Analysis of Twitter’s Impact On Cryptocurrency Trading

Cole Stankov
cstankov@sfu.ca
301295209

Karan Sachdeva
ksa128@sfu.ca
301368050

1. Introduction

Since the creation of bitcoin in 2008 cryptocurrencies have been the subject of much skepticism. However, as the technology matures it is becoming a rival to traditional currency across the world. Since the digital currency is available to purchase in many places, it has made it accessible for everyone. With the great success from both early and current investors into the currencies it has made people flock to invest drastically increasing the popularity as well the price. Due to the blockchain technology that is the backbone of the currency it promises secure and decentralized systems to help eliminate trust and security threats. Recently El Salvador has released they are approaching bitcoin as legal tender for the country.

The technology has also attracted many retailers pondering the idea of including certain crypto coins as a medium of exchange for companies such as Tesla. Although, when a tweet from Tesla’s owner, Elon Musk was published stating that they are suspending vehicle purchases using Bitcoin due to concerns of fossil fuels for Bitcoin mining, the digital currency, along with many others dropped drastically. Bitcoin’s price supposedly dropped $20,000 from that tweet alone. That is what leads us to the question, how much influence do famous or influential people have on crypto trading decisions through their tweets. In this report we attempted to predict whether the price of the top two cryptocurrencies, Etherum and Bitcoin, would increase or decrease with respect to tweets from influential people as well as analyze the data we retrieved.

2. Problem Statement

After seeing the supposed impact that Elon Musk tweets had on the cryptocurrency market we knew a deeper analysis of the situation must be made. In this project we will investigate the effects that influential people on twitter have on the cryptocurrency market. With that information we hope to see if we are able to predict the price change based on the sentiment analysis of the tweets.

3. Data Collection

We used three datasets in total for our analysis. The datasets are a daily Ethereum price dataset, a daily Bitcoin price dataset and one for the tweets of influential people for cryptocurrencies. Both the Ethereum and Bitcoin datasets were retrieved from yahoo finance

[1][2]. For both datasets the features were the date (‘Date’), the opening price (‘Open’), the price high for the day (‘High’), the price low for the day (‘Low’), the closing price (‘Close’), the price before the closing price (‘Adj Close’), and the total sum of trades (‘Volume’). The dataset included those features daily from September 15th 2014 for Bitcoin and from August 5th 2015 for Ethereum which was 2170 rows in total.

As for the twitter dataset, we initially researched people who are considered influential on twitter in the cryptocurrency community. What we found was a blog listing the top forty cryptocurrency twitter users [4] as well as a yahoo finance page listing seven users [3]. Some more research was conducted and it appeared that similar sites to the above listed the same people therefore, we decided to use all users along with a few extras such as Elon Musk totalling in 54 users. Once we found our list of users we scraped their tweets using a program called twint.

When we started scraping the users tweets we noticed two of them did not contain any tweets so those users were subsequently removed resulting in 52 users in total. We limited the tweets we wanted to scrape by searching for keywords in the tweets. The keywords that we used were ‘btc’, ‘bitcoin’, ‘eth’, ‘ethereum’, ‘crypto’, ‘cryptocurrency’ and ‘blockchain’. After scraping the tweets the data was saved to a tab separated csv which contained thirty-two features and 168,775 rows in total. The vast majority of those tweets occurred in 2021 as you can see from figure 3.2.

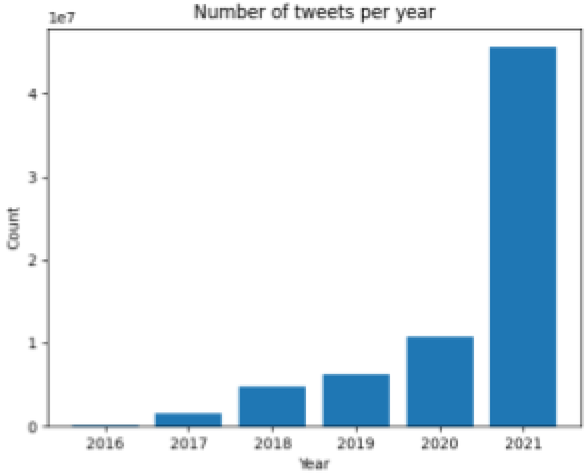

Figure 3.2: Number of Tweets Per Yea

4. Data Cleaning

As mentioned above the twitter dataset contained thirty-two features, however, the vast majority of the features were not relevant for the scope of this project. We ultimately decided on only using four features which are the following: 'date', 'name', 'tweet' and 'likes_count'. When we were attempting to process the twitter data we noticed that some of the tweets contained tabs. This resulted in problematic parsing of the dataset. Conclusively, we decided to remove those two lines all together. After, we examined the dataset for any ‘NaN’ values but there were zero in total.

As for the cryptocurrency datasets, we began with scanning for ‘NaN’ values. We found four rows that contained ‘NaN’ values for each dataset. After a closer examination on the rows we found that the only not ‘NaN’ values in the row were the dates. The dates for both the Ethereum and Bitcoin datasets were the same, which are the following: ‘2020-04-17’, ‘2020-10-09’, ‘2020-10-12’ and ‘2020-10-13’. All four rows in both datasets containing the “NaN” values were dropped. The price change was then calculated by taking the ‘Close’ values and subtracting them by the ‘Open’ values. If the price was increased then we set the ‘Price_change’ column to ‘1’ if it decreased it would be set to ‘-1’ and if there was no change it was set to ‘0’.

Lastly, after we ran the sentiment analysis on the twitter dataset which is discussed in section 5.1, the twitter dataset was joined to both the Ethereum and the Bitcoin datasets by Date. The two processed datasets were then saved in the processed folder.

5. Data Analysis

5.1 Sentiment Analysis

After the tweets were scraped we conducted a sentiment analysis on each tweet using the Vader sentiment analysis library. Vader is a lexicon and rule-based sentiment analysis tool that is specifically tuned for sentiment expressed in social media. We originally had decided to calculate the sentiment score for each tweet and assign a score of either ‘Positive’, ‘Negative’ or ‘Neutral’. However, we later changed the process to assign the polarity score which ranges from -1, which is considered a negative tweet to 1 which is considered a positive tweet. The scores were then assigned to a new column called ‘sentiment’. Due to having substantially more twitter data in comparison to the Ethereum and Bitcoin data as well as certain twitter users having considerably more likes on their tweets (As can be seen in figure 5.1), we decided to take the weighted average of the sentiment score in respect to the likes count associated with each tweet for a given day. To calculate the average sentiment score according to likes per day we used the following formula.

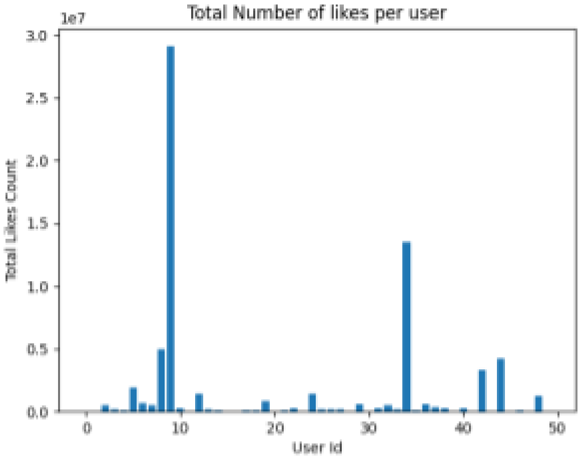

Figure 5.1 total number of likes per user

$$\frac{\sum_{grouped\ by\ Date} (Likes\ count * sentiment)}{\sum_{grouped\ by\ Date} likes\ count}$$

After the weighted average was calculated for each date all columns except for the weight average sentiment score were dropped. This allowed for a simpler merging process with the twitter and cryptocurrencies datasets.

5.2 Statistical Testing

After getting the weighted average of sentiment scores for all the tweets, we decided to conduct some statistical tests on our data. This would help us get an insight of what our data looks like and how it is distributed. The tests we conducted are as follows:

5.2.1 McNemar's Test

The McNemar test is used to determine if there are differences on a dichotomous dependent variable between two related groups. In our case, we used the McNemar test to determine whether there was any kind of similarity between the price change patterns of Bitcoin and Ethereum. The dependent variable was "coin price change", which has two categories: "increased" and "decreased" for two groups of Bitcoin and Ethereum. A 2 * 2 contingency table was created with the above categorical data in it and passed into the test function.

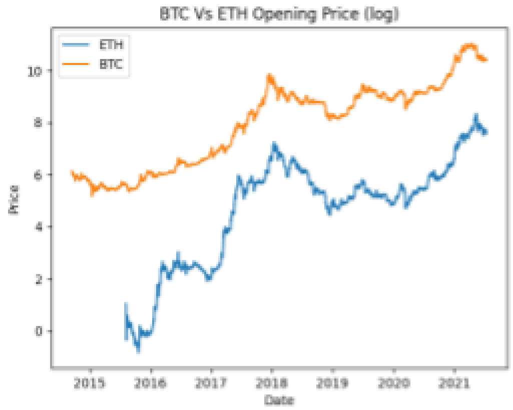

Figure 5.2: Bitcoin Vs Ethereum Opening Price

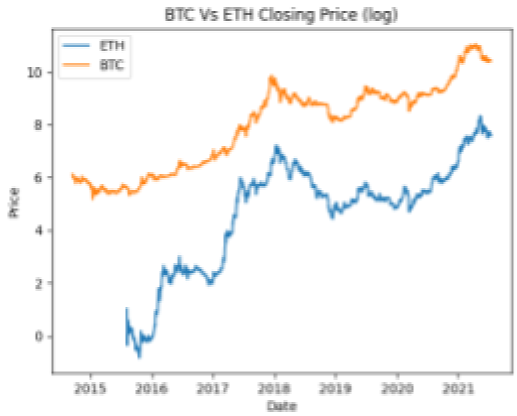

Figure 5.3: Bitcoin Vs Ethereum Closing Price

5.2.2 Chi-Squared Tests

Next, we decided to conduct two chi-squared tests (one for each currency) on sentiment scores versus the price change behavior. The sentiment scores were assigned a value of -1, 0 or 1 depending on if they were negative, neutral or positive. Similarly, The price change behavior was assigned a value of -1 or 1 depending on if it decreased or increased respectively. The aim was to find out if both the distributions were independent or not.

5.2.3 Comparison of Price Change Behavior For Each Date

After testing the overall price change pattern, we decided to dig deeper and find out price change patterns on a day to day basis of BTC and ETH. We found the number of days where both the stocks either decreased or increased along with the rest of the days on which the price change was opposite.

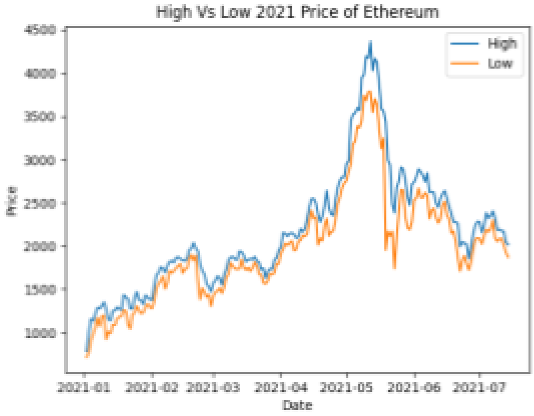

Figure 5.4: Ethereum High Vs Low Price 2021

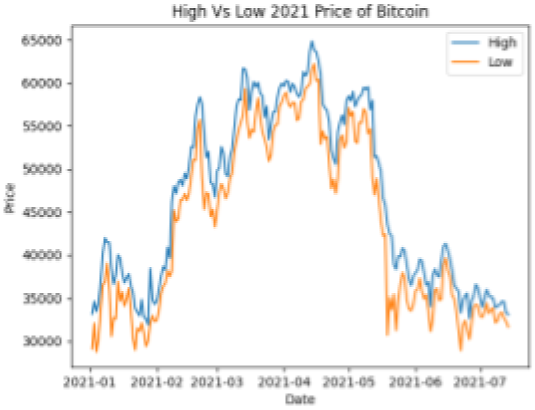

Figure 5.5: Bitcoin High Vs Low Price 2021

5.2.4 Comparison of Sentiment and Price Change For Each Date

Inspired by the previous day to day comparison, we decided to analyze the relationship between sentiment scores and price change behavior in a similar fashion. The number of days where positive and negative sentiment was seen with an increase and decrease in stock price respectively was recorded along with the days with an opposite scenario. This comparison would help us get an idea if twitter sentiment that day impacts stock prices in any manner.

5.3 Machine Learning

Machine learning has many applications, one of which is to forecast time series. One of the most interesting (or perhaps most profitable) time series to predict are, arguably, stock prices[5]. In our case, we built four kinds of machine learning models for our Ethereum and Bitcoin datasets to predict if the stock price will increase or decrease based on the features of tweet sentiment, opening price, maximum price, minimum price and the volume of the respective type of shares that day. In order to optimally solve the machine learning problem, we decided to conduct hyperparameter tuning on our models and get accurate predictions. This was accomplished with the help of GridSearchCV function from the sklearn library. The entire dataset was split into train and test data using the train_test_split function from the sklearn library.

The four machine learning models were based upon Linear Regression and three classifiers namely Gaussian Naive Bayes Classifier, Random Forest Classifier and the Multilayer Perceptron Classifier. After building the models, we used them to predict prices of a certain set of data and compared the results from the model from the real results we already had. For our linear regression model we altered our approach slightly in comparison to the other models. For the other models we used them to predict if the price would increase or decrease, however, with the linear regression model we used it to predict the price the coin would be. In doing so we had to drop the maximum price and the minimum price features leaving behind only the tweet sentiment, opening price and the Volume.

6. Results

6.1 McNemar's Test

After comparing the “price change” feature for the two groups of BTC and ETH, the result was the following:

```
p-value for McNemar's Test: 9.242595204427927e-274
```

After getting such a small p-value it became evident that the price for Bitcoin and Ethereum changes in a similar fashion.

6.2 Chi-Squared Tests

After comparing the sentiment scores with the price change behavior we got the following results for Bitconin and Ethereum:

```
p-value for chi squared test (ETH):  5.117679909512467e-112
p-value for chi squared test (BTC):  1.0515549685179357e-89
```

Again after getting such small p-values, it was concluded that the sentiment scores and price change behavior had similar distributions. This further peaked our interest in the relationship of tweets and stock trading and motivated us to build several models and make predictions.

6.3 Comparison of Price Change Behavior For Each Date

After comparing price change patterns on a day to day basis of BTC and ETH we found that **71.75%** of the days, both the stocks shared similar behavior of an overall increase or decrease. That left us with the remaining **28.25%** of the days when both the stocks had opposite price change patterns.

6.4 Comparison of Sentiment and Price Change For Each Date

Similar to the last day to day comparison, we analyzed the relationship between sentiment scores and price change behavior for both the stocks. For Ethereum, in approximately **51.1%** of the days, positive and negative sentiment was seen with an increase and decrease in stock price respectively. That leaves us with the remaining **48.8%** where sentiment and price change behavior do not have a similar pattern. For Bitcoin, in approximately **54%** of the days, positive and negative sentiment was seen with an increase and decrease in stock price respectively. That leaves us with the remaining **46%** where sentiment and price change behavior do not have a similar pattern.

6.5 Machine Learning

After the hyper tuning of the models using GridsearchCV the models were then run using the processed data on both the Ethereum and the Bitcoin datasets. Each of the three models (Random forest, multilayer perceptron and Gaussian Naive Bayes) were run five times in total to collect the accuracy. The accuracy was then plotted in a graph to help visualize if our model is overfitting the data. The following is the results for the models overfitting:

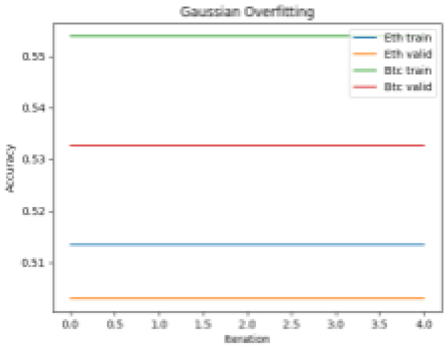

Figure 6.1: Gaussian overfitting

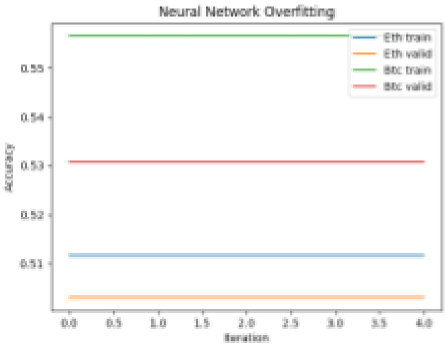

Figure 6.2: Gaussian Overfitting

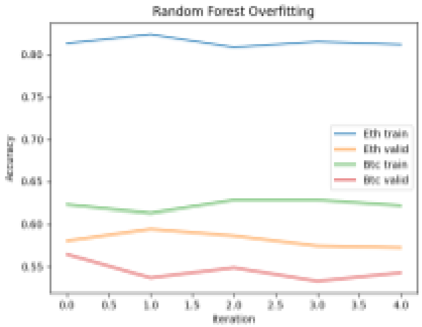

Figure 6.3: Random Forest Overfitting

As can be seen from the above graphs (Ethereum train accuracy: blue, Ethereum validation accuracy: yellow, Bitcoin train: green and Bitcoin validation: red) there appears to be a substantial amount of overfitting for the random forest classifier specifically for the Ethereum dataset. The other two models have less overfitting however, for the Bitcoin dataset it appears that all three models contain a little more overfitting.

The model's evaluation metrics were then calculated using sklearn's classification report. Although the most important metric to evaluate is the accuracy of our models, the precision and recall of the model is also interesting to examine as well. The following is the first evaluation metrics for the models on each dataset.

Models	Accuracy	Label	Precision	Recall	f1-Score
Random Forest	ETH: 0.58	Increased	Eth: 0.59 Btc: 0.70	Eth: 0.51 Btc: 0.13	Eth: 0.55 Btc: 0.21
	BTC: 0.56	Decreased	Eth: 0.57 Btc: 0.55	Eth: 0.65 Btc: 0.95	Eth: 0.61 Btc: 0.70

MLP	ETH: 0.50	Increased	Eth: 1.00 Btc: 1.00	Eth: 0.00 Btc: 0.00	Eth: 0.00 Btc: 0.00
	BTC: 0.53	Decreased	Eth: 0.50 Btc: 0.53	Eth: 1.00 Btc: 1.00	Eth: 0.67 Btc: 0.69
GaussianNB	ETH: 0.50	Increased	Eth: 0.50 Btc: 0.51	Eth: 0.00 Btc: 0.13	Eth: 0.01 Btc: 0.21
	BTC: 0.53	Decreased	Eth: 0.50 Btc: 0.54	Eth: 1.00 Btc: 0.89	Eth: 0.67 Btc: 0.67

As we can see from the classification report the accuracy for all the models regardless of the dataset is around 50% where random forest tends to have the higher values of 56% (BTC) and 58% (ETH). The other metric that we found interesting was the recall metric which tells us how much the measure of the model predicted the true positive. As we can see from the random forest models for the Ethereum dataset it was able to predict both increased and decreased correctly close to the same rate (Decreased: 65%, Increased: 51%). Whereas for the bitcoin dataset it was able to successfully predict when the data was going to decrease a lot more in comparison to increase.

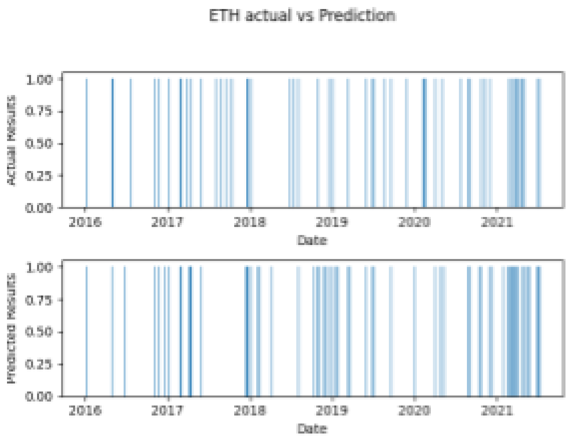

Figure 6.4: Ethereum predictions vs Actual values (Random Forest)

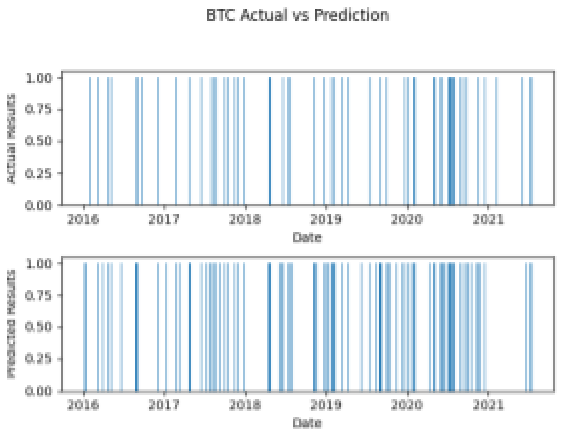

Figure 6.5: Bitcoin Predictions Vs Actual Values (Random Forest)

As for our linear regression model, it had substantial results for predicting the price of the coin for that given day. For both dataset the linear regression was able to predict the price 99% of the time for both validation and training data.The following is the linear regression predictions versus the actual closing price for both datasets:

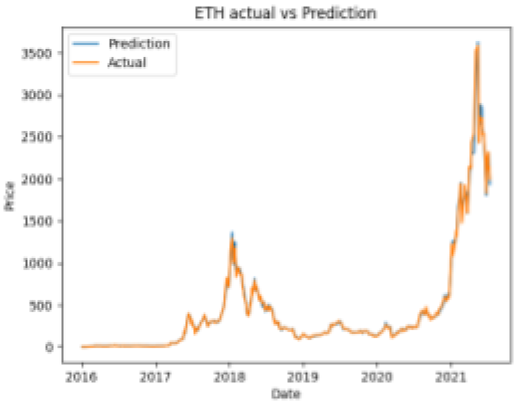

Figure 6.6: Ethereum Predictions Vs Actual Price (Linear Regression)

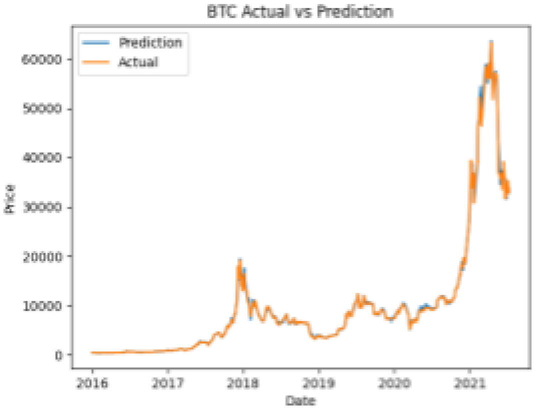

Figure 6.7: Bitcoin Predictions Vs Actual Price (Linear Regression)

7. Conclusion

We analyzed our data using various techniques such as sentiment analysis, statistical testing and machine learning in order to conclude whether or not twitter has an impact on cryptocurrency stock trading. After evaluating the results, we came to the conclusion that we don't have enough proof to make that statement with 100% confidence. We saw some relation between the sentiment and price behavior after conducting the testing and building the models but it is not enough. Even though the Chi-Squared and McNemar's tests showed promising results at start, the results from the rest of the tests and machine learning models were not conclusive enough.

8. Reflection

After reviewing our completed project we believe that we accomplished what we set out to complete in our problem statement. One thing we wish we could have added on to the project if we were given more time is allow the models to run on predicting the price similar to what we did with the linear regression model. The limitations that we faced werre largely to do with the tweet data. We found it difficult to find relative influential people who are as relevant in the cryptocurrency world as Elon Musk is. The twitter accounts that we did use have nowhere near as many likes and followers that Elon does.

9. Project Experience Summary

Cole Stankov

- Worked in a team of two to predict cryptocurrency behavior based upon twitter sentiment
- Collected, visualized and cleaned Bitcoin, Ethereum and twitter data to help analyze data

- Utilized Sklearn's Gaussian Naive Bayesian, random forest and MLP classifiers to build models to predict the price change of Bitcoin and Ethereum.
- Hypertunned our models using gridsearchCv to achieve the ideal parameters for each model given the processed data.

Karan Sachdeva

- Worked in a team of two to predict cryptocurrency behavior based upon twitter sentiment
- Collected, cleaned and analyzed Bitcoin, Ethereum and Twitter data with help of various Python Libraries
- Conducted sentiment analysis of various tweets to assign sentiment scores which in turn were used to predict stock price behavior
- Used machine learning classifiers namely Naive Bayesian, MLP and Random forest to build accurate models

10. Github Link

- https://github.com/cstankov/CMPT353-CryptoAnaylsisProject

11. References

[1] Ethereum USD (ETH-USD) Price, News, Quote & History. (2021, August 11). Retrieved From https://finance.yahoo.com/quote/ETH-USD/history/

[2] Bitcoin USD (BTC-USD) Price, News, Quote & History. (2021, August 11). Retrieved From https://finance.yahoo.com/quote/BTC-USD/history/

[3] 7 Bitcoin And Cryptocurrency Accounts To Follow On Twitter. (n.d.). Retrieved From https://finance.yahoo.com/news/7-bitcoin-cryptocurrency-accounts-twitter-134519804.html

[4] Here Are 40 Crypto Twitter Accounts That Really Matter. (2019, January 31). Retrieved From https://consensys.net/blog/news/i-read-crypto-twitter-for-hours-daily-here-are-the-40-accounts-that-really-matter/

[5] Machine Learning Techniques applied to Stock Price Prediction Retrieved From https://towardsdatascience.com/machine-learning-techniques-applied-to-stock-price-prediction-6c1994da8001