Opening Price Prediction, it is easy to say that ARIMA is better than the other 2 Models without a doubt. This accuracy can be cross checked when we have a look at the plot made for ARIMA.

Prophet Model

Parameter	Closing Price
Mean Absolute Error	849
Mean Squared Error	7594413
Root Mean Squared Error	2755
R2 Score	0.953

Prophet being the newest of Models out there, it was used just for an experiment to see how it performs given the volatility of Bitcoin. From what we can see based on the forecast plot, the model performs well when the price of the coin is stable. But when outbursts occur, the model cannot handle them. If we change the seasonality of the

model to track the prices monthly / weekly or quarterly, maybe the model would still perform better but as compared to the other 2, in this particular use case, it gives us the worst fit with an R2 Score of 0.95 and the Mean Absolute Error is high as compared to the other 2.

Project Results

Model	Forecast Error	Bias	MAE	MSE	RMSE	R^2 Score
LSTM Opening	137	137	578	1025314	1012	0.997
LSTM Closing	1684	1684	1937	10002410	3162	0.973
ARIMA	16	16	735	538471	1262	0.995
Prophet	-	-	849	7594413	2755	0.953

Impact of the Project Outcome

In this Project for Data Mining, our end goal was to identify which Time Series Model can be used to predict Price Movements with the highest of accuracies. Bitcoin being a very volatile trading instrument, we thought it would not be that easy to test the data points that we have at hand but Models such as LSTM and ARIMA gave some outstanding results when it came to the implementation of the Model. Since Bitcoin is comparatively new to be regularly traded, a lot of factors go in if we want to create a prediction based automatic trading algorithm for Bitcoin. But the fact that all the 3 Models give us good results means that if we put in more research in these models and start understanding how they can be improved, the scope for further improvement in Training these Models will improve drastically.