Data Storytelling and Visualization

Bitcoin price prediction using ARIMA and LSTM methods that incorporate Tweet sentiment and Google trend analytics

Project Report

Manmeet Kumar Chaudhuri, 11011780

Table of Contents

1. Introduction

1.1 Cryptocurrencies

Cryptocurrencies are a subset of virtual currencies that use cryptography for security. These are decentralized and open source currencies and hence function on a peer-to-peer basis. Cryptocurrencies mostly use a very complex cryptographic algorithm, that requires connected network of computers to conduct computationally expensive mathematical operations. Cryptocurrencies have a built-in implementation of cryptography in their design. At present, people are using cryptocurrencies to implement a new form of economy, because of its cheapness, online, and anonymous means of exchange. A list of cryptocurrencies and their prices can be found at https://coinmarketcap.com, which lists more than 2646 cryptocurrencies of varying types. Cryptocurrencies feature certain computer protocols that are out of any government control. These currencies are unregulated and highly volatile. As a result, it can quickly devalue overnight. These currencies have aggressive swings in their prices, as it is largely based on public perception. It is therefore very hard to make related risk assessment at any moment. With the increase of the prices of cryptocurrencies, mining has also turned into a very advantageous business for the people.

One of the most valuable and decentralized cryptocurrencies is Bitcoin, which was introduced by Satoshi Nakamoto on October 31, 2008. It captured around 70% of the total market capitalization. Bitcoin's greatest innovation is blockchain, which was introduced to solve the issue of double spending as well as to disrupt the control of centralized parties in the transaction of values. The blockchain is the technology in which a record of any financial and economic transactions made in any cryptocurrency are maintained using cluster of computers that are linked in a peer-to-peer network. In simple terms, it is a powerful technology, which has the capacity to maintain permanent records of commercial transactions, transfer of assets and contracts, financial records, and intellectual property. Blockchain is completely a public ledger that is made up of blocks and any node connected on the Bitcoin network can process and clear a transaction by posting the transaction. A pictorial representation of a transaction using cryptocurrency is provided below:

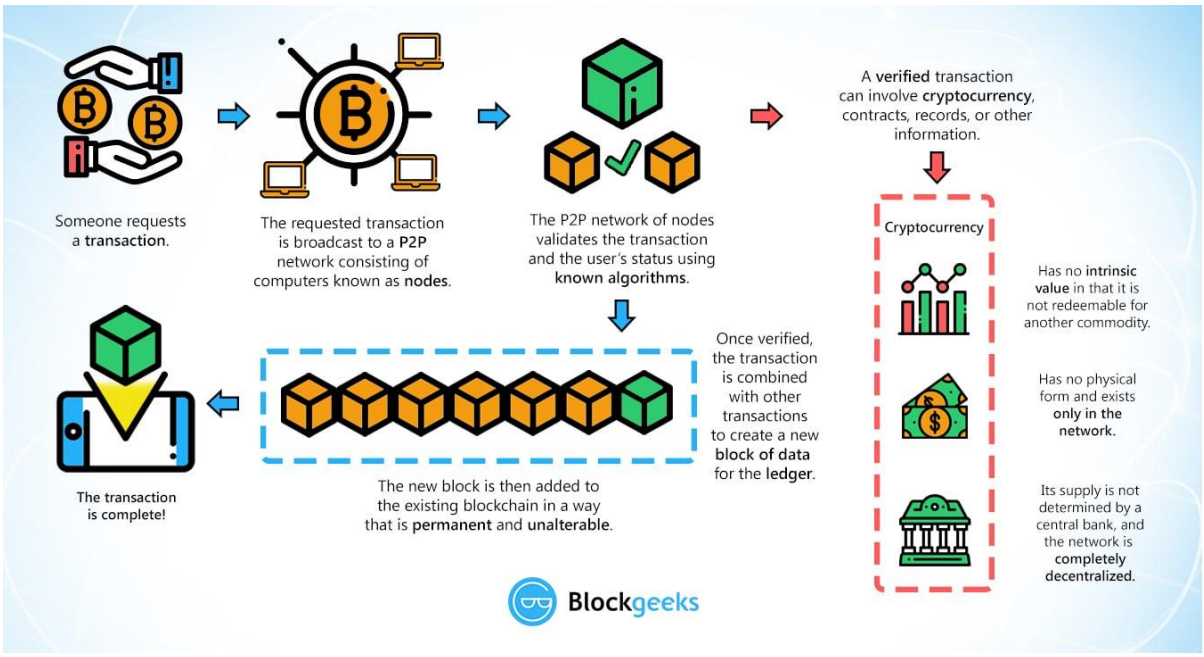


Fig 1.1 Flowchart of a transaction using Cryptocurrency

While the prices of cryptocurrencies have gone up since 2016 with great fluctuation, the enthusiasm of people to invest more and more in these virtual currencies stays more or less constant. People all over the world buy cryptocurrency to protect themselves against the devaluation of their national currency. Mostly in Asia, a vivid market for Bitcoin remittance has emerged, and the Bitcoin using darknets of cybercrime are flourishing.

1.2 Prediction

These virtual currencies are nowadays used in official cash flows and exchange of goods as a result, in recent years, various physical approaches and modelling techniques have been introduced by researchers and scholars to model the price of cryptocurrencies and to analyze the spontaneity of the market for making real decision support systems. These techniques include, but are not limited to, various dynamic topic modelling, machine learning, data mining, and text mining approaches. Moreover, to study the cryptocurrency market, agent based artificial financial market and genetic programming for finding attractive technical patterns have also been proposed. In addition, as cryptocurrencies are correlated, the cross correlation between price changes of various cryptocurrencies using random matrix theory and minimum spanning tress have also been studied. In recent years, different machine learning algorithms and techniques had also been taken into account to generate abnormal profits by exploiting the inefficiency of the cryptocurrency market.

1.3 Use case

One of the use cases for this project is as follows:

Retail investors want to allocate a portfolio of their investment/trades in cryptocurrency esp. Bitcoin. They want a Bitcoin price prediction system/app that will help them in making investment/trading decision. The output from this project would be one of the many factors that can be considered for making investment/trading decisions.

2. Motivation

2.1 Machine Learning

Cryptocurrencies' major issue of unpredictability is stifling efforts to bring blockchain finance to the mainstream as price fluctuation remains too high for many practical economic and monetary applications. Indeed, the cryptocurrency volatility index remains around 8% which is significant compared to the Foreign Exchange (Forex) market which experiences volatilities around 1%. This uncertain behavior persists even as the cryptocurrency market cap currently exceeds $270Bn. In this unstable environment, all too often compared to the subprime mortgage bubble of 2007, algorithmic trading seems to be the brightest light at the end of the tunnel.

While human speculation fuels high trading volumes on exchanges, it cripples the stability of cryptocurrencies. Fortunately, algorithmic trading allows traders to remain less biased and less prone to human error by predicting prices based on predetermined metrics. This is not to say algorithms are infallible as shown by the failure of the Nobel Prize winning Black-Scholes equation of 1997.

Nevertheless, deep learning algorithms are constantly praised for their abilities to bypass the curse of dimensionality, which provides a unique opportunity to investigate such an important economic

challenge. Moreover, being able to forecast more accurately cryptocurrency exchange rates is not without significant financial value.

2.2 Social Media and Google Trends

Nowadays, Web 2.0 services such as blogs, tweets, forums, chats, email etc. are widely used as communication media, with satisfying results. Sharing knowledge is an important part of learning and enhancing skills. Through the use of social media services, team members have the opportunity to acquire more detailed information about their peers' expertise [7]. Social media data represents a collective indicator of thoughts and ideas regarding every aspect of the world. It has been possible to assist to deep changes in habits of people in the use of social media and social network. Twitter, an online social networking website and microblogging service, has become an important tool for businesses and individuals to communicate and share information with a rapid growth and significant adoption. In addition, Twitter has rapidly grown as a mean to share ideas and thoughts on investing decisions.

Google Trends is a search trends feature that shows how frequently a given search term is entered into Google's search engine relative to the site's total search volume over a given period of time. Google Trends allows one to measure interest in a particular topic across search, from around the globe, right down to city-level geography.

A combination of Tweets and Google Trends data provide qualitative and quantitative information on requisite topics/keywords that can be incorporated in machine learning algorithms to further refine the outcome/model.

2.3 Data Visualization

We are an inherently visual world, where images speak louder than words. Data visualization is especially important when it comes to big data and data analyzation projects. Dashboards shall be prepared using various visualization techniques and other interaction techniques, such as brushing and linking, etc. Visualizations would be visually pleasing based on good usage of the design elements within the gestalt principles.

3. Problem Statement

The aim of this project is to perform exploratory data analysis on cryptocurrencies and to develop a machine learning pricing algorithm that is capable of modelling cryptocurrencies to offer next-day pricing prediction. Also, the results of the exploratory data analysis and machine learning algorithm shall be visualized by using requisite design principles.

More specifically, the objectives to achieve this are as follows:

- Automated and real time data collection: Cryptocurrency pricing data, tweets related to Bitcoin keyword and Google trends data on Bitcoin Keyword are extracted from the relevant websites on daily basis using python scripts.

- Exploratory data analysis: Exploratory data analysis is performed on the Cryptocurrency data, Tweeter data and Google trends data to generate useful insights from the data. Exploratory data analysis is used to identify a suitable cryptocurrency i.e. Bitcoin for further pricing prediction.
- Sentiment analysis of Tweets: Sentiment analysis of tweets is carried out to calculate the sentiment scores of each of the tweets.
- Time series forecasting: Traditional time series forecasting method shall be used to predict the next day Bitcoin price based on the historic Bitcoin price data.
- Machine learning prediction: A machine learning/deep learning framework/methodology is developed that predicts the next day Bitcoin price using the historic Bitcoin price data, sentiment of tweets containing Bitcoin keyword and Google trends data on Bitcoin keyword.
- Comparison of forecasting methods: The accuracy of traditional time series forecasting method and Lachine learning model are compared using confusion matrix, RMSE, MAPE, etc.
- Real time next-day price prediction: The best model shall be used to predict the real time next-day Bitcoin price prediction.
- Data visualization: The results of all the above-mentioned analysis shall be presented with the help of effective data visualization techniques. Tableau is used extensively for data visualization.

4. Literature Review

4.1 Related work

Cryptocurrency is a new digital asset in finance, which has extremely high volatility as compared to almost all other financial instruments. As a result of its high volatility and price fluctuations, a very limited number of articles, to the best of our knowledge, exists in the literature that deal with predicting the price fluctuations.

A strand of literature has examined whether Bitcoin returns are predictable, with Urquhart (2016) indicating that Bitcoin returns are predictable and therefore are contrary to the Efficient Market Hypothesis. This finding has been further supported by Nadarajah and Chu (2017), Tiwari et al. (2018), Kjuntia and Pattanayak (2018) and Caporale et al. (2018) amongst others. All of these papers examine the relationship between returns and whether there are patterns or correlations that may be exploitable by investors. A recent paper by Urquhart(2018) shows that the attention of Bitcoin, captured by Google Trends, can be explained by the previous days realized volatility and volume, indicating that the high volatility and trading volume experienced by Bitcoin number of times the term ‘Bitcoin’ has been searched for in the Google search engine, and therefore is a good measure of attention from uninformed individuals who want to find out more information about Bitcoin. However well-informed investors, who have knowledge of the cryptocurrency, will not be searching for it in the Google search engine but instead may be tweeting about it. These tweets may involve commenting on news posts related to Bitcoin or making predictions of the future price direction of Bitcoin or just giving an opinion of the popular cryptocurrency. Hence, it can be postulated that the sentiment of Bitcoin tweets is also a stronger measure of investor attention apart from Google Trends.

There is a growing literature examining the impact of Twitter on financial markets, such as Piñeiro Chousa et al. (2016), Sun et al. (2016) and Piñeiro Chousa et al. (2018) who all find significant relationships between Twitter and financial markets.

McNanny et al. (2018) have used recurrent neural network (RNN), long short time memory (LSTM) network and ARIMA model to predict the direction of Bitcoin price in USD. RNN and LSTM are two deep learning pipelines, that outperformed the ARIMA forecasting model. Root mean squared error (RMSE) is used to evaluate and compare the regression accuracy and an 80/20 holdout validation strategy is used to instrument the validation of models. As a result, the accuracy and RMSE obtained using ARIMA model are 50.05% and 53.74%. Using RNN (LSTM) model, the accuracy and RMSE obtained are 50.25% (52.78%) and 5.45% (6.87%). Hence, this report shall further use the LSTM model for Bitcoin

4.2 Contribution

We add to the literature by studying whether the sentiment of tweets involving the term 'Bitcoin' along-with the Google trends data on 'Bitcoin' keyword can predict the returns/price of Bitcoin. If the study is able to produce significant results, the approach can be extended to other cryptocurrencies as well and would help in explaining/predicting the volatility in the prices of cryptocurrencies.

In this research, we chose regression machine learning due to continuous values of Bitcoin price. For deep learning-based regression models, Keras library was used to create LSTM model.

5. Method/Architecture

5.1 Data collection

Cryptocurrency data: Data on cryptocurrencies shall be collected from the developer API and webpages of www.coinmarketcap.com. Data is be stored in .csv format and is automatically updated on daily basis to collect the latest price information. We have selected top 10 currencies by market capitalization for further study. These cryptocurrencies and their features selected for data collection are as follows:

S. No.	Cryptocurrency	Time Period
1	Bitcoin	28th April 2013 – until now
2	Ethereum	7th August 2015 – until now
3	Ripple	4th August 2013 – until now
4	Bitcoin Cash	23rd July 2017 – until now
5	Litecoin	28th April – until now
6	Tether	25th February 2015 – until now
7	Binance Coin	25th July 2017 – until now
8	EOS	1st July 2017 – until now
9	Bitcoin SV	9th November 2018 – until now
10	Monero	21st May 2014 – until now

Fig 5.1 Data on cryptocurrency and the time period of data

S. No.	Features	Unit
1	Open	USD
2	High	USD
3	Low	USD
4	Close	USD
5	Volume	USD
6	Market Cap	USD

Fig 5.2 Variables for each of the cryptocurrency dataset

Tweeter data: Tweets with 'Bitcoin' keyword are collected using Tweeter Developer API. Only four features are collected for each tweet i.e. tweet ID, Date, Time and Tweet. Tweet. Tweeter data is collected from 24th August 2018 until now.

Google trends data: Google trends data on 'Bitcoin' keyword is collected from https://trends.google.com. All the data is collected daily by running a python script and appended to the older data. Google trends data is collected for a period from 24th August 2018 until now.

5.2 System Design

System design is presented on Figure 1 and it consists of four components: Retrieving Twitter data, pre-processing and saving to database (1), stock data retrieval (2), Google trends data retrieval (3), model building (4) and predicting future stock prices (5). Each component is described later in this text.

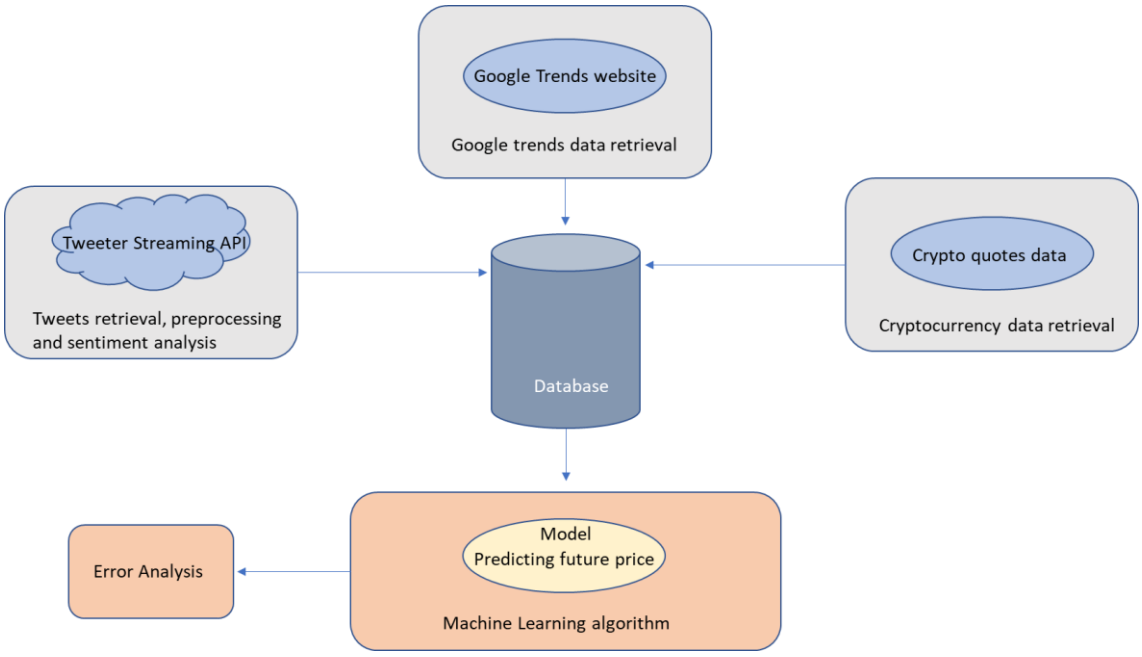


Fig 5.3 System design for forecasting Closing Price

1. Retrieving Twitter data, pre-processing and saving to database: This component is responsible for retrieving, pre-processing data and conducting sentiment analysis on the tweets.
2. Cryptocurrency data retrieval: Pricing data is gathered on daily basis. Afterwards, it is used for estimating future prices.
3. Google trends data retrieval: Google trends data on 'Bitcoin' keyword is gathered daily and is used in the model for predictions.
4. Model building: This component is responsible for preparing machine learning models using ARIMA and LSTM for price forecasting.
5. Predicting future stock prices: This component combines results of sentiment detection of tweets with past intraday cryptocurrency data and Google trend data to estimate future price values.

6. Analysis

6.1 Exploratory Data Analysis of Cryptocurrency data

Data shall be analyzed to perform comparative study of cryptocurrencies on various features and to provide data visualizations of these features through suitable visualization techniques. Dataset for each of the cryptocurrency dataset has the following raw features/columns:

- Date: It is the date for which various features are collected
- Open: It is the opening price (in USD) of the cryptocurrency
- High: It is the highest price achieved by a cryptocurrency on a particular day
- Low: It is the minimum price achieved by a cryptocurrency on a particular day
- Close: It is the last price (in USD) achieved by the cryptocurrency on a given day
- Volume: It is the total amount (in USD) of cryptocurrency that is traded on a particular day
- Market Capitalization: It is the total worth of a cryptocurrency on a given day. It is the multiplication of number of units of a cryptocurrency and the closing price of that cryptocurrency on a given date.

Feature wise data analysis of the each of the cryptocurrency is conducted in the following sections.

6.1.1 Price and Market Capitalization:

These are the raw features in the cryptocurrency dataset. Market capitalization is a measure of the value of security/cryptocurrency. Cryptocurrency market capitalization is both a quick way to gauge a coin's value. A lot can be learnt about crypto just by checking market caps. A high or low market cap can reveal a coin that is resistant to volatility, or vulnerable. Coins with small market caps consequently rock more when big news hits the market, or "whales" (large buyers) take positions. Also, normally the currency with higher price is generally less volatile as compared to the currency with lower price.

A dashboard prepared in Tableau containing the important pricing and market capitalization information is shown below:

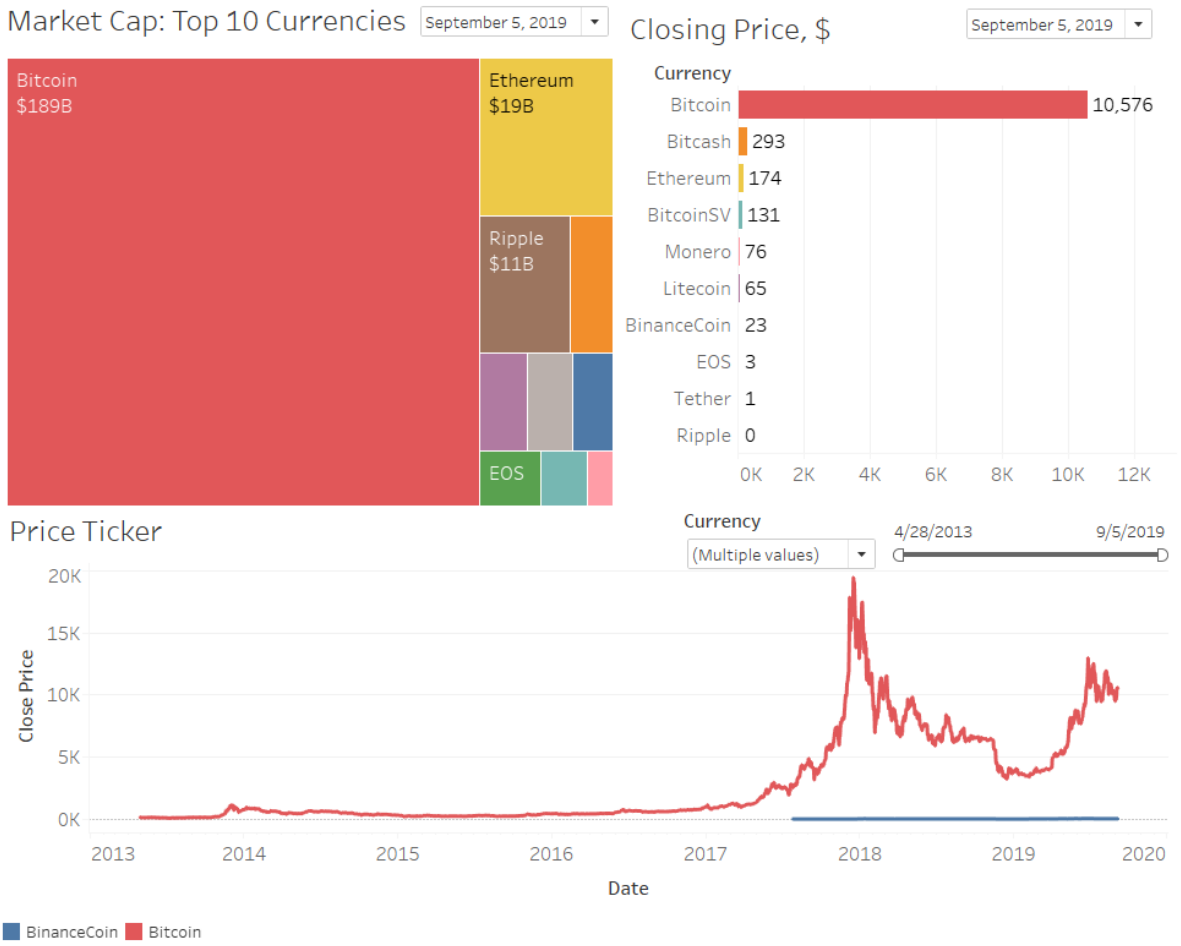


Fig 6.1 Tableau dashboard for Price and Market capitalization

Design elements in the dashboard are discussed in detail as follows:

S. No.	Visualization Feature	Type	Remarks
1	Data type by Task taxonomy	Temporal data	Data on cryptocurrency is time series data
2	Action target pair	Present features	Features are presented in the Market Cap display and Closing Price display
		Present trend	Trend is presented through the Price ticker
		Discover outliers	Outliers are identified in the Price ticker
3	Bertin's visual variables	Position	Position is used to display price information of respective currencies
		Size	Size is used to differentiate the market value and type of cryptocurrency
		Hue	Hue is used to differentiate the market value and type of cryptocurrency
4	Visualization technique	Standard 2D display	All the displays are in two dimensions
5	Visual displays	Tree maps	Market cap of cryptocurrencies

		Horizontal bar chart	Closing price of cryptocurrencies
		Line chart	Price ticker containing historic price
6	Marks	Area	Market cap display
		Line	Closing price display
		Points	Price ticker display
7	Channels	Area, color	Market cap display
		Length, color	Closing price display
		Length, color	Price ticker display
8	Interaction technique	Interactive filtering	Market capitalization and closing price can be filtered as per a selected date
		Brushing and linking	All the three visuals above are linked as per the currency type. Any one currency can be selected and then, all the displays provide information on that selected currency only.
		Focus and context via brushing	The price ticker can be changed by changing the range of dates

That isn't inherently surprising – the crypto markets are among the most volatile the world has ever seen. But holders of tokens with small market caps are at risk of being crushed by larger traders. If several whales conspire to sell at the same time, the price of a token can crash to nothing instantly. This would be much tougher with Bitcoin and Ethereum, which have large volume and are not easily manipulated. By market cap, Bitcoin is still the biggest cryptocurrency, with the current market value exceeding an enormous $189 billion, while Ethereum has a market cap of $19 billion, reclaiming its position as the second largest crypto after Bitcoin.

A healthy market cap is indicative of a strong coin. Always weigh market cap with some of the other metrics before making an investment decision. Based upon the pricing and market capitalization data, the top 3 strong coins are as follows:

1. Bitcoin
2. Ethereum
3. Bitcash

6.1.2 Daily Volatility and Volume/Market Cap ratio

These are the derived features from the cryptocurrency dataset. Volatility is defined as the statistical measure of dispersion of an asset's price. Simply put, volatility describes the extent to which an asset's price fluctuates over time. An investment is considered volatile if its prices move aggressively up or down daily. Volatility is a vital concept to understand since it measures risks. Different individuals possess a different level of risk tolerance, and this affects their choice of investments. For novice/beginner investors, it makes sense to choose less volatile cryptocurrency to begin with.

The Volume to Market Cap Ratio is also useful because a consistently high ratio indicates high liquidity. A sudden increase in the ratio could indicate a short-term change in liquidity associated with a pump or dump. Volume/Market cap is also an indirect measure of volatility as high volume/market ration indicates more trading and more volatility.

A dashboard prepared in Tableau containing the important daily volatility and volume/market capitalization information is shown below:

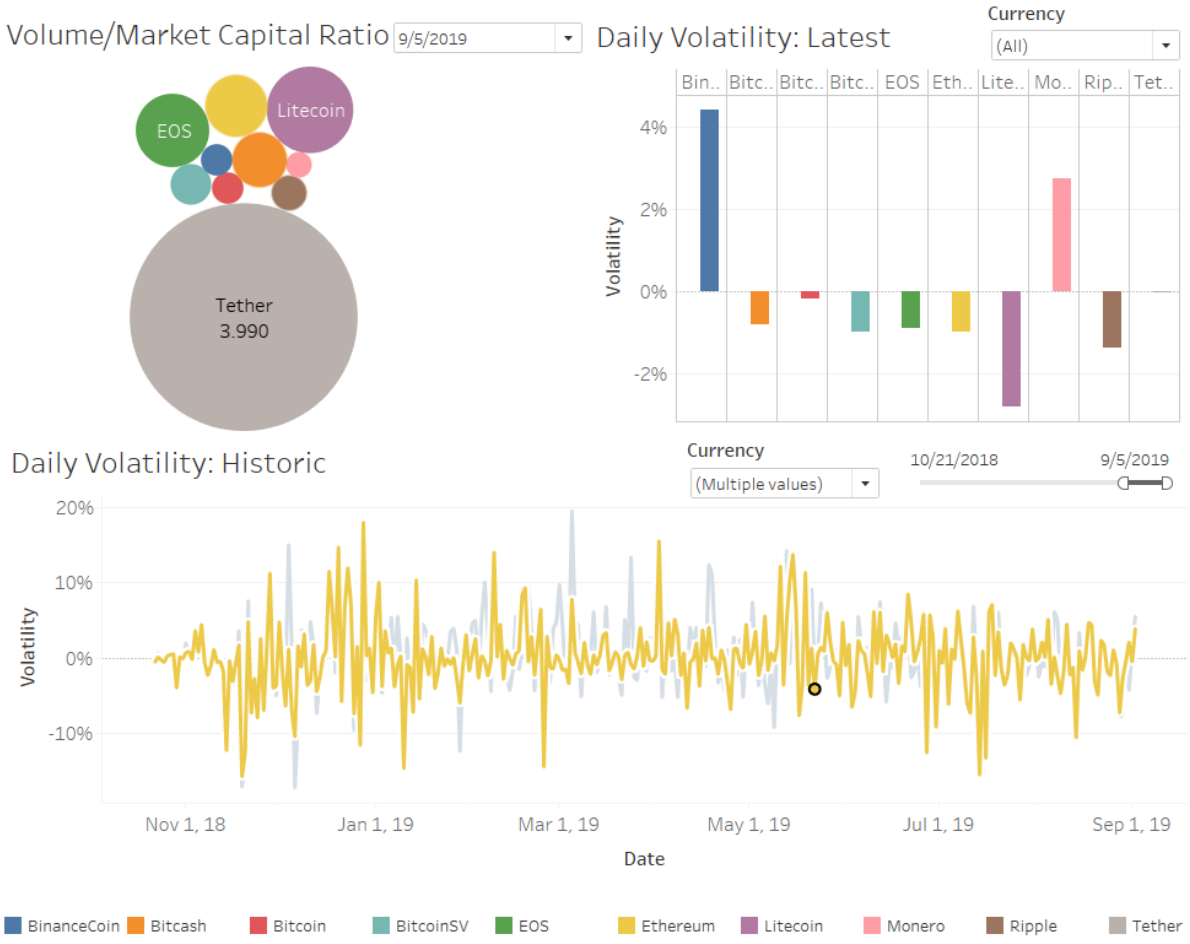


Fig 6.2 Tableau dashboard for Daily volatility and Volume/Market Cap ratio

Design elements in the dashboard are discussed in detail as follows:

S. No.	Visualization Feature	Type	Remarks
1	Data type by Task taxonomy	Temporal data	Data on cryptocurrency is time series data
2	Bertin's visual variables	Position	Position is used to display the daily volatility of respective currencies
		Size	Size is used to differentiate the volume/market capitalization ratio and type of cryptocurrency
		Hue	Hue is used to differentiate types of cryptocurrency
3	Visualization technique	Standard 2D display	All the displays are in two dimensions
4	Visual displays	Packed bubbles	Volume/Market cap ratio of cryptocurrencies
		Vertical bar chart	Daily volatility: Latest of cryptocurrencies
		Line chart	Daily volatility: Historic is depicted using line charts
5	Marks	Area	Volume/Market cap ratio display
		Line	Daily volatility: Latest display
		Points	Daily volatility: Historic display

6	Channels	Area, color	Volume/Market cap ratio display
		Length, color	Daily volatility: Latest display
		Length, color	Daily volatility: Historic display
7	Interaction technique	Interactive filtering	Volume/Market cap ratio can be filtered as per a selected date
		Brushing and linking	All the three visuals above are linked as per the currency type. Any one currency can be selected and then, all the displays provide information on that selected currency only.
		Focus and context via brushing	Daily volatility: Historic can be changed by changing the range of dates

From the above dashboard, it can be inferred that Bitcash and Bitcoin SV have the highest historic volatility. Based upon the volume/market capitalization and historic daily volatility data, the top 3 strong coins are as follows:

1. Bitcoin
2. Binance coin
3. Monero

These three currencies appear to be the most stable currencies among the choose 10 currencies.

6.1.3 Returns

Return is the derived feature in our cryptocurrency dataset. Returns over a period is the relative increase or decrease in the price at the end of the period over the price at the start of the period. Mathematically,

$$R_t = (P_t - P_{t-n})/ P_{t-n}$$

Where, R_t is the return over a period of n days, P_t is the price at the end of the period and P_{t-n} is the price at the beginning of the period.

Returns are the direct measure of the performance of a cryptocurrency over a period. Higher the returns, higher is the probability of money chasing those assets.

A dashboard prepared in Tableau containing return since launch of respective cryptocurrencies and the returns over a period of time is shown below:

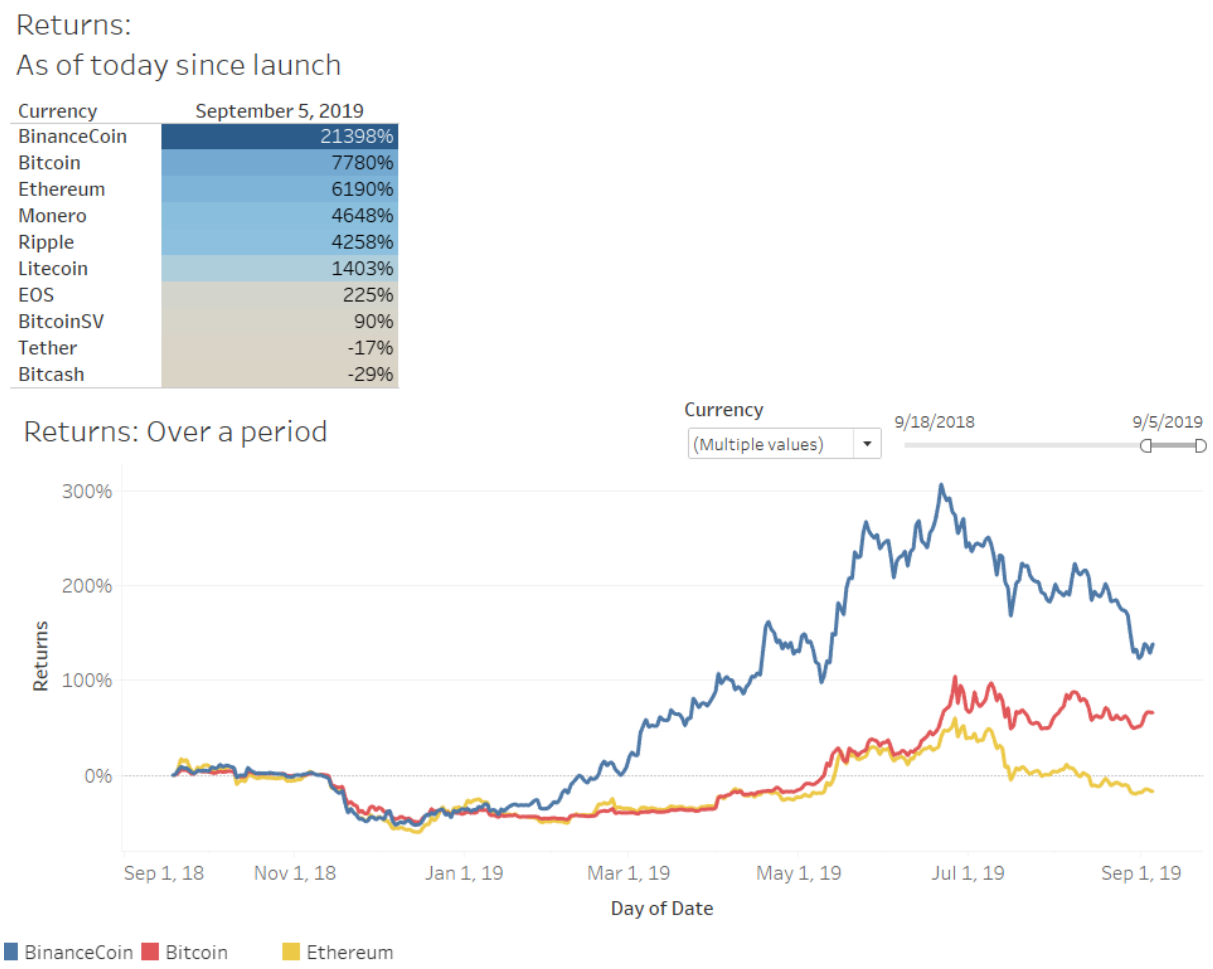


Fig 6.3 Tableau dashboard for Returns

Design elements in the dashboard are discussed in detail as follows:

S. No.	Visualization Feature	Type	Remarks
1	Data type by Task taxonomy	Temporal data	Data on cryptocurrency is time series data
2	Action target pair	Present features	Features are presented in the Returns since launch display
		Present trend	Trend is presented through the Returns over a period display
3	Bertin's visual variables	Position	Position is used to display the return over a period for respective currencies
		Hue	Hue is used to differentiate types of cryptocurrency and value of returns since launch
4	Visualization technique	Standard 2D display	All the displays are in two dimensions
5	Visual displays	Table	Returns as of today since launch is shown in tabular format
		Line chart	Returns: over a period is depicted using line charts
6	Marks	Points	Returns as of today since launch
		Points	Returns: over a period

7	Channels	Color	Volume/Market cap ratio display
		Length, color	Returns: over a period
8	Interaction technique	Interactive filtering	Returns as of today since launch can be filtered as per a selected date
		Brushing and linking	All the two visuals above are linked as per the currency type. Any one currency can be selected and then, all the displays provide information on that selected currency only.
		Focus and context via brushing	Returns: over a period can be changed by changing the range of dates

From the above dashboard, it can be inferred that Binance coin (21398%) and Bitcoin (7780%) have given the highest returns since the launch of the respective currencies. Bitcash and Tether have delivered negative returns of -17% and -29% since the launch respectively. Based upon the volume/market capitalization and historic daily volatility data, the top 3 strong coins are as follows:

1. Binance coin
2. Bitcoin
3. Ethereum

The exploratory data analysis based upon the various raw and derived features can be summarized in the following way:

Ranking	Market Capitalization	Volatility	Returns
1	Bitcoin	Bitcoin	Binance Coin
2	Ethereum	Binance Coin	Bitcoin
3	Bitcash	Monero	Ethereum

It can be seen that Bitcoin is the most suitable currency for investment/trading for beginners based upon the selected parameters. We shall now focus on developing a pricing model for Bitcoin that can help with predicting the price movements. Descriptive statistics of the Bitcoin dataset is conducted and the result is provided below:

	Open*	High	Low	Close**	Volume	Market Cap	Return
count	2323.000000	2323.000000	2323.000000	2323.000000	2.080000e+03	2.323000e+03	2323.000000
mean	2802.081855	2883.750771	2714.659148	2806.376737	3.343819e+09	4.737238e+10	19.910788
std	3664.272985	3792.185208	3515.893500	3667.152777	5.986651e+09	6.345289e+10	27.323969
min	68.500000	74.560000	65.530000	68.430000	2.857830e+06	7.784112e+08	-0.490000
25%	363.160000	375.475000	356.045000	363.795000	3.164785e+07	5.006118e+09	1.715000
50%	659.170000	671.510000	645.710000	659.260000	1.110510e+08	9.851765e+09	3.910000
75%	4380.515000	4463.100000	4270.960000	4379.645000	4.603353e+09	7.343775e+10	31.635000
max	19475.800000	20089.000000	18974.100000	19497.400000	4.510573e+10	3.270000e+11	144.280000

Co-relation analysis is also conducted on all variables present in the Bitcoin dataset.

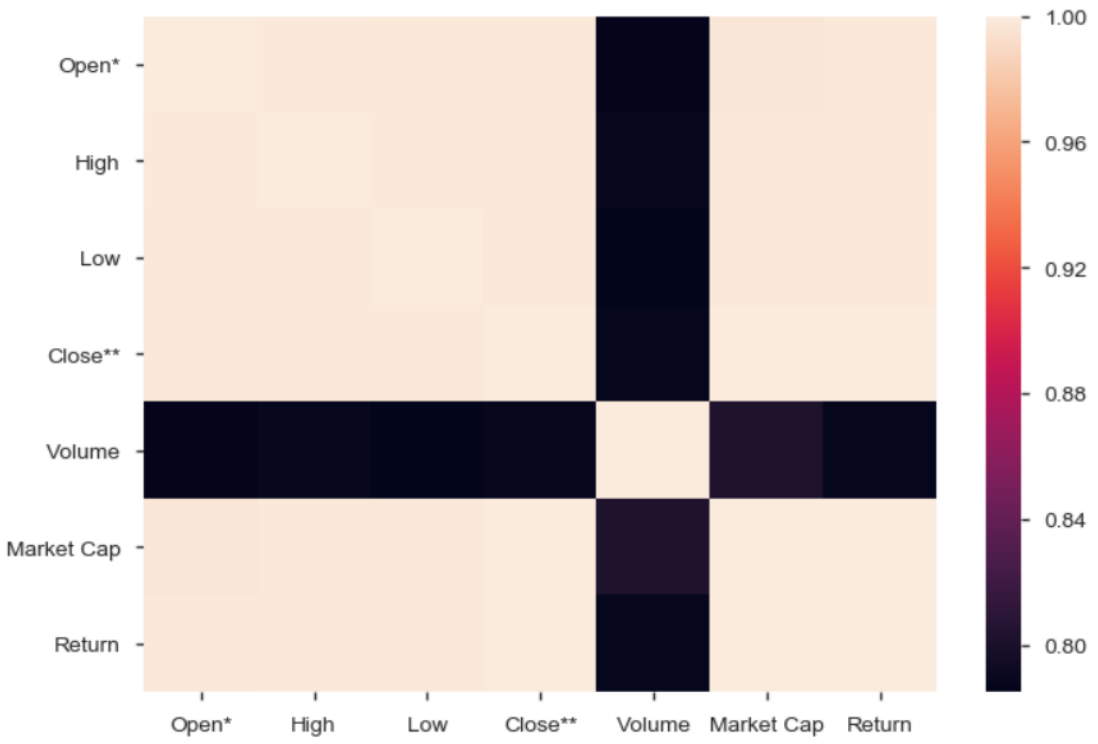


Fig 6.4 Correlation matrix on various variables

6.2 Google Trend Data Analysis

This provides a proxy metric for the general interest there is in cryptocurrencies at any given time, which could have a relationship with cryptocurrency prices over time as general interest increases and decreases. Google trends data is indexed to 100, where 100 is the maximum search interest for the time. Google trends data for Bitcoin keyword is collected since 24th August 2018. The data visualization of the trend value is prepared in Tableau and presented below.

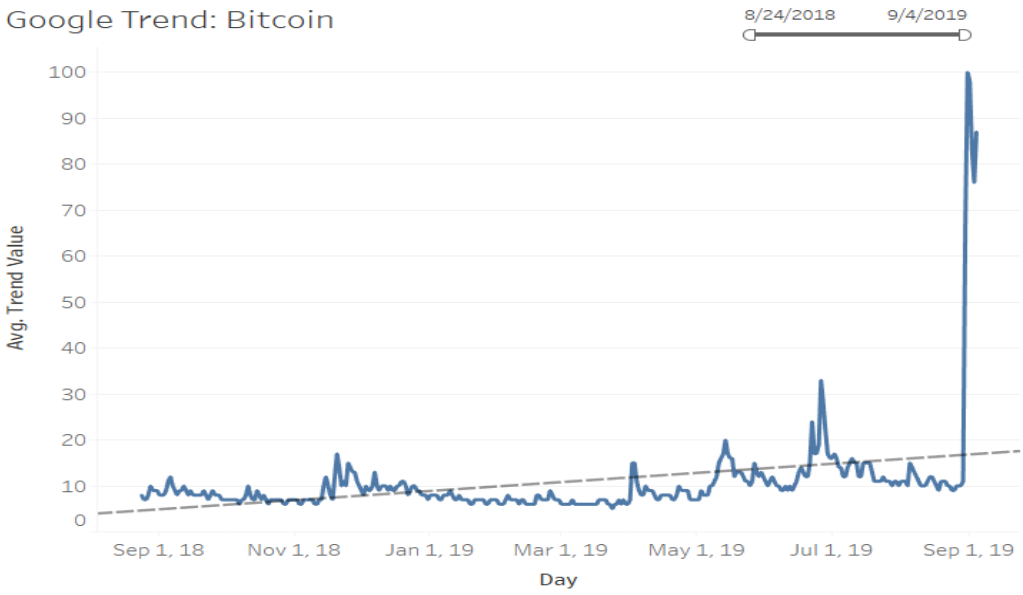


Fig 6.5 Google trend data on Bitcoin keyword

It can be seen that the trend is of upward nature. However, an anomaly can be observed wherein the last couple of datapoints from 31st August are very high compared to the average trend. This is a highly

unpredictable behavior and the reason for such a surge in google searches for Bitcoin should be further explored. Descriptive statistics of the Google trend dataset is conducted and the result is provided below:

	Trend_Value
count	377.000000
mean	10.671088
std	10.174961
min	5.000000
25%	7.000000
50%	9.000000
75%	11.000000
max	100.000000

6.3 Tweets Sentiment Analysis

Tweets from verified account only were collected for this study. Tweets are collected from 24th August onwards. Sentiment analysis is the act of extracting and measuring the subjective emotions or opinions that are expressed in text. There are multiple ways to do this. VADER (Valence Aware Dictionary for sEntiment Reasoning) sentiment analysis was used to analyse the collected tweets. VADER analysis provides several benefits including the fact that it not only classifies text as positive, negative, or neutral, but also measures the intensity, or polarity, of words used. For our purposes, we also benefit from the fact that the words and scores used in VADER are specifically tuned to microblog and social media contexts. To eliminate noise from the analysis we first clean the collected tweets. The end goal of this analysis is to apply sentiment analysis to collected tweets so that it can be determined if the tweets are generally positive or negative in their opinions of cryptocurrencies. An analysis of variation of number of verified tweets on daily basis is presented below:

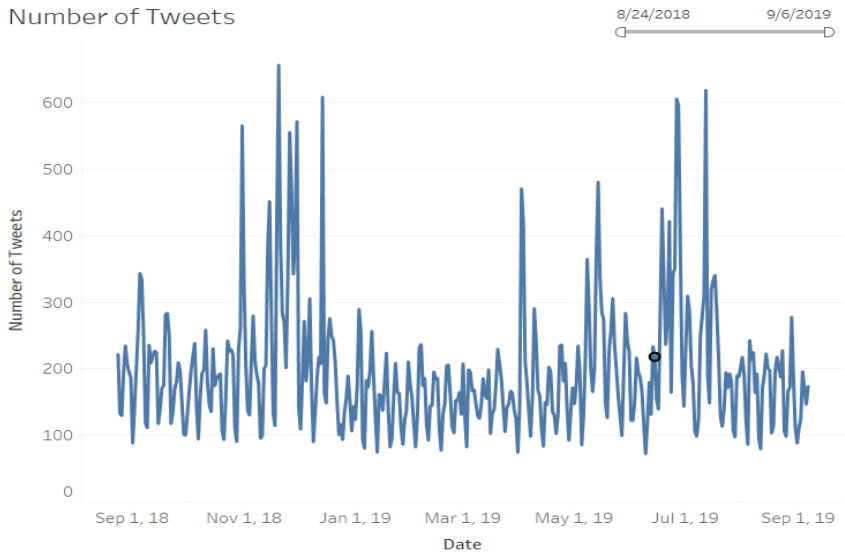


Fig 6.6 Daily count of number of verified tweets on Bitcoin keyword

From the above chart, it can be observed that the daily count of number of verified tweets normally range between 100 to 300. There are some days wherein the number of tweets has reached 600 as well.

Tweets contain a large amount of noise, such as hashtags, URLs, and emotions. These characters make Twitter sentiment analysis a challenging assignment. Preprocessing of the data is a very important step as it decides the efficacy of the other steps down in line for sentiment analysis. For this we used regular expressions. Broadly speaking, regular expressions are a collection of patterns that can be used to identify certain kinds of text and to clean text of erroneous patterns. Regular expressions were used to remove the # tags, quotes and question marks were also removed as it causes biased results for sentiment analysis. The https links were removed as well using regular expressions.

VADER provides a compound sentiment score between -1.0 and 1.0 for the text fed to it. With the sentiments returned by VADER, the individual tweet sentiment scores are grouped into time-series. For each group, the sentiment mean is taken on the underlying tweets to indicate the average sentiment over a day. The compound sentiment scores of tweets are provided below:

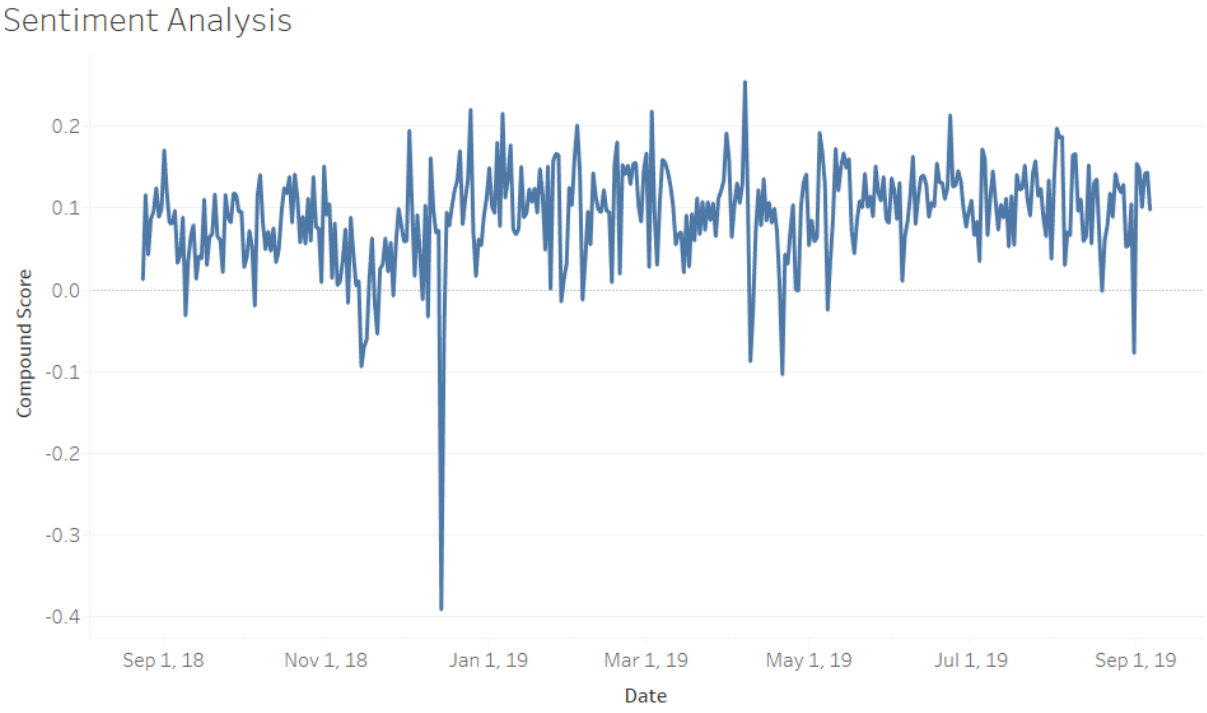


Fig 6.7 Daily variation of sentiment score of tweets

It is observed that 94% of the tweets have positive sentiment with compound scores between 0.0 and 0.2. Only around 6% of the tweets have negative sentiment. Overall, there is an affirmative environment around Bitcoin in social media.

7. Forecasting Models

Over the years, various forecasting models have been developed in the literature, but they have produced minimum accuracy in forecasting the bitcoin price. In this section, we shall discuss the forecasting methods using traditional ARIMA method and the latest LSTM method.

7.1 Prediction using ARIMA Method

Time series forecasting is quite different from other machine learning models because;
1. It is time dependent. So, the basic assumption of a linear regression model that the observations are independent doesn't hold in this case.

2. Along with an increasing or decreasing trend, most time series have some form of seasonality trends, i.e. variations specific to a particular time frame.

In this section, the most successful time series model i.e. ARIMA (Auto regression integrated moving average) price of Bitcoin. ARIMA is a class model that captures a suite of different standard temporal structures in time series data which include trend, seasonality, cycles, errors and nonstationary data. This allows it to exhibit dynamic temporal behaviour in a time sequence. The underlying principle in ARIMA model is to estimate the trend and seasonality in the series and remove those from the series to get a stationary series. Then statistical forecasting techniques can be implemented in this series. The final step would be to convert the forecasted values into the original scale by applying trend and seasonality constraints back. The various steps used in the method are as follows:

7.1.1 Plotting the time series

After importing the data as panda dataframe. The closing price of Bitcoin is plotted against the time. The variation in the price with time is shown below. Also, the dataset is broken down to identify the seasonality, trend, etc. in the time series data.

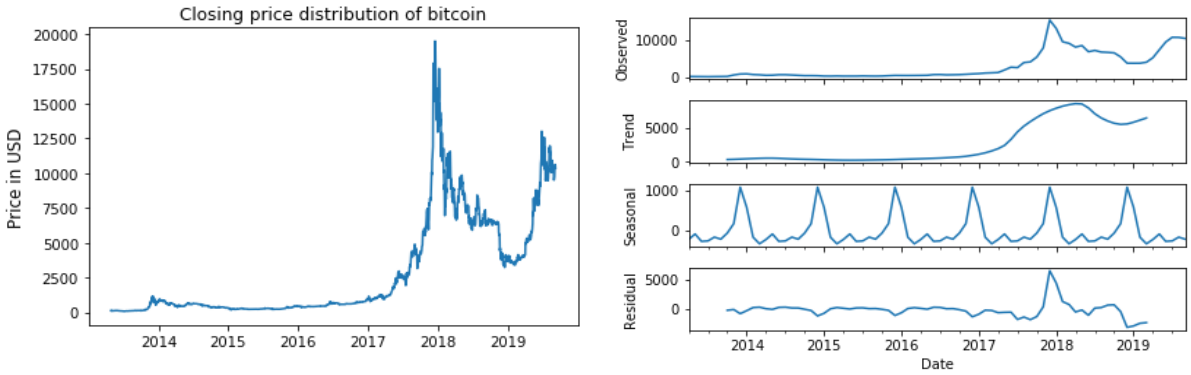


Fig 7.1 Historic price of Bitcoin and decomposition of price into various components

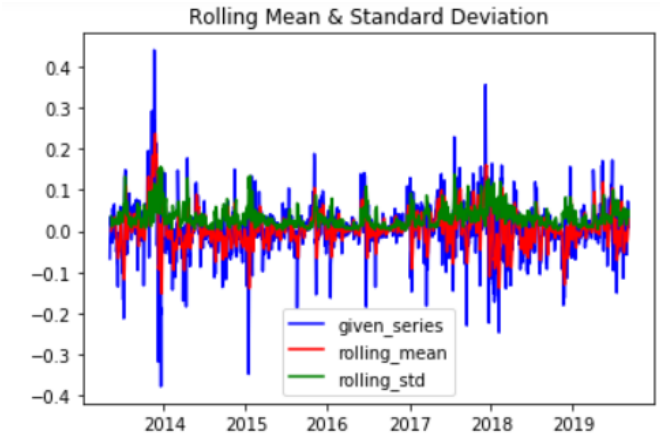
7.1.2 Testing for Stationarity

The Augmented Dicky Fuller test is used to find the presence of stationarity in the data. P-values from the test comes out to be 0.611. Since, the p value is greater than 0.05 the time series is non stationary. The data needs to be transformed to make it stationary.

7.1.3 Log transformation of series, and removal of trend and seasonality with differencing

Log transformation is used to un-skew highly skewed data. Thus, helping in forecasting process. Then, differencing is used to make the time series stationary in which the current value is subtracted with the previous values. Due to this the mean is stabilized and hence the chances of stationarity of time series are increased.

Augmented Dicky Fuller test is done on the transformed data. As the p-value is smaller than 0.05 now, the time series data is stationary. Time series forecasting model, ARIMA can be applied on the dataset for predicting the prices. A pictorial summary of the transformed data and the test is provided below.




```
Results of Dickey-Fuller Test:
Test Statistic                -8.570713e+00
p-value                        8.212241e-14
#Lags Used                     2.700000e+01
Number of Observations Used    2.289000e+03
Critical Value (1%)           -3.433210e+00
Critical Value (5%)           -2.862804e+00
Critical Value (10%)          -2.567443e+00
dtype: float64
```

Fig 7.2 Dickey Fuller Test of Bitcoin price

7.1.4 Applying ARIMA Model

Typically, an ARIMA model has two components: an autoregressive (AR) component and a moving average (MA) component. The AR component models association between the value of a variable at a specified time with its value in previous time(s), and the MA component models association between values of error term of a variable at a specified time with its error term value in previous time(s). The integrated (I) component comes into consideration when the time series becomes stationary after the first or second difference.

A standard notation is used for ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used. Parameters are defined as follows:

p: The number of lag observations included in the model, also called the lag order.
d: The number of times that the raw observations are differenced, also called the degree of differencing.
q: The size of the moving average window, also called the order of moving average.

Model parameters p, d and q are selected based upon the ACF (Auto correlation function), PACF (Partial auto correlation function) and unit root tests. Steps involved in the ARIMA method for this dataset are as follows:

a. Define the model by calling ARIMA() and passing in the p=0, d=1, and q=10 parameters.
b. The model is prepared on the training data by calling the fit() function.
c. Predictions can be made by calling the predict() function and specifying the index of the time or times to be predicted.

The Bitcoin pricing data from 24th August 2018 is used in the pricing model. Next-day Bitcoin price is predicted from 8th August 2019 until 6th September 2019. The period between 8th August and 6th September is used to test the model. The plot of actual and predicted price of Bitcoin as per the ARIMA model is provided below. It can be seen that the predicted pricing trend closely follows the actual pricing trend. A discussion on the accuracy of the model is provided in the next section.

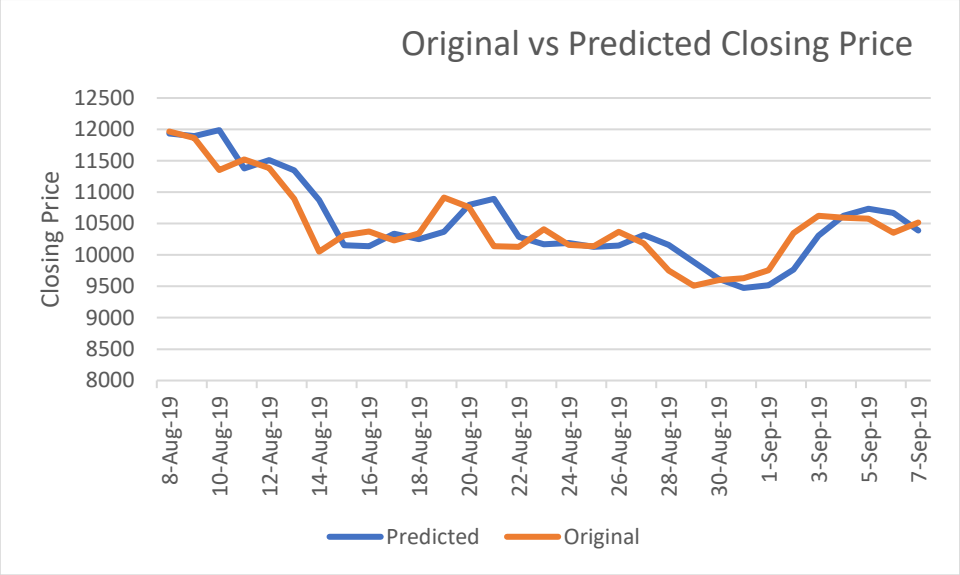


Fig 7.3 Original vs Predicted price using ARIMA model

7.2 Prediction using LSTM Method

The Long Short-Term Memory network or LSTM address the vanishing gradient problem common in the recurring neural network. This is a type of recurrent neural network used in deep learning because very large architectures can be successfully trained. LSTM allow the network to continue to learn over many time steps by upholding a more constant error. This allows the network to learn long-term reliance. An LSTM cell contains forget and remember gates which allow the cell to decide what information to block or pass based on its strength and importance. As a result, weak signals can be blocked which prevents vanishing gradient.

For Deep Learning backend system, we choose Tensor-flow, and Keras as the front-end layer of building neural networks fast. Pandas is used extensively for data related tasks, Numpy is utilized for matrix/vector operations and for storing training and test data sets, Scikit-learn (also known as: sklearn) is used for performing the min-max normalization. Lastly, Plotly is used for displaying the charts.

7.2.1 Splitting data into training and test data

The Bitcoin pricing data from 24th August 2018 is used in the pricing model. We have also used the Tweeter data and Google trends data from 24th August onwards in the LSTM Model. Next-day Bitcoin price is predicted from 8th August 2019 until 6th September 2019. The period between 8th August and 6th September is used to test the model.

7.2.2 Scenarios for modeling

We have used the following four scenarios in the next day price prediction:

Scenario 1: Effect of Tweet sentiment, Google trend data, volume and historic closing price on next day Bitcoin price

Scenario 2: Effect of Google trend data, volume and historic closing price on next day Bitcoin Price

Scenario 3: Effect of volume and historic closing price on next day Bitcoin Price

Scenario 4: Effect of historic closing price on the next day Bitcoin Price

We shall be predicting the price for all the above mentioned four scenarios and compare the results. We just run a quick correlation between all the variables in the combined dataset and the result is as follows:

	Close	Volume	Sentiment Score	Trend_Value
Close	1.000000	0.684179	0.134867	0.362308
Volume	0.684179	1.000000	0.210239	0.315712
Sentiment Score	0.134867	0.210239	1.000000	0.002935
Trend_Value	0.362308	0.315712	0.002935	1.000000

It is observed that Volume has a high correlation with Price. Sentiment score seems to have low correlation with the Closing Price.

7.2.3 Turning data into tensors

LSTM expects that the input is given in the form of a 3-dimensional vector of float values. A key feature of tensors is their shape, which in Python is a tuple of integers representing the dimensions of it along the 3 axes.

Furthermore, in LSTM the input layer is by design, specified from the input shape argument on the first hidden, these three dimensions of input shape are: Samples, Window size, Number of features. Hence, the data is turned suitably into tensors for all the four scenarios under consideration.

7.2.4 Architecture of the Network

We used the Sequential API of Keras, rather than the functional one. The overall architecture is as follows:

- 1 LSTM Layer: The LSTM layer is the inner one, and all the gates, mentioned at the very beginning are already implemented by Keras, with a default activation of hard-sigmoid [Keras2015]. The LSTM parameters are the number of neurons, and the input shape as discussed in the next section.
- 1 Dropout Layer: Typically, this is used before the Dense layer. As for Keras, a dropout can be added after any hidden layer, in our case it is after the LSTM.
- 1 Dense Layer: This is the regular fully connected layer.
- 1 Activation Layer: Because we are solving a regression problem, the last layer should give the linear combination of the activations of the previous layer with the weight vectors. Therefore, this activation is a linear one. Alternatively, it could be passed as a parameter to the previous Dense layer.

7.2.5 Hyperparameters

We used these hyperparameters based upon various trials and guesses.

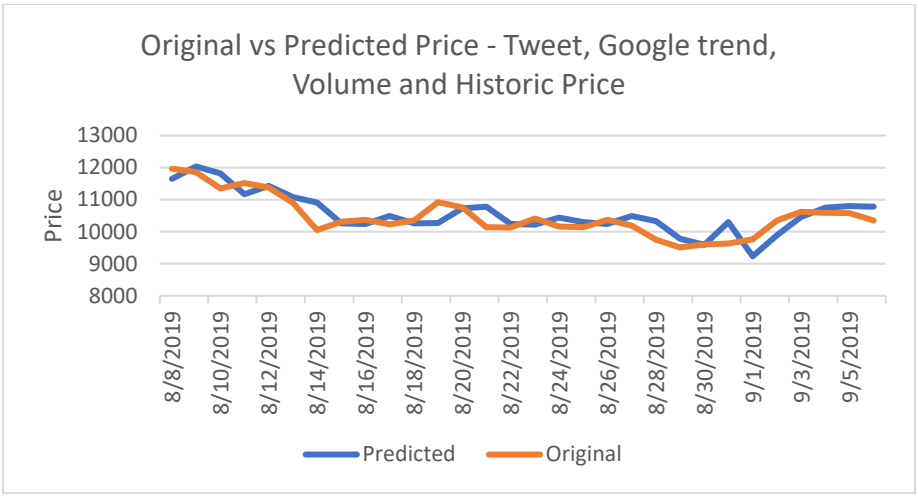
Parameters	Value
Optimizer	adam
Loss function	Mae (mean absolute error)
Activation function	ReLu
Number of neurons in hidden layer	100
Epochs	1000
Batch size	100

Fig 7.4 Hyperparameters used in the LSTM model

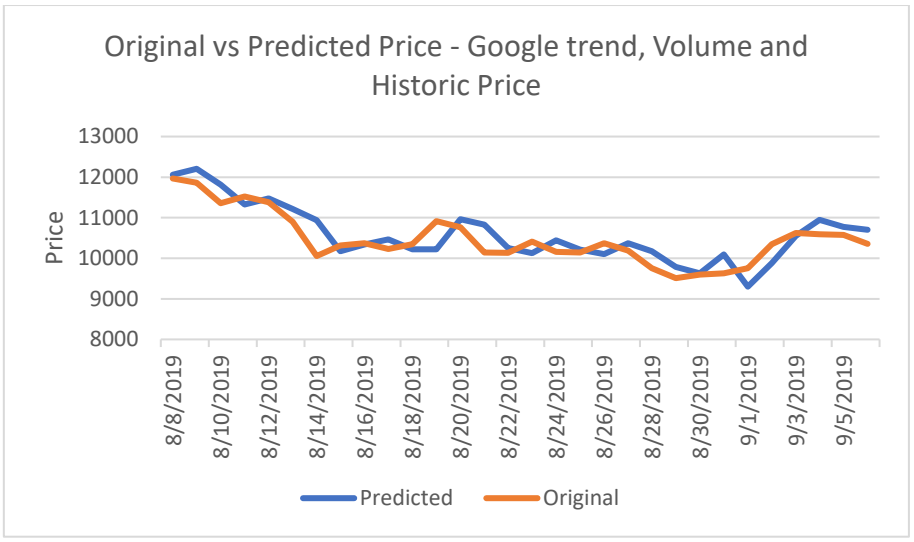
7.2.6 Applying LSTM Model

The plot of actual and predicted price of Bitcoin as per the LSTM model in all the four scenarios are provided below. More detailed analysis of these findings shall be done in the next section. We shall try to find out the scenario that lead to finding out the best price prediction.

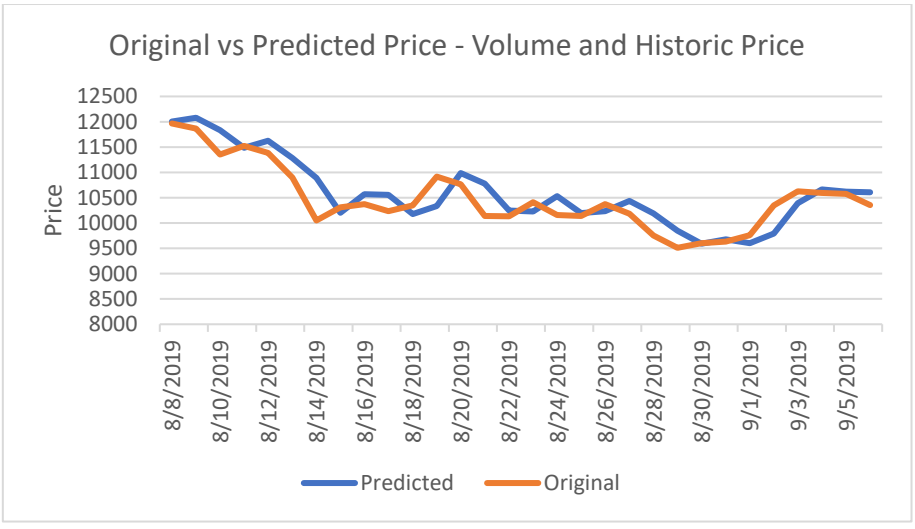
Scenario 1:



Scenario 2:



Scenario 3:



Scenario 4:

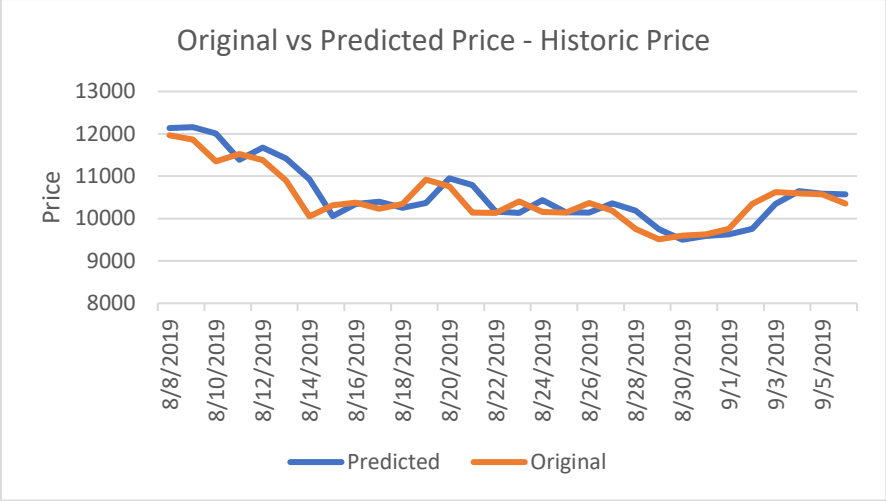


Fig 7.5 Original vs Predicted price under various scenarios using LSTM model

8. Result and Evaluation

The performance of both the models i.e. ARIMA and LSTM on the test set for the prediction of exact prices, was evaluated using the Root Mean Squared Error metric together with the Mean Absolute Percentage Error as error metrics, since MAPE is the preferred metric for forecast accuracy classification due to its simplicity (Hyndman & Koehler, 2006):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}i - yi)^2}{n}}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|true - predicted|}{|true|}$$

Moreover, the predicted values were stored in a Pandas dataframe and were compared to the real values in terms of percentage deviation. Subsequently, the direction (up, down) of the predicted variables were compared to the direction of the real Bitcoin price. The True Positive class corresponds to a correct prediction of the market going up were as True Negative corresponds to a correct prediction of the market not going up. The F1-score could be calculated using the precision and recall metric as follows:

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision}$$

Where precision is the model's ability to return only relevant instances (up/up):

$$Precision = \frac{TP}{TP + FP}$$

And recall is the model's ability to classify all relevant instances (up/up and down/down):

$$Recall = \frac{TP}{TP + FN}$$

		Class = Up	Class = Not up
True	Class = Up	True Positive	False Negative
	Class = Not up	False Positive	True Negative

The evaluation of ARIMA and LSTM model on the performance parameters discussed above is presented below [Test data from 8th August 2018 to 7th September 2019):

Parameter	ARIMA	LSTM			
		Scenario 1	Scenario 2	Scenario 3	Scenario 4
RMSE	331.62	367.36	357.52	327.19	342.92
MAPE, %	2.38	2.86	2.82	2.48	2.53
Accuracy, %	45%	43.3	50	60	50
Precision	0.37	0.35	0.45	0.5	0.41
Recall	0.43	0.38	0.38	0.58	0.38
F1-Score	0.40	0.37	0.41	0.54	0.40

Fig 8.1 Evaluation matrix for pricing using ARIMA and LSTM model

It is observed that ARIMA model has the lowest RMSE and MAPE. Whereas the LSTM model with historic Volume and Price data has the highest Accuracy, Precision, Recall and F1 score. LSTM model with Tweet Sentiment data, Google trend data, Bitcoin Volume and Historic Price data has the highest RMSE and MAPE. It also has the lowest Accuracy, Precision, Recall and F-1 score.

9. Discussions and Conclusions

9.1 Summary

With this study, we have been able to achieve a number of tasks that were identified in the beginning of this project. We started off with the collection of data from the internet/developer API in real time to develop the database for the project. We carried out the exploratory data analysis of the cryptocurrency data to select a suitable currency for carrying out the price prediction. Dashboards of the exploratory analysis is prepared in Tableau by keeping in mind the visualization techniques. Based upon various raw and derived variables, we selected Bitcoin for price prediction. We selected the most traditional method i.e. ARIMA method and most advanced method i.e. LSTM for price forecasting. In ARIMA, we only used the historic Bitcoin price for forecast. In LSTM, we also incorporated Tweet sentiment and Google trend data related to the keyword 'Bitcoin' in the forecasting model. We developed a system that forecasts the next day Bitcoin price. Real time forecasting is done for a period between 8th August 2019 and 6th September 2019. The forecasting results are compared in performance measures like accuracy, RMSE, MAPE, etc. Based upon the result of price forecast by ARIMA and LSTM, the following can be concluded:

- ARIMA method is more robust than the LSTM Method for predicting the price of the Bitcoin
- LSTM model with historic volume and price data is more robust than other methods for predicting the up/down movement in Bitcoin price.
- Tweet sentiments are not effective in predicting the next day Bitcoin price as is evident from the values of evaluation parameters for Scenario 1 vis-à-vis other Scenarios
- Google trend data on 'Bitcoin' keyword is also not useful in predicting the next day Bitcoin price forecast. Again, it is evident from the evaluation parameters for Scenario 2 vis-à-vis other Scenarios

- Historic price and historic volume of the Bitcoin are the best variable to predict the next day Bitcoin price
- Perhaps, Tweet sentiments and Google trend data are a result of the Bitcoin price changes/fluctuation and not the other way around

9.2 Real time next day Bitcoin price dashboard

A real time dashboard that depicts the next day Bitcoin price has been successfully built in Tableau. This dashboard gives a user a good estimate of future price movements in Bitcoin Price. A snapshot of the dashboard is provided below.

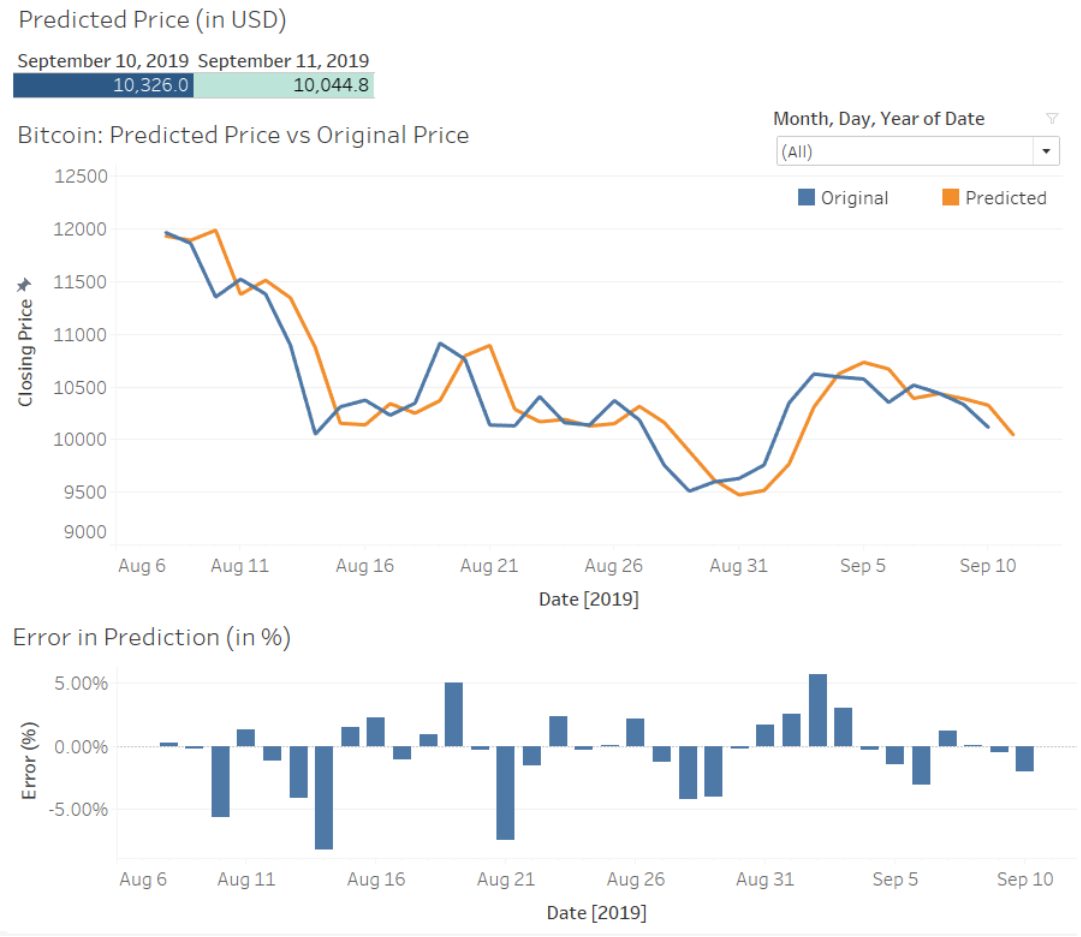


Fig 8.1 Tableau dashboard showing the Original and Predicted price of Bitcoin

9.3 Shortcomings

However, using deep neural networks, has provided us with a better understanding of Bitcoin, and LSTM architecture. The work in progress, includes implementing hyperparameter tuning, in order to get a more accurate network architecture. Also, other features can be considered (although from our experiments with Bitcoin, more features have not always led to better results). Microeconomic/Macroeconomic factors might be included in the model for a better predictive result. Furthermore, cryptocurrency prices are influenced by numerous other factors like: general news, sentiment on other internet forums or harmful events in the world of cryptocurrencies such as hacks. Regarding the data that was used in these analyses, the density of the data can be considered a shortcoming. When analyzing traditional stock or cryptocurrency markets, researchers often use day to day historical price data for these analyses since traditional stocks tend not to fluctuate as much. However, with the volatility of the cryptocurrency market, researchers might prefer more dense data

with smaller time intervals as input. Additionally, sentiment on internet forums can change by the hour so the same shortcoming of density might apply to the sentiment input data.

9.4 Future Research

To investigate if the simple LSTM model is capable of producing higher accuracy scores, future research can use denser data for cryptocurrency prices as well as Reddit sentiment. Instead of using day to day time intervals, future researchers can use smaller time intervals, like hour to hour, which leads to an increase in training and test data. Furthermore, with more computational power, the performance of a more complex model with more hidden layers and longer sequences should be considered. This increase in the complexity of the model may lead to a model that is able to take market lag and momentum into account.

10.References

- Kjuntia, S., Pattanayak, J., 2018. Adaptive market hypothesis and evolving predictability of bitcoin.
- Nadarajah, S., Chu, J., 2017. On the inefficiency of bitcoin
- Piñeiro Chousa, J., López-Cabarcos, A.A., Pérez-Pico, A.M., 2016. Examining the influence of stock market variables on microblogging sentiment
- Piñeiro Chousa, J., López-Cabarcos, A.A., Pérez-Pico, A.M., Ribeiro-Navarrete, B., 2018. Does social network sentiment influence the relationship between the S&P 500 and gold returns?
- Sun, A., Lachanski, M., Fabozzi, F.J., 2016. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. Int. Rev.
- Tiwari, A.K., Jana, R., Das, D., Roubaud, D., 2018. Informational efficiency of bitcoin–an extension.
- Urquhart, A., 2016. The inefficiency of Bitcoin.
- Urquhart, A., 2018. What causes the attention of Bitcoin?
- Shen, D., Urquhart, A., Wang, P., 2019. Does twitter predict Bitcoin?

11.Evaluation: Data Visualization Checklist

This section contains the feedback received from end user on the various visualisations used in this project. A data visualisation checklist is prepared for this purpose and attached to this report.

We also asked a few questions related to the visualisations from some prospective end users to gauge the reactive time and understand the effectiveness of our visualisation techniques. It is observed that users are able to identify and respond to specific questions in the desired time frame.

infogram

Data Visualization Checklist

Sanika Medankar

Conception

Before you begin visualizing:	Yes	No
Does your data have a purpose?	✓	
Do you know your target audience?	✓	
Is your data clean and organized?	✓	
Did you pick the right chart type?	✓	

Design

Once you've begun visualizing:	Yes	No
Are your axes labeled appropriately?	✓	
Is your text legible?	✓	
Is your title simple and informative?	✓	
Did you start the Y-axis at 0?	Depends	
Were you careful not to overload your chart?	✓	
Were you selective with your color scheme?	✓	
Did you use color to highlight key points?	✓	

Additions

Would your data visualization benefit from:	Yes	No
A legend?	✓	
A grid?	✓	
Interactivity?	✓	

Final Thoughts

Last but not least!	Yes	No
Can you understand the visualization in under 30 seconds?	✓	
Did you emphasize the most important data?	✓	
Did you avoid excessive color and 3D imagery?	✓	

infogram

Data Visualization Checklist

Jasmine Kaur

Conception

Before you begin visualizing:	Yes	No
Does your data have a purpose?	✓	
Do you know your target audience?	✓	
Is your data clean and organized?	✓	
Did you pick the right chart type?	✓	

Design

Once you've begun visualizing:	Yes	No
Are your axes labeled appropriately?	✓	
Is your text legible?	✓	
Is your title simple and informative?	✓	
Did you start the Y-axis at 0?	Depends	
Were you careful not to overload your chart?	✓	
Were you selective with your color scheme?	✓	
Did you use color to highlight key points?	✓	

Additions

Would your data visualization benefit from:	Yes	No
A legend?	✓	
A grid?	✓	
Interactivity?	✓	

Final Thoughts

Last but not least!	Yes	No
Can you understand the visualization in under 30 seconds?	✓	
Did you emphasize the most important data?	✓	
Did you avoid excessive color and 3D imagery?	✓	