

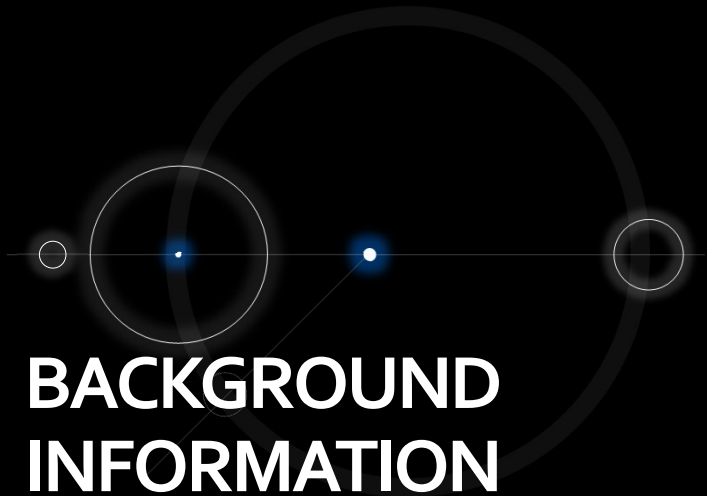
CRYPTOCURRENCY PRICE DIRECTION PREDICTION

ID:WQD180093
NAME: L CHONG MING KEAT

OBJECTIVE

To Predict Cryptocurrency price direction with:

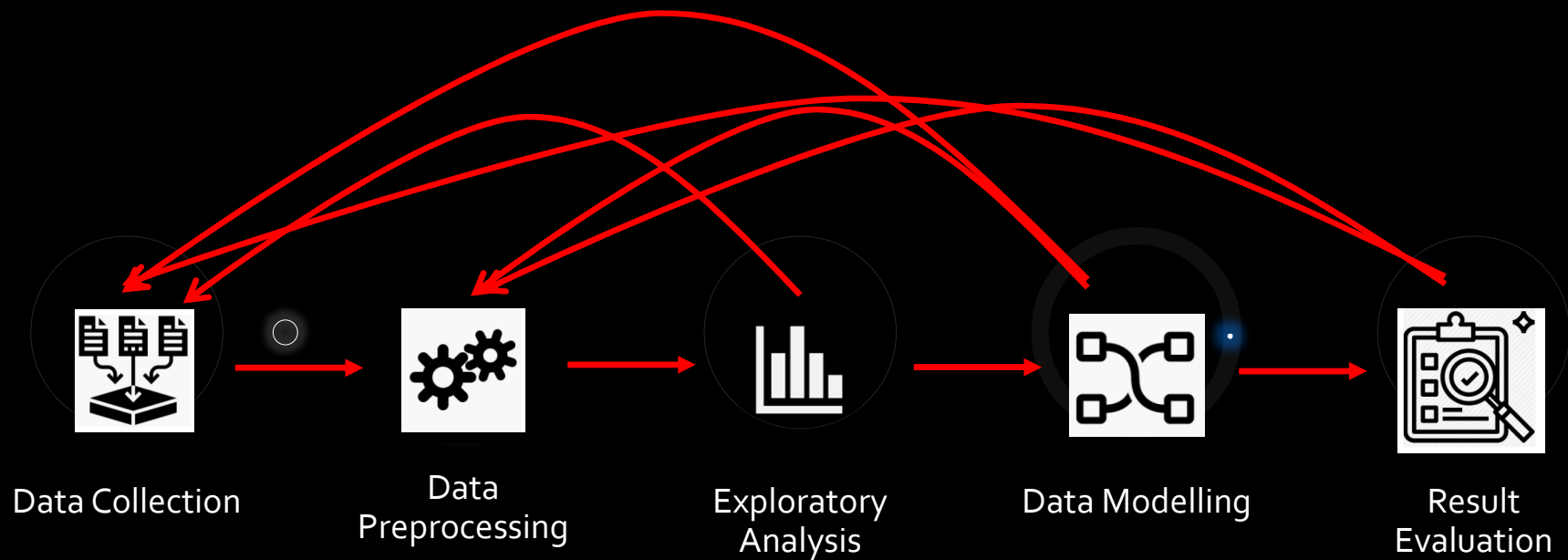
- 1.) Historical Price Information
- 2.) Crypto News Headline Sentiment
- 3.) Tweets Volume
- 4.) Reddits Volume
- 5.) Google Trend



BACKGROUND INFORMATION

Bitcoin is the largest cryptocurrency in terms of market capitalization. Around 130bil (2019) around 67% of the total crypto market. Bitcoin is opted to be the primary focus on the crypto price direction prediction in our study.

The Process



The process is not one way !
Revisiting on Data Collection and Preprocessing is often required.

DATA COLLECTION



Bitcoin Price

Source: Coinmetrics

Method: Download



Tweets Volume

Source: Bitinfocharts

Method: Web-crawling



News Headline

Source: Investing

Method: Web-crawling



Reddit Volume

Source: Redditmetrics

Method: Web-crawling



Google Trend

Source: Google

Method: Download



BITCOIN PRICE

- The interval for the price data collection is in daily.
- The target duration for the study is 1 year (23/11/2018 -23/11/2019)
- Data obtained contain total 40 Columns and 365 Rows
- Full Data description can be view at <https://coinmetrics.io/community-data-dictionary/>

Name	Model Role	Measurement Level	Description
Date	TIMEID	INTERVAL	Date
AdrActCnt	INPUT	INTERVAL	The sum count of unique addresses that were active in the network
BlkCnt	INPUT	INTERVAL	The sum count of blocks created that day that were included in the main (base) chain.
PriceUSD	INPUT	INTERVAL	The fixed closing price of the asset as of 00:00 UTC the following day
VtyDayRet 6od	INPUT	INTERVAL	The 6oD volatility, measured as the deviation of log returns



Tweets & Reddit Volume

- The interval for the tweet and reddit data collection is in daily.
- Data obtained contain total 2 Columns and 365 Rows (Date and Tweets Volume)

Name	Model Role	Measurement Level	Description
Date	TIMEID	INTERVAL	Date
Volume	INPUT	INTERVAL	The total volume of tweets/reddit comment posted on that day

Google Trend

- The interval for the google trend data collection is in monthly and daily.
- Data obtained contain total 2 Columns and 365 Rows (Date and Search Index)

Crypto News

- The interval for the crypto news collection is in daily.
- Data obtained contain total 2 Columns and 365 Rows (Date and Tweets Volume)

Data Preprocessing

Google Trend Background Information

- Google trend data provides information on how popular given search terms are relative to other search terms at any given time.
- This provide a proxy metric for the general interest there is in crypto at any given time.
- Google does not provide search volumes but search volume index.
- Search volume index is calculated by dividing each data point by the total searches within a geographic region and time range.
- It can be ranging from 0-100.
- When the trend data queried more than 90 days, weekly SVI will be return instead of daily.
- To adjust the value to daily index, we follow the method by Erick Johansson.

<http://erikjohansson.blogspot.com/2014/12/creating-daily-search-volume-data-from.html>

Data Preprocessing

Google Trend Adjusting

- Step 1: Collect daily search data from Google Trends and combine it into one array.
- Step 2: Collect weekly search data over the same time period
- Step 3: Adjust the daily data based on the weekly data

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

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

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

<http://erikjohansson.blogspot.com/2014/12/creating-daily-search-volume-data-from.html>

Data Preprocessing

Cryptocurrency News Headline Sentiment

- Total 18356 row of headline news is crawled.
- The News headline are less spatial compare to tweets comment.
- Removed short words = 3
- Lower casing
- Removed numbers
- Removed stopwovrds
- Removed punctuation mark
- Lemmatization
- Python Package Vader (Valence Aware Dictionary for sEntiment Reasoning) is apply to each headline to obtain sentiment score.
- The sentiment score will be aggregate based on the average of each individual day.

Data Preprocessing

Bitcoin Data

- Original 40 columns of data is reduce to 18 based on correlation study, attribute with 1.0 correlation is prune and consider redundant.

	AdrActCnt	Blk SizeByte	Blk SizeMean 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
Blk SizeByte	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
Blk SizeMean Byte	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

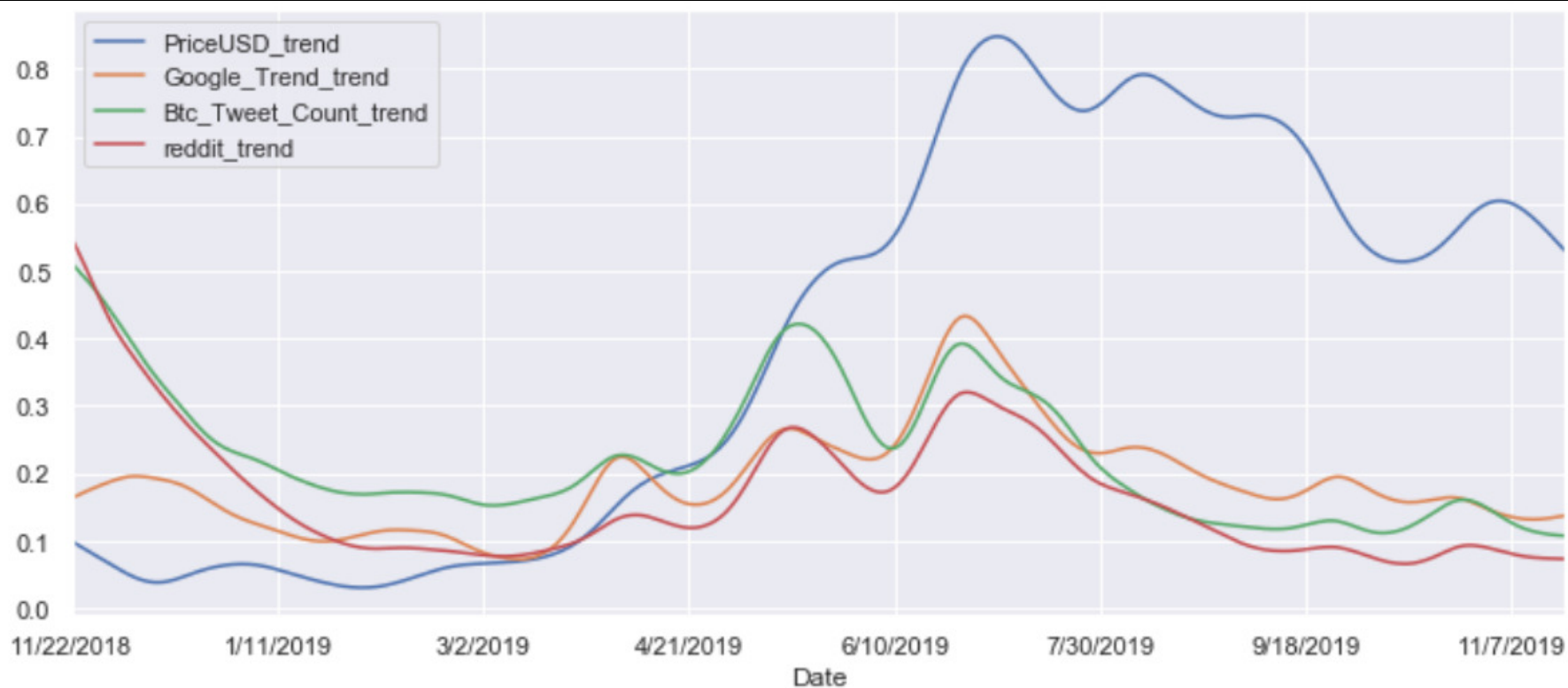
EXPLORATORY



EXPLORATORY



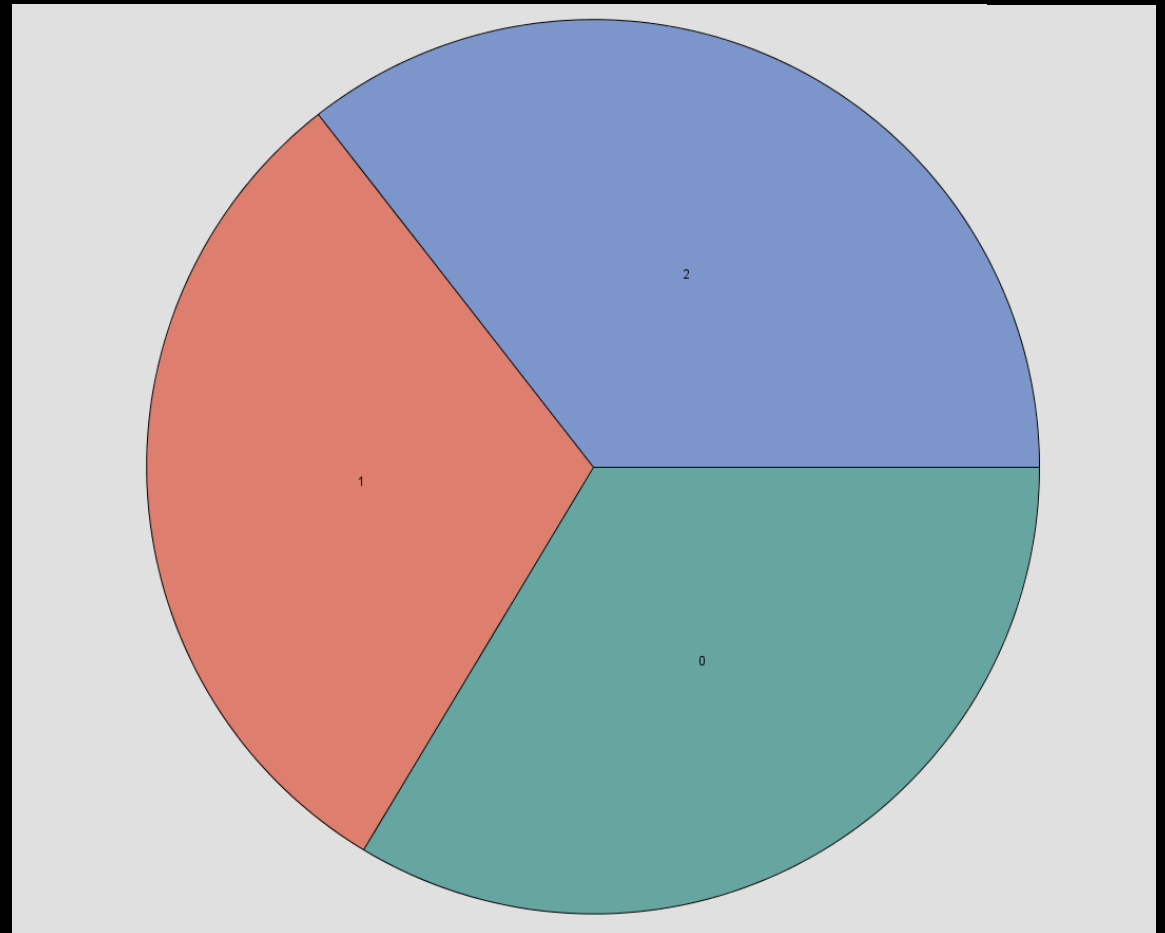
EXPLORATORY



EXPLORATORY

Pie Chart For Price Direction

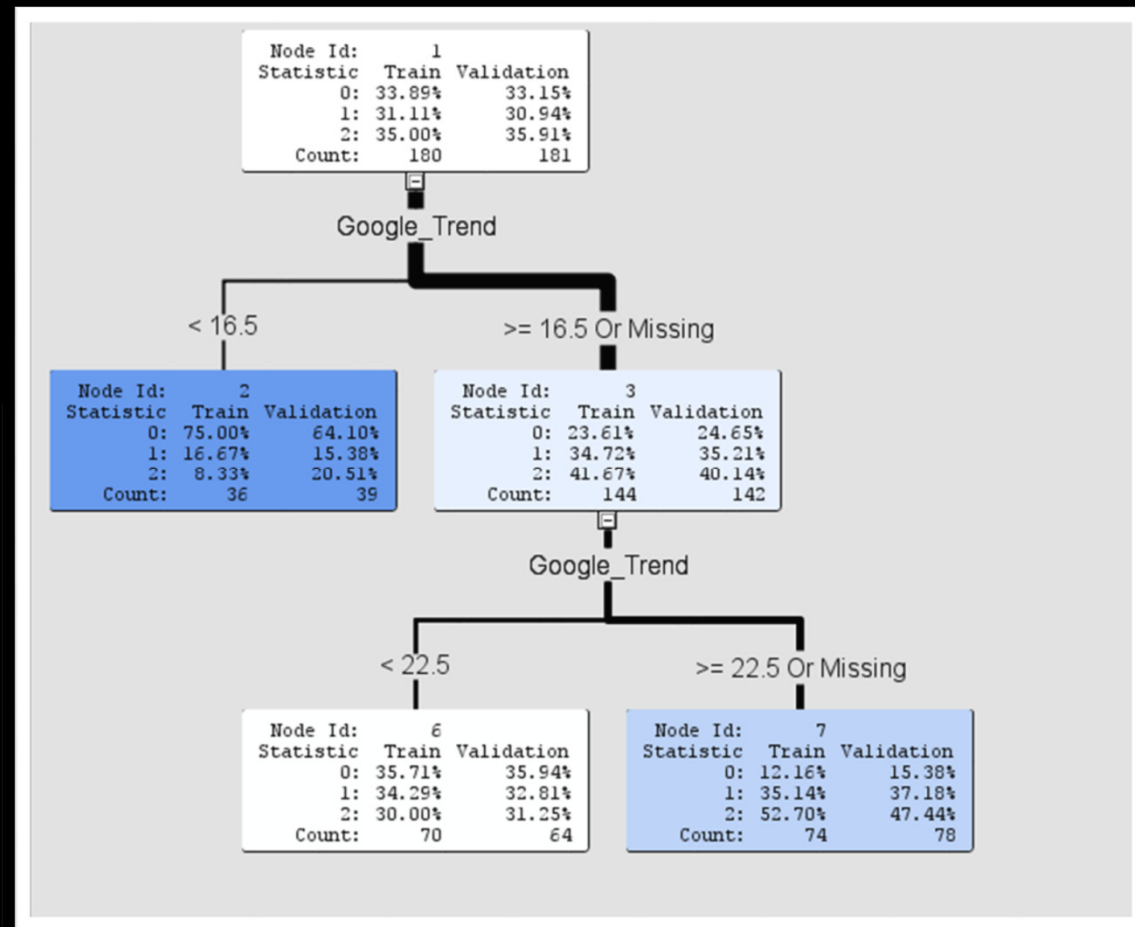
- 0: No price changes (Green 34%)
- 1: Price drop (Red 31%)
- 2: Price increase (Blue 35%)



EXPLORATORY

Decision Tree

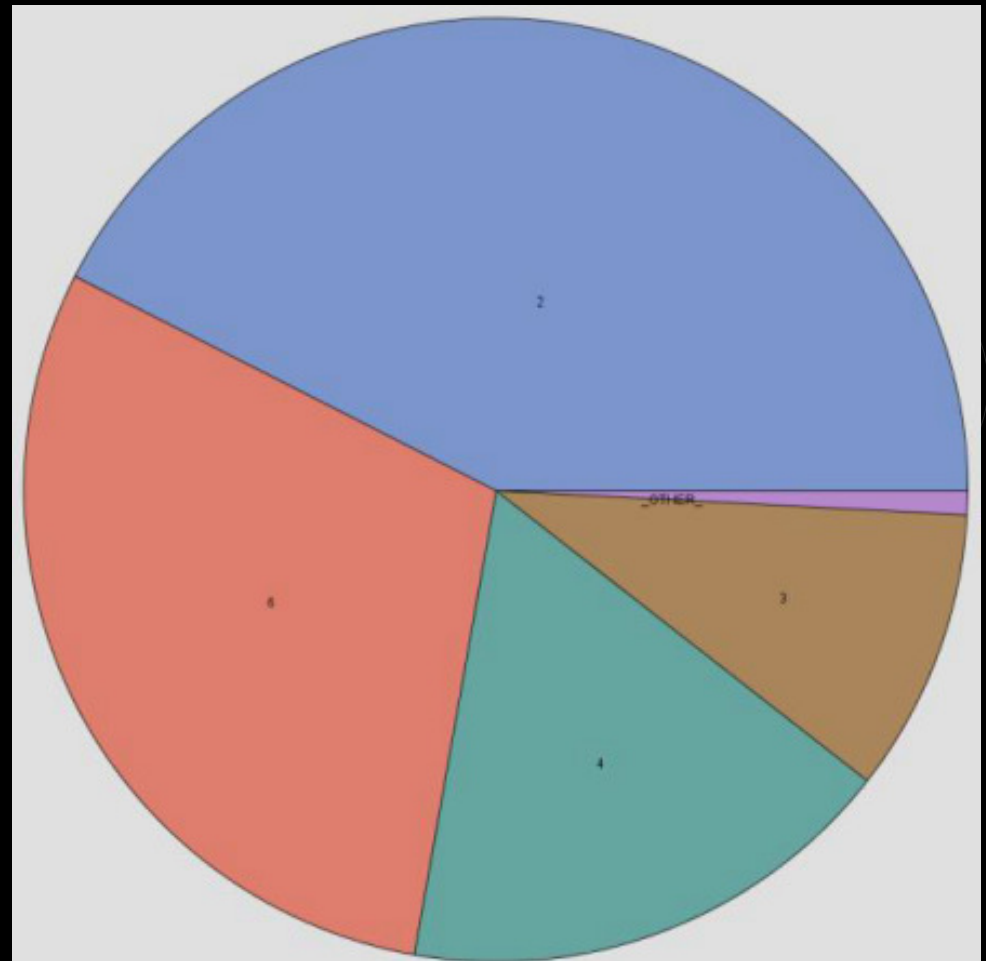
- With Lag 2 parameter, google trend index has high information gain for predict the price direction.
- Google Trend <16.5 the train and validation has 75% and 65% to be category 0 indicate no price change.
- Google Trend >22.5 has 53% to 48% to be category 2 and 88% to 85% the price will be change category 1 or 2



EXPLORATORY

Clustering Analysis

- Total of 6 segment is created with SAS clustering profile node.
- Segment 2 has 155 records, segment 3 has 36 records, segment 6 has 109 records and segment 4 has 62 records other has total of 3 records



EXPLORATORY

Clustering Analysis

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	comments PerDay
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

For segment 1 there's only single observation, however:

- The price_diff is highest for the record period with 1137usd (18%) increase on that single day
- Google Trend search volume was highest 100 on same day
- Tweets volume and Reddit volume was highest on same day 42k tweets and 4.4k comments

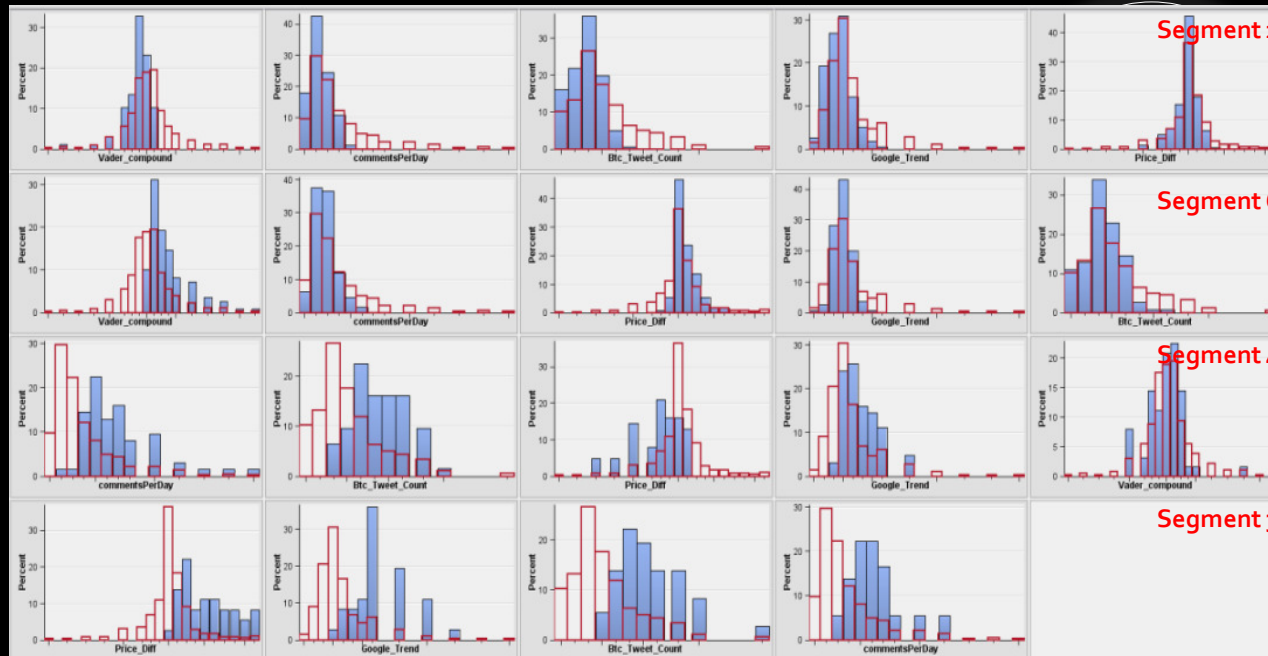
For segment 5 there's two observation recorded:

- The price_diff has significant drop on that two days (26th Jun -1753 and 16th July -1455)
- Google Trend search volume was not as high as first segment 61
- Tweets volume and Reddit volume also relatively lesser than segment 1 but slightly higher than rest of the segment

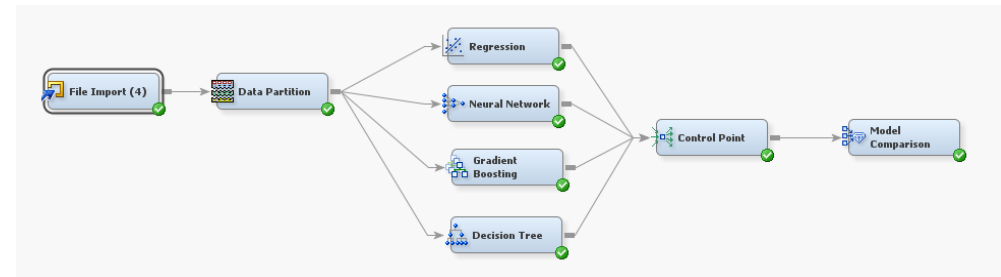
EXPLORATORY

Clustering Analysis

Mean Histogram (Segment Profile)



MODELLING AND EVALUATION



MODELLING AND EVALUATION

Input and Model

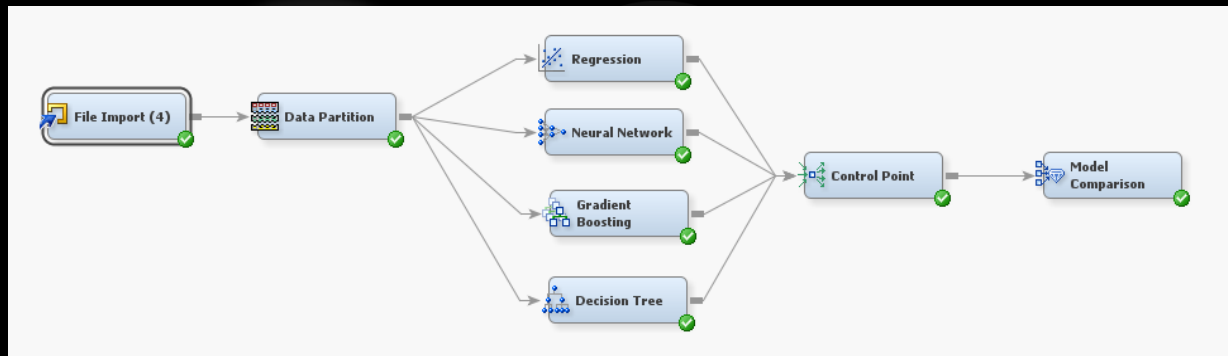
- Bitcoin prices, Tweets and Reddit volume, Google Trend, News sentiment score are combined in single table with date as index.
- The price direction is set as target for training and prediction. The level is set a nominal.
- Date is input as timeID
- Remaining input for level are set as interval.

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

MODELLING AND EVALUATION

Input and Model

- 4 different machine learning model is select for training and evaluate the data (Logistic Regression, Neural Network, Gradient Boosting and Decision Tree).
- 3 separate dataset with lag of (1 day, 2day ,3day) is insert in SAS for sensitivity study.



MODELLING AND EVALUATION

Result

- 4 different machine learning model is select for training and evaluate the data (Logistic Regression, Neural Network, Gradient Boosting and Decision Tree).
- Highest Accuracy observe is 66.48% by decision tree model in lag 2 dataset and 69% for Recall
- Gradient Boost with lag 3 is highest performance in overall category; 65%Accuracy, 51% precision, 65% Recall and 57% F1 score.

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

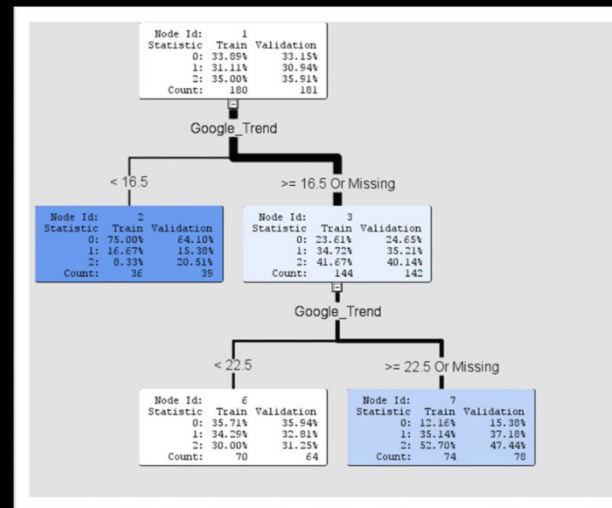
Lag 2									
Model	Node	FN	TN	FP	TP	Accuracy	Precision	Recall	F1 score
Regression	Train	19	85	32	44	71.67	57.89	69.84	63.31
Regression	Validate	37	73	44	28	55.49	38.89	43.08	40.88
Neural	Train	19	90	27	44	74.44	61.97	69.84	65.67
Neural	Validate	40	69	48	25	51.65	34.25	38.46	36.23
Boost	Train	11	109	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

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	72.17
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

MODELLING AND EVALUATION

Discussion

- In the earlier exploratory session, decision tree diagram with lag 2 is shown, google trend play high important role for the information gain.
- As for the importance factor for gradient boost (lag 3), we can observed that google_trend is also significant to the contribution of model in both validation and training. Tweet and Reddit volume has less contribution.



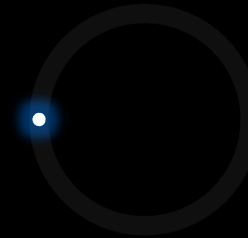
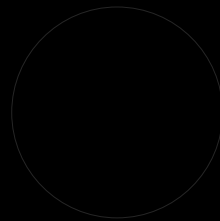
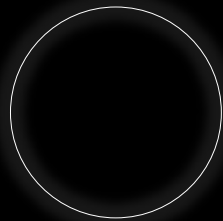
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
VtyDayRet30d		6	1	0	0
Google_Trend		5	0.992489	0.887611	0.894329
BlkSizeByte		6	0.965631	0.241078	0.249658
ROI1yr		6	0.882235	0.815528	0.924389
commentsPerDay		4	0.807034	0	0
AdrActCnt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

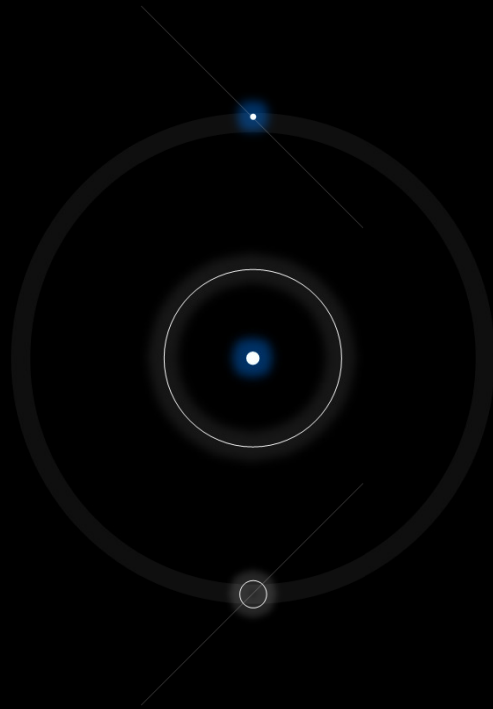
Conclusion

- The tweet and reddit volume fluctuation can cause the price changing in both way, it's difficult to detect the exact influence.
- Google trend search index prove to be relatively high importance in prediction model as compare to the tweets and reddit volume as well as news sentiment.
- Highest accuracy is obtained by decision tree model with lag 2, and overall more robust model would be gradient boost with lag 3.
- Failed to observe any groundbreaking observation.
- Bitcoin price is highly volatile and the model not able to predict single spike event which will have significant impact in the overall return of investment.

Reference


- Salač, A. (2019). *Forecasting of the cryptocurrency market through social media sentiment analysis* (Bachelor's thesis, University of Twente).
- Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, 1(3),





THANK YOU

CHONG MING KEAT 

wqd180093@siswa.um.edu.my 

https://github.com/JechtChong80/WQD7005_DataMining 