

Predicting Cryptocurrency Movements with News Sentiment Analysis Using K-Nearest Neighbour

WQD 7005 Data Mining Milestone 6

Name: Lim Kaomin, Leslie

Matric No: WQD 180076

1) Introduction

With the advent of Internet Age since the late 1990s, many wonderful innovations and technologies had emerged as a product of what the internet offered. Among those technologies are Cryptocurrencies which is defined as a virtual currency or digital currency that can be used in electronic payment systems, substituting physical or fiat currencies that we were so familiar. Bitcoin, which was regarded as one of the most popular Cryptocurrency was created in 2009 as the first decentralized cryptocurrency. Since then, numerous other cryptocurrencies have been created. They are frequently called “altcoins” as a blend of bitcoin alternative. Bitcoin and its derivatives use decentralized control as opposed to centralized electronic money/centralized banking systems. The decentralized control is related to the use of bitcoin’s blockchain transaction database in the role of a distributed ledger (A.O.Victor,2017). In short, a cryptocurrency is a virtual coinage system that functions much like a standard currency, enabling users to provide virtual payment for goods and services free of a central trusted authority. Cryptocurrencies rely on the transmission of digital information, utilizing cryptographic methods to ensure legitimate, unique transactions. Bitcoin took the digital coin market one step further, decentralizing the currency and freeing it from hierarchical power structures. Instead, individuals and businesses transact with the coin electronically on a peer-to-peer network (R.Farell, 2015).

1.1) History of Cryptocurrencies

Cryptocurrencies has a history dating back to the year 1998 when the first known cryptocurrency ‘b-money’ was published anonymously by Wei Dai. Shortly after, another cryptocurrency called ‘bit god’ was created by Nick Szabo but was never implemented. Although bit gold is considered the first precursor to bitcoin, cryptocurrency pioneer David Chaum’s company DigiCash (a company founded in 1989 which attempted to innovate digital currency), Wei Dai’s b-money (a conceptual system published in 1998 which Satoshi cites it in the Bitcoin white paper), and “e-gold” (a centralized digital currency that started in 1996) are all notable early mentions (cryptocurrencyfacts..com).The first cryptocurrency, Bitcoin, was launched in 2009 by under the pseudonym Satoshi Nakamoto. It caught wide attention beginning in 2011, and various altcoins – a general name for all other cryptocurrencies post-Bitcoin – soon appeared (R.Farell, 2015).

Litecoin was released in 2011, gaining modest success and enjoying the highest cryptocurrency market cap after Bitcoin until it was overtaken by Ripple on October 4th, 2014. Litecoin modified Bitcoin’s protocol, increasing transaction speed with the idea that it would be more appropriate for day-to-day transactions. Ripple, launched in 2013, introduced an entirely unique model to that used by Bitcoin and currently maintains the second highest market capitalization of approximately \$255 million (April 22). Another notable coin in the evolutionary chain of cryptocurrency, Peercoin, employs a revolutionary technological development to secure and sustain its coinage. Peercoin merges the proof-of-Work (PoW) technology used by Bitcoin and Litecoin along with its own mechanism, proof-of-stake (PoS), to employ a hybrid network security mechanism (R.Farell, 2015).

Outline

- 1) Introduction
 - 1.1) History of Cryptocurrencies
- 2) Analysis Goals
- 3) Data Pipeline
- 4) Research Methodology
- 5) Data Collection
 - 5.1) Cryptocurrency Web Scraping
 - 5.2) News Headlines Web Scraping
- 6) Data Pre-Processing
 - 6.1) Data pre-processing for the cryptocurrency prices
 - 6.2) Data pre-processing of news headlines
 - 6.2.1) News Sentiment Score Analysis
 - 6.2.1.1) Lowercasing
 - 6.2.1.2) Special characters
 - 6.2.1.3) Stopwords
 - 6.2.1.4) Performing Sentiment Score with TextBlob
- 7) Data Visualization
- 8) Feature Engineering
 - 8.1) Cryptocurrency Dataset
 - 8.2) News Sentiment Score Dataset
- 9) Preparation of Dataset Before Modelling
 - 9.1) Merging Datasets
 - 9.2) Create Three Separate Datasets
 - 9.2.1) Original Dataset
 - 9.2.2) Standard Scaled Dataset
 - 9.2.3) Feature Selected Dataset
 - 9.3) Training and Testing Datasets
- 10) Modelling
 - 10.1) K Nearest Neighbour Classifier
 - 10.2) K Nearest Neighbour Regression
- 11) Results and Performance Evaluation
- 12) Discussion
- 13) Conclusion
- 14) References

2) Analysis Goals

Due to the trade wars on going between US and China, stagnating economies, political crisis and low interest rates or negative interest rates in countries in the European Union and Japan. Many investors are turning to digital assets namely digital currencies or cryptocurrencies to secure their wealth. Cryptocurrencies are gaining popularity as many people are buying it as a form of investment. However, just like stock prices, cryptocurrency prices are extremely volatile and prices depends on the market sentiment. Thus, predicting cryptocurrency price movement to determine which cryptocurrency would be worth buying or selling prove important for the many people seeking to navigate the market. Hence, this project aims to apply machine learning techniques to predict cryptocurrency prices by analysing news sentiment analysis.

Analysis Goals:

- i) Predict effectively cryptocurrency price movements with machine learning algorithm
- ii) Predict cryptocurrency prices as close as possible to the actual price value
- iii) Determine if analysing news sentiment improves the machine learning model

3) Data Pipeline

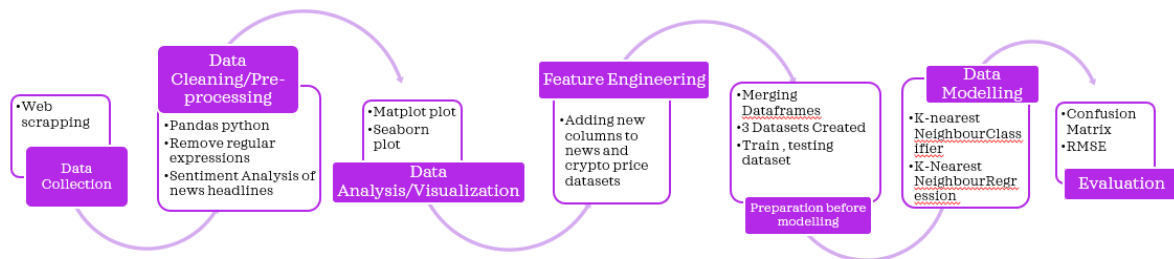


Figure 1: Data Pipeline

4) Research Methodology

The entire research methodology used in this project which covers from Data Collection to Results and Performance Evaluation was done in Python Programming. The software used to code were done in Jupyter Notebook.

The *Data Collection* would cover the process of how the data were collected. Two websites were used as the data source, one website contained the list of cryptocurrencies with the prices while a second website which contained news headlines. The data collection from the websites was done using web scraping.

Data Pre-processing steps involved the process of data cleaning such as removing regular expressions in numerical values and data normalizations of the news headlines. News sentiment Scoring was also done in this process using the TextBlob package.

Data Visualization was done using packages such as matplotlib and seaborn plot to gain a better understanding of the datasets.

Feature Engineering aims to create useful features that can enhance the machine learning model.

Preparation of Datasets before Modelling involves merging the cryptocurrency price dataframe and news scoring dataframes, creating three separate datasets for different results and splitting training and test dataset.

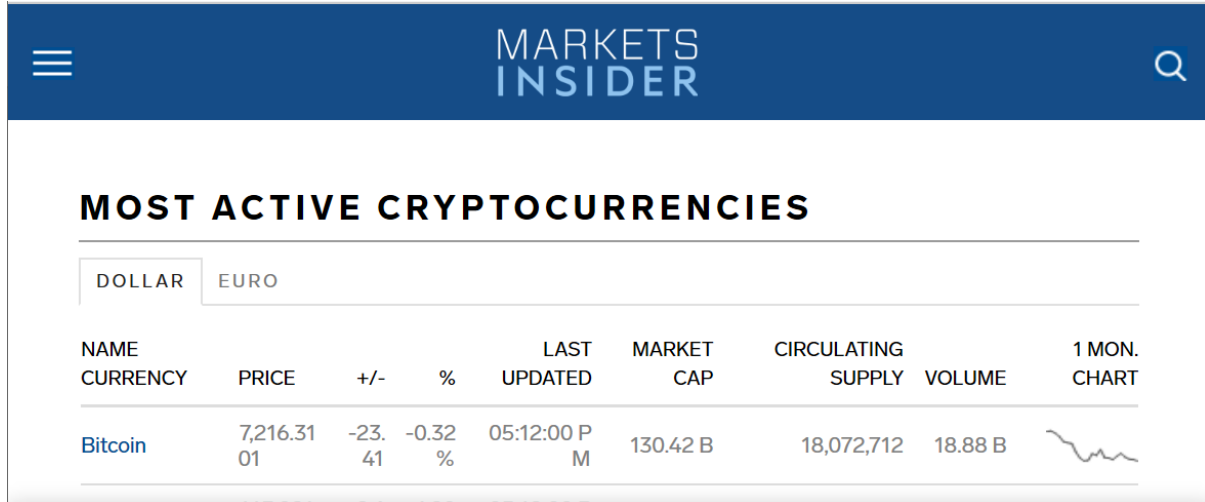
Modelling was done using the K Nearest Neighbour Classifier (KNN Classifier) for 'Up' or 'Down' labels and K Nearest Neighbour Regression (KNN Regression) for predicting prices.

Evaluation was done to evaluate the model's performance. Confusion Matrix was used for the categorical variables and Root-Mean-Square-Error was used to evaluate the continuous variables.

5) Data Collection

5.1) Cryptocurrency Web Scraping

The cryptocurrency price data are scraped from website <https://markets.businessinsider.com/cryptocurrencies>



The screenshot shows the 'MARKETS INSIDER' website header with a hamburger menu on the left and a search icon on the right. Below the header, the section 'MOST ACTIVE CRYPTOCURRENCIES' is displayed. There are two tabs: 'DOLLAR' (selected) and 'EURO'. The table below lists various cryptocurrencies with columns for Name, Currency, Price, +/-, %, Last Updated, Market Cap, Circulating Supply, Volume, and a 1 Mon. Chart.


NAME	CURRENCY	PRICE	+/-	%	LAST UPDATED	MARKET CAP	CIRCULATING SUPPLY	VOLUME	1 MON. CHART
Bitcoin		7,216.3101	-23.41	-0.32%	05:12:00 PM	130.42 B	18,072,712	18.88 B	

Figure 2: Front page of the website markets.businessinsider.com/cryptocurrencies

The website contains a list of many cryptocurrencies however this project would only focus on 20 cryptocurrencies such as Bitcoin, Ethereum, Ripple and so on. From the website, data that were crawled mainly are Open price, Previous Close, Date, Time, Daily High, Daily Low, 52 Week High, 52 Week Low, Max Supply, Market Cap, Volume, and Circulation Supply. The data scraped range from the date 6th October 2019 to 5th December 2019 which was a period of 45 days.

5.2) News Headlines Web Scraping

The news headlines were scraped from website as well <https://cointelegraph.com/tags/cryptocurrencies>

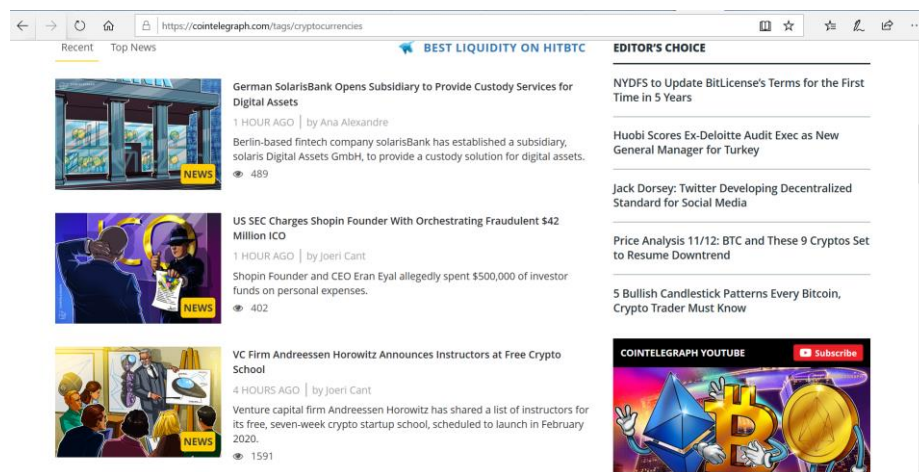


Figure 3: Front page of the website cointelegraph.com/tags/cryptocurrencies

The website mainly contains news about cryptocurrencies thus making it perfect as a source of news headlines sentiment analysis to be applied later. The news headlines were mainly scraped for their article date, article headlines and sub headlines. The date range of the news ranges from 4th October 2019 to 5th December 2019.

The data from the websites are scraped using the BeautifulSoup package from Python. The process was mainly done in Jupyter Notebook.

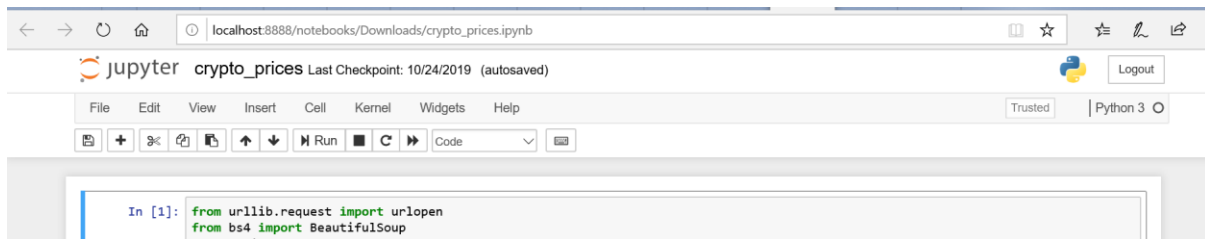


Figure 4: Interface of Jupyter Notebook

6) Data-Preprocessing

Data pre-processing involves a collection of steps which helps to purify the data and extract the useful and remove the insignificant information. Data obtained from real-world is incomplete, inconsistent and it also contains numerous errors. Thus, to counter this issue with the data, we are using data pre-processing which aids in removing discrepancies in names and related problems.

6.1) Data pre-processing for the cryptocurrency prices.

```
data = pd.read_csv('cryptocurrency_prices.csv')
data.head()
```

	CryptoCurrency Type	Trade Time(US Time)	Daily High	Trade Date	Daily Low	Open	52-Week High	Prev.Close	52-Week Low	Market Cap	Volume	Circ.Supply	Max.Supply	Price Direction
0	BTC	06:13AM	8,231.15	6/10/2019	7,931.36	8,168.93	13,829.07	8,168.24	3,178.33	141.67 B	43.36 B	17.78 M	21.00 M	Up
1	ETH	06:13AM	177.6656	6/10/2019	172.4018	177.1539	362.819	177.0681	82.427	18.46 B	15.88 B	106.65 M	NaN	Up
2	XRP	06:13AM	0.2569	6/10/2019	0.2478	0.2543	0.5655	0.2543	0.2198	10.76 B	3.57 B	42.57 B	100.00 B	Down
3	BCC	06:13AM	226.3218	6/10/2019	218.3572	222.8465	642.8143	222.8102	74.7157	3.97 B	3.42 B	17.86 M	21.00 M	Up
4	LTC	06:13AM	57.3156	6/10/2019	55.4602	56.8784	145.8824	56.8571	22.662	3.49 B	5.60 B	62.40 M	84.00 M	Up

Figure 5: Cryptocurrency price dataset

From the dataset head, it can be seen the columns names with their respective values. Viewing the dataset shape would also be useful to get an idea on the number of rows and columns of the dataset. Viewing the info of the dataset to understand the dataset better such as if the values are a string, float or integer.

```
data.shape
(900, 16)

data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 900 entries, 0 to 899
Data columns (total 16 columns):
CryptoCurrency Type    900 non-null object
Trade Time(US Time)    900 non-null object
Daily High              900 non-null object
Trade Date              900 non-null object
Daily Low               900 non-null object
Open                   900 non-null object
52-Week High            900 non-null object
Prev.Close              900 non-null object
52-Week Low             900 non-null object
Market Cap              900 non-null object
Volume                 900 non-null object
Circ.Supply             900 non-null object
Max.Supply              501 non-null object
Price Direction         900 non-null object
Price Change(Open-Prev.Close) 900 non-null float64
Price Change(Daily High-Daily Low) 900 non-null float64
dtypes: float64(2), object(14)
memory usage: 112.6+ KB
```

Figure 6

From the dataset head it can be observed that columns Daily High, Daily Low, Open, 52-Week High, Prev.Close and 52-Week Low are objects, this is due to comma values. Thus, the commas must be removed before converting the string object into float. The reason float was used because it has decimal points thus integer might not be suitable.

The removal of commas can be done using regular expressions package called 're', and then the conversion of object to float can be done.

```
col = ['Daily High', 'Daily Low', 'Open', '52-Week High', 'Prev.Close', '52-Week Low']
data[col] = data[col].replace({' ','.'}, regex=True)
data[col] = data[col].astype(float)
data.head()
```

	CryptoCurrency Type	Trade Time(US Time)	Daily High	Trade Date	Daily Low	Open	52-Week High	Prev.Close	52-Week Low	Market Cap	Volume	Circ.Supply	Max.Supply	Price Direction
0	BTC	06:13AM	8231.1500	6/10/2019	7931.3600	8168.9300	13829.0700	8168.2400	3178.3300	141.67 B	43.36 B	17.78 M	21.00 M	Up
1	ETH	06:13AM	177.6656	6/10/2019	172.4018	177.1539	362.8190	177.0681	82.4270	18.46 B	15.88 B	106.65 M	NaN	Up
2	XRP	06:13AM	0.2569	6/10/2019	0.2478	0.2543	0.5655	0.2543	0.2198	10.76 B	3.57 B	42.57 B	100.00 B	Down
3	BCC	06:13AM	226.3218	6/10/2019	218.3572	222.8465	642.8143	222.8102	74.7157	3.97 B	3.42 B	17.86 M	21.00 M	Up
4	LTC	06:13AM	57.3156	6/10/2019	55.4602	56.8784	145.8824	56.8571	22.6620	3.49 B	5.60 B	62.40 M	84.00 M	Up

Figure 7: Remove commas and converting to integers

Viewing to see if the conversion successfully took place.

```
data.dtypes
CryptoCurrency Type    object
Trade Time(US Time)    object
Daily High              float64
Trade Date              object
Daily Low               float64
Open                   float64
52-Week High            float64
Prev.Close              float64
52-Week Low             float64
Market Cap              object
Volume                 object
Circ.Supply             object
Max.Supply              object
Price Direction         object
Price Change(Open-Prev.Close) float64
Price Change(Daily High-Daily Low) float64
dtype: object
```

Figure 8: Checking data types

6.2) Data pre-processing of news headlines.

```
data = pd.read_csv('cryptocurrency_news.csv')
```

```
data.shape
```

```
(675, 3)
```

```
data.head()
```

	Article Headlines	Article Date	Article Content
0	Crypto News From the Spanish-Speaking World: S...	2019-10-06	Cointelegraph en Español presents a weekly di...
1	Crypto News From the German-Speaking World: Se...	2019-10-06	This week's selected cryptocurrency and block...
2	Liechtenstein's Parliament Unanimously Approve...	2019-10-05	Liechtenstein's Parliament unanimously passed...
3	Former US Army Interpreter Gets 30 Years for D...	2019-10-05	Former Iraqi interpreter for the U.S. militar...
4	Tether and Bitfinex Expect a Market Manipulati...	2019-10-05	Tether and its affiliate exchange Bitfinex an...

Figure 9: News Headlines Dataset

Drop any duplicates from the dataset

```
#Drop duplicates from dataset  
data = data.drop_duplicates()
```

```
data.shape
```

```
(522, 3)
```

Figure 10: Drop duplicates from News Headlines dataset

Checking for data type in news headlines dataset

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 522 entries, 0 to 670  
Data columns (total 3 columns):  
Article Headlines    522 non-null object  
Article Date         522 non-null object  
Article Content      522 non-null object  
dtypes: object(3)  
memory usage: 16.3+ KB
```

Figure 11: Check data type

Check for any missing values in news headlines dataset.

```
data.isnull().any()
```

```
Article Headlines    False
Article Date         False
Article Content       False
dtype: bool
```

Figure 12: Check missing values

The news headlines dataset does not have much pre-processing to do as it does not have any missing values and the dataset contains only string values.

6.2.1) News Sentiment Score Analysis

What sentiment analysis is used for

Sentiment analysis is useful for quickly gaining insights using large volumes of text data. In addition to the customer feedback analysis use case. Another two examples of where sentiment analysis can be useful. One example is stock trading companies who trawl the internet for news. Here, the algorithms can detect particular companies who show a positive sentiment in news articles. In this example, this can mean a significant financial return, as this may trigger people to buy more of company stocks. Having access to this type of data, may mean that traders have the opportunity to make decisions before the market has had time to react. For the purpose of this project, TextBlob would be used to perform sentiment analysis on news headlines.

6.1.2.1) Lowercasing

Before moving forward to calculate the sentiment scores for each review it is important to pre-process the textual data. Lowercasing helps in the process of normalization which is an important step to keep the words in a uniform manner.

```
#Change Article Content to lower case
data['Article Content'] = data['Article Content'].str.lower()
data.head()
```

	index	Article Headlines	Article Date	Article Content
0	0	Crypto News From the Spanish-Speaking World: S...	2019-10-06	cointelegraph en español presents a weekly di...
1	1	Crypto News From the German-Speaking World: Se...	2019-10-06	this week's selected cryptocurrency and block...
2	2	Liechtenstein's Parliament Unanimously Approve...	2019-10-05	liechtenstein's parliament unanimously passed...
3	3	Former US Army Interpreter Gets 30 Years for D...	2019-10-05	former iraqi interpreter for the u.s. militar...
4	4	Tether and Bitfinex Expect a Market Manipulati...	2019-10-05	tether and its affiliate exchange bitfinex an...

Figure 13: Lowercasing Article Content column from News Headlines dataset

6.2.1.2) Special characters

Special characters are non-alphabetic and non-numeric values such as {!,@#\$\$%^*()~;:/<>|+_-[]?}. Dealing with numbers is straightforward but special characters can be sometimes tricky. During tokenization, special characters create their own tokens and again not helpful for any algorithm, likewise, numbers. Thus, these special characters might interfere in the sentiment scoring process and it must be removed.

```
#Remove special characters from Article Content
data['Article Content'] = data['Article Content'].str.replace('[^$\w\s]', '')
data
```

	index	Article Headlines	Article Date	Article Content
0	0	Crypto News From the Spanish-Speaking World: S...	2019-10-06	cointelegraph en español presents a weekly di...
1	1	Crypto News From the German-Speaking World: Se...	2019-10-06	this weeks selected cryptocurrency and blockc...
2	2	Liechtenstein's Parliament Unanimously Approve...	2019-10-05	liechtensteins parliament unanimously passed ...
3	3	Former US Army Interpreter Gets 30 Years for D...	2019-10-05	former iraqi interpreter for the us military ...
4	4	Tether and Bitfinex Expect a Market Manipulati...	2019-10-05	tether and its affiliate exchange bitfinex an...

Figure 14: Removing special characters of Article Content column from News Headlines dataset

6.2.1.3) Stopwords

Stop-words being most commonly used in the English language; however, these words have no predictive power in reality. Words such as I, me, myself, he, she, they, our, mine, you, yours etc.

```
#Remove stopwords from Article Content
from nltk.corpus import stopwords
stop = stopwords.words('english')
data['Article Content'] = data['Article Content'].apply(lambda x: " ".join(x for x in x.split() if x not in stop))
data
```

	index	Article Headlines	Article Date	Article Content
0	0	Crypto News From the Spanish-Speaking World: S...	2019-10-06	cointelegraph en español presents weekly diges...
1	1	Crypto News From the German-Speaking World: Se...	2019-10-06	weeks selected cryptocurrency blockchain news ...
2	2	Liechtenstein's Parliament Unanimously Approve...	2019-10-05	liechtensteins parliament unanimously passed b...
3	3	Former US Army Interpreter Gets 30 Years for D...	2019-10-05	former iraqi interpreter us military sentenced...
4	4	Tether and Bitfinex Expect a Market Manipulati...	2019-10-05	tether affiliate exchange bitfinex announce pr...

Figure 15: Remove stopwords of Article Content column from News Headlines dataset

6.2.1.4) Performing Sentiment Score with TextBlob

TextBlob is another excellent open-source library for performing NLP tasks with ease, including sentiment analysis. It also a sentiment lexicon (in the form of an XML file) which it leverages to give both polarity and subjectivity scores. Typically, the scores have a normalized scale as compare to AFINN. The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very

subjective. After performing the sentiment scores, the data frame would look as the figure given below. The data frame contains columns such as 'Sentiment_Score_25%_negative' which is the First Quartile for the negative score, 'Sentiment_Score_count_negative' which is the number of negative news headlines, 'Sentiment_Score_max_negative' and 'Sentiment_Score_min_negative' for max score and min score for the news headlines respectively.

	Article_Date	Sentiment_Score_25%_negative	Sentiment_Score_25%_neutral	Sentiment_Score_25%_positive	Sentiment_Score_50%_negative	Sentiment_Score_50%_positive
0	2019-10-04	-0.22500	0.0	0.1010	-0.2250	0.1010
1	2019-10-05	-0.07125	0.0	0.1405	-0.0555	0.1405
2	2019-10-06	-0.41450	0.0	0.2500	-0.3290	0.2500
3	2019-10-07	-0.20000	0.0	0.0810	-0.2000	0.0810
4	2019-10-08	NaN	0.0	0.1420	NaN	0.1420

Figure 16: News Sentiment Score Dataset

7) Data Visualizations

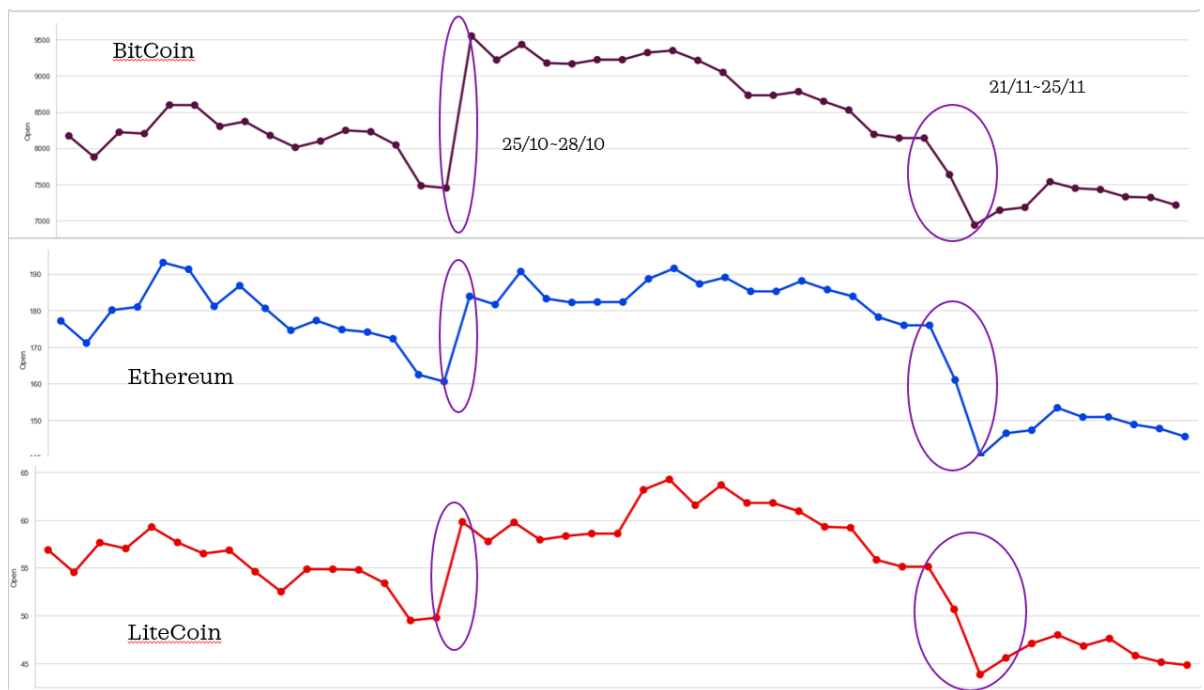


Figure 17: Line Chart of Open price for Bitcoin, Ethereum and Litecoin

A line chart of Bitcoin, Ethereum and Litecoin Open Price was plotted. The reason for choosing the three cryptocurrencies was just done arbitrarily due to the popularity of the cryptocurrencies. From the line chart, the Open Price for the three cryptocurrencies showed similar patterns on the date 25th October 2019 to 28th October 2019 whereby they showed an increase in price together. This was an indicator that something might have happened to excite the cryptocurrency markets. Similarly starting from 21st November 2019 to 25th November 2019, the trend showed a drop in prices which indicated that something might have happened that scared the cryptocurrency markets.

Article Headlines		Article Date
225	Lebanon Banks' Shutdown Is 'Most Potent Case' for Crypto: Nassim Taleb	2019-10-25
226	Crypto Lending Platform Nexo Lowers Rates on Instant Credit Lines	2019-10-25
227	Michael Novogratz's Galaxy Firm Is Launching New Bitcoin Funds	2019-10-25
228	Crypto Exchange Kraken Announces WebSockets Private API Is Live	2019-10-25
229	Number of Americans Owning Crypto Doubled in 2019: Finder	2019-10-25

**POSITIVE
NEWS
HEADLINES**

Article Headlines		Article Date
240	China: Forex Regulator Warns Against Illegal Crypto Cross-Border Flows	2019-10-28
256	Kraken Cryptocurrency Exchange to Add Support for OmiseGo and PAX Gold	2019-10-28
257	Privacy Vs. Security, Do Authorities Monitor Every Crypto Transaction?	2019-10-28
258	China to Be First to Launch Digital Currency, Says Think Tank Exec	2019-10-28
259	Bakkt Teases Launch of Consumer Payments App Scheduled for 2020	2019-10-28
260	Libra Might Become Unrecognizable by Navigating Regulatory Concerns	2019-10-28
261	EOS Holds Top Spot, Bitcoin 11th in China's Latest Crypto Rankings	2019-10-28

Figure 18: Positive news headlines on 25th October 2019 and 28th October 2019

Upon further investigation, it can be seen after querying for news headlines from 25th October 2019 to 28th October 2019, the main news that sent confidence to the cryptocurrency markets was the headline that reads 'China to be the First to Launch Digital Currency' which is the recognition of the Chinese government to use digital currency.

Article Headlines		Article Date
510	OneCoin Fugitive Cryptoqueen Allegedly Paid \$50 M to Lawyer to Launder Funds	2019-11-21
511	US Think Tank Releases Report on Investigation Into Illicit Transactions on the Dark Web	2019-11-21
512	PayPal CEO Hods Bitcoin and Only Bitcoin	2019-11-21
526	Trading App for Kraken Futures Is Now Available on iOS and Android	2019-11-21
527	Bithumb Quashes Shanghai Office Closure Rumors After Binance Denial	2019-11-21
528	Backfire in Argentina: Citizens Want BTC Over Peso Amid USD Crackdown	2019-11-21
529	Lawyer Found Guilty of Money Laundering for OneCoin's Cryptoqueen	2019-11-21
530	Yahoo Finance Adds CoinMarketCap's Crypto Prices to Its Website	2019-11-21
531	Binance Denies Police Raids, Vows Existence of Shanghai Offices	2019-11-21
532	New Bill Would Put Facebook's Libra Stablecoin Under US Securities Law	2019-11-21
533	Venezuela Cuts Petro's Backing from 5B Barrels of Oil to 30M: Reuters	2019-11-21
534	Google, Facebook Take on Banking Duties, Crypto Shrugged to the Side?	2019-11-21
535	Assassin's Creed Dev Becomes Block Producer for EOS-Based Gaming Startup	2019-11-21
536	Sacramento Kings CTO: Fans Quit Spending Bitcoin When the Price Hiked	2019-11-21
537	'Mystery Man' Gets Berlin Shop to Drop BitPay, Accept Bitcoin Directly	2019-11-21
538	Grayscale: 84% of Q3 Interest Came From Non-Crypto Hedge Funds	2019-11-21

**NEGATIVE
NEWS
HEADLINES**

Figure 19: Negative news headlines on 21st November 2019

Similarly, from 21st November 2019 to 25th November 2019, negative news headlines such as shown in Figure sent cryptocurrencies prices tumbling as the market loses confidence. Adverse news such as a particular digital currency trading company or the company shareholder was found involved in money laundering or corruption, country political crisis creates a low confidence environment for crypto investors and buyers.

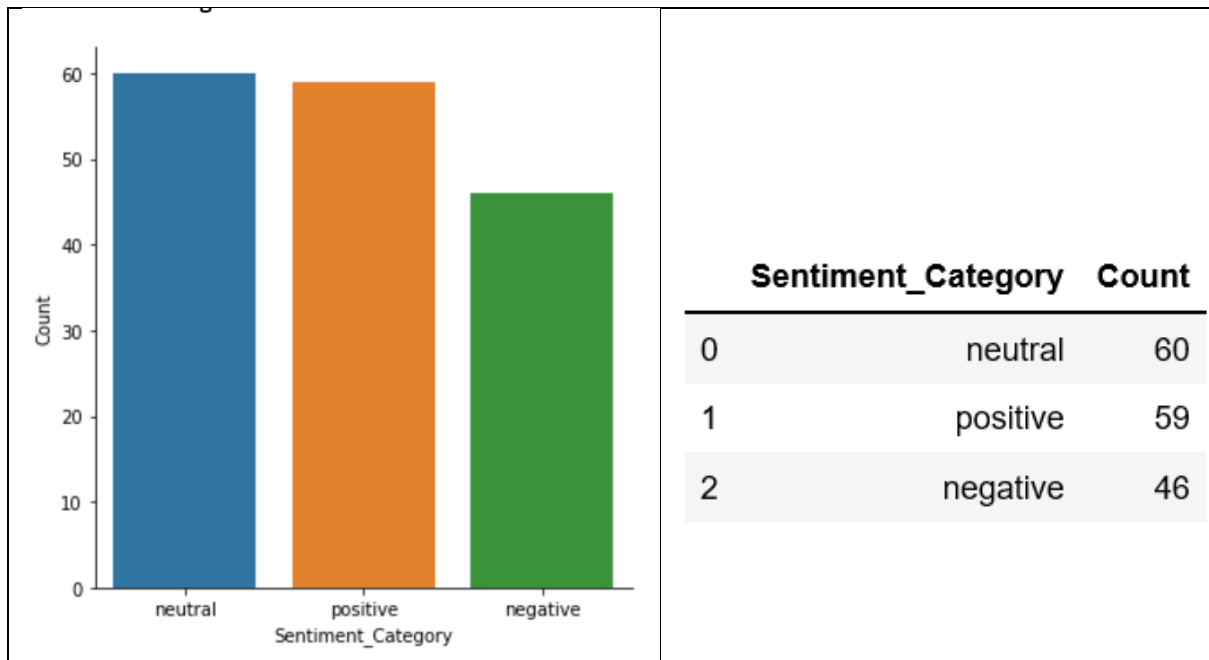


Figure 20: Bar Chart plot and table of total sentiment category

From the bar chart in Figure 20, it was shown that the dominant headlines throughout from 4th October 2019 to 5th December 2019 period belong to the category of 'Neutral' headline, followed by 'Positive' and 'Negative'. This showed that most of the headlines were in fact 'Neutral' type.

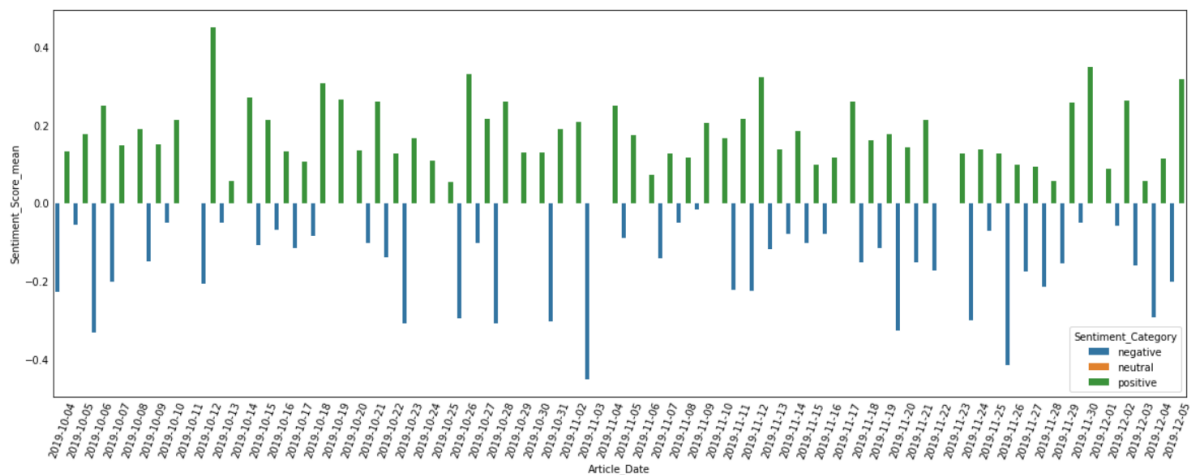


Figure 21: Bar Chart plot of sentiment scores

Figure shows the distribution of sentiment scores from 4th October 2019 to 5th December 2019. It can be observed that the positive news score peaked at 12th October 2019 however after reviewing the Cryptocurrency prices showed no large increase in the prices. The lowest negative sentiment score recorded was in 2nd November but it did not cause huge price falls in the cryptocurrency markets.

8) Feature Engineering

8.1) Cryptocurrency Dataset

Feature engineering aims to increase the predictive power of learning algorithms by creating features from raw data that help facilitate the machine learning process. Feature Engineering can be done better if one has domain knowledge which helps to identify useful features from the dataset and create useful features. This newly created features would help the machine learning model to

Daily High	Trade Date	Daily Low	Open	52-Week High	Prev.Close	52-Week Low	Market Cap	Volume	Circ.Supply	Max.Supply	Price Direction	Price Change(Open-Prev.Close)	Price Change(Daily High-Daily Low)
8,231.15	6/10/2019	7,931.36	8,168.93	13,829.07	8,168.24	3,178.33	141.67 B	43.36 B	17.78 M	21.00 M	Up	0.69	299.79
177.6656	6/10/2019	172.4018	177.1539	362.819	177.0681	82.427	18.46 B	15.88 B	106.65 M	NaN	Up	0.09	5.26
0.2569	6/10/2019	0.2478	0.2543	0.5655	0.2543	0.2198	10.76 B	3.57 B	42.57 B	100.00 B	Down	0.00	0.01
226.3218	6/10/2019	218.3572	222.8465	642.8143	222.8102	74.7157	3.97 B	3.42 B	17.86 M	21.00 M	Up	0.04	7.96
57.3156	6/10/2019	55.4602	56.8784	145.8824	56.8571	22.662	3.49 B	5.60 B	62.40 M	84.00 M	Up	0.02	1.86

Figure 22: Cryptocurrency price dataset after Feature Engineering

For example in the cryptocurrency dataset, the new feature 'Price Change(Daily High-Daily Low)' was created from the columns of difference between 'Daily High' and 'Daily Low'.

8.2) News Sentiment Score Dataset

Feature engineering was also done on the news headlines dataset. The feature engineering was done during the sentiment analysis stage whereby new columns were created. The new feature columns that were created are listed down below.

Some description of the newly created columns:

- 'Sentiment_Score_25%_negative' – 25% sentiment score for negative news
- 'Sentiment_Score_count_negative' – Number of negative news
- 'Sentiment_Score_max_negative' – Max negative news score
- 'Sentiment_Score_min_negative' – Min negative news score
- 'Sentiment_Score_std_negative' – Standard Deviation negative news score

```
Index(['Article_Date', 'Sentiment_Score_25%_negative',  
      'Sentiment_Score_25%_neutral', 'Sentiment_Score_25%_positive',  
      'Sentiment_Score_50%_negative', 'Sentiment_Score_50%_neutral',  
      'Sentiment_Score_50%_positive', 'Sentiment_Score_75%_negative',  
      'Sentiment_Score_75%_neutral', 'Sentiment_Score_75%_positive',  
      'Sentiment_Score_count_negative', 'Sentiment_Score_count_neutral',  
      'Sentiment_Score_count_positive', 'Sentiment_Score_max_negative',  
      'Sentiment_Score_max_neutral', 'Sentiment_Score_max_positive',  
      'Sentiment_Score_mean_negative', 'Sentiment_Score_mean_neutral',  
      'Sentiment_Score_mean_positive', 'Sentiment_Score_min_negative',  
      'Sentiment_Score_min_neutral', 'Sentiment_Score_min_positive',  
      'Sentiment_Score_std_negative', 'Sentiment_Score_std_neutral',  
      'Sentiment_Score_std_positive'],  
      dtype='object')
```

Figure 23: List of features created for News Sentiment Score Dataset

9) Preparation of Dataset Before Modelling

Preparation Steps:

- 1) Merging of Data Frames
- 2) Create three separate datasets used, as Original, Standard Scaled and Feature Selected
- 3) Training and Testing Dataset

9.1) Merging of Data Frames

Before the dataset can be passed into the machine learning model, several pre-processing steps still needed to be done.

From the news headlines dataset, missing values are filled up with the value zero. Certain columns such as 'Sentiment_Score_25%_neutral', 'Sentiment_Score_50%_neutral' and 'Sentiment_Score_75%_neutral' were also dropped because the columns only contains zero values.

```
news_df = news_df.fillna(0)
news_df = news_df.loc[:, (news_df != 0).any(axis=0)]
```

Figure 24: Data-preprocessing of news sentiment score dataset before merging

Merging the two datasets with pandas merge using the Trade Date and Article Date as the primary key.

```
merge_df = prices_df.merge(news_df, how='left', left_on=['Trade Date'], right_on=['Article Date'])
```

Figure 25: Merging Cryptocurrency dataset with News Sentiment Score dataset

After merging the two data frames, empty rows are created. Thus, filling up the missing values are done again and then checked if are there still any missing values that exist.

```
merge_df = merge_df.fillna(0)
merge_df.isnull().any()
```

Figure 26: Filling missing values after dataset merging

Furthermore, the target variable 'Price Direction' needed to be converted to a numerical value in order for the machine learning model to use it. The Figure 26 and 27 shows the 'Price Direction' column before and after conversion.

Daily High	Trade Date	Daily Low	Open	52-Week High	Prev.Close	52-Week Low	Market Cap	Volume	Circ.Supply	Max.Supply	Price Direction	C
3231.1500	6/10/2019	7931.3600	8168.9300	13829.0700	8168.2400	3178.3300	141.67 B	43.36 B	17.78 M	21.00 M	Up	
177.6656	6/10/2019	172.4018	177.1539	362.8190	177.0681	82.4270	18.46 B	15.88 B	106.65 M	NaN	Up	
0.2569	6/10/2019	0.2478	0.2543	0.5655	0.2543	0.2198	10.76 B	3.57 B	42.57 B	100.00 B	Down	
226.3218	6/10/2019	218.3572	222.8465	642.8143	222.8102	74.7157	3.97 B	3.42 B	17.86 M	21.00 M	Up	
57.3156	6/10/2019	55.4602	56.8784	145.8824	56.8571	22.6620	3.49 B	5.60 B	62.40 M	84.00 M	Up	

Figure 26: 'Price Direction' before conversion of categorical to numerical values.

Daily High	Trade Date	Daily Low	Open	52-Week High	Prev.Close	52-Week Low	Market Cap	Volume	Circ.Supply	Max.Supply	Price Direction
8231.1500	6/10/2019	7931.3600	8168.9300	13829.0700	8168.2400	3178.3300	141.67 B	43.36 B	17.78 M	21.00 M	1
177.6656	6/10/2019	172.4018	177.1539	362.8190	177.0681	82.4270	18.46 B	15.88 B	106.65 M	0	1
0.2569	6/10/2019	0.2478	0.2543	0.5655	0.2543	0.2198	10.76 B	3.57 B	42.57 B	100.00 B	-1
226.3218	6/10/2019	218.3572	222.8465	642.8143	222.8102	74.7157	3.97 B	3.42 B	17.86 M	21.00 M	1
57.3156	6/10/2019	55.4602	56.8784	145.8824	56.8571	22.6620	3.49 B	5.60 B	62.40 M	84.00 M	1

Figure 27: 'Price Direction' after conversion of categorical to numerical values.

9.2) Create Three Separate Datasets

For this project, three separate datasets are used with is:

- 1) Original Dataset
- 2) Standard Scaled Dataset
- 3) Feature Selected Dataset

9.2.1) Original Dataset

The Original dataset was created by dropping columns that are not useful for prediction purposes. Columns that were dropped included 'Cryptocurrency Type', 'Trade Time', 'Trade Date', 'Article Date', 'Market Cap', 'Volume', 'Circ.Supply', and 'Max.Supply'. These columns were dropped either because the columns have constant values or are just irrelevant. The column 'Open' price was also dropped as this would be a target variable to be used during modelling.

Once the columns that are irrelevant were dropped, we already have our first dataset which will be called the Original Dataset.

	Daily High	Daily Low	52-Week High	Prev.Close	52-Week Low	Price Change(Daily High-Daily Low)	Sentiment_Score_25%_negative	Sentiment_Score_25%_positive	Sentiment_Score_50%_
0	8231.1500	7931.3600	13829.0700	8168.2400	3178.3300	299.79	0.0	0.0	
1	177.6656	172.4018	362.8190	177.0681	82.4270	5.26	0.0	0.0	
2	0.2569	0.2478	0.5655	0.2543	0.2198	0.01	0.0	0.0	
3	226.3218	218.3572	642.8143	222.8102	74.7157	7.96	0.0	0.0	
4	57.3156	55.4602	145.8824	56.8571	22.6620	1.86	0.0	0.0	

Figure 28: Original Dataset

9.2.2) Standard Scaled Dataset

Scaling on the dataset was also done in order to standardize the values in the dataset. This is due to large range of values between different cryptocurrency prices such as Bitcoin having 'Open' price of 7000 to 8000 but other cryptocurrencies having 'Open' price of less than 1000 or less than 100. The Figure shows the scaled dataset using standard scaler.

	Daily High	Daily Low	52-Week High	Prev.Close	52-Week Low	Price Change(Daily High-Daily Low)	Sentiment_Score_25%_negative	Sentiment_Score_25%_positive	Sentiment_Score_50%_ne
0	4.193910	4.241201	4.353429	4.271505	4.355725	2.858229	0.0	0.0	
1	-0.155316	-0.153241	-0.140833	-0.153115	-0.138064	-0.175465	0.0	0.0	
2	-0.251124	-0.250744	-0.261732	-0.251015	-0.257390	-0.229540	0.0	0.0	
3	-0.129039	-0.127213	-0.047386	-0.127789	-0.149257	-0.147654	0.0	0.0	
4	-0.220310	-0.219473	-0.213234	-0.219675	-0.224815	-0.210485	0.0	0.0	

Figure 29: Standard Scaled Dataset

9.2.3) Feature Selected Dataset

The Original dataset contains 24 columns to be crunched by the machine learning algorithm. Even though machine learning models improve with more data being crunched, however if the dataset dimensionality was too huge it could decrease the performance of the model, this is known as the ‘Curse of Dimensionality’. To overcome this problem, feature selection can be done to reduce the size of the dataset. Feature Selection selects the variables which can explain most of the variance of the dataset.

For this project, the number of columns chosen was done using Principal Component Analysis (PCA).

Principal component analysis is a statistical technique that analyses the interrelationships between the column variables and explain the variables in terms of a smaller number of variables, called principal components, with a minimum loss of information. The number of components was determined by plotting a PCA plot. From the Figure , the number of components chosen was 5.

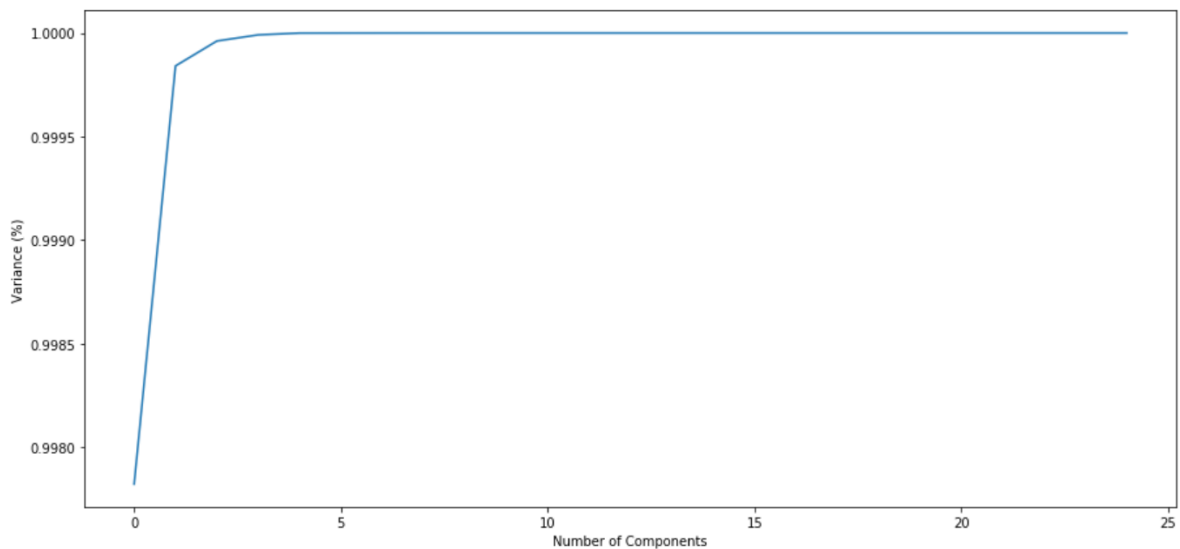


Figure 29: PCA Plot

After the number of components was specified, feature selection of which five columns should be chosen was done using `f_classif` and `SelectkBest` package. The `f_classif` calculates the ANOVA F-value. The F-value scores examine if, when we group the numerical feature by the target vector, the means for each group are significantly different. The `SelectkBest` then selects the columns based on the results from `f_classif`. Figure shows the Feature Selected Dataset.

	Daily High	Daily Low	52-Week High	Prev.Close	52-Week Low	Price Direction
0	8231.1500	7931.3600	13829.0700	8168.2400	3178.3300	1
1	177.6656	172.4018	362.8190	177.0681	82.4270	1
2	0.2569	0.2478	0.5655	0.2543	0.2198	-1
3	226.3218	218.3572	642.8143	222.8102	74.7157	1
4	57.3156	55.4602	145.8824	56.8571	22.6620	1

Figure 30: Feature Selected Dataset

9.3) Training and Testing Dataset

Figure shows how the training dataset and testing dataset are prepared by splitting with the ratio 80% to training set and 20% testing set.

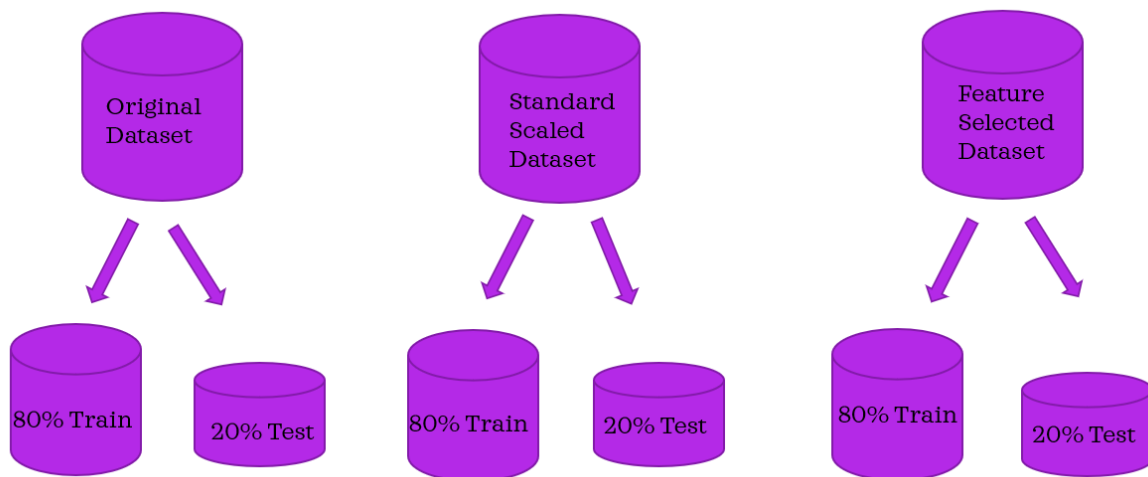


Figure 31: Splitting Train and Test Sets for 3 Datasets

10) Modelling

There would be two machine learning models used in this project with different targets. This project would not only focus on predicting the cryptocurrency prices movement but would also predict by how much the prices would rise or fall. The machine learning model K Nearest Neighbour Classifier (KNN Classifier) would predict if the cryptocurrencies prices moves 'Up' or 'Down' while K Nearest Neighbour Regression (KNN Regression) would predict how much would the prices move.

10.1) K Nearest Neighbour Classifier

K nearest neighbors is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common

amongst its K nearest neighbors measured by a distance function. If $K = 1$, then the case is simply assigned to the class of its nearest neighbor.

10.2) K Nearest Neighbour Regression

A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. KNN regression uses the same distance functions as KNN classification.

11) Results and Performance Evaluation

Using KNN Classifier to predict the labels 'Up' or 'Down' gives us the results as below.

	Original	Standard Scaled	Feature Selected
Accuracy Rate	73.34%	73.89%	71.67%

Table 1: Accuracy of KNN Classifier for each dataset

Table shows that Standard Scaled achieved the highest accuracy in terms of predicting the correct labels, followed by the Original dataset with 73.34% and Feature Selected dataset with 71.67%.

Using KNN Regression to predict how much difference between the Predicted Prices and Actual Prices yields us the results as below.

	Original	Feature Selected
Root-Mean-Square-Error	7.838	6.937

Table 2: Root-Mean-Square-Error of KNN Regression for each dataset

12) Discussion

From the results and performance evaluation, the Standard Scaled dataset achieved the highest in terms of accuracy while Feature Selected scored the lowest among the three datasets. The Standard Scaled dataset improved accuracy compared to the Original dataset is minimal thus suggesting that KNN Classifier was able to perform well without scaling of the values. Meanwhile, the Feature Selected dataset was initially thought to supposedly yield the highest accuracy after doing dimension reduction but fall short of expectations. This could be due the fact that Feature Selected dataset only have five columns of which none of the columns have sentiment news scores. It was observed that the Original and Standard Scaled dataset achieved better accuracy with the sentiment news scores columns proved that sentiment scores does help to improve model's accuracy.

Another take away from this study was that KNN Regression was able to predict the Price of Cryptocurrency with a small error. The Feature Selected dataset showed better results when KNN Regression was used to predict the cryptocurrency 'Open' price the next day with a root-mean-square-error of 6.937 as compared to using the Original dataset which yielded a value of 7.838. What is noticeable here was that KNN Regression was not applied on the Standard Scaled dataset as the scaled dataset would yield a small root-mean-square-error thus giving a misleading result that it a small error.

From the discussion, the Analysis Goals was assessed if the goals of this project were met. First Analysis Goal, i) Predict effectively cryptocurrency price movements with machine learning algorithm. The accuracy rate was on average 72.96% using three separate datasets with KNN Classifier, thus the First Analysis Goal was met. Second Analysis Goal, ii) Predict cryptocurrency prices as close as possible to the actual price value. From the root-mean-square-error value obtained after applying KNN Regression to the datasets, the value was on average 7.3875 which was considered impressive given only 45 days of data to work with.

The Third Analysis Goal, iii) Determine if analysing news sentiment improves the machine learning model. The Feature Selected Dataset, which does have any sentiment score variables, accuracy rate was lower than the other two datasets which the KNN Classifier took into account sentiment news score. Thus, proving that sentiment score helped to improve the machine learning performance.

13) Conclusion

Predicting the future has always been a dream for mankind. Predicting price movements across the stock markets using statistical methods in the past has always been difficult as many external and internal factors causes the stock markets prices to be volatile. But now with the help of technologies such as machine learning, predicting future prices might be no longer a pipe dream but slowly becoming a reality.

From this project, KNN Classifier and KNN Regression showed promising potential to be applied in predicting future cryptocurrency movements and their prices. However, a more in depth study is needed as the data here consists of only 45 days of data with 900 rows. Furthermore, future research could focus on applying Google Trend results to observe how trending headlines on Google Trend could have effect on cryptocurrency prices.

14) References

- 1) <https://cryptocurrencyfacts.com/>
- 2) Victor, Alexander. (2017). INTRODUCING CRYPTOCURRENCY. READS Capital. 1. 2.
- 3) Farrell, Ryan, "An Analysis of the Cryptocurrency Industry" (2015).Wharton Research Scholars. 130
- 4) https://www.saedsayad.com/k_nearest_neighbors.htm
- 5) https://www.saedsayad.com/k_nearest_neighbors_reg.htm