Rochester Institute of Technology
B. Thomas Golisano College
of
Computing and Information Sciences

Master of Science in Information Sciences and Technology

~ Project Proposal Approval Form ~

Student Name: Vidhi Bimal Shah

Project Title: Influence of Elon Musk's Tweets on the price of Bitcoin and Dogecoin and Price Prediction

Project Area(s):	□ Application Dev.	□ Database	□ Website Dev.
(√ primary area)	□ Game Design	□ HCI	□ eLearning
	□ Networking	□ Project Mngt.	□ Software Dev.
	□ Multimedia	□ System Admin.	□ Informatics
	□ Geospatial	√ Other Data Analytics	

~ MS Project Committee ~

Name	Signature	Date

Prof. Charles Border
Chair

Prof. John-Paul Takats
Committee Member

Approved: _____ Date: ____/____/____ □ electronic copy received

Influence of Elon Musk's Tweets on the price of Bitcoin and Dogecoin and Price Prediction

By: Vidhi Shah

Project submitted in partial fulfillment of the requirements forthe degree of Master of Science in **Information Sciences and Technologies**

Rochester Institute of Technology

B. Thomas Golisano College of Computing and Information Sciences

Department of Information Science and Technology

Approved: _____ Date: ____/____/____ ☐ electronic copy received

Table of Contents

Contents

Approved: _____ Date: ____/____/____ ☐ electronic copy received

Influence of Elon Musk's Tweets on the price of Bitcoin and Dogecoin and Price Prediction

Abstract

Cryptocurrency's popularity has risen astronomically and has given birth to a revolutionary new method of payment and investing. These blockchain-backed securities have succeeded in luring not only the Wall Street sharks but also the retail investor. This meteoric rise has caused the market capitalization of cryptocurrencies and blockchains to skyrocket. Bitcoin was the first leading cryptocurrency ever created, and its market cap has crossed 783 billion dollars. The sum of all cryptocurrency's general market cap is estimated to cross 1087.7 billion dollars by 2023. Despite these staggering numbers, one problem that the cryptocurrency market has been grappling with is high volatility. Like traditional markets, the crypto market is prone to volatility due to news developments and speculation fueling price swings. However, due to the liquidity shortage in the crypto market, this effect is exaggerated, and any small news or speculation can cause price swings. Elon musk is ranked among the wealthiest men and richest men across the globe. The biggest promoter of volatility in the cryptocurrency market is Elon Musk due to his 58 million followers on Twitter and his mystifying tweets regarding individual cryptocurrencies like Bitcoin and Dogecoin. Significantly higher than usual trading volumes have been observed following several of his tweets. The researchers suggest that social media activity can influence these movements; therefore, influence persons such as Elon Musk can significantly influence cryptocurrency. The project aims to examine related Elon Musk's activities on Twitter with respective impact on the cryptocurrency market. The project's goal is to examine Elon Musk's tweets have any influence on Bitcoin and Dogecoin and price prediction. To accomplish this, the project proposes various models such as Auto Regressive Model, Moving Average Model, and Auto Regressive Integrated Moving Average Model. Based on various metrics such as log-likelihood, Corrected Akaike Information Criterion and Bayesian Information Criterion, best model is chosen for forecasting the future.

Keywords: Twitter, Elon Musk, Bitcoin, Dogecoin, Auto Regressive Integrated Moving Average Model.

1. Introduction

Bitcoin's emergence has paved the way for a radical technological revolution in banking and investment markets. This decentralized electronic currency system can change how we exchange money for goods and services using peer-to-peer networks and complex cryptography solutions. There is no central administration like a bank or a financial institution managing this currency; instead, the transactions are verified by a distributed network of computers updating a distributed central ledger.

The ominous nature of web 2.0 has drastically changed how internet users are served data and communicate with each other. Web 2.0 is built on three core pillars of innovation social, cloud, and mobile. Several web 2.0 services like tweets, posts, blogs, chats, etc., are widely accepted as primary forms of communication, and a large amount of information is exchanged in this process. The data collected from different social media platforms collectively represents the current sentiment and serves as an indicator of thoughts and ideas around the world. Twitter is the most widespread and extensively used to share thoughts and ideas, including investment decisions concisely. The tweet data can be used to extract information about investment trends and what/who are the impacting factors on the prices of stocks or cryptocurrencies. Elon Musk, who has the 17th highest followers on the platform, is thought of as an influential personality, especially for stock prices like Tesla and cryptocurrencies like Bitcoin and Dogecoin. This paper aims to find any relation between his tweets and the costs of these two cryptocurrencies using his tweets and forecast the price of these cryptocurrencies.

Various uncertain factors such as political issues, economic issues that are impacted at the local or global level influence the global equities. Thus, it becomes essential to provide accurate price predictions, which becomes complicated. Moreover, the advent of communication networks such as Twitter, Instagram, Facebook, etc., has made the exchange of experiences and information simple. Twitter users are creating vast amounts of Bitcoin/Dogecoin-related tweet data every day. Thus, using Natural Language Processing, Time Series Analysis, Machine Learning, etc., the continuously generated real-time data from social networking sites can be helpful to research cryptocurrencies.

On January 29, 2021, Elon Musk changed the Twitter bio #bitcoin, which resulted in the rise of the Bitcoin price from about $32000 to over $38000 in just a matter of hours, increasing the market capitalization by $111 billion [2]. The relevance of his tweets for financial markets has already become noticeable in various contexts. Recently Elon Musk suggested that Bitcoin was almost being fully accepted. He further in one of his interviews, stated that he was late to the party and that he is a supporter of Bitcoin.

Further, he claims that his post related to cryptocurrency, more specifically about Dogecoin, are just but jokes. Although he suggests that these are only but jokes, his tweets directly impact the trends in the cryptography market. Based on this, this project aims to showcase Elon Musk's tweets impact on the cryptocurrency market developments.

The proposed project focuses on providing a methodology for price predictions for Bitcoin and Dogecoin and investigating the impact of Elon Musk's Twitter activity on the cryptocurrency market fluctuations. Also, sentiment analysis of social media tweets is beneficial in terms of market perspective and providing feedback.

Problem Statement:

The potential impact on the cryptocurrency market affects the currency and has a massive effect on the investors. There is always a sense of losing money that the investors have, and helping them get rid of this fear, will encourage them to invest, resulting in potential profits. Also, the Twitter sentiments of influential individuals tend to make the stock market more volatile. Hence having a methodology that predicts the price of the cryptocurrency and analyzing the tweets of a significant person on the price helps the investors make wise decisions.

Goals and Objectives:

The motive of this proposal is to present a methodology that predicts the price of cryptocurrencies, namely Bitcoin and Dogecoin, along with analyzing the impact of Elon Musk's Twitter activity affects the price. It is imperative to explore the social media activity of well-known and influential individuals. This methodology brings out and examines Elon Musk's tweets that are labeled with the name of these cryptocurrencies and explore remarkable features mapped with the coexisting prices to build prediction arcs to predict the prices soon.

Goal 1 – Understanding the personality of Elon Musk.

Objective 1: Download the tweets associated with the Twitter username elonmusk.

Objective 2: Create a line graph showing the tweet counts over the years.

Objective 3: Analyzing his popularity by understanding the likes, replies, and retweets.

Objective 4: Extracting the most frequent words used and representing in a word cloud.

Objective 5: Calculate the percentage of tweets with the words Bitcoin and Dogecoin, and sort the tweets based on the likes.

Goal 2 – Predicting the future price of Bitcoin.

Objective 1: Analyzing the bitcoin price and volume over time.

Objective 2: Determine if the time series data has a trend.

Objective 3: State the null and alternate hypothesis for Augmented Dicky Fuller Test.

Objective 4: Determine the p-value, critical value, and ADF Statistic.

Objective 5: Determine which ARIMA (Auto Regression Integrated Moving Average) model is appropriate to forecast the cost for the time-series data.

Objective 6: Split the data into training and testing datasets, apply the model created to predict price.

Goal 3 – Determine if there is any correlation between Elon Musk's tweets and the price of Bitcoin/Dogecoin.

Objective 1: Analyze the points in time when he tweeted about Bitcoin/Dogecoin.

2. Literature review

Exploring related work for this paper shows there are several precedents where we can see that the crypto and stock markets are vulnerable, and that vulnerability is predictable based on influential events.

In this paper, Satyabrata et al. [1] analyze the stock price movements by feeding the relevant event-related tweets. They have collectively used over 200,000 tweets spread over five days and the financial stock market data from yahoo finance. Most frequent keywords have been identified. Each keyword's usage trend is tracked to find a positive relationship between stock prices and the average sentiment score for specific heavily used keywords. They have successfully used Keyword and Sentiment analysis to find that tracking the usage trend of certain keywords shows that sentiments are an impactful factor for the financial market.

This paper presented by Lennart Arte [2] explains how Elon Musk's tweets, when shared with over 44.7 million followers on Twitter, tend to affect the short-term cryptocurrency market. Six separate events started up by one of Elon Musk's tweets when abnormally high trading volumes were observed between 2020 and 2021 have been selected for the analysis in this paper. Unusually high returns of up to 18.99% for Bitcoin and 17.31% for Dogecoin are observed at different points of time frames [2]. The report does a great job of constructing a case for proving that social media activity from influential individuals is an impactful and effective factor contributing to cryptocurrency price fluctuation.

Adebiyi et al. [3] have successfully and accurately predicted by building an autoregressive integrated moving average (ARIMA) model for time series prediction. The extensive process of creating the ARIMA model and training the model by collecting historical stock market data as a training data set from the Nigeria Stock Exchange (NSE) and New York Stock Exchange (NYSE) has been described in this paper. Conclusively, it is observed in this paper the ARIMA model built by the authors of this paper makes a strong case for itself in terms of short-term stock price prediction.

In this paper [4], Bollen et al. have tried to investigate the extent to which the aggregated average mood states calculated from large-scale Twitter data feeds are related to the value of the Dow Jones Industrial Average (DJIA) all over the extensive time interval. The authors have successfully done sentiment analysis using third-party software by analyzing many tweets to calculate the mood state for that period. They have tried to

accurately predict the future price and fluctuations of the stock market by analyzing six different commonly disregarded moods.

In this paper [4], Bollen et al. have tried to investigate the extent to which the aggregated average mood states calculated from large-scale Twitter data feeds are related to the value of the Dow Jones Industrial Average (DJIA) all over the extensive time interval. The authors have successfully done sentiment analysis using third-party software by analyzing many tweets to calculate the mood state for that period. They have tried to accurately predict the future price and fluctuations of the stock market by analyzing six different commonly disregarded moods.

The recent meteoric rise of Bitcoin price from around a dollar in 2010 to 18000 in 2017 has successfully lured in Dr. Fiaidhi and the team to employ the ARIMA model to predict Bitcoin price for sub-periods of the stretch [5]. They have used empirical data of Bitcoin prices from 2013 to 2019 to predict Bitcoin prices for 2020. They have tweaked the model parameters to configure an ARIMA model which has the lowest MSE (Mean Squared Error) of prediction.

3. Initial Data Exploration and Visualization

This section is divided into various sections explaining the overview of the dataset. Initially, it provides information about the dataset, attributes of the data followed by basic visualizations using bar charts. The next section describes the generation of word clouds and the distribution of reviews. Then, data cleaning and preprocessing of the data are explained which will be helpful in model building. Lastly, it includes a description of metadata and attributes selection from metadata.

3.1 Data Acquisition

The dataset used for this project is available on yahoo finance which has all cryptocurrencies listed. For this project, I have used the historical data of Bitcoin from 2012 to 2021 and Dogecoin from 2014 to 2021. It contains the Date, Open, High, Low, Close, Adjusted Volume, and Volume.

For tweets related to Elon Musk, I created an account as a Twitter developer and filled a form to request the permission. The approval process from Twitter took a couple of days and then they granted me API keys and Access tokens to access the API to retrieve tweets. This dataset consisted of date and time of the tweet, username, user_id, tweet, no. of replies, no. of retweets, no. of likes and some additional fields that are not being used for my analysis.

3.2 Dataset Description

Table [1] contains a detailed description of the cryptocurrency dataset used in this Capstone project.

Variable	Definition & Relevance
Date	Date of the record
Open	Price at which the stock began
Close	Price at which the stock closed

High	Maximum price on that day
Low	Minimum price on that day
Adjusted Close	Stock's closing price including other factors like dividends, stock splits and new stock offerings
Volume	No. of shared traded in a stock

Table [1] shows the description and relevance of each attribute in the dataset.

Table [2] contains a detailed description of the relevant fields for tweets dataset used in this Capstone project.

Variable	Definition & Relevance
Date, time	Date and time of the tweet
User_id	Elon Musk's tweeter user_id
User_ name	Elon Musk
tweet	The text of the tweet
replies_count	No. of replies on the tweet
retweets_count	No. of retweets on the tweet
likes_count	No. of likes on the tweet

Table [2] shows the description and relevance of each attribute in the dataset.

3.3 Dataset cleaning

The tweets had some special characters, that needed to be replaced by a space. The replace() method is used to replace the special characters by a space. Some tweets had hyperlinks, which had to be removed. Here the python's built-in package re is used to work with regular expression. The sub function is used to replace one or many strings. Also, there were several mentions and hashtags used in the text, which were substituted by a space. Various emojis and punctuation were used and were simply replaced by space. The cleaned data is then used for further processing and analysis.

3.4 Data Exploration and Analysis

a) Understanding Elon Musk's personality

For this, it is very important to understand how his tweets look, the progression over time and understanding the overall feel of how, what and when he tweets. This analysis helps in recognizing and analyzing the tweets that have the word Bitcoin and Dogecoin in them.

i) Tweet Count Evolution

Elon Musk barely used the Twitter platform a few times in a year during the beginning of his tweeting journey. From the below graph, 2015 was the breaking point when he started tweeting more. Everybody knew Musk before 2015 and got even more famous when he put out an autobiography of his personal life. In 2015, he announced Tesla's Powerwall battery. The frequency of his tweets over the years is shown in the **Figure [1]**. The tweet count evolution shows an upward trend over the years which is depicted in the below graph.

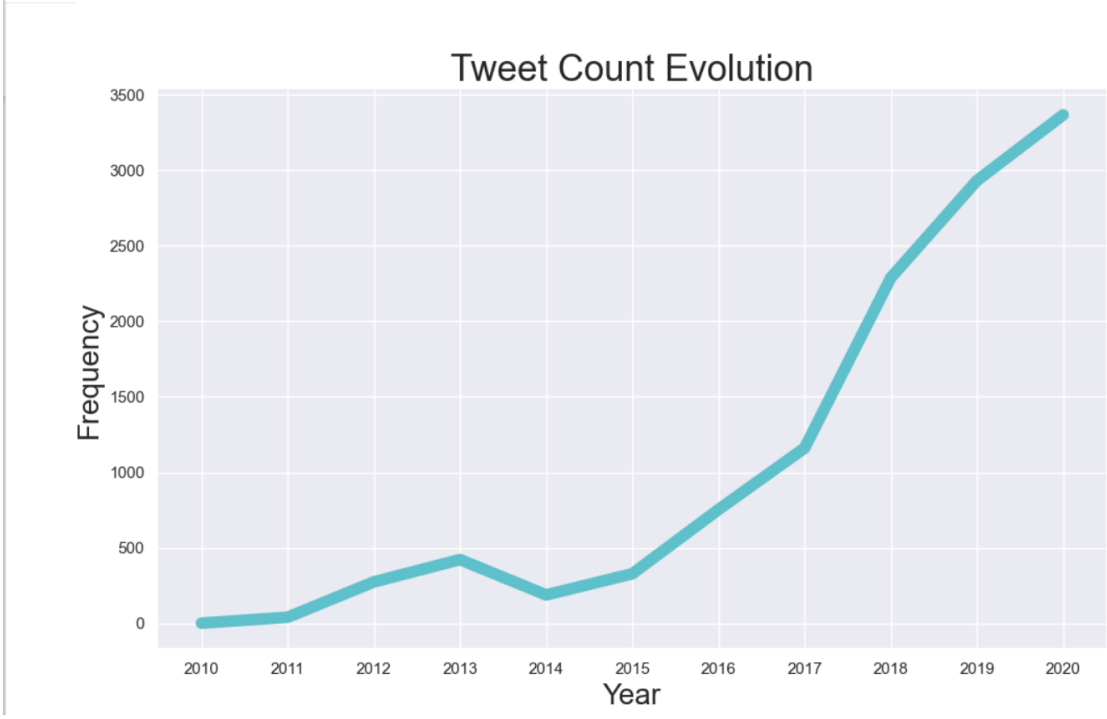


Figure [1] Tweet Count Evolution

ii) Popularity

Elon Musk's popularity grew as the number of tweets increased as shown in **Figure [2]**. The below graph shows gradual increase in the replies which shows he started communicating more with the larger audience. Also, the number of likes show an upward trend with its peak in 2020. However, there is a nominal change in the retweets he has done in his tweeting journey.

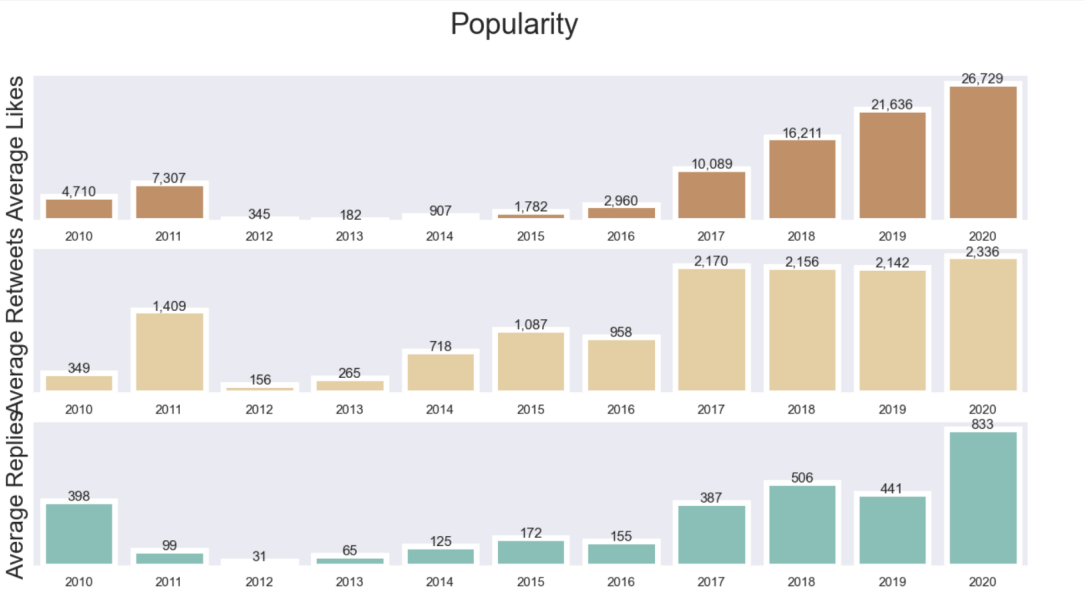


Figure [2] Popularity

iii) Most Frequent Words

Word Cloud is an image that is composed of words which is used in a particular text, where the frequency of the words is measured by the size and so is the importance of the word in the overall set. The more often the word appears in the data, it appears bigger and bolder in the word cloud. The visual representation of the frequency of the words is an advantage of the word cloud. The popularity of word cloud has increased since the human brain finds it appealing to visualize text.

The word cloud is generated based on the tweets in the dataset and the most frequently occurring words are appearing in the word cloud. In **Figure [3],** we can see that he talks a lot about Tesla, Mars, rocket, cars, launch etc. But we can also see the words yes, yeah, great, thank which makes his super enthusiastic in his tweets.

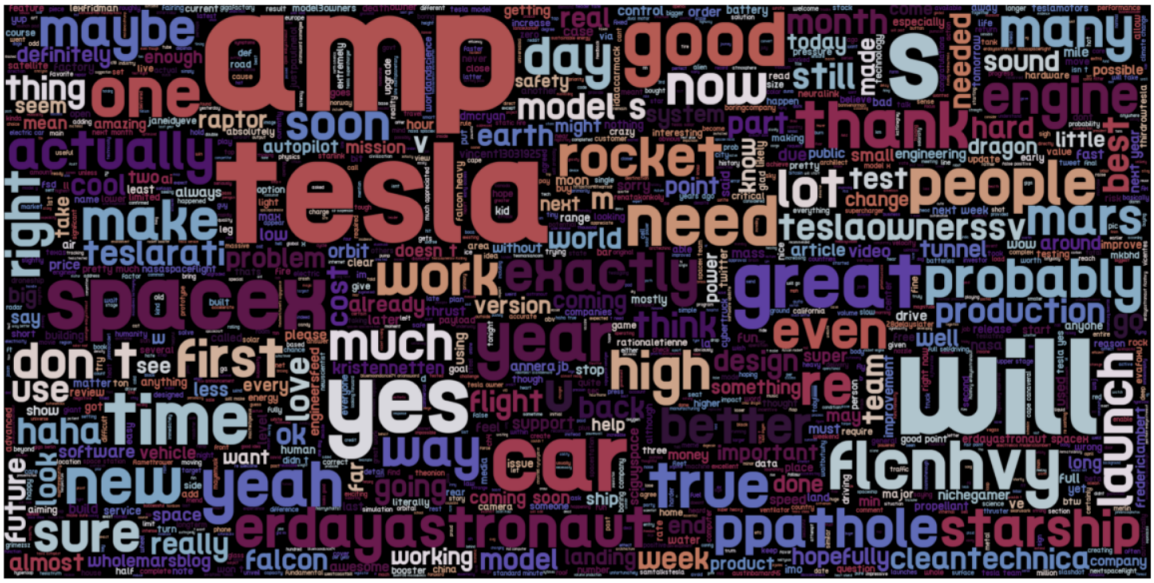


Figure [3] Word Cloud

iv) Bitcoin and Dogecoin in Tweets

In this step, in order to calculate the percentage of tweets that contain the word Bitcoin and Dogecoin, we first find the number of tweets that contains the word and then divide it by the total number of tweets. Then we calculate the most liked tweets that have the word Bitcoin and so we do the same for Dogecoin. I have displayed the top 10 most liked tweets as showed in **Figure [4]**.

```
% of tweets about Bitcoin: 0.127%
% of tweets about Dogecoin: 0.135%

 Most liked BITCOIN tweets:
1. you can now buy a tesla with bitcoin
2. btc bitcoin is an anagram of tbcthe boring company what a coincidence
3. bitcoin is my safe word
4. tesla is using only internal amp open source software amp operates bitcoin nodes directly bitcoin paid to tesla will be retained as bitcoin not converted to fiat currency
5. pay by bitcoin capability available outside us later this year
6. bitcoin is almost as bs as fiat money
7. bitcoin is not my safe word
8. maisiewilliams toss a bitcoin to ur witcher
9. business to be clear i am not an investor i am an engineer i don't even own any publicly traded stock besides tesla however when fiat currency has negative real interest only a fool wouldn't look elsewhere bitcoin is almost as bs as fiat money the key word is "almost"
10. business tesla's action is not directly reflective of my opinion having some bitcoin which is simply a less dumb form of liquidity than cash is adventurous enough for an sampp500 company

 Most liked DOGECOIN tweets:
1. dogecoin is the people's crypto
2. bought some dogecoin for lil x so he can be a toddler hodler
3. spacex is going to put a literal dogecoin on the literal moon
4. if major dogecoin holders sell most of their coins it will get my full support too much concentration is the only real issue imo
5. dogecoin rulz httpstcoflwwugaglu
6. origin of doge day afternoon the ancient romans sacrificed a dogecoin at the beginning of the doge days to appease the rage of sirius believing that the star was the cause of the hot sultry weather
7. dogecoinrich
8. dogecoin i just set up some little doge mining rigs with my kids it was fun
9. sadiaslayy dogecoinbets
10. dogecoin value may vary httpstcouwerahpv63
```

Figure [4] Bitcoin and Dogecoin in Tweets

Approved: _____ Date: ____/____/____ ☐ electronic copy received

b) Bitcoin Evolution

Here we look at the Bitcoin data to understand the trend of the price and volume over time.

i) Price over time

We notice that the price of Bitcoin was really low in the start until 2017. After 2017, the price started gradually increasing and in 2018, we can see that there was a sudden peak. It gradually started falling and was fluctuating until 2019. Later on we see sudden increase in the price of Bitcoin from **Figure [5].**

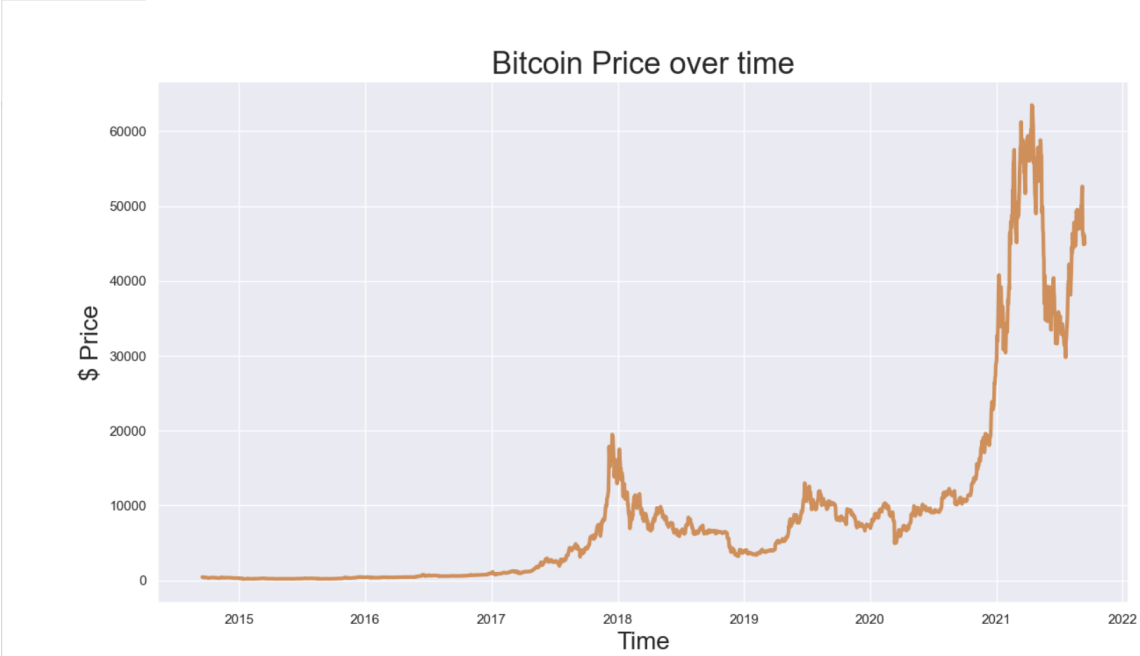


Figure [4] Bitcoin price over time

ii) Volume over time

The graph analogous to the previous one. We can see from **Figure [5]** that the volume of this cryptocurrency has increased significantly from 2019. We also see some peaks in the volumes in 2020 and a sudden peak in 2021.

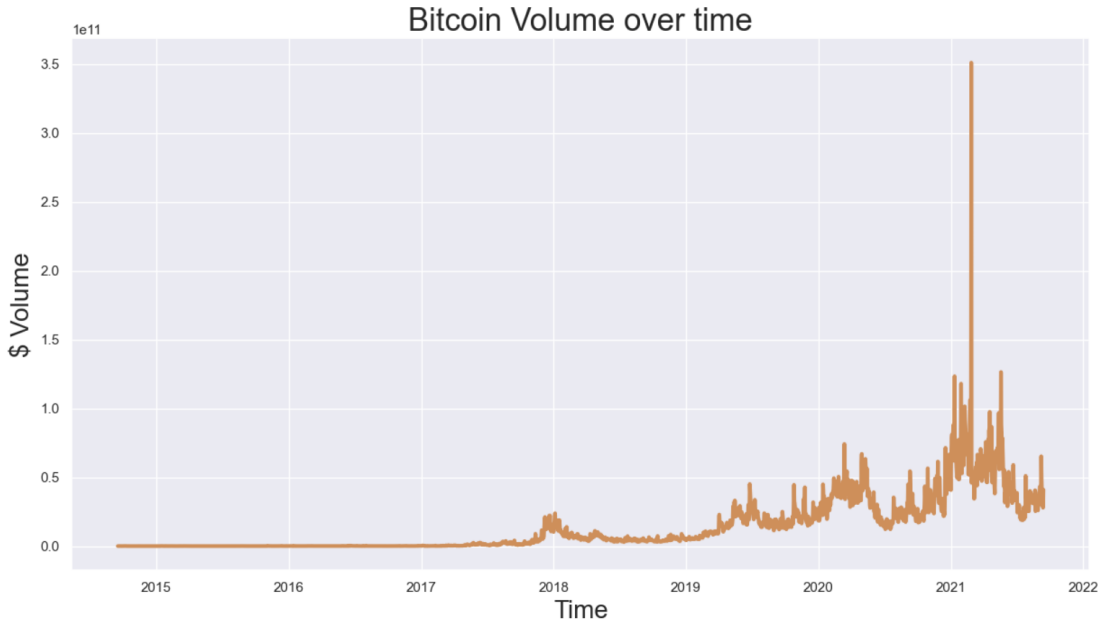


Figure [5] Bitcoin Volume over time

4 Methodology

The dataset used in the analysis is time-based; hence using time series analysis and identifying if the data has a trend becomes of utmost importance. Since the time series data may or may not have a trend, the Augmented Dicky Fuller Test is used for determination. The null hypothesis states that the time series data is not stationary, and on the other hand, the alternative theory states that the data is stationary. The statistical properties of the system do not change over time, which means that the data is stationary [7]. This does not imply that the values for each data point must be the same, but the comprehensive behavior of the data should remain constant. If the data has a trend, meaning it is stationary, we remove it and analyze it.

ADF determines the strength and weakness of time series as defined by the trend in the cryptocurrency change. The test returns the p-value, value of the test statistic, and the critical value cutoffs. We reject the null hypothesis when the statistic is lower than the p-value and conclude that the data is stationary.

From **Figure [6]**, we can see that the p-value is greater than the significance level of 0.05, also the statistic is higher than the any of the critical values. This means that we do not reject the null hypothesis and conclude that the time series data is non-stationary.

```
ADF Stastistic:  4.694853
p-value: 0.898845
The graph is non stationary! (it has a trend)
Critical values:
        1%: -3.432
        5%: -2.862
        10%: -2.567
```

Figure [6] Test Stationary for Time Series Data

From the plots of rolling mean and standard deviation, we notice that it is not constant, which makes it prove that the data is non-stationary. **Figure [7]** shows the plot.

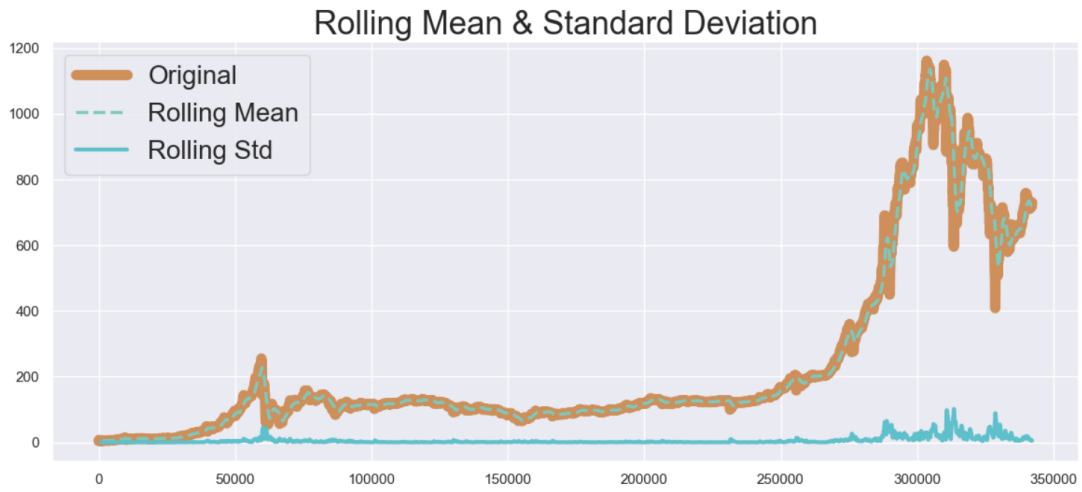


Figure [7] Rolling Mean & Standard Deviation of non-stationary data

On the other hand, from **Figure [8]**, we can see that the statistic is lower than the significance level of 0.05 and critical value. This means that we do reject the null hypothesis and conclude that the time series data is stationary.

```
ADF Stastistic:  -3.434271
p-value: 0.077653
The graph is stationary! (it doesn't have a trend)
Critical values:
        1%: -3.432
        5%: -2.862
        10%: -2.567
```

Figure [8] Test Stationary for Time Series Data

Figure [9] shows the plot of rolling mean and standard deviation indicating that the series data is stationary.

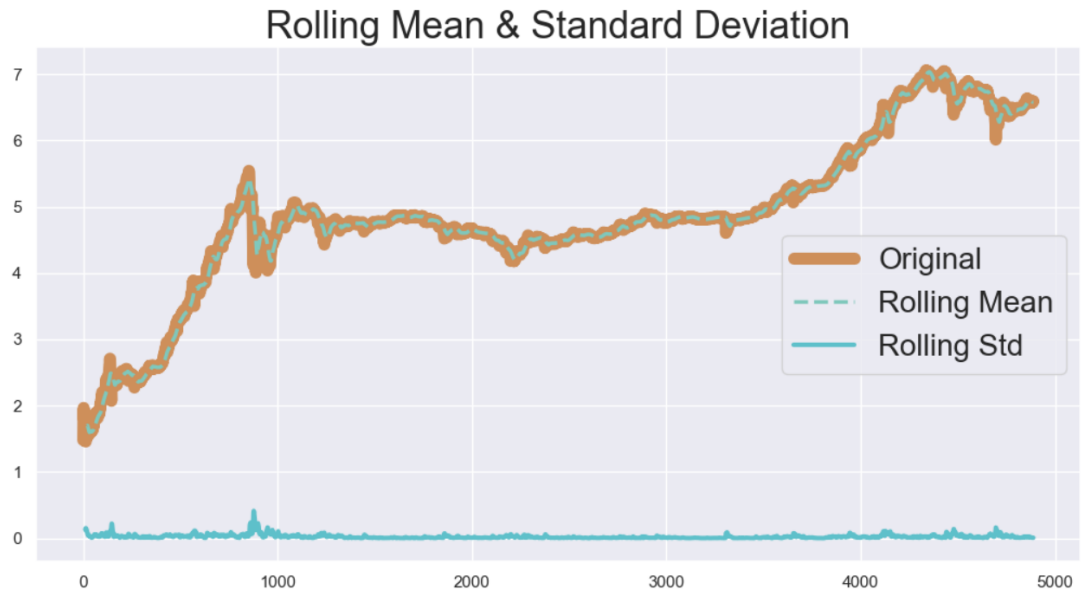


Figure [9] Rolling Mean & Standard Deviation of stationary data

Auto-regressive Integrated Moving Average was chosen for use as a portion of the linear regression prototype to a certain the next values in consideration of the past behavior of the target. The advantage of using ARIMA is that it does not apply the knowledge of exogenous values pinned on them, although they depend on previous target results for future prediction [8]. ARIMA can be split into MA, I, and ARR. AR applies the idea of target regression on its previous variables without lagging on it. The incorporated

element in ARIMA deals with stationarity enhancement of data. (p, d, q) of ARIMA represents the standard notation of ARIMA models [8]. These datasets can be substituted with numbered values to highlight the type of model implied. Parameter 'p' represents the lag rank of AR, which is the integer of lags in Y, which must be inclusive in the model, 'd' represents the differencing needed to convert the data into stationarity. Lastly, 'q' is the order of MA, the integer of lag that predicts future error. The equation below summarizes the ARIMA model,

$$Y = B0 + B1 * E\_lag1 + B2 * E\_lag2 + \ldots + Bn * E\_lag$$

A part of the economic and financial time series is usually characterized by autoregressive (AR) models. Among the main examples in finance, we have the valuation of prices and dividends, real exchange rates, interest rates and interest rate differentials [10]. ARIMA is an alternative and highly useful model in the modeling of series in finance, which we will denote by MA for its acronym in English (moving-average).

The models considered in the previous sections, AR, and MA, can be very useful in the modeling of certain data series in various fields of knowledge. However, in practice, specifically in finance, it may be necessary to consider models whose orders can lead to complications, motivated by the large number of parameters that are required to describe in such a way [10]. Adequate dynamic structure of the data. One way to overcome this problem is to consider a type of process that combines the properties of the AR and MA models in a more compact expression, which allow the reduction of parameters to be considered. This process is known as the moving average autoregressive process. ARMA models are obtained as a combination of autoregressive and moving average models.

The generic characteristics of the predictions made with ARIMA models include:

- In AR (p) models. The prediction as the time horizon increases, tends to the mean of the process.
- In the MA (q) models. If the time horizon of the prediction is greater than the order of the process, then the prediction is equal to the mean.
- In the ARMA (p, q) models. For periods higher than the order of the moving averages process, the prediction function behaves like that of an AR (p) process and, therefore, tends to the mean.
- In the ARIMA models (p, d, q). The prediction no longer tends to the mean but will be a straight line with a slope equal to the mean of the process that is obtained by making the necessary transformations so that the series is stationary.

Once the model has been identified, meaning the values of p, d and q are known, it is important to understand which model fits best. Using probabilistic statistical measures, we choose the model that not only fits the model performance on the training data but also the complexity of the model. The complexity of the model is defined by how well the trained model is behaving after the training. On the contrary, model performance is defined how well the model is performing on the training dataset. Advantage of using this probabilistic approach is that all the data can be fit to the model and the final model is used for prediction. However, the downside to this approach is that each metric must be carefully derived from each model as a general statistic cannot be applied to a varied range of models.

The prediction of subsequent actions based on earlier actions can be made using the autoregression model. This model is used for forecasting whenever there is some association connecting values in the time series. Various indicators such as Log-likelihood, Akaike, and Bayesian Information Criterion would be the base that would tell how much information is lost, which in turn measures the model's performance. The less information lost, the better is the performance. Hence, which regression model to choose depends on these criterion values. The moving-average model determines that the output variable depends directly on the term's current and past values. The ARIMA model uses its own lags for a given time series data to forecast the future value [6].

Hence, we use three criterions on which we measure the model performance, and they are log likelihood, Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). The AIC, and BIC are calculated using the log likelihood for the model which comes from maximum likelihood estimation used for finding and optimizing the parameters of a model is response to the training dataset. In likelihood estimation, the conditional probability of the data needs to be maximized given a specific probability distribution. The name log likelihood is derived from how frequently the word log is used in the likelihood function. Log likelihood comprises of various statistical computations such as the MSE (mean squared error) for regression model.

AIC or the Akaike Information Criterion is one of the methods for selecting the model and is derived from a framework and is usually not interpreted as the approximation to the likelihood. We simply choose the model that gives the smallest value of all the models built. The following formula computes the AIC.

$$AIC = -2/N * LL + 2 * k/N \text{ [11]}$$

Where N = total count in the training dataset.

LL = Log Likelihood.
K = no. of parameter

BIC or Bayesian Information Criterion also uses the log likelihood for its computation. Like AIC, the model with lowest BIC value is chosen. However, AIC puts more emphasis on model performance when compared to BIC and helps in selecting more complex models. The following formula computes the BIC.

$$BIC = -2 * LL + \log(N) * k \quad [11]$$

Where log () = base e natural log
LL = Log Likelihood.
N = total count in the training dataset.
K = no. of parameter

Figure [10] shows the values for Auto Regression Model – AR (1) +I (1) [1,1,0]

```
log-likelihood (smaller the better): -12,765.828

Corrected Akaike Information Criterion (AICc) (smaller the better): 25,537.656

Bayesian Information Criterion (BIC) (smaller the better): 25,556.069
```

Figure [10] Auto Regression Model

Figure [11] shows the values for Moving Average Model – I (1) + MA (1) [0,1,1]

```
log-likelihood (smaller the better): -12,765.81

Corrected Akaike Information Criterion (AICc) (smaller the better): 25,537.62

Bayesian Information Criterion (BIC) (smaller the better): 25,556.033
```

Figure [11] Moving Average Model

Figure [12] shows the values for Auto Regression Integrated Moving Average Model – AR (1) +I (1) + MA (2) [1,1,2]

```
log-likelihood (smaller the better): -12,753.391

Corrected Akaike Information Criterion (AICc) (smaller the better): 25,516.782

Bayesian Information Criterion (BIC) (smaller the better): 25,547.471
```

Figure [12] Auto Regression Integrated Moving Average Model

From all the above computations, we notice that the ARIMA fits the data the best, since the values of AIC and BIC are lower as compared to others.

5 Price Prediction of Bitcoin

In this step we will be performing the price prediction of Bitcoin based on the chosen model. The first step here is selecting a small portion of the dataset, splitting the dataset into training, and testing set. We will be predicting the future for 10 observations for ease and calculate the error rate. The error is calculated by subtracting the actual value from the predicted value over the actual value. We consider the absolute value and multiply by 100 to calculate the deviation of the predicted price. Below screenshots show the predicted, and actual price along with the error.

```
Step 342159. |  Predicted: [58405.07572754] | Actual: 58327.94864 | ERROR: [0.13223007]
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            4     M =           12
 This problem is unconstrained.

At X0         0 variables are exactly at the bounds

At iterate    0    f= -4.56919D+00    |proj g|=  4.00266D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    4      4     18      1     0     0   1.741D-03  -4.569D+00
 F =  -4.5691936001132101
```

Approved: _____ Date: ____/____/____ ☐ electronic copy received

```
Step 342160. |  Predicted: [58364.47090533] | Actual: 58348.52898400001 | ERROR: [0.02732189]
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            4     M =           12
 This problem is unconstrained.

At X0         0 variables are exactly at the bounds

At iterate    0    f= -4.56934D+00    |proj g|=  3.85983D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    4      4     15      1     0     0   1.728D-03  -4.569D+00
  F =  -4.5693383847284172
```

Approved: _____ Date: ____/____/____ ☐ electronic copy received

```
Step 342161. |  Predicted: [58349.97126892] | Actual: 58364.562976000016 | ERROR: [0.02500097]
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            4     M =           12
 This problem is unconstrained.

At X0         0 variables are exactly at the bounds

At iterate    0    f= -4.56948D+00    |proj g|=  4.00635D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    4      4     19      1     0     0   1.750D-03  -4.569D+00
  F =  -4.5694833885626212
```

Approved: _____ Date: ____/____/____ ☐ electronic copy received

```
Step 342162. |  Predicted: [58363.99269246] | Actual: 58467.79953600001 | ERROR: [0.17754532]
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            4     M =           12
 This problem is unconstrained.

At X0         0 variables are exactly at the bounds

At iterate    0    f= -4.56956D+00    |proj g|=  3.89215D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    4      4     19      1     0     0   1.725D-03  -4.570D+00
  F =  -4.5695563908195309
```

Approved: _____ Date: ____/____/____ ☐ electronic copy received

```
Step 342163. |  Predicted: [58435.09404786] | Actual: 58724.34408000002 | ERROR: [0.49255558]
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            4     M =           12
 This problem is unconstrained.

At X0         0 variables are exactly at the bounds

At iterate    0    f= -4.56914D+00    |proj g|=  3.90025D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    4      4     15      1     0     0   1.737D-03  -4.569D+00
  F =  -4.5691353786075295
```

Approved: _____ Date: ____/____/____ ☐ electronic copy received

```
Step 342164. |  Predicted: [58635.76362955] | Actual: 58720.45584000003 | ERROR: [0.14422948]
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            4     M =           12
 This problem is unconstrained.

At X0         0 variables are exactly at the bounds

At iterate    0    f= -4.56923D+00    |proj g|=  3.91599D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    4      4     15      1     0     0   1.739D-03  -4.569D+00
  F =  -4.5692334035480817
```

Approved: _____ Date: ____/____/____ ☐ electronic copy received

```
Step 342165. |  Predicted: [58682.82476934] | Actual: 58639.31992800003 | ERROR: [0.07419056]
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            4     M =           12
 This problem is unconstrained.

At X0         0 variables are exactly at the bounds

At iterate    0    f= -4.56937D+00    |proj g|=  3.89574D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    4      4     17      1     0     0   1.728D-03  -4.569D+00
  F =  -4.5693669070355982
```

Approved: _____ Date: ____/____/____ ☐ electronic copy received

```
Step 342166. |  Predicted: [58656.16381403] | Actual: 58725.07906400003 | ERROR: [0.11735233]
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            4     M =           12
 This problem is unconstrained.

At X0         0 variables are exactly at the bounds

At iterate    0    f= -4.56948D+00    |proj g|=  3.85469D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    4      4     15      1     0     0   1.725D-03  -4.569D+00
  F =  -4.5694812136143543

CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH
Step 342167. |  Predicted: [58705.81426661] | Actual: 58509.816583999986 | ERROR: [0.33498256]
```

The average error is 0.16948986365682422. This shows the overall model performance. **Figure [13]** shows the graph of the predicted vs actual values.

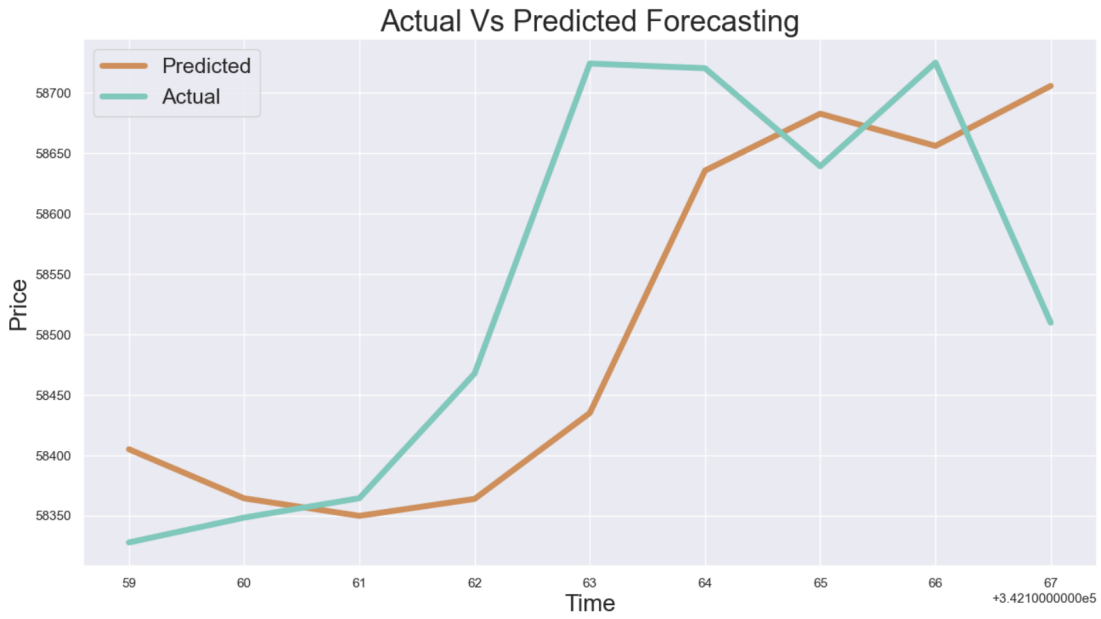


Figure [13] Actual Vs Predicted Forecasting

6 Correlation

Here we look at the points where he tweeted about Bitcoin. **Figure [14]** shows some tweets that have the word Bitcoin in it.

1. pay by bitcoin capability available outside us later this year
2. tesla is using only internal amp open source software amp operates bitcoin nodes directly bitcoin paid to tesla will be retained as bitcoin not converted to fiat currency
3. you can now buy a tesla with bitcoin
4. btc bitcoin is an anagram of tbcthe boring company what a coincidence
5. business to be clear i am not an investor i am an engineer i don't even own any publicly traded stock besides tesla however when fiat currency has negative real interest only a fool wouldn't look elsewhere bitcoin is almost as bs as fiat money the key word is "almost"
6. business tesla's action is not directly reflective of my opinion having some bitcoin which is simply a less dumb form of liquidity than cash is adventurous enough for an sampp500 company
7. bitcoin is almost as bs as fiat money
8. bitcoin is my safe word
9. toss a bitcoin to ur witcher
10. i still only own 025 bitcoins btw
11. pretty much although massive currency issuance by govt central banks is making bitcoin internet money look solid by comparison
12. bitcoin how much for some anime bitcoin httpstcoitqrslfncb
13. bitcoin
14. bitcoin2020conf
15. bitcoin is not my safe word
16. wanna buy some bitcoin

Figure [14] Tweets with word Bitcoin

Figure [15] shows some correlation between the tweets and price. However, this isn't much clear if they are much dependency of his tweets, hence let's take a closer look at only 3 tweets to confirm the correlation.

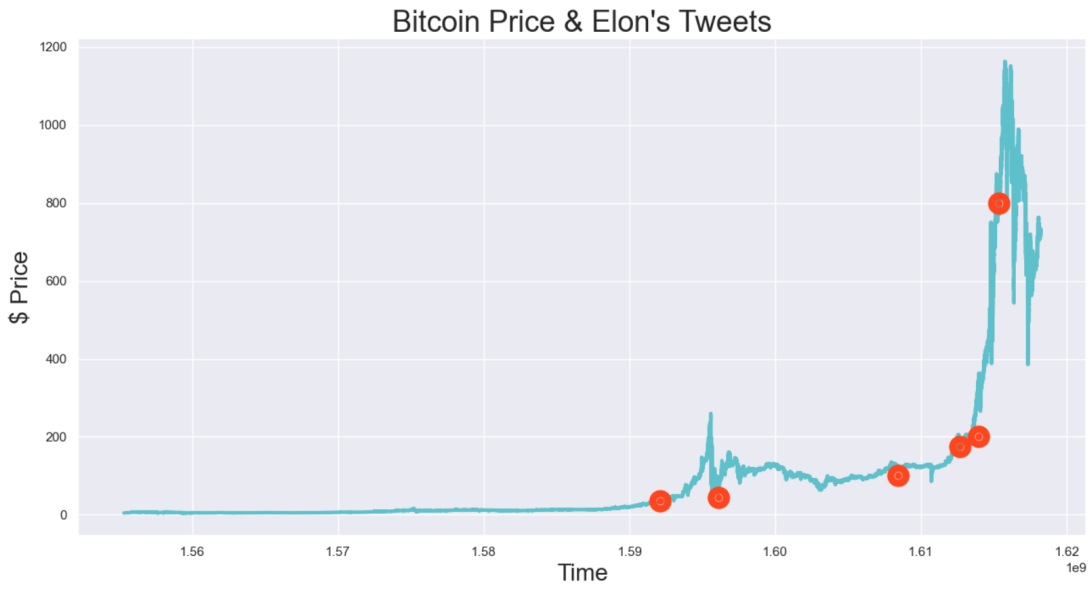
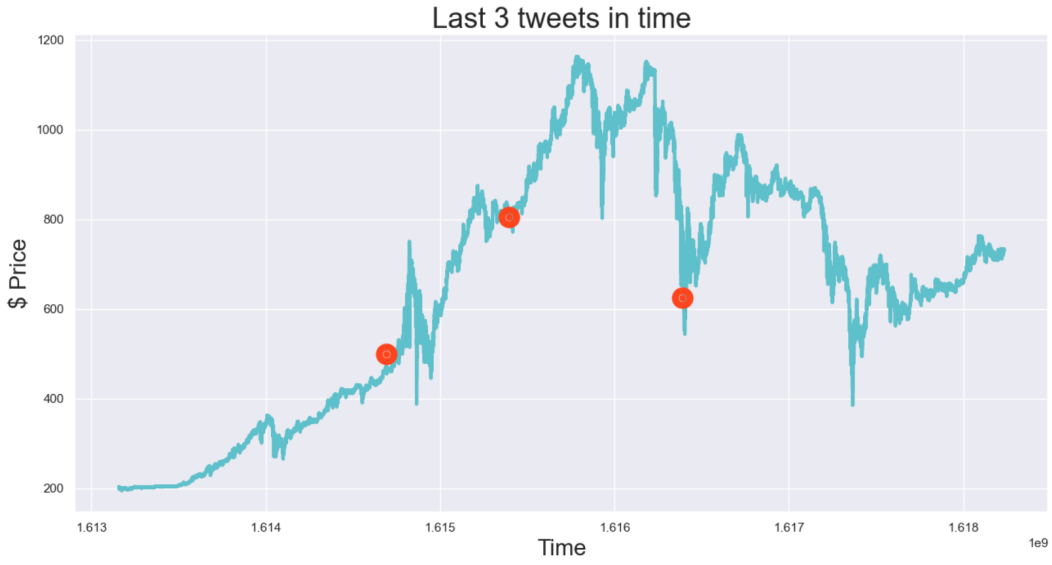


Figure [15] Correlation between tweets and price

Figure [16] shows some sudden peaks within a few days of making his tweets.



Approved: _____ Date: ____/____/____ ☐ electronic copy received

7 Limitations

There are a few limitations to this project. The project only considers tweets of Elon Musk and does not take into consideration other important Twitter accounts. Also, this project is focused on two specific cryptocurrencies and does not consider the whole stock market.

8 Conclusions and Future Scope

Price prediction relies greatly on Twitter sentiments, and thus the analysis of tweets is an important field. There are many news updates about cryptocurrency, and most research use Twitter for sentiment analysis. Sometimes, the short-term fluctuations in the prices are not concerning, but intense moments affect the users who use the currency for transactions or stash. This project aims to provide a strategy to investigate the impact of Elon Musk's tweets on the cryptocurrency and predict the price of Bitcoin and Dogecoin using Time Series Analysis.

References:

[1] Aich, Satyabrata, et al. "Analyzing stock price changes using event related Twitter feeds." 2017 19th International Conference on Advanced Communication Technology (ICACT). IEEE, 2017.

[2] Ante, Lennart. "How Elon Musk's Twitter Activity Moves Cryptocurrency Markets." Available at SSRN 3778844 (2021).

[3] Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. "Stock price prediction using the ARIMA model." 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE, 2014.

[4] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of computational science 2.1 (2011): 1-8.

[5] Ayaz, Zeba, et al. "Bitcoin Price Prediction using ARIMA Model." (2020).

[6] https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

[7] https://towardsdatascience.com/why-does-stationarity-matter-in-time-series-analysis-e2fb7be74454

[8] Lai, Yuchuan, and David A. Dzombak. "Use of the autoregressive integrated moving average (ARIMA) model to forecast near-term regional temperature and precipitation." *Weather and Forecasting* 35.3 (2020): 959-976.

[9] Singh, S. N., and Abheejeet Mohapatra. "Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting." Renewable energy 136 (2019): 758-768.

[10] Fattah, Jamal, et al. "Forecasting of demand using ARIMA model." International Journal of Engineering Business Management 10 (2018): 1847979018808673.

Approved: _____ Date: ____/____/____ ☐ electronic copy received

[11] https://machinelearningmastery.com/probabilistic-model-selection-measures/