BITCOIN PREDICTION USING SOCIAL NETWORKING TRAFFIC AND GOOGLE TREND

NAME:

CHONG MING KEAT WQD180093

Lecturer:
Dr. TEH YING WAH .

Course:
WQD7005 Data Mining

DATE:
21 December 2019

Contents

List of Figures

Chapter 1

Introduction

In 2019, the market capitalization of Bitcoin has reach 130 billion dollars. It makes the Bitcoin now is on par with Netflix in term of market capitalization [1]The volatility and fluctuation caught the public attention in 2017, at its extreme the price of one Bitcoin experienced 2000% of increase from $863 to $17,550 in a single year.Due to it lucrative profit, many researcher has started to study the reason of bitcoin fluctuation with others factor such as tweets sentiment, news sentiment, financial report and etc. Recently, Google trend has emerge as tools for researcher to understand google search trend of a topic for any time-frame, it provide great insight to extract hidden pattern or meaning from the search trend. Several research utilizing google trend [2] [3] has show some success on predicting micro-economics such as automobile sales and unemployment rate.

Problem Statement Nowadays, Bitcoin has been consider as one of the most popular investment product in the world. It spark the interest of researcher to study factors affecting the price direction. However, most of the research direction are concentrate on social media network sentiment. The recent rising tool Google trend has show promising result in other field. Therefore, the inclusion of Google Trend in the factor maybe valuable to overall prediction.

Objectives The aim of this project is to predict the daily Bitcoin price direction with primary factor (Google Trend) and secondary factor (Tweets Volume,Reddit Volume, News Headline Sentiment and Historical Price).

Scope of Work In this project, we would first introduce the workflow and process of the overall data mining project. Each processes will be presented and discuss in detail. Lastly, we will provide recommendation and conclusion based on the results and finding.

Chapter 2

Data Mining Workflow

In order to extract useful knowledge from the data set, it is important to oversee and ensure that the analysis outcome is well-suited for the desired project objective. Numerous data mining process models are designed to provide a systematic framework for the above-mentioned purpose. Among the commonly used models, Cross-Industry Standard Process for Data-Mining (CRISP-DM) has always been the top model used by the data analysts since 2007 [4]. For this reason, CRISP-DM model is select as a reference for process of this data mining project refer to figure 2.1. As shown in the figure, data mining is not typical waterfall process instead it require reiteration and refining in each process and often require revisiting first two process which are Data Collection and Data Prepossessing. As it is well known that 70-80 % of time will be spend in these two process and it is no different in this project.

Two major tools are use through out the whole process, Python are heavily use in data-collection and data preprocessing stage, some data visualization are using Python matplotlib library. SAS enterprises miner are mainly responsible for exploratory analysis thru clustering and decision tree, the project has utilize the machine learning tools available in SAS enterprise for data modelling and result evaluation.

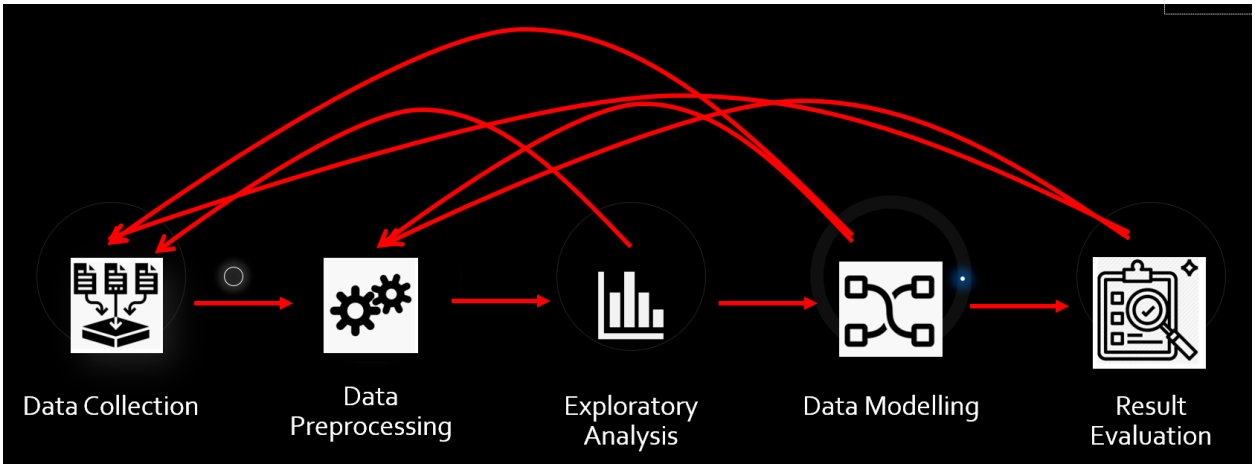


Figure 2.1: DATA MINING PROCESS

2.1 Data Collection

To complete our project objective which is to predict the Bitcoin price direction base on primary factor and secondary factor, the method and source for the data collection is presented as table 2.2.

Data	Provider	Link	Method	Script
Bitcoin Historical Price	Coinmetrics	https://coinmetrics.io/data-downloads/btc.csv	Download	-
Tweets Volume	Bitinfochart	https://bitinfocharts.com/comparison/tweets-btc.html#log	Web-crawling	Tweets_Volume_Crawl.ipynb
News Healine	Investing	https://www.investing.com/news/cryptocurrency-news	Web-crawling	Crawl Bit_News.ipynb
Reddit Volume	Redditmetrics	https://subredditstats.com/r/bitcoin	Web-crawling	Reddit_Vol_Crawl.ipynb
Google Trend	Google	https://trends.google.com/trends/?geo=US	Download	-

Figure 2.2: DATA COLLECTION SUMMARY

2.1.1 Web Crawling

Python programming is responsible for the web-crawling for Tweets volume,News headline and reddit volume. The package use for the web-crawling are Beautifulsoup and Selenium. It is worth mentioning the volume data are presented as interactive plot in the website and Selenium is particular useful for crawling JavaScript object embedded in the webpage, Selenium will render the object with any web-browser set-up by user, and so Beautifulsoup would able to crawl the html source after javascript object is rendered by web-browser. Refer figure 2.4 for Reddit volume graphical plot in web-page

As for the News web-page, the news headline is contain in individual sub-table and in order to crawl history news headline, the script has include function to use Selenium package to click on each page until all news for desire data range is crawl and total of 18356 row of headline news is crawled for this project.

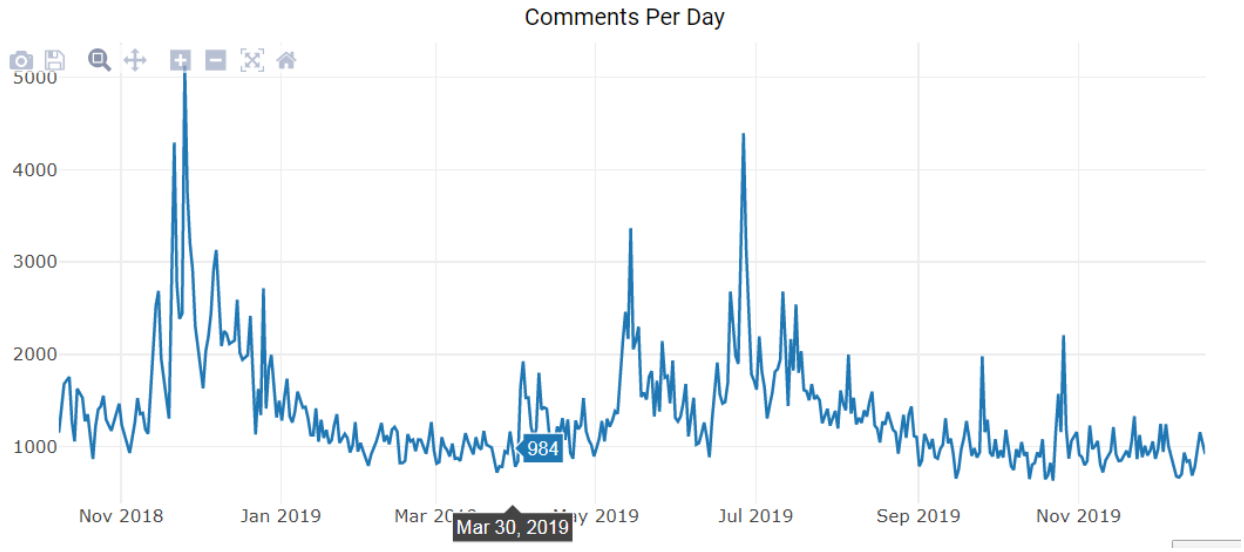


Figure 2.3: Reddit Volume

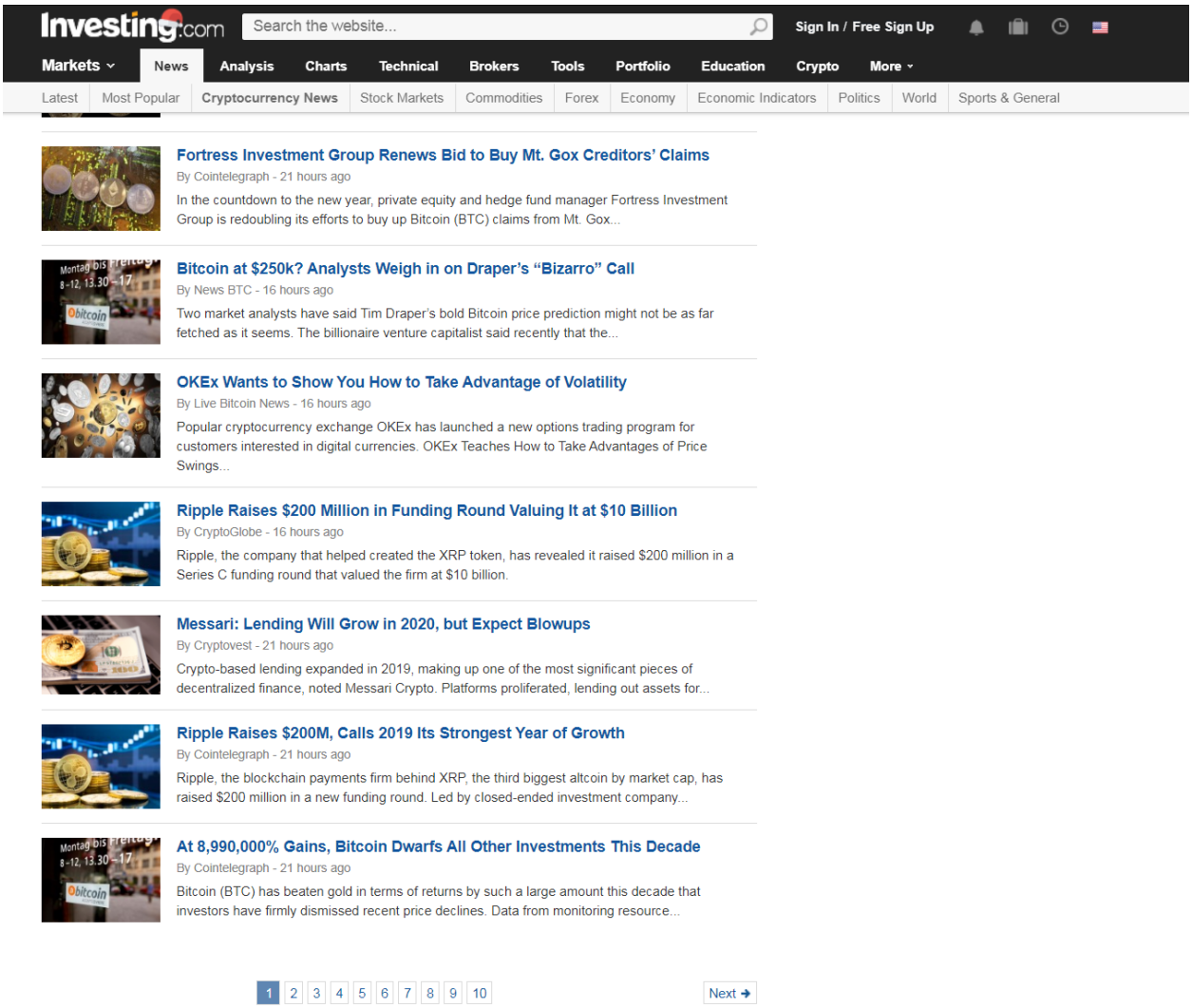


Figure 2.4: News Web Page

2.1.2 Metadata

This section is to present the metadata of the crawled data from the data collection process. For all the data, the time range chosen for this study is from November 2018 to November 2019, the time interval is in daily, therefore total of 365 rows of data is obtained for each sources.

Bitcoin Price

Total 40 attribute has been obtained obtained from the source link,full description of each attribute is presented below refer to figure 2.5

Short name	Metric name	Interval	Definition
AdrActCnt	Addresses, active, count	1 day	The sum count of unique addresses that were active in the network (either as a recipient or originator of a ledger change) that day. All parties in a ledger change action (recipients and originators) are counted. Individual addresses are not double-counted if previously active.
BlkCnt	Block, count	1 day	The sum count of blocks created that day that were included in the main (base) chain.
BlkSizeMeanByte	Block, size, mean, bytes	1 day	The mean size (in bytes) of all blocks created that day.
CapMVRVCur	Capitalization, MVRV, current supply	1 day	The ratio of the sum USD value of the current supply to the sum "realized" USD value of the current supply.
CapMrktCurUSD	Capitalization, market, current supply, USD	1 day	The sum USD value of the current supply. Also referred to as network value or market capitalization.
CapRealUSD	Capitalization, realized, USD	1 day	The sum USD value based on the USD closing price on the day that a native unit last moved (i.e., last transacted) for all native units.
DiffMean	Difficulty, mean	1 day	The mean difficulty of finding a hash that meets the protocol-designated requirement (i.e., the difficulty of finding a new block) that day. The requirement is unique to each applicable cryptocurrency protocol. Difficulty is adjusted periodically by the protocol as a function of how much hashing power is being deployed by miners.
FeeMeanUSD	Fees, transaction, mean, USD	1 day	The USD value of the mean fee per transaction that day.
FeeMedUSD	Fees, transaction, median, USD	1 day	The USD value of the median fee per transaction that day.
FeeTotUSD	Fees, total, USD	1 day	The sum USD value of all fees paid to miners that day. Fees do not include new issuance.
IssContNtv	Issuance, continuous, native units	1 day	The sum of new native units issued that day. Only those native units that are issued by a protocol-mandated continuous emission schedule are included.
IssContPctAnn	Issuance, continuous, percent, annualized	1 year	The percentage of new native units (continuous) issued on that day, extrapolated to one year (i.e., multiplied by 365), and divided by the current supply on that day. Also referred to as the annual inflation rate.
IssTotUSD	Issuance, total, USD	1 day	The sum USD value of all new native units issued that day.
NVTAdj	NVT, adjusted	1 day	The ratio of the network value (or market capitalization, current supply) divided by the adjusted transfer value. Also referred to as NVT.
NVTAdj90	NVT, adjusted, 90d MA	1 day	The ratio of the network value (or market capitalization, current supply) to the 90-day moving average of the adjusted transfer value. Also referred to as NVT.
PriceBTC	Price, BTC	1 day	The fixed closing price of the asset as of 00:00 UTC the following day (i.e., midnight UTC of the current day) denominated in BTC.
PriceUSD	Price, USD	1 day	The fixed closing price of the asset as of 00:00 UTC the following day (i.e., midnight UTC of the current day) denominated in USD. This price is generated by Coin Metrics' fixing/reference rate service.
SplyCur	Supply, current	All time	The sum of all native units ever created and visible on the ledger (i.e., issued) as of that day. For account-based protocols, only accounts with positive balances are counted.
TxCnt	Transactions, count	1 day	The sum count of transactions that day. Transactions represent a bundle of intended actions to alter the ledger initiated by a user (human or machine). Transactions are counted whether they execute or not and whether they result in the transfer of native units or not (a transaction can result in no, one, or many transfers). Changes to the ledger mandated by the protocol (and not by a user) or post-launch new issuance issued by a founder or controlling entity are not included here.
TxTfr	Transactions, transfers, count	1 day	The sum count of transfers that day. Transfers represent movements of native units from one ledger entity to another distinct ledger entity. Only transfers that are the result of a transaction and that have a positive (non-zero) value are counted.
TxTfrValAdjNtv	Transactions, transfers, value, adjusted, native units	1 day	The sum of native units transferred that day removing noise and certain artifacts.
TxTfrValAdjUSD	Transactions, transfers, value, adjusted, USD	1 day	The USD value of the sum of native units transferred that day removing noise and certain artifacts.

Figure 2.5: Bitcoin Data Dictionary from [5]

Google Trend

Google trend data provides information on how popular given search terms are relative to other search terms at any given time.Therefore,this provide a proxy metric for the general interest there is in Bitcoin at any given time. Nevertheless,Google does not provide search volumes but search volume index. The search volume index is calculated by dividing each data point by the total searches within a geographic region and time range, and tt can be ranging from 0-100. Refer to figure 2.6 for Bitcoin google search trend over time. It can be observed that the peak is at 2017 December where the Bitcoin hit all time high at 17000 dollar per Bitcoin, it's understandable there's some correlation between bitcoin price with search trend, however, it is possible that it may only indicate direct correlation but has no relationship for future predictability.

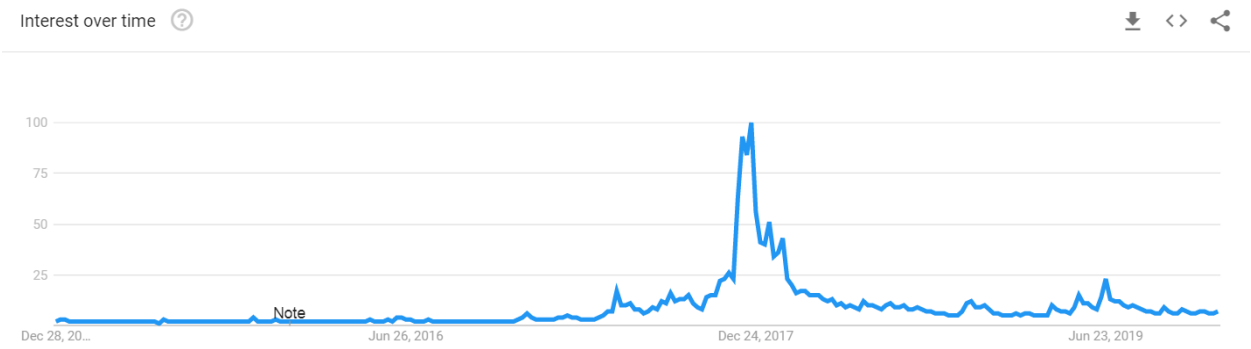


Figure 2.6: Bitcoin Search in Google Trend

Tweets and Reddit Volume

For Tweets volume, the data collected are showing total tweets with keywords of Bitcoin is posted on the day until 00.00 UTC time. Bitinfocharts is freely provide the total number of tweet since starting of April 2014. The Reddit volume are counted with sum of comment with Bitcoin as keyword posted per day, this information are freely available at subredditstats web page.

Chapter 3

Data Preprocessing

3.0.1 Bitcoin News Headline Sentiment

Proper data preparation is the key to accurate results in data mining analysis especially unstructured data. For this project, the major data prepossessing task are around news headline and google trend data. News headline are unstructured data crawled from the website, before merging all the data for as input for SAS entreprise Miner it is necessary to convert the News headline as sentiment score the interval measure for training the model. Generally, the data preparation stage can be summarized into four steps which are data preprocessing, tokenizing, filterating and words reduction, as supported by a number of existing studies [6], [7], [8], [9]. Text data may contain lots of data sources such as HTML files, emails, text documents, posts or short notes [10], which resulted in the need to standardize the format of data in data pre-processing step. The purpose is to convert all components to a chosen format and also to standardize the data, say in lowercase alphabets, as some of the tools are case sensitive, such as those written in Python programming language. This process is time-consuming and complex, and has motivated the engineers to design text processing tools to increase the efficiency. Among others, Natural Language Toolkit (NLTK) for instance is one of the most widely mentioned Natural Language Proccesing (NLP) library for Python [11].

Filtering

This step removes stopwords, namely the words that bring no information to the analysis. This is a necessary step as the unwanted words will highly affect the consistency of the analysis result, as described in the study by [12]. Stopwords (such as “the”, “a”, “an”, “in”) are very common and may appear frequently in most texts. Without removing these words, they mostly will appear to be the most frequent used words in a text which could make the analysis outcome meaningless.

Words Reducing

This is a process to convert infinite tenses and nouns to their singular forms so that different words bringing the same meaning can be approached as one item. There are two basic methods to do this, lemmatization and stemming, which convert the words to their proper root form and the canonical form of original words respectively.

Sentiment Analysis

Python package VADER from NLTK is use to obtain the sentiment score for each cleaned news headline, VADER is lexicon and rule-based sentiment tools that tailor for sentiment analysis in social media network, however, it's also popular to be apply for News sentiment analysis. The output of sentiment analysis will provide 4 output positive score, neutral score, negative score, polarity score and compound score. All the these features will be retain for training the model. Compound score is basically an overall score for whole sentence by dissecting each sentence and sum up the score for complete sentence. Also, due to structures of dataset is in daily interval therefore the score will be aggregate based on mean value for each. Refer figure 3.1 for the final result of sentiment analysis. All the relevant codes for News cleaning and sentiment analysis can be find under jupyter notebook "Crawl Bit News.ipyb" in my GitHub.

	Date	Headline	tidy_Text	polarity	Vader_neg	Vader_neu	Vader_pos	Vader_compound
222	dec 13 2019	wall st. to washington: bakkt launches new pro...	wall washington bakkt launches products ...	0.000000	0.000	1.000	0.000	0.0000
223	dec 12 2019	us health insurance giant piloting blockchain ...	health insurance giant piloting blockchain s...	0.133333	0.000	0.745	0.255	0.3400
224	dec 13 2019	bitcoin (btc) flirts with key resistance, bear...	bitcoin flirts with resistance bears con...	0.000000	0.000	0.746	0.254	0.1779
225	dec 12 2019	hybrid ai firm cindicator launches crypto fund...	hybrid firm cindicator launches crypto fund b...	-0.200000	0.000	0.721	0.279	0.4767
226	dec 12 2019	shopin founder pleads guilty to orchestrating ...	shopin founder pleads guilty orchestrating fr...	-0.500000	0.545	0.455	0.000	-0.7184
227	dec 12 2019	blockfi to offer first interest-bearing crypto...	blockfi offer first interest bearing crypto a...	0.250000	0.000	0.727	0.273	0.4588
228	dec 13 2019	ethereum based defi forecast to hit $5 billion...	ethereum based defi forecast billion	0.000000	0.000	1.000	0.000	0.0000
229	dec 12	ripple (xrp) price targets fresh lows, btc &	ripple price targets fresh lows consol...	0.300000	0.222	0.494	0.284	0.1280

Figure 3.1: News headline sentiment

3.0.2 Google Trend Preprocessing

When the data queried for Google Trend is more than 90 days the search volume index result will be return in weekly interval which is not align with others data we have collected so far, therefore, preprocessing is require to convert this weekly interval data into daily in order to merge with other dataset for further analytical process. To covert this data, we use method details by [13]. The first step are collect daily search data from Google Trend and combine it into one array, second collect weekly search data over the same time period and lastly adjust the daily data based on weekly data with multiplying daily data with ratio of weekly divide by daily for first day of the week. Diagrammatic explanation is provided below:

- Step 1: Collect daily search data from Google Trends and combine it into one array.

Day	bitcoin: (Worldwide)
11/21/2018	95
11/22/2018	71
11/23/2018	74
11/24/2018	67
11/25/2018	100
11/26/2018	100
11/27/2018	90
11/28/2018	84
11/29/2018	74
11/30/2018	68
12/1/2018	59
12/2/2018	53
12/3/2018	66
12/4/2018	67
12/5/2018	63
12/6/2018	72

(a) Step 1

- Step2: Collect weekly search data over the same time period

Week	bitcoin: (Worldwide)
11/18/2018	53
11/25/2018	55
12/2/2018	45
12/9/2018	41
12/16/2018	44
12/23/2018	38
12/30/2018	33
1/6/2019	35
1/13/2019	31
1/20/2019	28
1/27/2019	29
2/3/2019	31
2/10/2019	28
2/17/2019	32
2/24/2019	30
3/3/2019	27

(b) Step 2

Figure 3.2: Google Trend Preprocessing Step 1 & 2

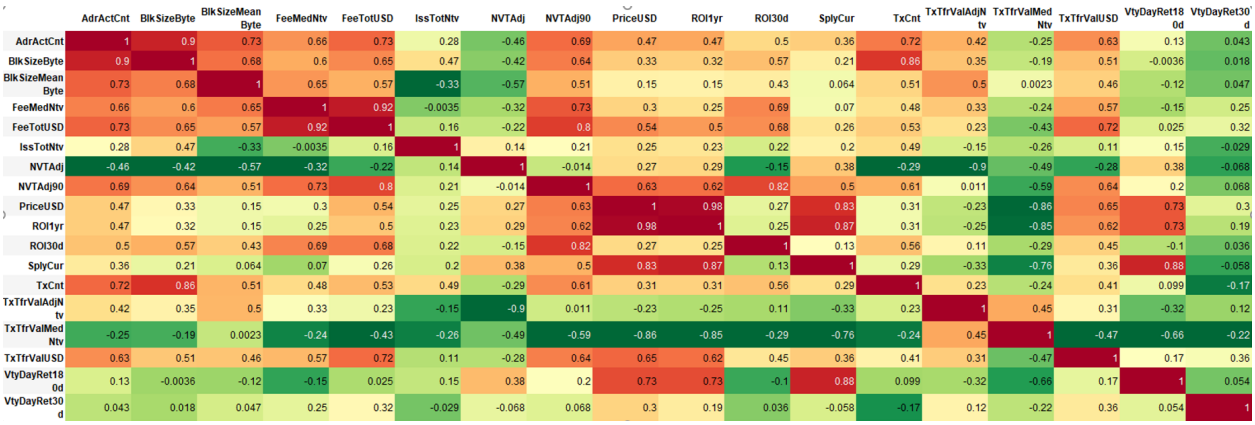
Step3: Adjust the daily data based on the weekly data;
Multiplying the daily data with ratio of first day of week (weekly/daily)

Day	bitcoin: (Worldwide)			
11/21/2018	95	53	0.557895	53
11/22/2018	71			40
11/23/2018	74			41
11/24/2018	67			37
11/25/2018	100	55	0.55	55
11/26/2018	100			55
11/27/2018	90			50
11/28/2018	84			46
11/29/2018	74			41
11/30/2018	68			37
12/1/2018	59			32
12/2/2018	53	45	0.849057	45
12/3/2018	66			56
12/4/2018	67			57
12/5/2018	63			53
12/6/2018	72			61
12/7/2018	89			76
12/8/2018	68			58

Figure 3.3: Google Trend Preprocessing Step 3

3.0.3 Bitcoin Price Data

Mentioned in earlier, total 40 attribute of Bitcoin price are obtain from the sources. It's necessary to skim through the data to reduce the redundant attribute to avoid needless data input for our training later. Therefore, we try to inspect the correlation by plotting the heatmap, the attribute that has very high correlation with each other will be trim off as its redundant for model training refer figure 3.4 for heatmap plotted in Python.Final 18 attribute is retained and 22 attribute is drop after the correlation analysis.



	AdrActCnt	BlkSizeByte	BlkSizeMean Byte	FeeMedNtv	FeeTotUSD	IssTotNtv	NVTAdj	NVTAdj90	PriceUSD	ROI1yr	ROI30d	SplyCur	TxCnt	TxTfrValAdjNtv	TxTfrValMedNtv	TxTfrValUSD	VtyDayRet180d	VtyDayRet30d
AdrActCnt	1	0.9	0.73	0.66	0.73	0.28	-0.46	0.69	0.47	0.47	0.5	0.36	0.72	0.42	-0.25	0.63	0.13	0.043
BlkSizeByte	0.9	1	0.68	0.6	0.65	0.47	-0.42	0.64	0.33	0.32	0.57	0.21	0.86	0.35	-0.19	0.51	-0.0036	0.018
BlkSizeMeanByte	0.73	0.68	1	0.65	0.57	-0.33	-0.57	0.51	0.15	0.15	0.43	0.064	0.51	0.5	0.0023	0.46	-0.12	0.047
FeeMedNtv	0.66	0.6	0.65	1	0.92	-0.0035	-0.32	0.73	0.3	0.25	0.69	0.07	0.48	0.33	-0.24	0.57	-0.15	0.25
FeeTotUSD	0.73	0.65	0.57	0.92	1	0.16	-0.22	0.8	0.54	0.5	0.68	0.26	0.53	0.23	-0.43	0.72	0.025	0.32
IssTotNtv	0.28	0.47	-0.33	-0.0035	0.16	1	0.14	0.21	0.25	0.23	0.22	0.2	0.49	-0.15	-0.26	0.11	0.15	-0.029
NVTAdj	-0.46	-0.42	-0.57	-0.32	-0.22	0.14	1	-0.014	0.27	0.29	-0.15	0.38	-0.29	-0.9	-0.49	-0.28	0.38	-0.068
NVTAdj90	0.69	0.64	0.51	0.73	0.8	0.21	-0.014	1	0.63	0.62	0.82	0.5	0.61	0.011	-0.59	0.64	0.2	0.068
PriceUSD	0.47	0.33	0.15	0.3	0.54	0.25	0.27	0.63	1	0.98	0.27	0.83	0.31	-0.23	-0.86	0.65	0.73	0.3
ROI1yr	0.47	0.32	0.15	0.25	0.5	0.23	0.29	0.62	0.98	1	0.25	0.87	0.31	-0.25	-0.85	0.62	0.73	0.19
ROI30d	0.5	0.57	0.43	0.69	0.68	0.22	-0.15	0.82	0.27	0.25	1	0.13	0.56	0.11	-0.29	0.45	-0.1	0.036
SplyCur	0.36	0.21	0.064	0.07	0.26	0.2	0.38	0.5	0.83	0.87	0.13	1	0.29	-0.33	-0.76	0.36	0.88	-0.058
TxCnt	0.72	0.86	0.51	0.48	0.53	0.49	-0.29	0.61	0.31	0.31	0.56	0.29	1	0.23	-0.24	0.41	0.099	-0.17
TxTfrValAdjNtv	0.42	0.35	0.5	0.33	0.23	-0.15	-0.9	0.011	-0.23	-0.25	0.11	-0.33	0.23	1	0.45	0.31	-0.32	0.12
TxTfrValMedNtv	-0.25	-0.19	0.0023	-0.24	-0.43	-0.26	-0.49	-0.59	-0.86	-0.85	-0.29	-0.76	-0.24	0.45	1	-0.47	-0.66	-0.22
TxTfrValUSD	0.63	0.51	0.46	0.57	0.72	0.11	-0.28	0.64	0.65	0.62	0.45	0.36	0.41	0.31	-0.47	1	0.17	0.36
VtyDayRet180d	0.13	-0.0036	-0.12	-0.15	0.025	0.15	0.38	0.2	0.73	0.73	-0.1	0.88	0.099	-0.32	-0.66	0.17	1	0.054
VtyDayRet30d	0.043	0.018	0.047	0.25	0.32	-0.029	-0.068	0.068	0.3	0.19	0.036	-0.058	-0.17	0.12	-0.22	0.36	0.054	1

Figure 3.4: Bitcoin Data Corr Plot

3.0.4 New Features

A new column is created for the price direction. Price direction is categorical data, therefore, we set the category with the following rules; if the price has increase compare to previous day it will be label as 2 indicate of up, vice versa 1 is indicate of price down. Also we presume that the price changing within 2% are trivial relative to price movement of Bitcoin, threfore price changing in 2% is labeled as 0 for indicate no price change.

3.0.5 Merged Dataset & Normalizing

On the last part of preprocessing, all the data are merged together into single dataframe and ready to input to SAS enterprise miner for further exploratory analysis and data modelling. Also, all dataframe are normalize with min-max value for of 0 to 1 standardization.

Chapter 4

Exploratory Analysis

4.0.1 Time-series Plot With Python

Firstly, the time series plot of Bitcoin price against various primary and secondary factor are plotted to study the potential relationship.

Figure 4.1: BTCPrice VS Google Trend

Figure 4.2: BTCPrice VS Tweet Volume

Figure 4.3: BTCPrice VS Reddit Volume

Figure 4.4: BTCPrice VS News Sentiment

Refer to 4.1 4.2, we can observed that there's a spike of activity at first of April for both google trend and Tweet Volume, for the same day Bitcoin has experience 20% growth from around 4000 usd to 4800 usd. Base on the source from NASDAQ, the possible reason speculate around Bloomberg was probably contribute by April Fool's. However, there also the theory saying that the trading activity is iced and stagnant for quite substantial time, and a single 1 million trading activity has trigger the rise. Furthermore, after this gain the Bitcoin has start the Bull market for continuous 2 month and reached the peak at 12000 USD,according to the source [14] the significant price raised was believe contributed by the News of Facebook launching new Crytocurrency Libra coin.

Beside, news from Forbes [15] as early as 25th of March has reported the possibility of bitcoin price to have sustained growth with the News title of "The Next Bitcoin Price Rally Could Be Sparked By A Surprising Source", it's reported in the News that central bank may consider stock up bitcoin as digital gold. Also, refer to figure 4.4 we can observed there is fluctuation of sentiment at early of March, this may also contribute the Bull market for Bitcoin in the coming 2 months.

To study the relationship of all combining parameter, a graph with noise smoothing and isolated trend is plotted in figure 4.5. From the graph, the trend for Google SVI, Tweets volume and Reddit comment has similar pattern at mid of March to end of July, we can see a triple hill shape for all three parameters. Also, we can see a fluctuation happens for the trend if Bitcoin price has spike on drop or raise, it understand that News and buyer may be interested to know what factors cause that and therefore, more discussion or comment is posted when that happens, and in other hand the trend are consistently downward when the Bitcoin price is stagnant or with minimal movement.

In summary, the observed of radical movement of news at early march may have trigger the price rally of Bitcoin on April, and there are particular two major news we found that probably contributed to the bulls market, first central bank considering to stock Bitcoin and second Facebook launching of Libra coin. Also, we observed the similar moving trend for all primary and secondary factor for the period between mid of March to end of July.

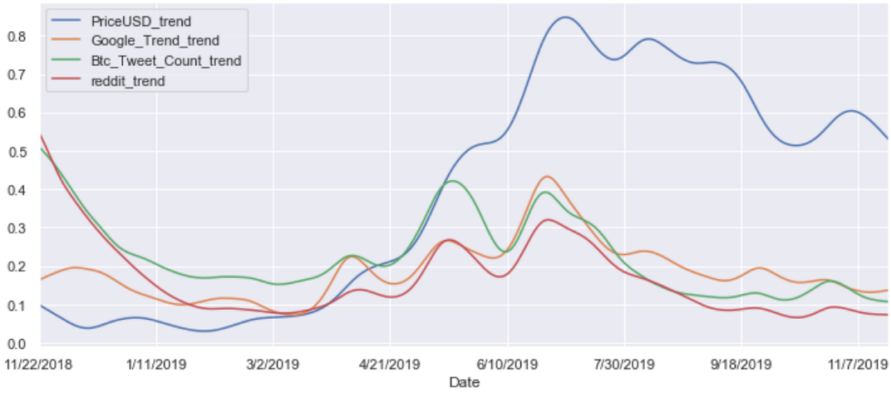


Figure 4.5: Combined Trend

4.0.2 Exploratory Analysis with SAS Enterprise Miner

Pie Chart For Price Direction Category

3 different category 0 : No Change, 1 : Price Down, 2 : Price Up is plotted in the pie chart 4.7 for distribution study. 34% for segment 0 green color, 31% for segment 1 red color and 35% for segment 2 blue, the overall distribution are quite uniform and this will be selected as our targeted attribute for prediction model.

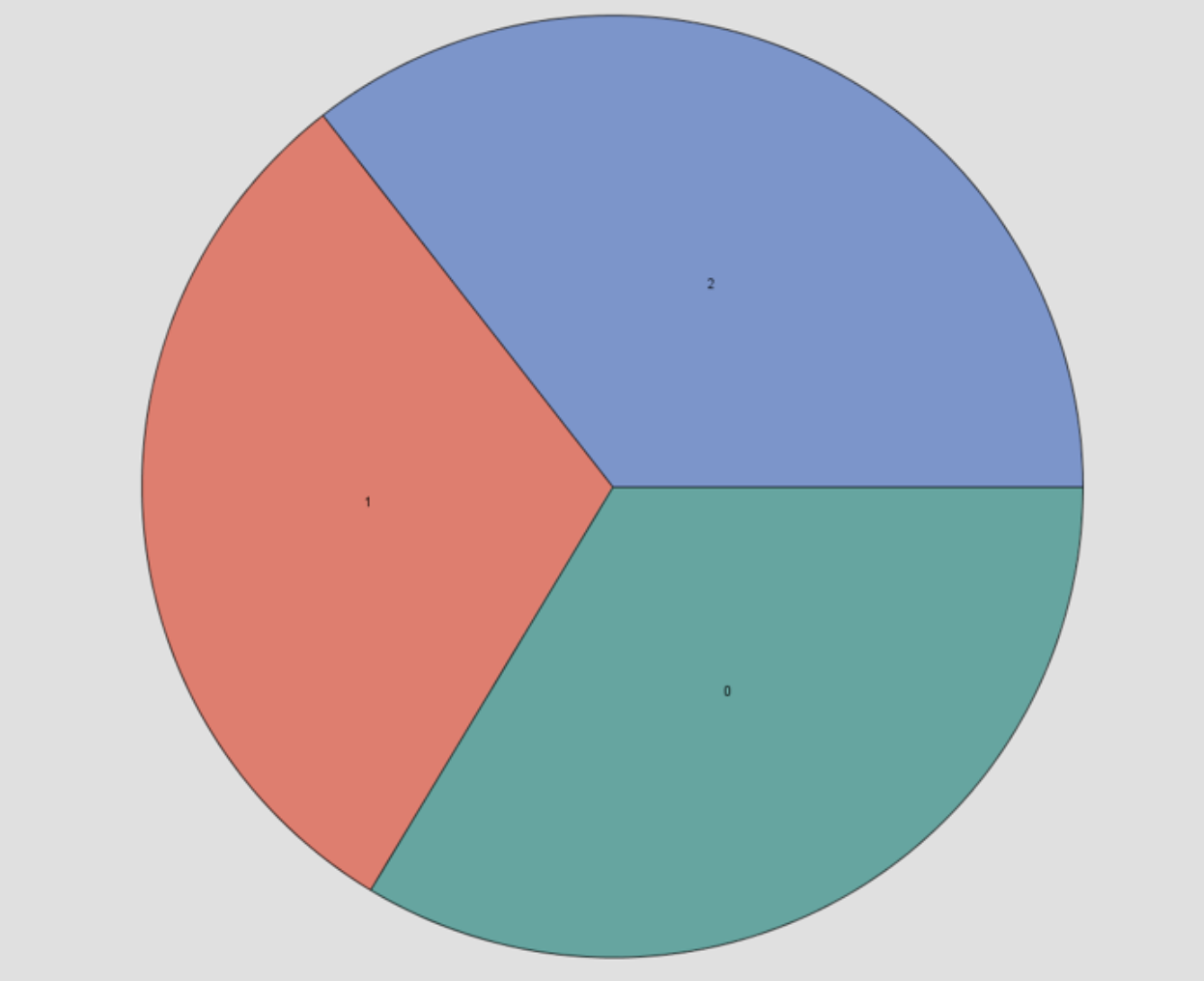


Figure 4.6: Pie Chart for Price Direction

Decision Tree Study

The decision tree was plotted in SAS enterprise Miner with auto branches option. 50% 50% was selected for partitioning the dataset to training and validation. Refer to Figure 4.7, the result from decision tree indicate that the Google Trend has high information gain for price direction prediction. From the first branches, if the SVI is lower than 16.5 the price direction has 75% probability to be no change in training and 65% probability in validation. For the second branches, it tell us that if the SVI is more than 22.5 there are 52% and 47% of chance the price will experience increase and we can deduce that there are 87% and 84% probability the price will have movement if SVI is more than 22.5.

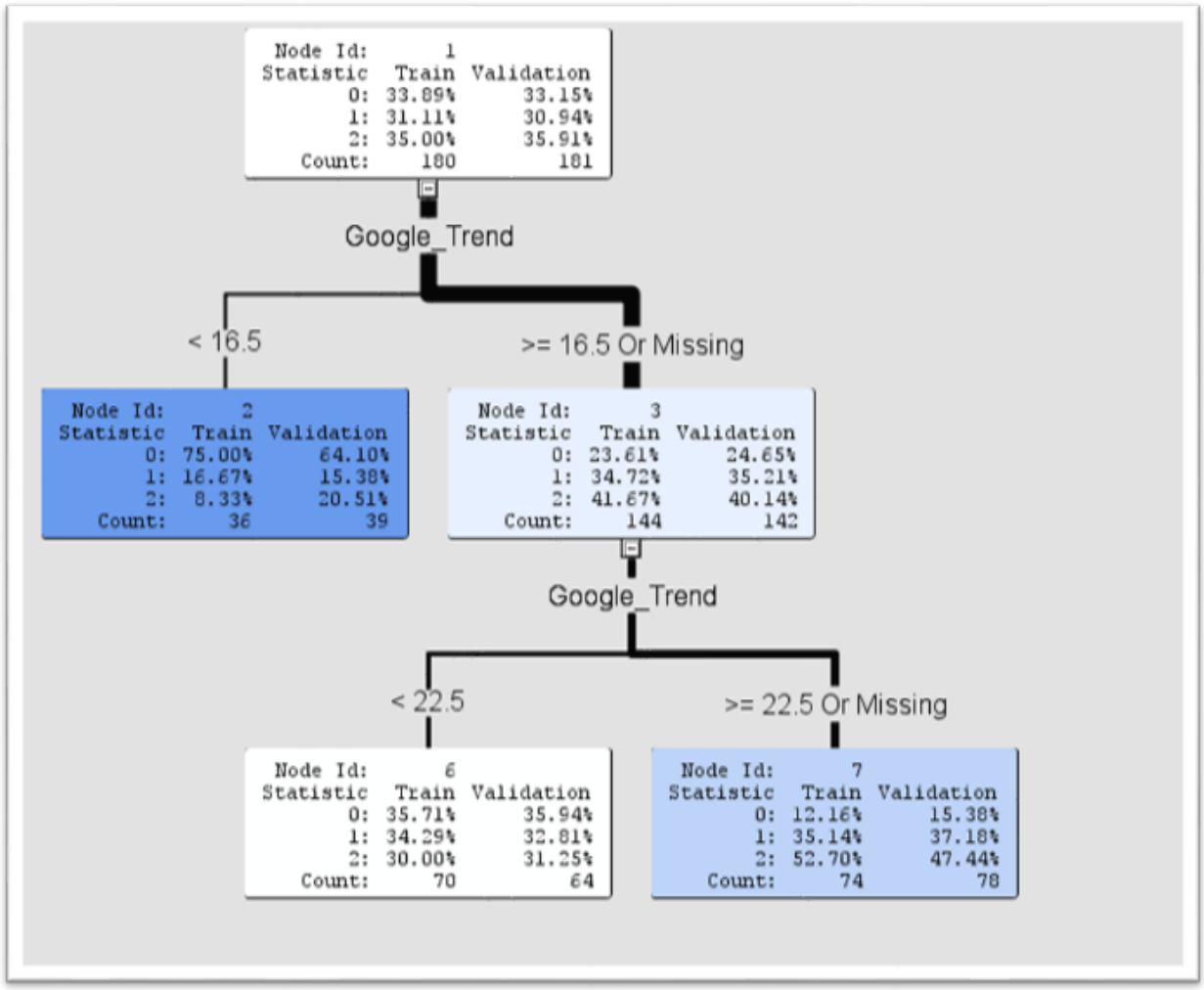


Figure 4.7: Decision Tree Result

Clustering Analysis

The clustering Analysis is perform in SAS enterprise miner with the auto clustering profile node as per figure 4.8. The clustering segment pie chart are presented at figure 4.9,total 6 segment has been created for the dataset. Segment 2 blue color has 155 records, segment 3 (brown) has 36 records, segment 6 (red) has 109 records, segment 4 (green) has 62 record and others (magenta color) has 3 records.

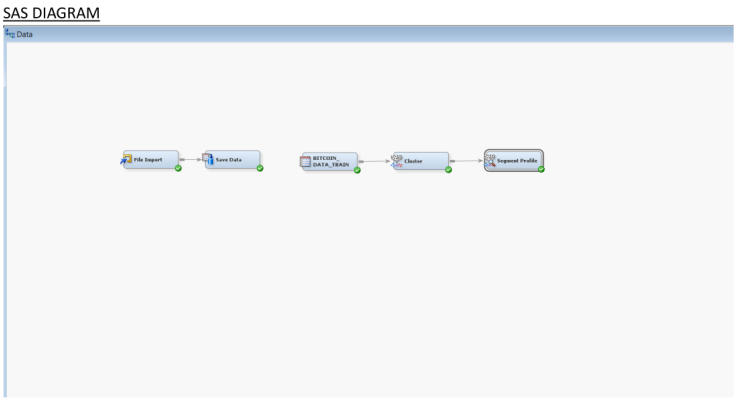


Figure 4.8: SAS Diagram for Clustering

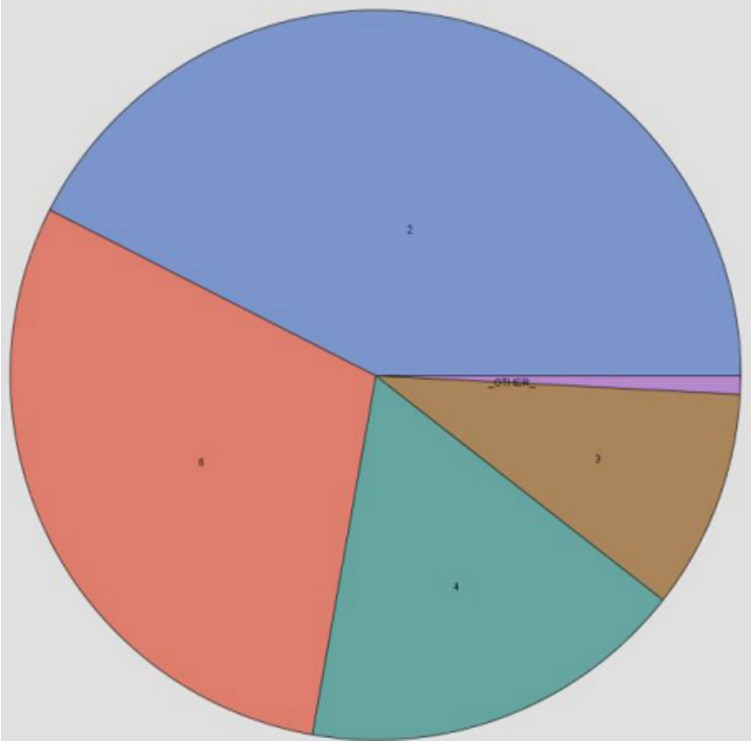


Figure 4.9: Clustering Segment Profile

Refer to figure 4.10 for the mean statistics of clustering segment profile.There are two small cluster as show in the table, 1 observation for segment 1 and 2 observation for segment 5 but we can draw some interesting information from these small clustering. For segment 1,The price difference is highest for whole record period with 1137 USD (18%) increase on that single day, and Google Trend search volume was highest at 100 on same day, Tweets volume and Reddit volume was highest on same day with 42k tweets and 4.4k comments respectively.

As for segment 5 there's total 2 observation recorded, the Bitcoin price has significant drop on that two days (26Th Jun -1753 and 16Th July -1455), google trend index is record at 61 not as high as segment 1, Tweets volume and Reddit volume also relatively lesser than segment 1 but slightly higher than rest of the segment.

Segment 2 has the largest count of frequency 155, we can observe from the comparison the histogram of all the variable are quite fit with the distribution of full dataset. Therefore, we can call segment 2 as "Typical Day". This data belong to this segment are likely to be neutral with no up or down price.

Segment 6 has recorded 109 count, for this cluster the sentiment score is on higher side compare to mean value, the Reddits comment volume is slightly higher than average, also the price difference for Bitcoin is slight higher than average value. This segment has small profit with high sentiment and higher Tweet and Reddit traffic than the average. We can call this segment "Small Earning Day"

For segment 4, the Reddit comments are higher than previous segment same goes for Tweets volume and the bitcoin price is on down trend. We can call this "Loss Day".

Last for segment 3 we have total 36 record. For this segment, the bitcoin price is at high gaining, google trend,Tweet and Reddit volume are significant higher than mean value. This is different compare to segment 4, We believe the significant price increase has attract the interest of people, therefore, there are high activity observed in the data on same day. We can call this "High gain Day".

In summary, the clustering result has provide us meaningful insight to all the 6 different cluster. If the price is experience unusual high increment it will be reflected in social media traffic and same goes to price going down, but this the recorded traffic are not as high compare to price is going up. Although, we can see high correlation between Bitcoin price with social networking traffic, we believe this is expected and this relationship may not be helpful to achieve our project goal to forecast the price direction with reasonable lead time.

Mean Statistic

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean -Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Btc_Tweet _Count	Google_Trend	Price_Diff	Vader_compound	commentsPerDay
0.673144	0.011596	.	1	1	.	0	3	8.186764	41687	100	1137.197	0.051293	4394
0.673144	0.011596	.	2	155	0.536853	4.340412	6	1.473069	17412.37	19.12903	-32.526	-0.01516	1070.815
0.673144	0.011596	.	3	36	0.954884	3.938202	4	2.751408	25536.78	40.69444	512.7471	0.03609	1993.771
0.673144	0.011596	.	4	62	0.902601	5.254091	6	2.572448	23723.8	28.24194	-270.007	0.02365	2105.847
0.673144	0.011596	.	5	2	1.581712	2.500907	4	5.414135	26479.5	61	-1604.45	-0.02832	2847.5
0.673144	0.011596	.	6	109	0.588294	4.458332	2	1.473069	18429.57	20.45872	79.37627	0.102329	1178.426

Figure 4.10: Clustering Segment Mean Statistics

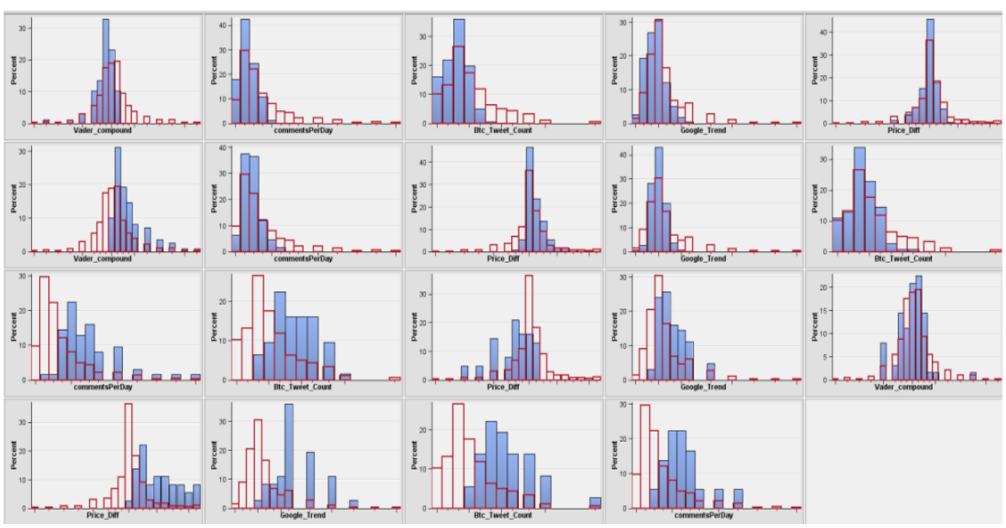


Figure 4.11: Clustering Segment Mean Histogram

Chapter 5

Modelling and Evaluation

5.0.1 Data Input

The modelling and evaluation process are fully performed in SAS enterprise miner. As mentioned earlier, the merged dataset with 35 attribute and 365 rows is import into SAS data library, details of the input metadata can refer to figure 5.1. The highlighted price_dir is set as target role for modelling prediction. The level of Price_dir is set to nominal and Date is input as timeID.The remaining input for level are set as interval.

NAME	ROLE	LEVEL	REPORT	ORDER	DROP	LOWERLIMIT	UPPERLIMIT
AdrActCnt	INPUT	INTERVAL	N		N	null	null
BlkSizeByte	INPUT	INTERVAL	N		N	null	null
BlkSizeMeanByte	INPUT	INTERVAL	N		N	null	null
BTC_Diff	INPUT	INTERVAL	N		N	null	null
Btc_Tweet_Count	INPUT	INTERVAL	N		N	null	null
comment_dif	INPUT	INTERVAL	N		N	null	null
commentsPerDay	INPUT	INTERVAL	N		N	null	null
Date	TIMEID	INTERVAL	N		N	null	null
FeeMedNtv	INPUT	INTERVAL	N		N	null	null
FeeTotUSD	INPUT	INTERVAL	N		N	null	null
Goo_T_Dir	INPUT	INTERVAL	N		N	null	null
Google_Diff	INPUT	INTERVAL	N		N	null	null
Google_Trend	INPUT	INTERVAL	N		N	null	null
IssTotNtv	INPUT	INTERVAL	N		N	null	null
NVTAdj	INPUT	INTERVAL	N		N	null	null
NVTAdj90	INPUT	INTERVAL	N		N	null	null
polarity	INPUT	INTERVAL	N		N	null	null
Price_Diff	INPUT	INTERVAL	N		N	null	null
Price_dir	TARGET	NOMINAL	N		N	null	null
PriceUSD	INPUT	INTERVAL	N		N	null	null
Reddit_dir	INPUT	INTERVAL	N		N	null	null
ROI1yr	INPUT	INTERVAL	N		N	null	null
ROI30d	INPUT	INTERVAL	N		N	null	null
SplyCur	INPUT	INTERVAL	N		N	null	null
Tweet_Dir	INPUT	INTERVAL	N		N	null	null
TxCnt	INPUT	INTERVAL	N		N	null	null
TxTfrValAdjNtv	INPUT	INTERVAL	N		N	null	null
TxTfrValMedNtv	INPUT	INTERVAL	N		N	null	null
TxTfrValUSD	INPUT	INTERVAL	N		N	null	null
Vader_compound	INPUT	INTERVAL	N		N	null	null
Vader_neg	INPUT	INTERVAL	N		N	null	null
Vader_neu	INPUT	INTERVAL	N		N	null	null
Vader_pos	INPUT	INTERVAL	N		N	null	null
VtyDayRet180d	INPUT	INTERVAL	N		N	null	null
VtyDayRet30d	INPUT	INTERVAL	N		N	null	null

Figure 5.1: SAS Column Metadata

5.0.2 SAS Machine Learning Modelling

Four different machine learning model is selected for training and evaluate the data (Logistic Regression, Neural Network, Gradient Boosting and Decision Tree) and model comparison node will provide comprehensive result for all the model and provide the model with highest accuracy score. Refer to 5.2 for the SAS machine learning diagram. Moreover,three separate dataset with lag of 1,2 and 3 days is included in the modelling for sensitivity study.

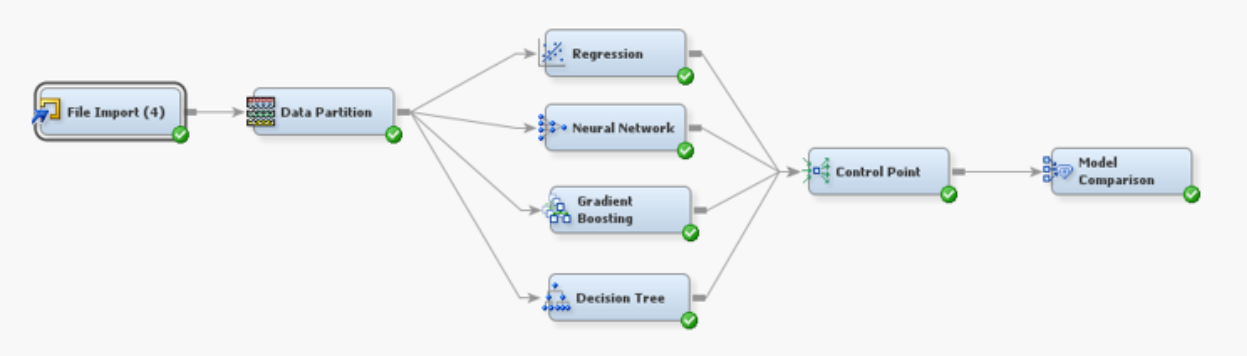


Figure 5.2: SAS Diagram for Machine Learning Modelling

5.0.3 Result

The complete result are presented in figure 5.3,5.4,5.5. In general, decision tree and gradient boost has respectable accuracy compare to other models. We can see that on dataset lag 1 and lag 2 gradient boost model has 90% of accuracy but it fall to 59% for validation dataset, it indicate there's highly non-linear behaviour for the Bitcoin price and it cause some over-fitting on the gradient boosting model. Nevertheless, gradient boosting has most robust result compare to the rest especially at lag 3 dataset, it score 65% accuracy and 57% F1 score. In term of accuracy, decision tree with lag 2 dataset has highest accuracy of 66.5% but only 49% F1 score. From the result of gradient boosting refer to figure 5.6, the table show number of splitting and importance factor contribute to the prediction for both training and validation set. The result has show that Google Trend and ROI1yr (Return Of Investment for 1 year period) has high ratio of validation to training importance, meaning these variable is highly important for both training set and validation set prediction. On the other hand,variable like Reddit comment and Tweets volume has 0.76 and 0.81 importance factor at training set but only 0.26 in validation set, the sentiment score are not in the variable list at all.

Lag 1									
Model Node		FN	TN	FP	TP	Accuracy	Precision	Recall	F1 score
Regression	Train	16	86	31	47	73.89	60.26	74.60	66.67
Regression	Validate	46	85	33	19	56.83	36.54	29.23	32.48
Neural	Train	18	88	29	45	73.89	60.81	71.43	65.69
Neural	Validate	42	85	33	23	59.02	41.07	35.38	38.02
Boost	Train	11	109	8	52	89.44	86.67	82.54	84.55
Boost	Validate	40	83	35	25	59.02	41.67	38.46	40.00
Tree	Train	13	42	75	50	51.11	40.00	79.37	53.19
Tree	Validate	13	45	73	52	53.01	41.60	80.00	54.74

Figure 5.3: Prediction Result for dataset lag 1

Lag 2									
Model Node		FN	TN	FP	TP	Accuracy	Precision	Recall	F1 score
Regression	Train	19	85	32	44	71.67	57.89	69.84	63.31
Regression	Validate	37	73	44	28	55.49	38.89	43.08	40.88
Neural	Train	19	90	27	44	74.44	61.97	69.84	65.67
Neural	Validate	40	69	48	25	51.65	34.25	38.46	36.23
Boost	Train	11	109	8	52	89.44	86.67	82.54	84.55
Boost	Validate	35	77	40	30	58.79	42.86	46.15	44.44
Tree	Train	32	95	22	31	70.00	58.49	49.21	53.45
Tree	Validate	36	92	25	29	66.48	53.70	44.62	48.74

Figure 5.4: Prediction Result for dataset lag 2

Lag 3									
Model Node		FN	TN	FP	TP	Accuracy	Precision	Recall	F1 score
Regression	Train	19	86	31	44	72.22	58.67	69.84	63.77
Regression	Validate	37	76	40	28	57.46	41.18	43.08	42.11
Neural	Train	17	92	25	46	76.67	64.79	73.02	68.66
Neural	Validate	41	77	39	24	55.80	38.10	36.92	37.50
Boost	Train	18	102	15	45	81.67	75.00	71.43	73.17
Boost	Validate	23	76	40	42	65.19	51.22	64.62	57.14
Tree	Train	12	77	40	51	71.11	56.04	80.95	66.23
Tree	Validate	20	66	50	45	61.33	47.37	69.23	56.25

Figure 5.5: Prediction Result for dataset lag 3

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
VtyDayRet30d		6	1	0	0
Google_Trend		5	0.992489	0.887611	0.894329
BlkSizeByte		6	0.965631	0.241078	0.249658
ROI1yr		6	0.882235	0.815528	0.924389
commentsPerDay		4	0.807034	0	0
AdrActCnt		5	0.803931	0	0
NVTAdj		5	0.769806	0	0
Btc_Tweet_Count		6	0.759244	0.266059	0.350426

Figure 5.6: Gradient Boosting Information Table

Chapter 6

Discussion and Recommendation

6.0.1 Discussion

As observed during exploratory analysis, the Tweet and Reddit volume fluctuation show correlation with the Bitcoin price direction in the same day, however, the modelling result is indicate that the movement of Tweet and Reddit volume is likely only showing the reflection of the Bitcoin price movement on the same day and has little importance to use for predict future Bitcoin price direction. In contrast, the result has show Google trend is highly importance in both decision tree model (refer figure 4.7) and gradient boost model (refer figure 5.6) to predict the Bitcoin price direction. For example,the decision tree result showing that there are 87% and 84% probability the price will have movement if Google Trend SVI is more than 22.5. Also, the Google trend has show high ratio of validation to training importance, as compare to other secondary factors. In conclusion, our model has achieve respectable 66% accurate score to predict the Bitcoin price movement, the primary factor (Google Trend) has contribute significant to the prediction as compare to others secondary factor (Tweets volume,Reddit volume and News Headline sentiment).

6.0.2 Recommendation

Although the model has achieved 66% accuracy, but the F1 score has only 57%. We believe that the accuracy may be improve by providing additional volume of data compare to current 365 rows. Currently, the study interval is fix at November 2018 to November 2019, it's possible to extend the study interval for 2,3 and 4 years to study the overall effect and improvement on the prediction. However, this is not outside of scope of this project as some of the data is only available since late 2018 for example the Tweets and Reddit volume, also during early age of Bitcoin, we believe it has not been saturate in social networking service not until 2017 the explosion of bitcoin news is spread viral due to its radical price movement.

Furthermore, we understand Bitcoin price is highly volatile, in our current model the prediction are only limited to up,down and stagnant it is unable to predict single spike event which will have significant impact on the overall price. We think that it is more beneficial if the model is able to capture the anomaly price movement event such as the event that discussed on clustering analysis, the cluster in segment 1 and segment 5 is experienced 30-40% price movement for this 3 days. Nevertheless, this is not cover in the scope of this project.

Bibliography

[1] S. Nakaboto, *Satoshi nakaboto: 'bitcoin's market cap is now on par with netflix's'*, https://thenextweb.com/hardfork/2019/12/04/satoshi-nakaboto-bitcoins-market-cap-is-now-on-par-with-netflixs/, 2019.

[2] H. Choi and H. Varian, "Predicting the present with google trends," *Economic Record*, vol. 88, pp. 2–9, 2012.

[3] M. Ettredge, J. Gerdes, and G. Karuga, "Using web-based search data to predict macroeconomic statistics," *Communications of the ACM*, vol. 48, no. 11, pp. 87–92, 2005.

[4] G. Piatetsky, *Crisp-dm, still the top methodology for analytics, data mining, or data science projects*, Oct. 2014. [Online]. Available: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[5] coinmetrics, *Data dictionary*, https://coinmetrics.io/community-data-dictionary/, 2019.

[6] A. Hotho, A. Nürnberger, and G. Paass, "A brief survey of text mining," *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, vol. 20, pp. 19–62, Jan. 2005.

[7] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919*, 2017.

[8] N. R. S. Falguni N. Patel, "Text mining: A brief survey," *International Journal of Advanced Computer Research*, vol. 2, no. 4, pp. 243–248, 2012.

[9] *All you need to know about text preprocessing for nlp and machine learning*, https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html, Accessed: 2019-12-7.

[10] G. Miner, J. Elder IV, A. Fast, T. Hill, R. Nisbet, and D. Delen, *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.

[11] *5 heroic tools for natural language processing*, https://towardsdatascience.com/5-heroic-tools-for-natural-language-processing-7f3c1f8fc9f0, Accessed: 2019-12-7.

[12] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013, ISSN: 1877-0509.

[13] E. Johansson, *Creating-daily-search-volume-data-from*, http://erikjohansson.blogspot.com/2014/12/creating-daily-search-volume-data-from.html/, 2019.

[14] TheStar, *Bitcoin climbs to one-year high on facebook crypto pact report*, https://www.thestar.com.my/business/business-news/2019/06/18/bitcoin-climbs-to-oneyear-high-on-facebook-crypto-pact-report/, 2019.

[15] B. Bambrough, *The next bitcoin price rally could be sparked by a surprising source*, https://www.forbes.com/sites/billybambrough/2019/03/26/the-next-bitcoin-price-rally-could-be-sparked-by-a-surprising-source/#2c664a337380, 2019.