NewsBit

Investment decision support tool built using machine learning models

Karan Raja - Kaidi Wu - Linhan Cai - Yunhua Su - Zhiwei Gu

Capstone Advisor: Dr. Lee Fleming

Team Background

Karan Raja

B.S. in Finance
M.Eng in IEOR

Yunhua Su

B.S. in CEE
M.Eng in CEE

Linhan Cai

B.S. in Data Science
M.Eng in IEOR

Zhiwei Gu

B.S. in Applied Math
M.Eng in IEOR

Kaidi Wu

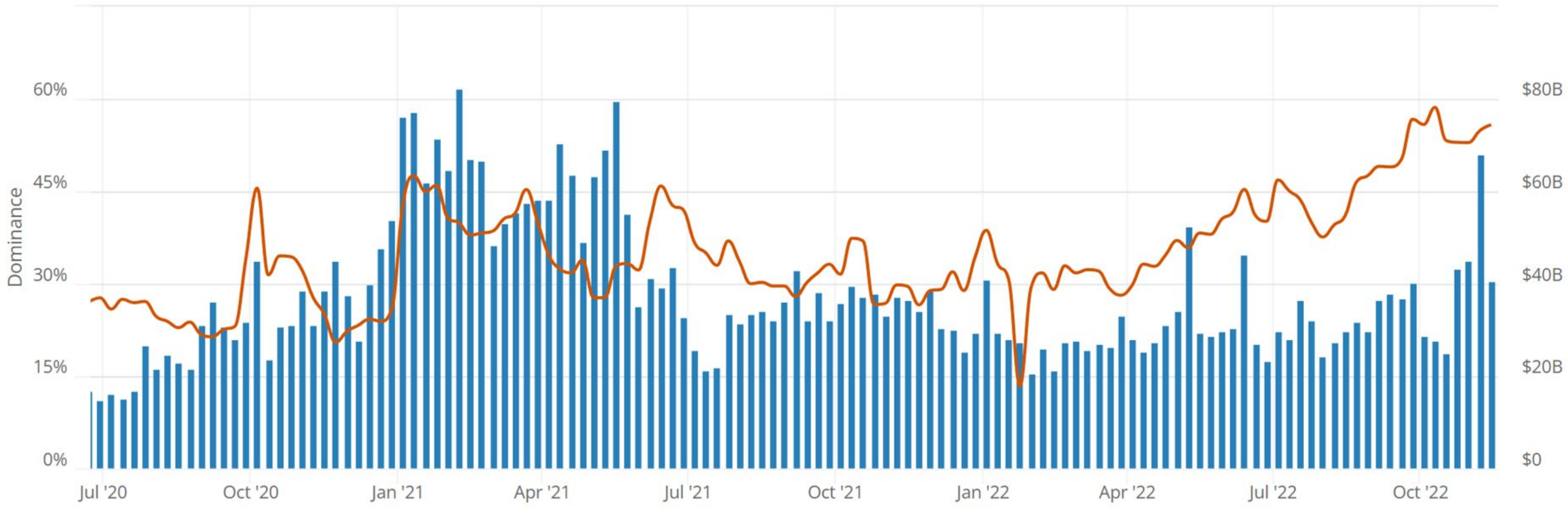
B.S. in System Engineering
M.Eng in IEOR

Capstone Advisor: Dr. Lee Fleming

“Unlike traditional assets, there is no means to measure the true value of bitcoin. Hence, we want to identify the factors that most influence the price of Bitcoin. To do this, we will use socio economic indicators and apply several machine learning models to identify the most effective price prediction strategy”

—NewsBit Goal

Bitcoin daily trade value since July 2020 was between $18B and $82B



2. Crypto Slate; Bitcoin Trade Volume, Value and Dominance - Accessed February 2022

“There are over 50M retail traders that make up 17% of crypto trading volume”

GlassNode

3. Glassnode: Portion of Bitcoin Supply Held by Retail Investors Reaches All-Time High

Retail Investors

- Most are millennial
- Approximately $150,000 annual income
- Strong social media influence
- Little to no background in investing / personal finance
- Invest for themselves

Institutional Investors

- Manage between million to trillion USD for clients
- Build strategies based on fundamental analysis
- Several years of experience
- Invest on behalf of institutions such as pension funds, endowments, and HNIs

"Over 80% of new crypto investors, inevitably lost money on their initial investment."

— Bank of International Settlements

9. GizModo: Majority of People who invest in crypto inevitably lose their money, study shows

“Since no framework exists to determine the price of Bitcoin, there is significant scope to use statistical learning models to develop a framework which encompasses a mix of financial, economic, and social indicators”

—NewsBit Thesis

Executive Summary

“Retail investors need tools that are easy to interpret and datasets that are easy to obtain”

01

Easy to access

All variables used can be easily accessible to retail investors - with the exception of social media and news article sentiment

02

Linear regression

With a high test-set accuracy and low MAE, Linear regression provides an accurate and easy to interpret method to support this decision.

03

Decision Tree Classifier

Decision tree classifiers are also an easy to interpret method providing high accuracy and low MAE

“Institutional investors need tools that optimize the risk-return function”

01

Public information

All variables used can be easily accessible to retail investors - with the exception of social media and news article sentiment. This can however, be purchased from companies that provide ‘social-listening’ analytics services

02

Gradient Boosting

With an accuracy of 0.637 and MAE of 0.066 this approach offers the best risk-reward tradeoff.

03

GBR with K-fold

Institutional investors have the infrastructure to support big data collection and processing in an efficient manner

04

Include more features

improve the performance of boosting models by including significant features - something that is possible for institutional investors

Over 40 indicators evaluated

Social Indicators

- News Media
  - Article Count
  - Article Sentiment
- Twitter
  - Tweet Count
  - Tweet Sentiment
- Reddit
  - Post count
  - Post sentiment
  - Post comment count

Financial Indicators

- Equities
  - S&P 500
  - DJI
  - Nasdaq
- Fixed Income
  - US 1Y Treasury
  - US 10 Y Treasury
- Commodities
  - Gold
  - Copper
  - Crude Oil
- Currencies
  - Euro-USD
  - Singapore-USD

Feature Engineering

- Lagged Features
  - 1d, 3d, 7d, 14d
- Moving Averages
  - 3d, 7d, 14d
- Time Based Features
  - Day of week

Summary of all models used

Classification

Determine if the price of bitcoin one week from today is higher (1) or lower (0) than the price today

	Interpretability	R2 / Accuracy	OSR2	MAE
Logistic	High	0.627	0.295	0.373
Decision Tree	High	0.699	0.151	0.209
Random Forest*	Low	0.335	0.376	0.346
Boosting*	Low	0.181	0.671	0.132
KNN	Low	0.65	F1 0.63	-

Gradient boosting offers the best test-set performance along with the lowest mean error

Regression

Determine the % change in price of bitcoin one week from today based on the price today

	Interpretability	R2 / Accuracy	OSR2	MAE
Linear	High	0.996	0.993	0.324
Decision Tree	High	0.804	0.208	0.254
Random Forest*	Low	0.479	0.491	0.051
Boosting	Low	0.219	0.637	0.066
SVM	High	0.339	0.33	8871.176
Time Series (AR)	Low	0.997	0.987	2336.706
Time Series (RF)	Low	0.731	0.704	11774.009
Neural Network	Low	0.908	-	3487.607

Linear Regression offers the best test-set performance while Boosting offers the lowest mean error

Lack of complete data is an inherent challenge

Possibility of bias based on the data collected

Incomplete information may have produced results that are not indicative of market performance

Access to social media data (Twitter / Reddit) is becoming increasingly restrictive

Abundant data available, but it is a time consuming to process it correctly to derive insights

Given the growing popularity of crypto activity, there is significant room for growth and improvement

Include economic indicators such as Real GDP, CPI, Unemployment rate, among others

Include additional social media channels such as StockWits, YouTube, and TikTok

Process video content and derive sentiment and include traditional media (CNN / Fox News)

Consider shorter time intervals (hours - minutes - seconds) to build a more robust system for traders

Opportunity to scale this across other esoteric assets such as wine, art, other collectibles

Appendix

Acknowledgements

Capstone Advisor
Prof. Fleming

Capstone Mentor
Nick Farell

Mentor
Prof. Grigas

Capstone Writing
Prof. Bauer

Tools Used

Project Planning and Coordination

GitHub

jupyter

CO

Data Collection

kaggle

Hugging Face

Data Processing

Selenium Python

TextBlob

Financial Market Indicators

Financial market indicators encompassed all asset classes with uniform information that is easy to collect for any user

These indicators are helpful in determining the overall investor sentiment as well as the affinity to absorb risk

Since bitcoin is a non-traditional investment, viewing the activity in traditional investment avenues was insightful

To some extent, financial market indicators are also helpful in determining the overall economic activity and sentiment

Social Indicators

Social media has become a new avenue for people to share information about themselves. It is also gaining popularity among people seeking and giving financial advice

The role social media platforms such as Reddit and Twitter play in influencing investment decision for retail investors has grown tremendously since 2021 - with influencers garnering millions of views on content directed towards portfolio management, 401k planning, and others

Social media sentiment gives us an idea of the engagement certain asset types garner among millennial and Gen-Z investors

Linear Model Approach + Results + Reflections

80% Training | 20% Test

17^{th} AUG 2021 — 22^{th} SEP 2022 — 4^{th} APR 2023

ModelOG	ModelOG2
0.993	0.960
0.994	0.939
30 Variables	9 Variables

- All values had $P > |t|$ values of less than 0.05
- All values had VIF scores less than 10

Variables:

- Price of bitcoin from one week prior (x1)
- Price of the SP500 (USD) (x2)
- Price of crude oil (USD) (x3)
- Price of gold (USD) (x4)
- Price of copper (USD) (x5)
- Price of corn (USD) (x6)
- USD to SGD exchange rate (x7)
- Average twitter sentiment for the day (x8)
- Intraday price movement (x9)

ModelOG2 Linear Equation

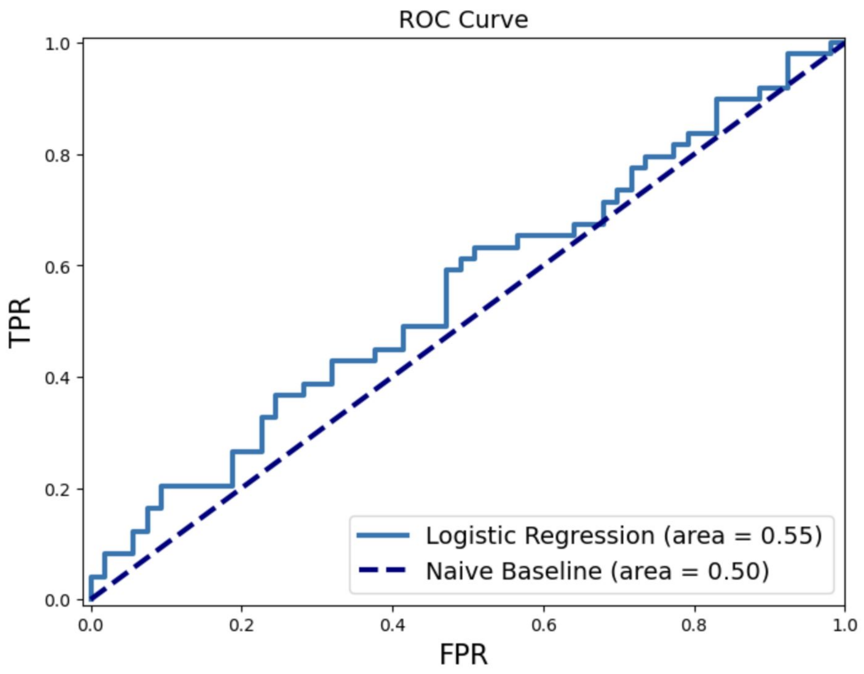
$$-0.0005 + 8.55x_1 + 6.88x_2 + 99.70x_3 - 11.21x_4 + 2582.60x_5 - 1122.40x_6 + 0.0002x_7 + 0.0003x_8 - 961.92x_9$$

- Positive correlation with the one-week historic BTC price and interday movement – historic prices can be informative
- Positive correlation with the price of SPY – an indication that traditional investing has influence on BTC price
- Positive correlation with the price of crude oil – BTC prices can be linked with oil supply
- Negative correlation with the price of gold – BTC prices can be linked with inflation
- Positive correlation with the price of copper – BTC price may be linked with industrial output
- Negative correlation with the price of corn – farm-payroll and activities influence the BTC price
- Positive correlation to the SGDUSD rate – BTC price is linked with the movement of the USD
- Positive correlation with the Twitter sentiment – BTC price is linked with social media trends and sentiment

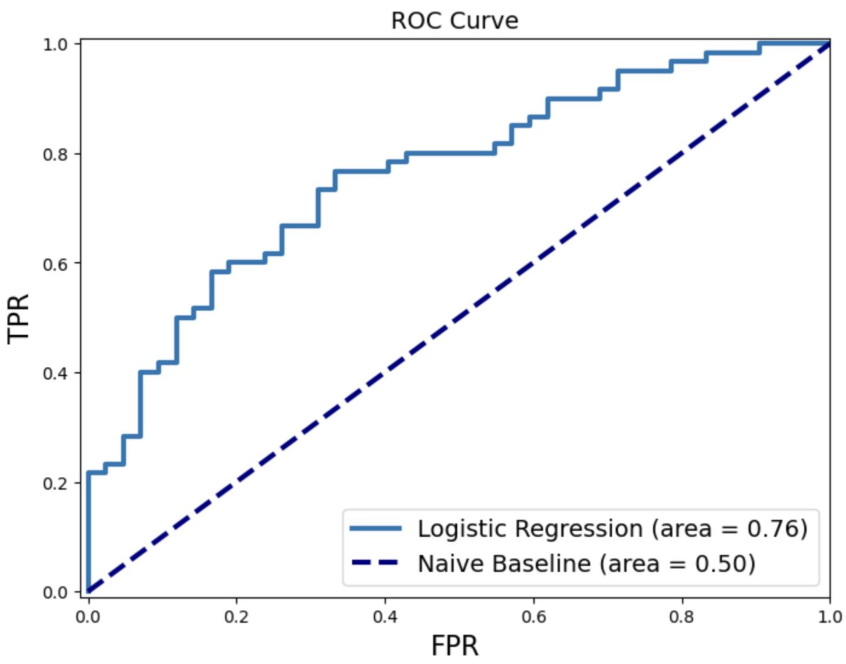
Logistic Model Approach + Results + Reflections

ROC Curve Graph

- Day-to-Day prediction performs better than predicting same-day closing price
- Significant variables suggest classifying next-day opening price relies more heavily on crypto information than media sentiment value



Prediction for current day closing price



Prediction for next day opening price

Classification for Current-day Price Movement

- Same train-test split as logistic regression
- Accuracy: 55.9%
- MSE: 0.441

```
Confusion Matrix :
 [[51  2]
 [43  6]]
```

Variables with P-value < 0.1

```
Bitcoin_Price_Previous_Day_Open
Litecoin_Price
SP500
UST_1Y_Maturity
UST_10Y_Maturity
Euro
All_Twitter_Posts
Positive_Tweets
Neutral_Tweets
```

Classification for Next-day Price Movement

- Same train-test split as logistic regression
- Accuracy: 62.7%
- MSE: 0.373

```
Confusion Matrix :
 [[37  5]
 [33 27]]
```

Variables with P-value < 0.1

```
Bitcoin_Price
Bitcoin_Price_Previous_Day_Open
Bitcoin_Price_One_Week_Prior_Open
Litecoin_Price
Crypto_Global_Ranking
```

Tree Based Models Approach + Results + Reflections

CART Model

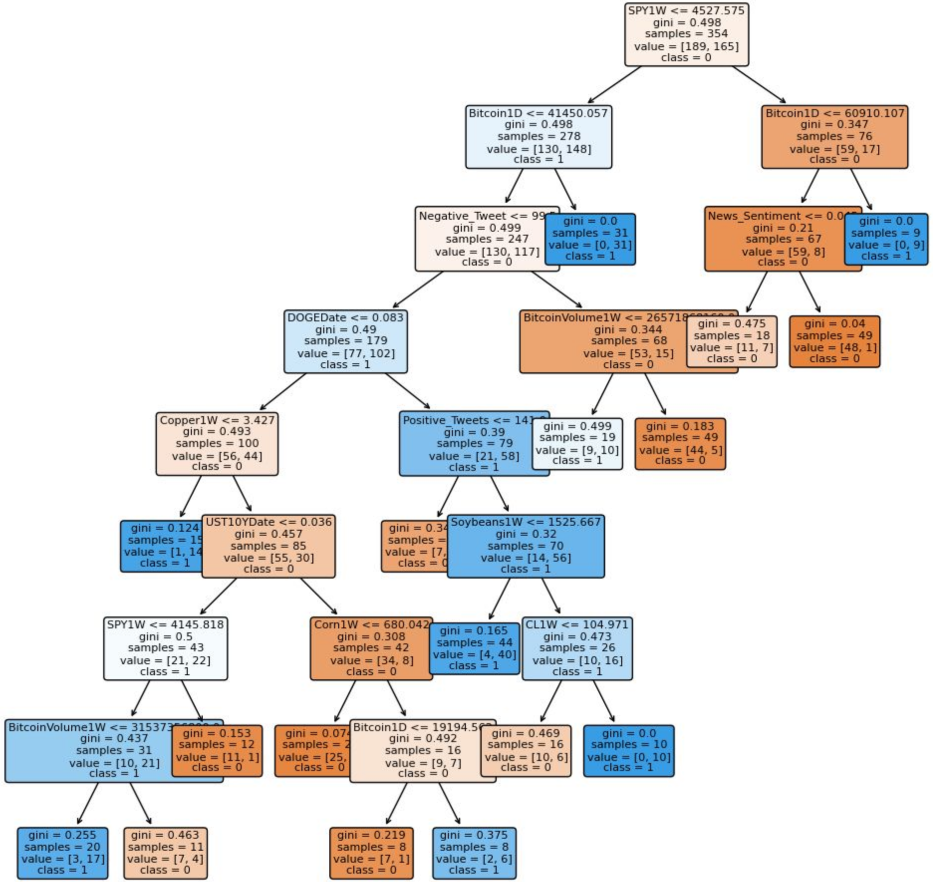
Classification

Bitcoin Price Movement – 1W

Test Size: 30%
Train Size: 70%

Baseline Accuracy: 46.6%
Model 93 Accuracy: 69.9%
TPR: 0.6761
FPR: 0.2805

Minimum Sample Size = 8
CCP Alpha = 0.008



Random Forest Model

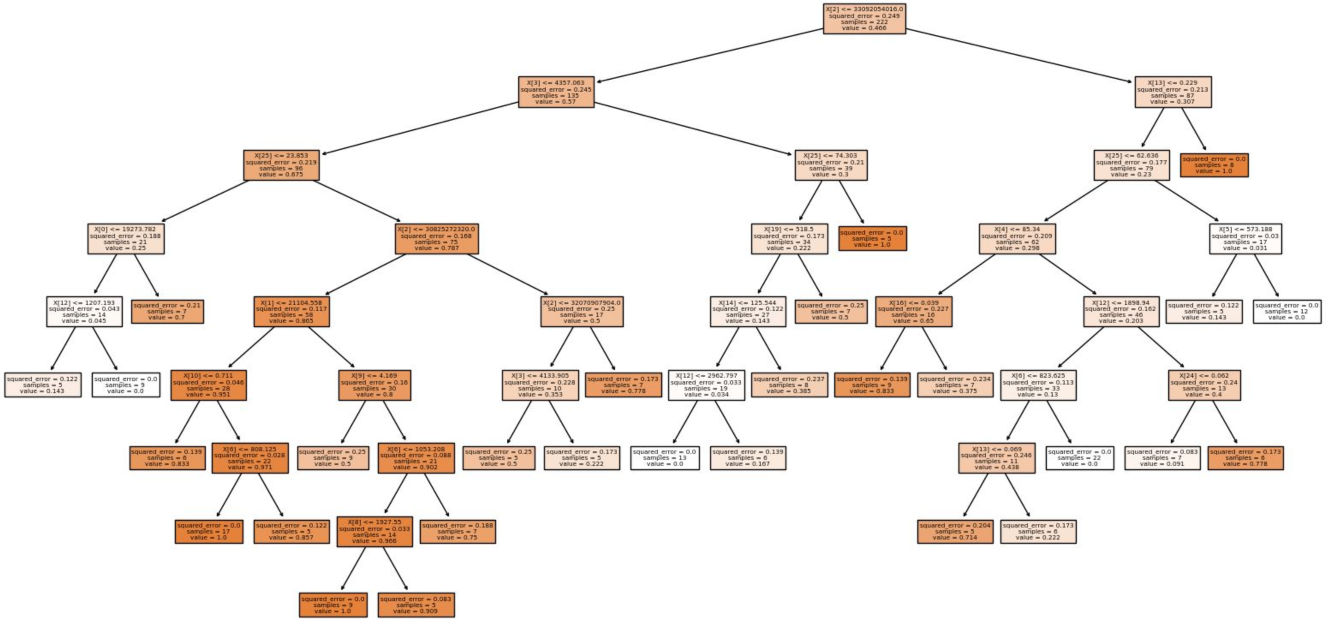
Classification

Bitcoin Price Movement – 1W

Test Size: 30%
Train Size: 70%

Max Features = 5
Minimum Sample Leaf = 5
N-Estimators = 500
Verbose = 2

OSR2 = 0.37573



CART Model

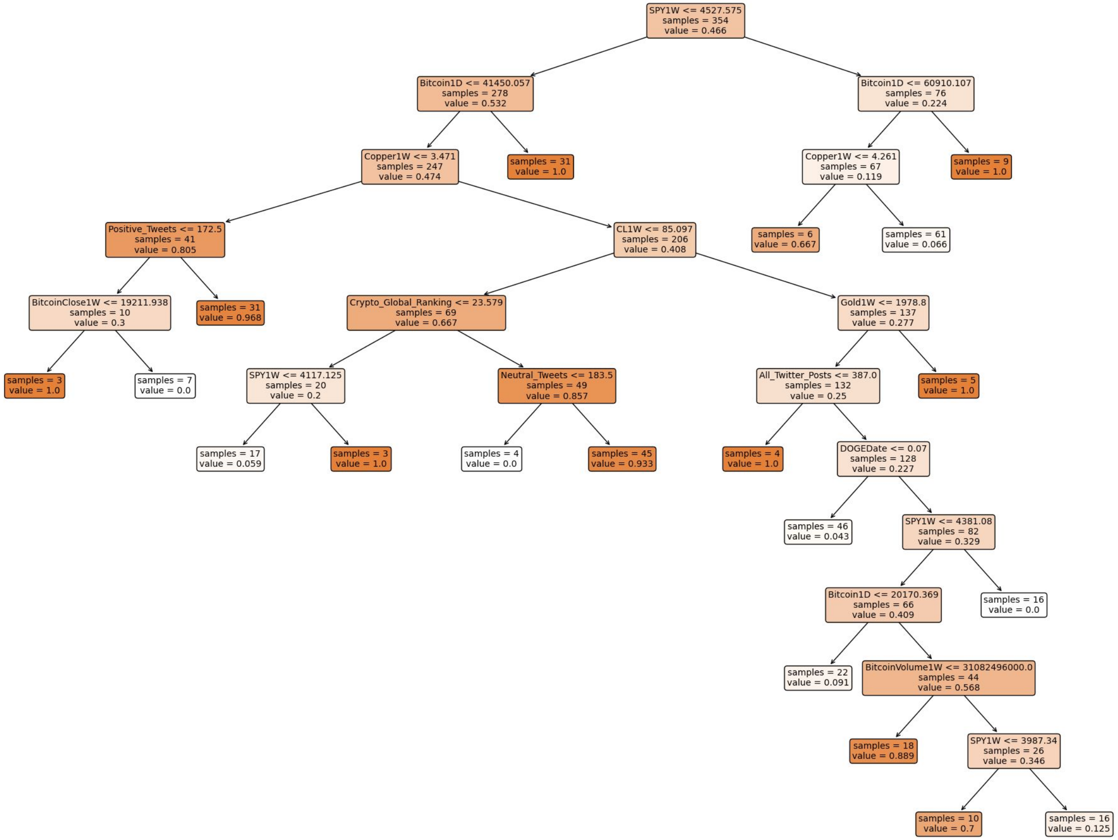
Classification

Bitcoin Price Movement – 1W

Test Size: 30%
Train Size: 70%

Baseline Accuracy: 46.6%
Model 93 Accuracy: 69.9%
TPR: 0.6761
FPR: 0.2805

Minimum Sample Size = 8
CCP Alpha = 0.008



Random Forest Model

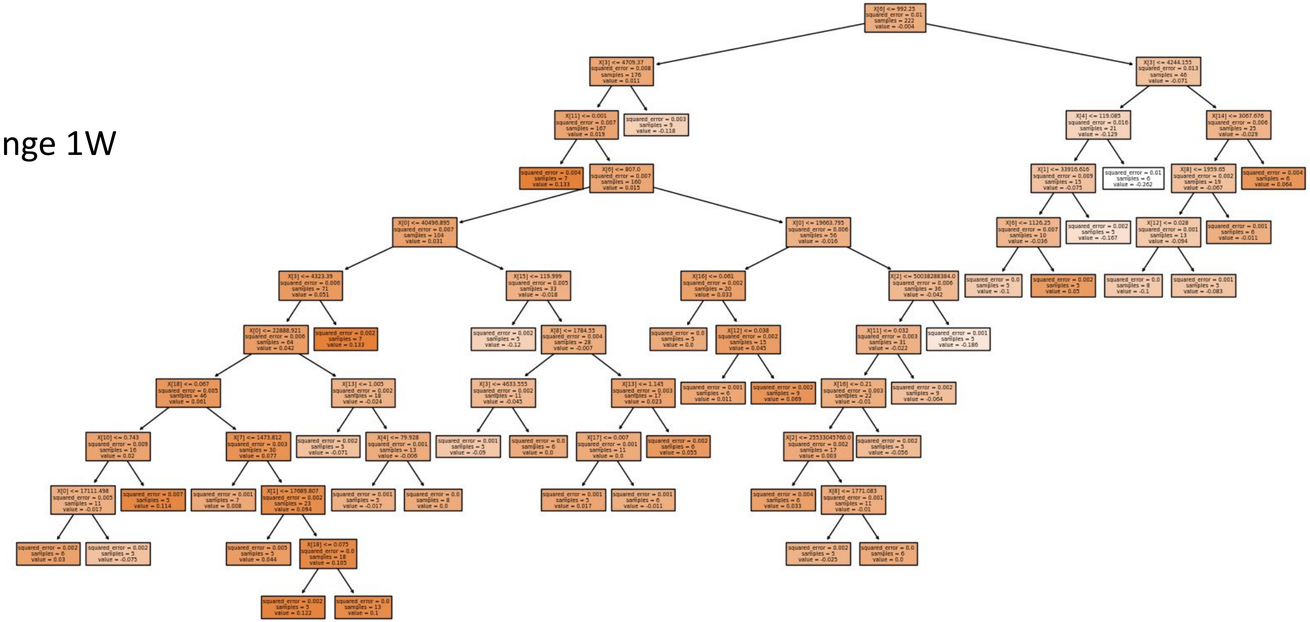
Regression

Bitcoin Price Movement – Change 1W

Test Size: 30%
Train Size: 70%

Max Features = 5
Minimum Sample Leaf = 5
N-Estimators = 500
Verbose = 2

OSR2 = 0.53702



Classification	RF – OSR2 Value	RFCV – OSR2 Value	GBR – OSR2 Value	GBRCV – OSR2 Value
Model 93	0.36484	0.37573	0.18055	0.47122

Regression	RF – OSR2 Value	RFCV – OSR2 Value	GBR – OSR2 Value	GBRCV – OSR2 Value
Model 93	0.45989	0.49115	0.21902	0.67122

K-Fold Cross Validation improved the performance of **ALL** models on the testing dataset
Marked improvement in the OSR2 values after implementing K-Fold cross validation

Applying K-Fold cross validation on the Gradient Boosting model tool considerable time and computing power
Split size = 5 data was divided into 5 equal sub-datasets
Random state = 333

- For classification, gradient boosting regression had the best performance amongst all model types
- Model 93 had the best performance for CART
    - This highlights that is important to identify suitable variables
    - Feature engineering is important to extract value from the selected variables

- Model 53 had the best performance for both RF and GBR
    - This indicates that while media sentiment was important, inclusion of additional relevant features is essential in building a robust model

- Despite problems with interpretability, it is recommended that people use GBR or RF over decision tree classifiers.

- The K-Fold cross validation approach improved the performance of Random Forest and GBR models
    - Improvement with RF was positive as seen in change of OSR2
    - Considerable improvement in GBR performance
    - This suggests that while the K-Fold cross validation approach takes time and processing power, it is essential in developing a suitable GBR model
    - The same may not be true for a RF

Classification

- Goal was to predict if the price will move up or down over one week (1 / 0)

- While Model93 performed best with the decision tree classifier, Model 53 performed best for both Random Forest and Gradient Boosting

- In all cases the use of cross validation improved the model performance and reduced the mean absolute error
- Model performance was unchanged when we tried to normalize the price functions using Log values

- Given the dataset size for the three models (13, 53, and 73) it took considerable processing time and power to complete the cross validation for the gradient boosting.
  - Model 13: 9,178s
  - Model 53: 12,322s
  - Model 73: 35,813s
  - Model 93: 5,797

- Decisions pertaining to the purchase of sale of Bitcoin is always time sensitive.
- While the goal is to maximize profits, an investor will always consider the opportunity to minimize losses first

- The Gradient Boosting with **Model 53** provides the optimum balance between high OSR2 and low MAE

Regression

- Goal was to estimate the percentage movement of price over one week –
  - Range of values were given as -0.5 to + 0.5 in models using dataset 2
  - Range of values were given as -0.3 to + 0.1 in models using dataset 23

- For regression we created a new column in the respective dataframe to determine the weekly change %

- The program does not accept the dataset if there is only one instance of a given value in the column
  - Only one day when the price changed by 0.8971%
  - Used rounding to get a more normalized set of values

- While Model93 performed best with the decision tree classifier, Model 53 performed best for both Random Forest and Gradient Boosting

- In all cases the use of cross validation improved the model performance and reduced the mean absolute error
- Model performance was unchanged when we tried to normalize the price functions using Log values

- Decisions pertaining to the purchase of sale of Bitcoin is always time sensitive.
- While the goal is to maximize profits, an investor will always consider the opportunity to minimize losses first

NN
Approach + Results + Reflections

Model Building

How we improve the model:

- Cleaning the out-liners
- Adding new layers
- Using a learning rate scheduler

01

Preprocessing

Scale the input features with StandardScaler

02

Adding layers

4 layers(RELU, purelin), learning rate=0.001, loss function: MSE

03

Output

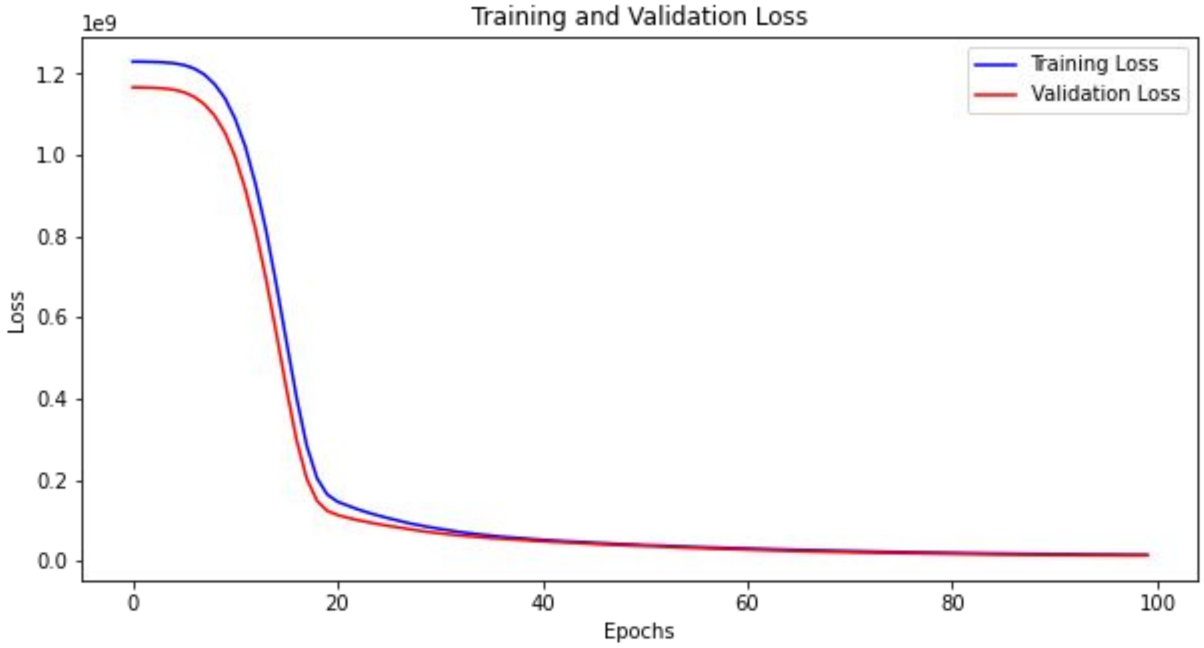
Predict

04

Evaluation

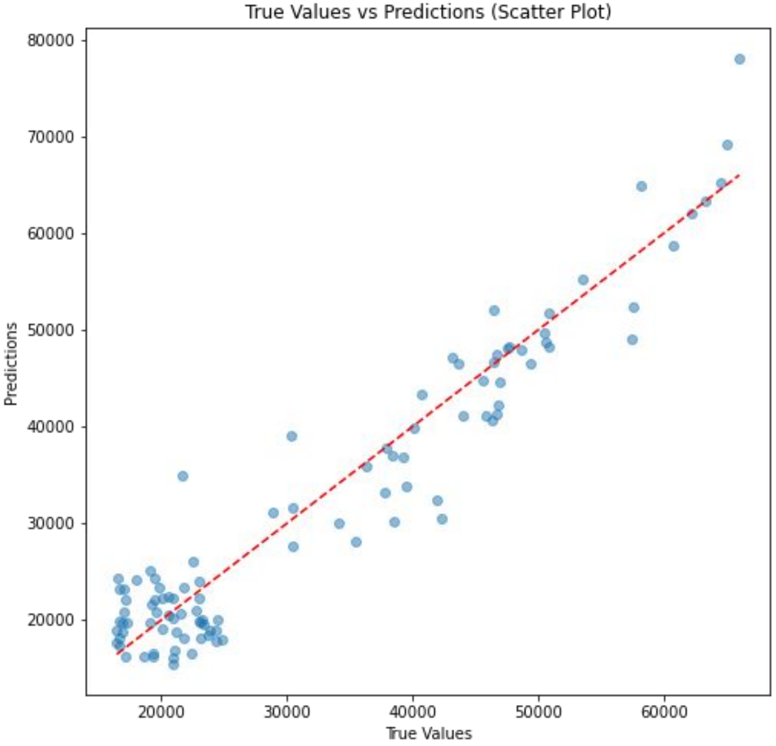
RMSE:4426.97
R^2 score: 0.91
MAE: 3487.60

Training and Validation Loss



The training and validation loss both decrease and converge to a similar value, it's an indication that the model is performing well.

Evaluations



3487.61

MAE

Bitcoin prices are in the range of [32000,52000] of units, an MAE of 3487.61 might be considered reasonably small.

0.91

R^2 score

The R^2 score ranges from 0 to 1, with higher values indicating a better fit.
The model can explain approximately 91% of the variance in the Bitcoin prices

Explanations

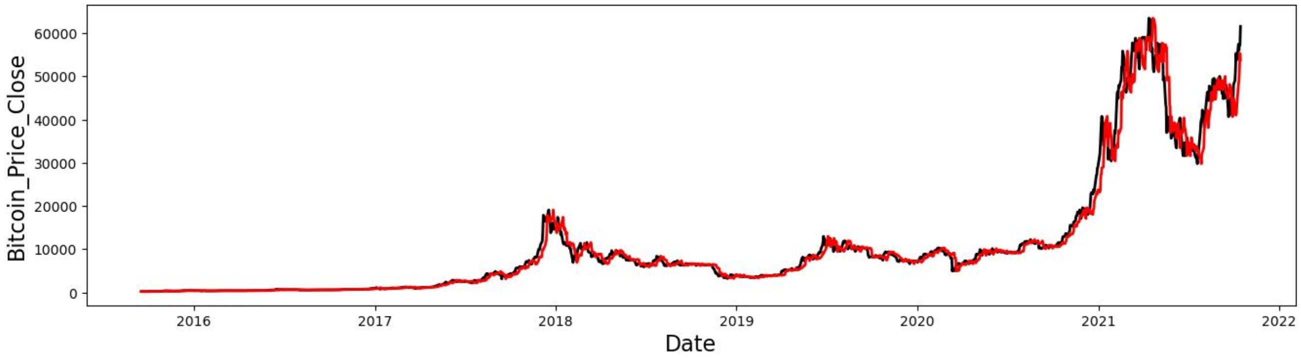
- This is a relatively naive feed forward neural network (FFNN) model, but still have a fair performance, which proves that the features we collected are helpful.
- The MAE was 3000+ which was larger than other models, but should be considered reasonably small because Bitcoin prices are in the range of [32000,52000] of units.
- The R^2 score ranges from 0 to 1, with higher values indicating a better fit. 0.91 means the model can explain approximately 91% of the variance in the Bitcoin prices

Time Series Approach + Results + Reflections

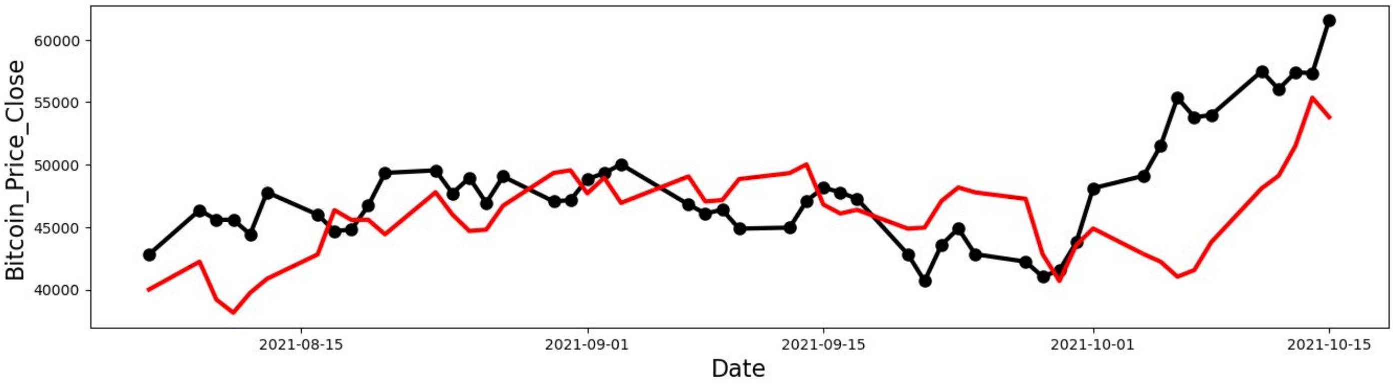
Time Series Models

Random Walk	Overfits, less sophisticated
Auto-Regressive Models	Good Result
Random Forest	Fair Result, good for long term
ARIMA	Bad due to non-stationary financial data

Random Walk

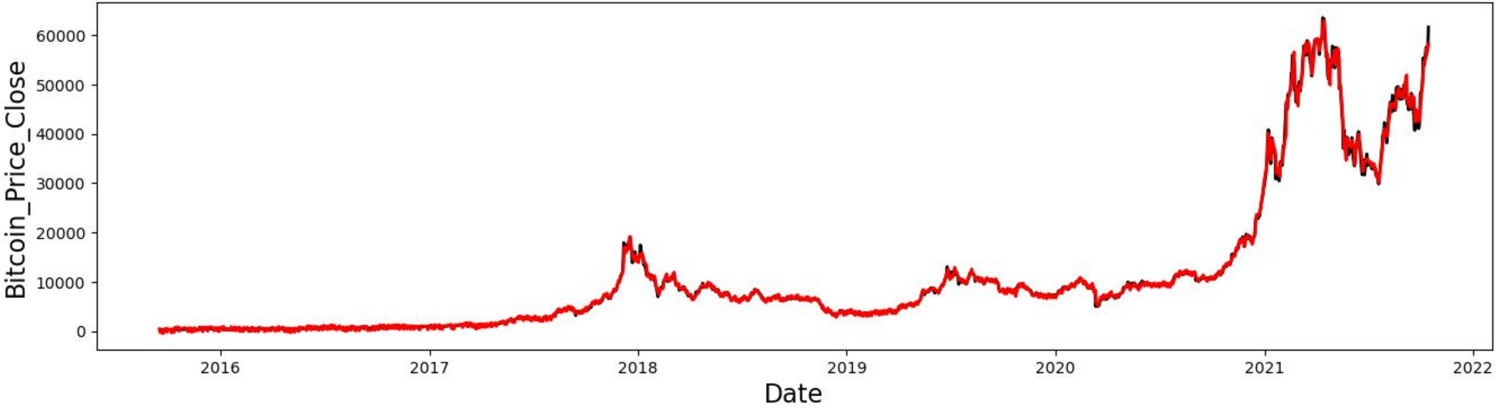


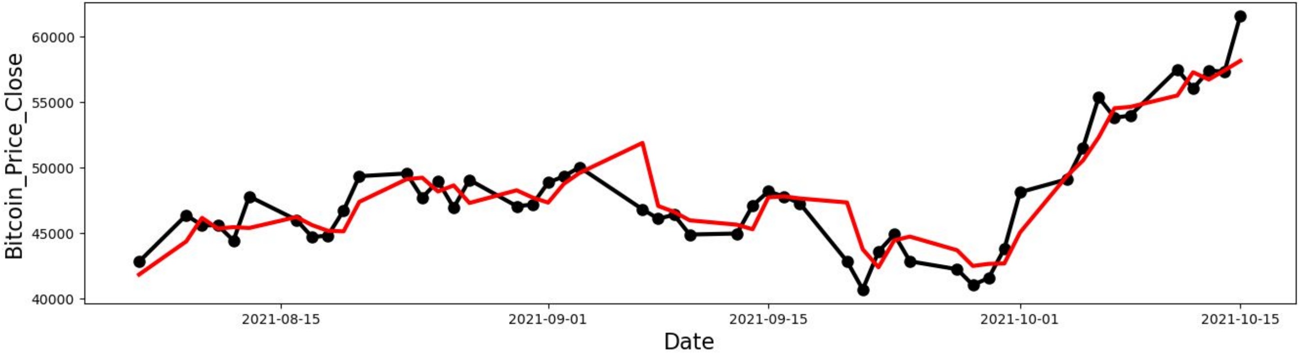
Bitcoin_Price_Close
60000
50000
40000
30000
20000
10000
0
2016
2017
2018
2019
2020
2021
2022
Date



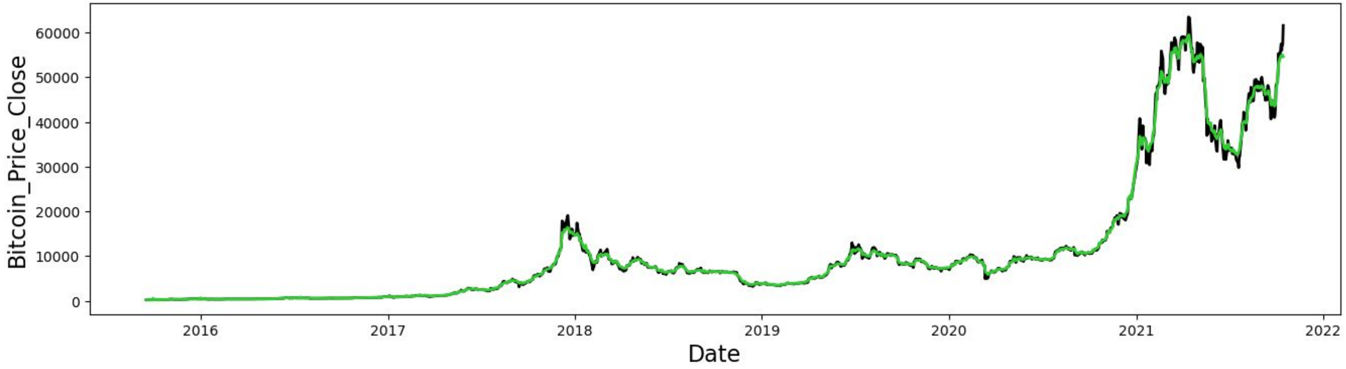
Bitcoin_Price_Close
60000
55000
50000
45000
40000
2021-08-15
2021-09-01
2021-09-15
2021-10-01
2021-10-15
Date

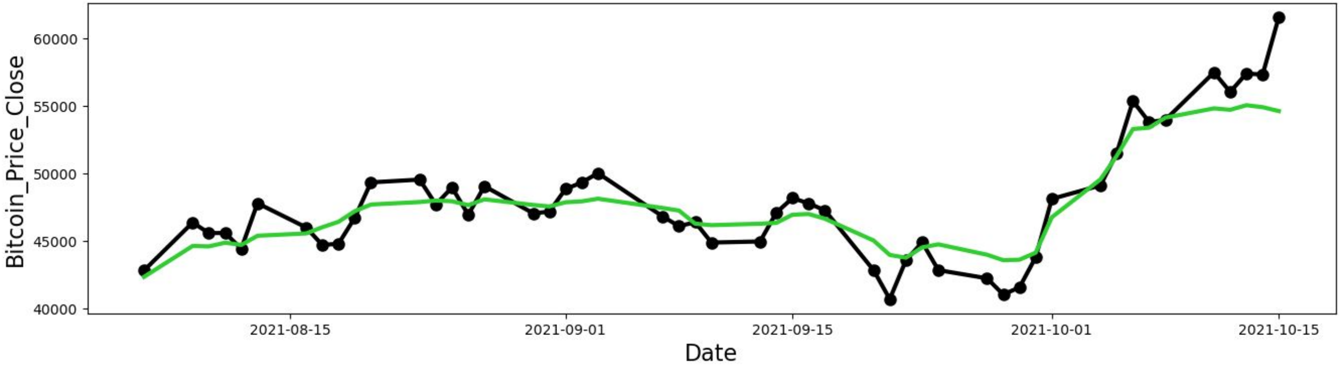
AR Models


Bitcoin_Price_Close
60000
50000
40000
30000
20000
10000
0
2016
2017
2018
2019
2020
2021
2022
Date

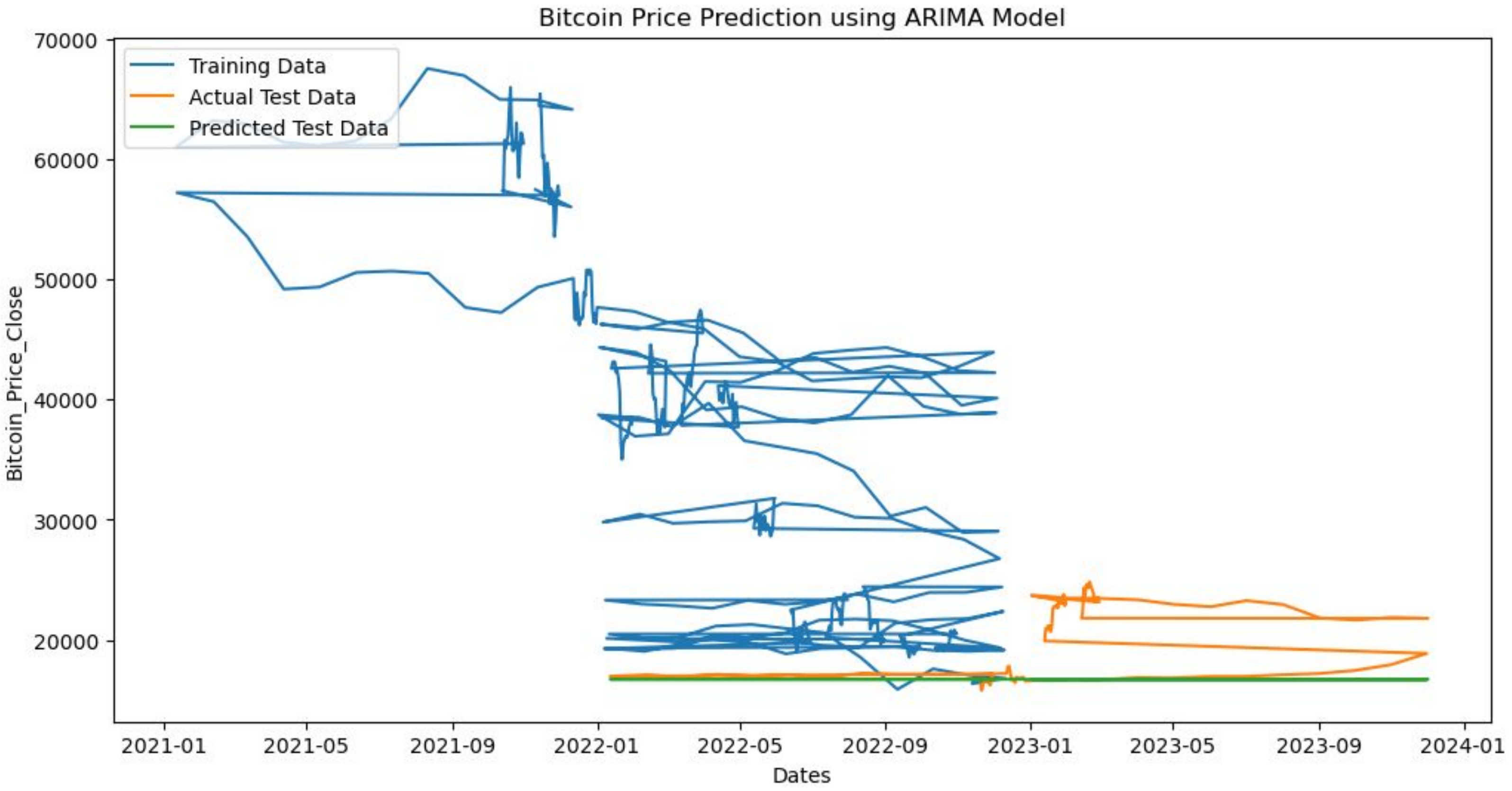

Bitcoin_Price_Close
60000
55000
50000
45000
40000
2021-08-15
2021-09-01
2021-09-15
2021-10-01
2021-10-15
Date

Random Forest





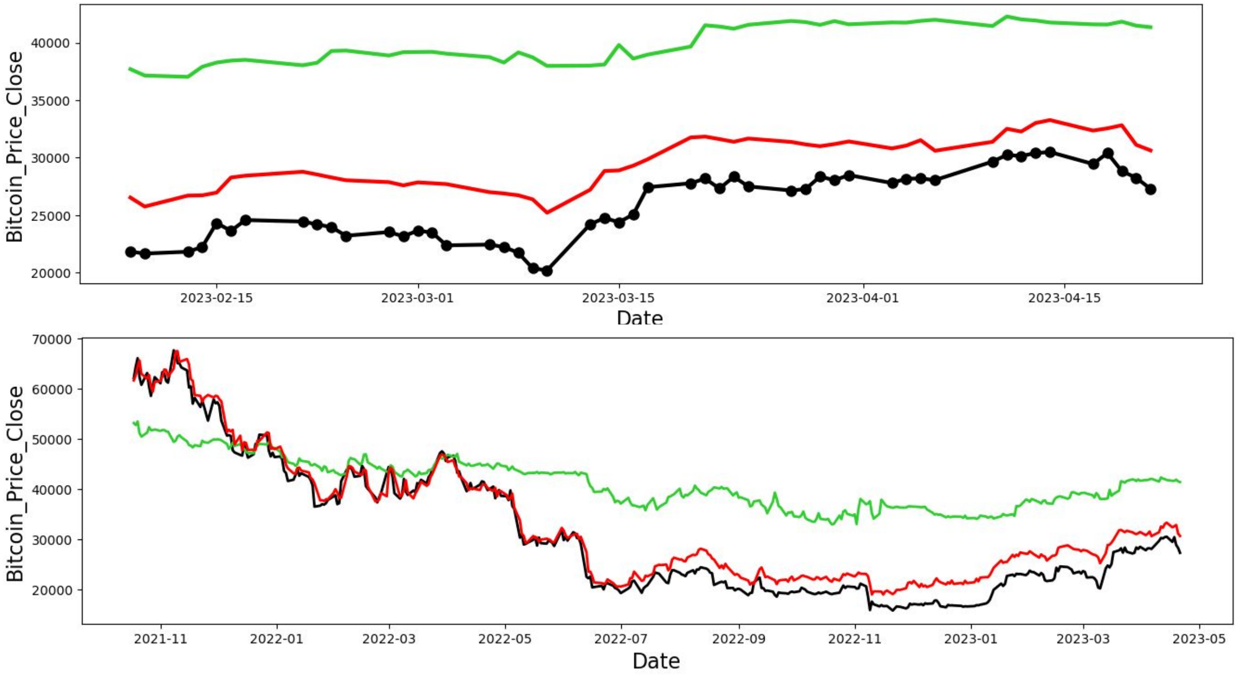
ARIMA (AutoRegressive Integrated Moving Average)



AR v.s. RF

Random Forest Model OSR2: 0.70356

Auto-regressive Model OSR2: 0.987

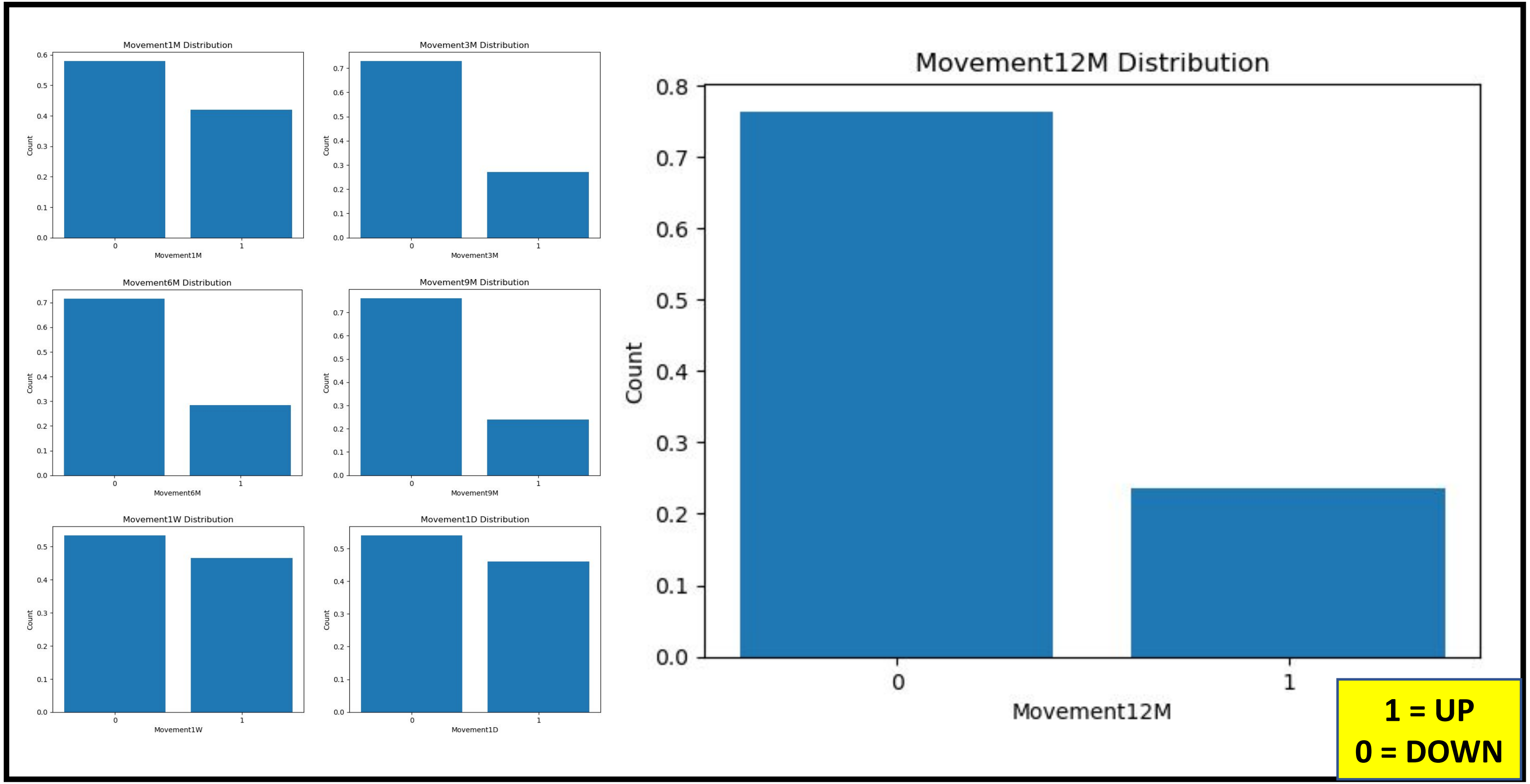


Explanations

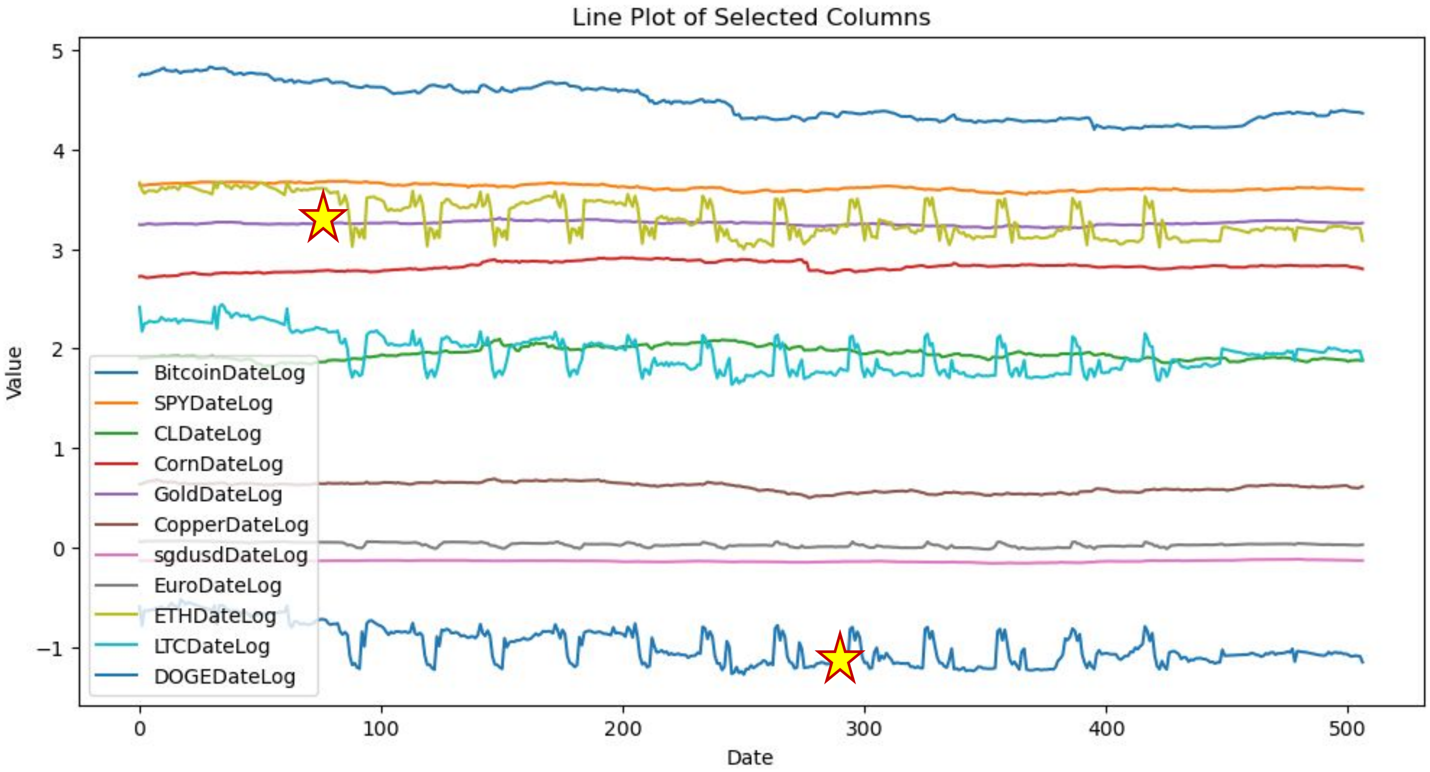
- Random walk is not sophisticated enough, also has the overfitting problem
- AR Model is sophisticated enough, however is slightly overfitted. But still good for short term use
- ARIMA could not handle this problem well. It only takes in stationary data, but in this case, Bitcoin-related financial data are not stationary, thus it returns bad results
- Random Forest appears to be the most reasonable model, however, it is not as accurate as the AR model
- In conclusion, AR and RF models are good and can be applied in both retail as well as institutional settings, but for institutional investors, the data is not in second-scale, thus might not be as useful to them as it is to the retail investors. The models are great for retail investor who want to know the trend of Bitcoin price in for the next week/month

KNN
Approach + Results + Reflections

Other details about the dataset



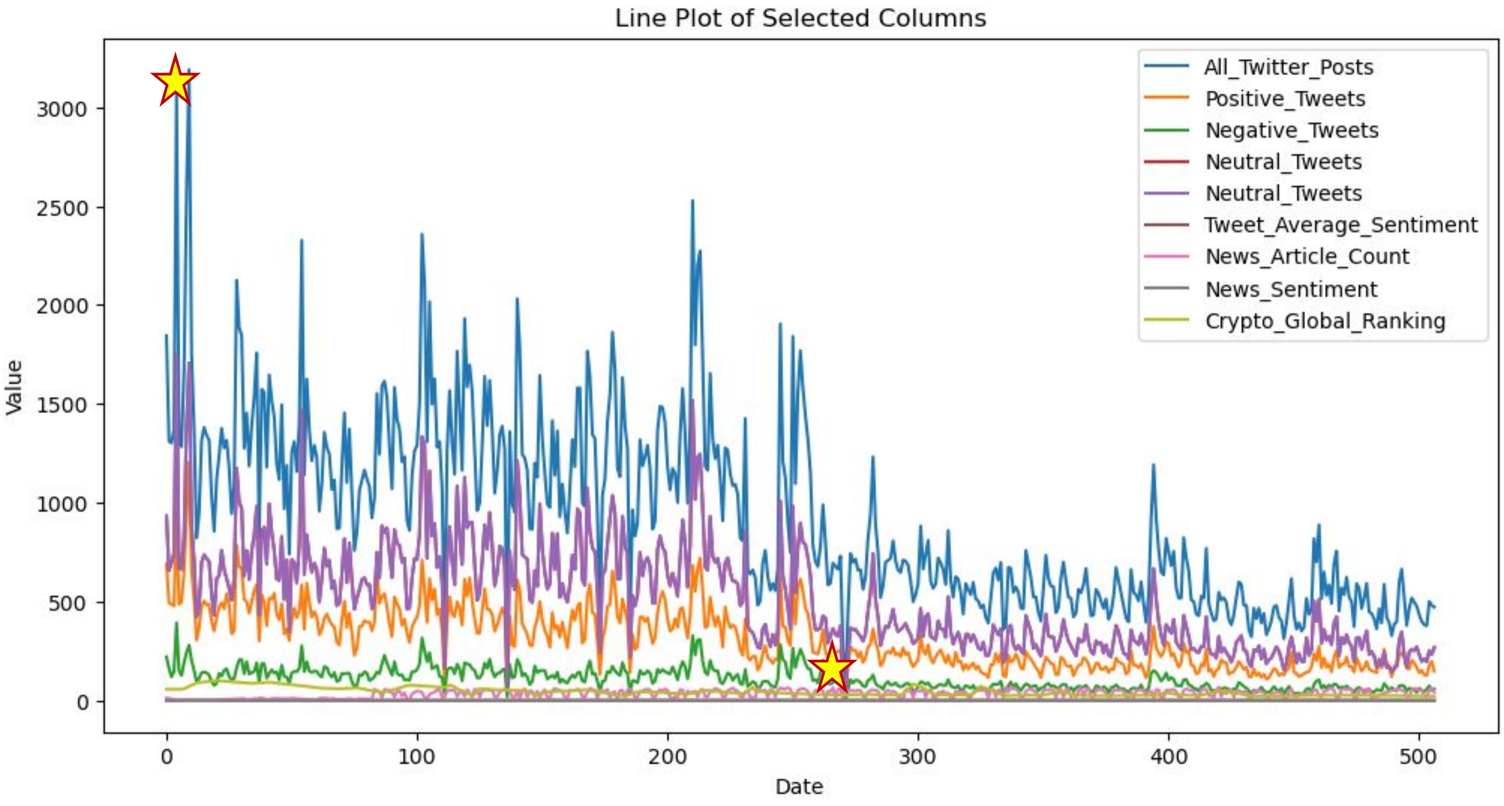
Baseline accuracy for 12M classification will be the lowest with 0.25 whereas the highest for 1W with 0.43



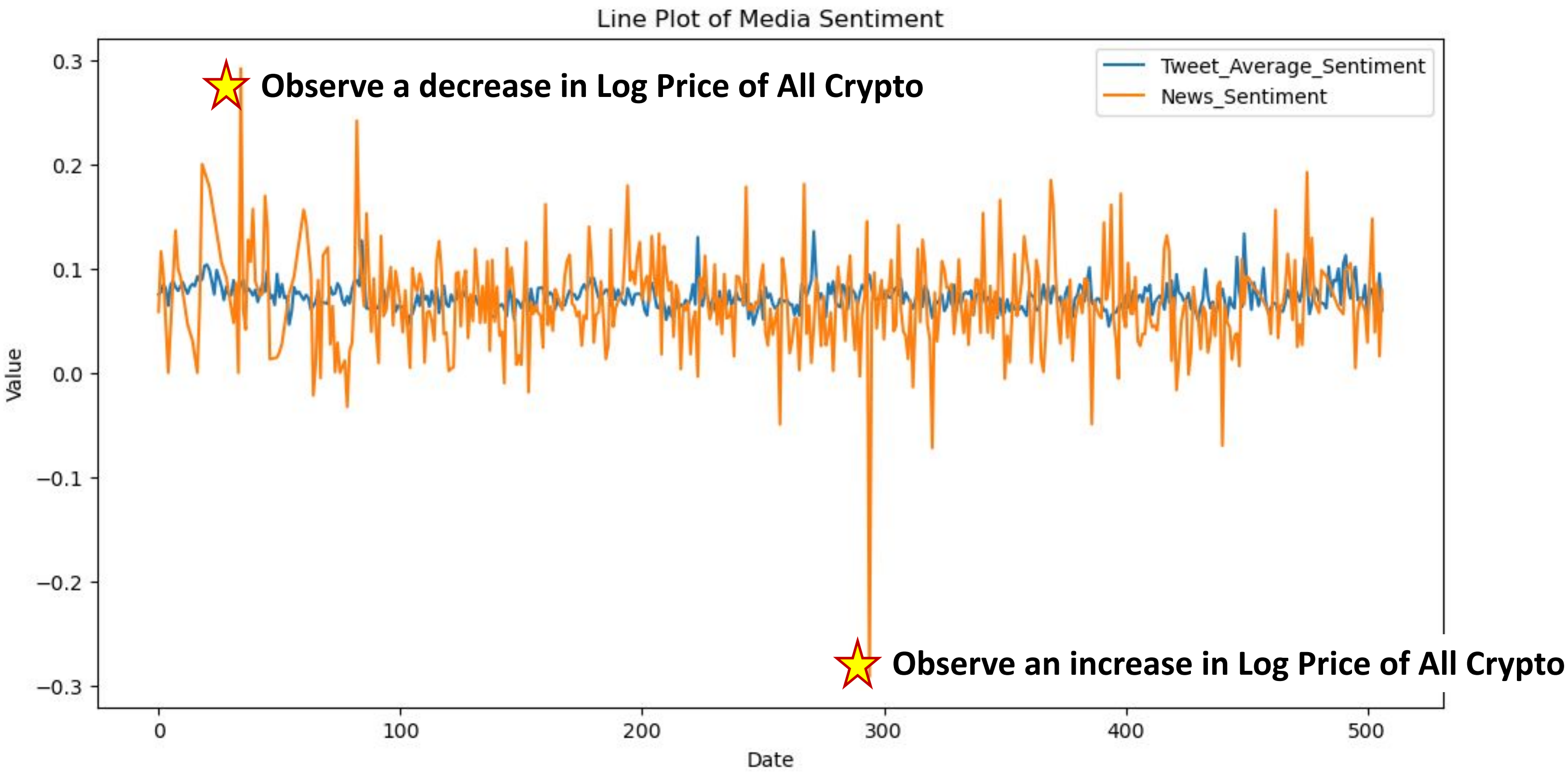
Max; Min

Price movement for all crypto coins are similar, but tempered for Bitcoin
Euro also has similar movement to crypto coins

Bitcoin Price - Distribution

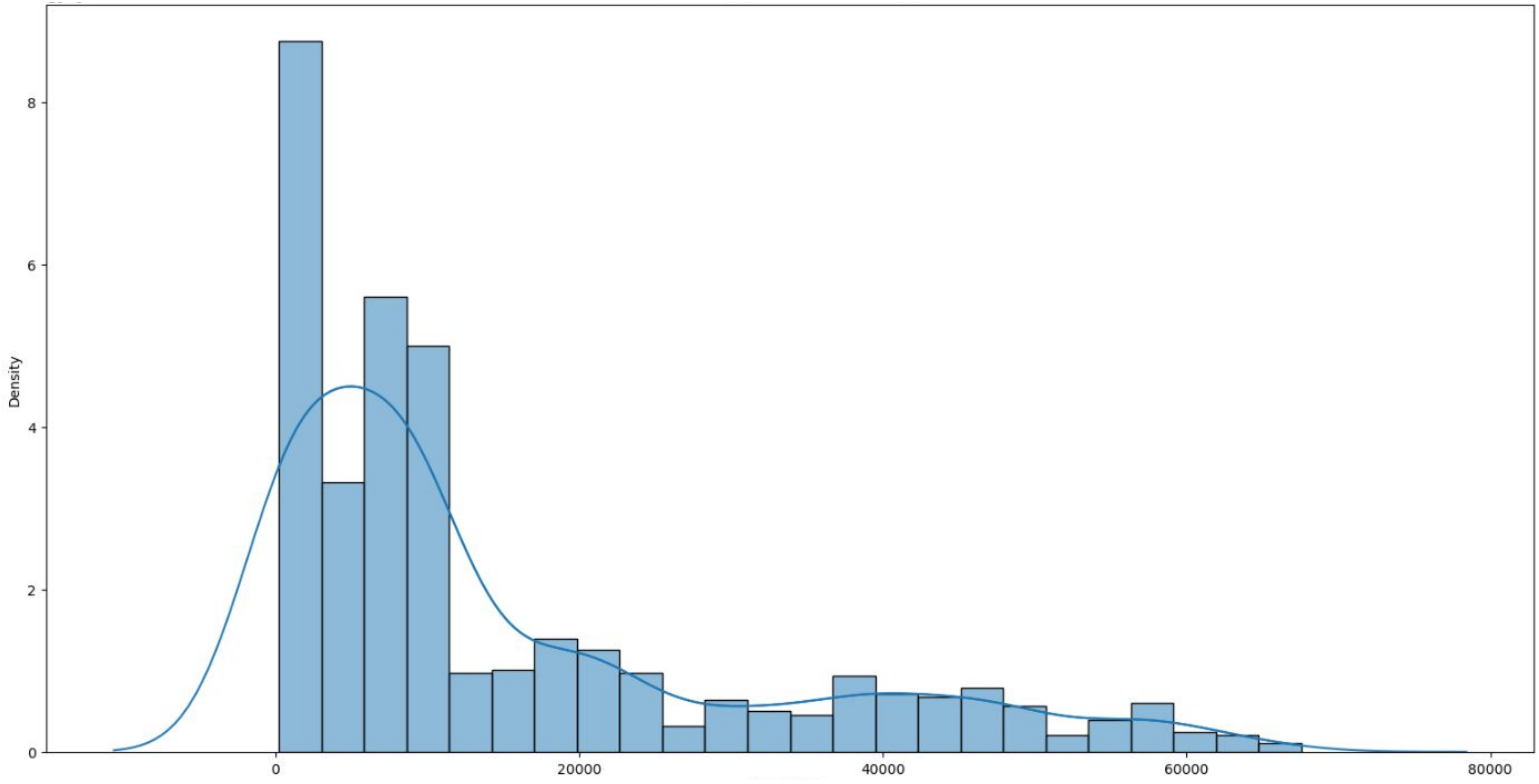


Max; Min

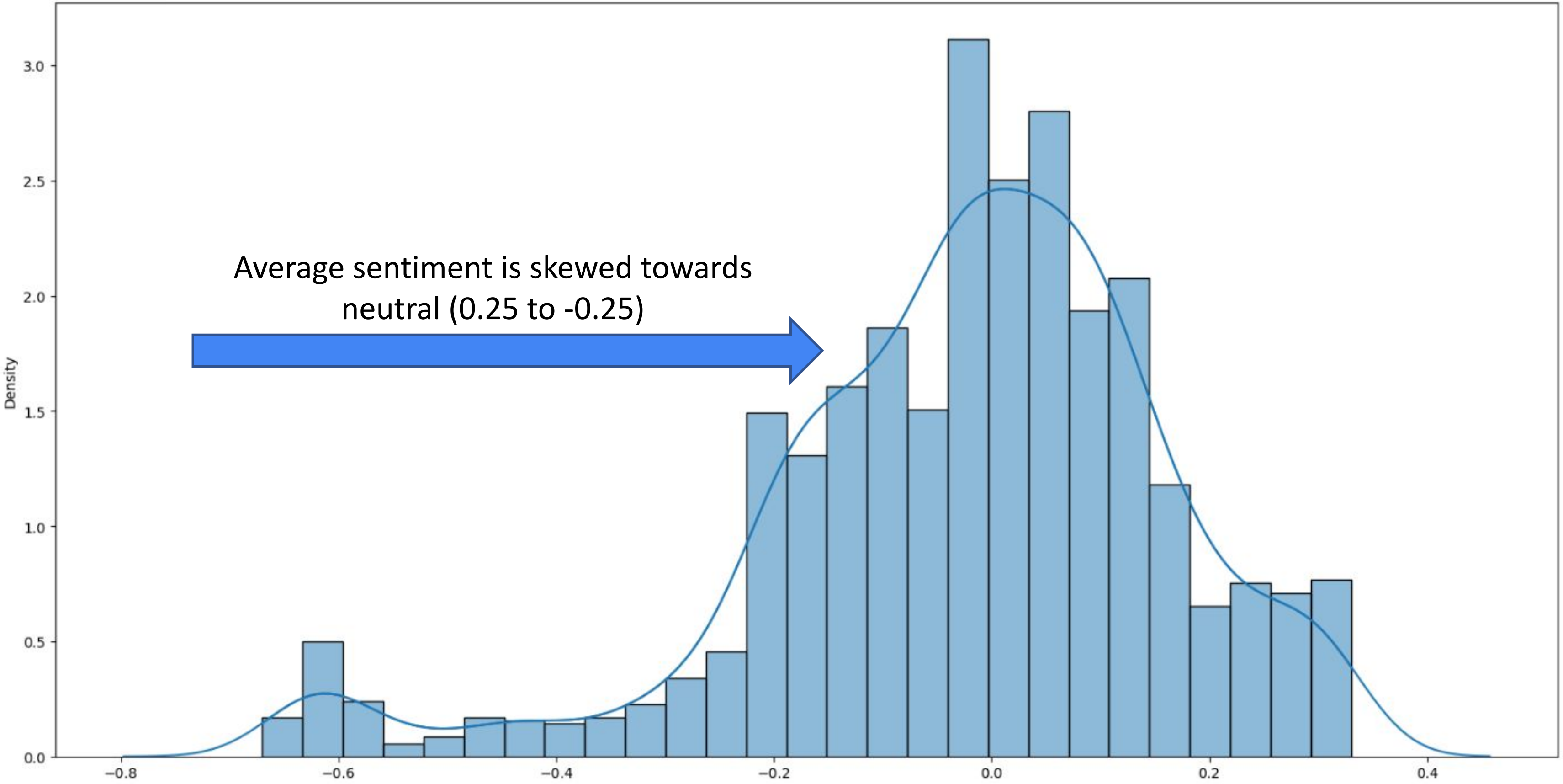

Line Plot of Media Sentiment
Observe a decrease in Log Price of All Crypto
Observe an increase in Log Price of All Crypto
Tweet_Average_Sentiment
News_Sentiment
Value
Date
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
100
200
300
400
500

Max; Min

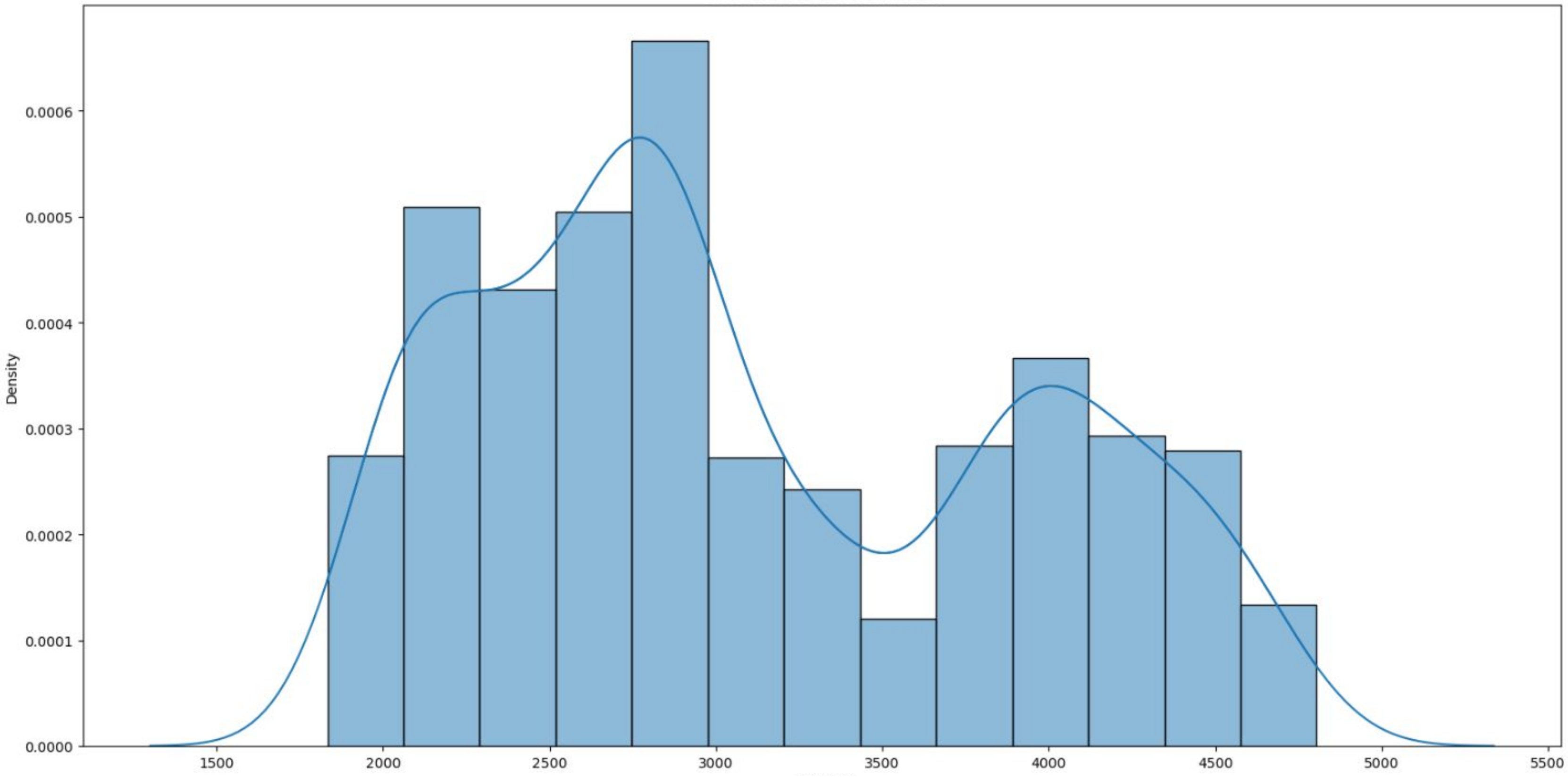
Bitcoin Price - Distribution



Average Sentiment - Distribution



SPY Price - Distribution



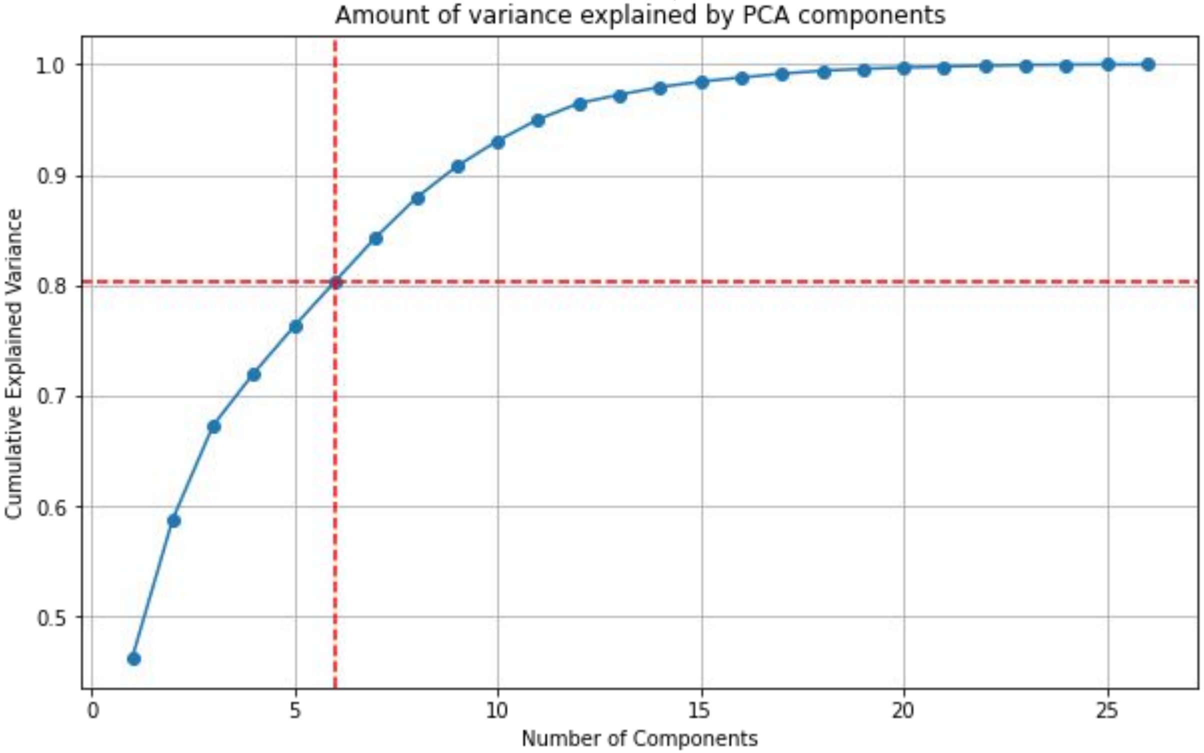
Key feature summary for Dataset 2

	Min	Max	Mean	Median	STD Dev	Variance	CAGR (%)
Bitcoin	227	67550	15066	8722	16383	268387119	90
SPY	1833	4805	3113	2900	803	644341	10
CL	-14	125	60	57	19	356	7
Corn	302	818	450	378	135	18252	8
Gold	1054	2054	1508	1418	278	77140	8
Copper	1.94	4.91	3.11	2.89	0.76	0.58	7
SGDUSD	0.69	0.77	0.73	0.73	0.02	0.00	1
Sentiment	-0.67	0.33	-0.03	-0.01	0.19	0.04	NA

Key feature summary for Dataset 23

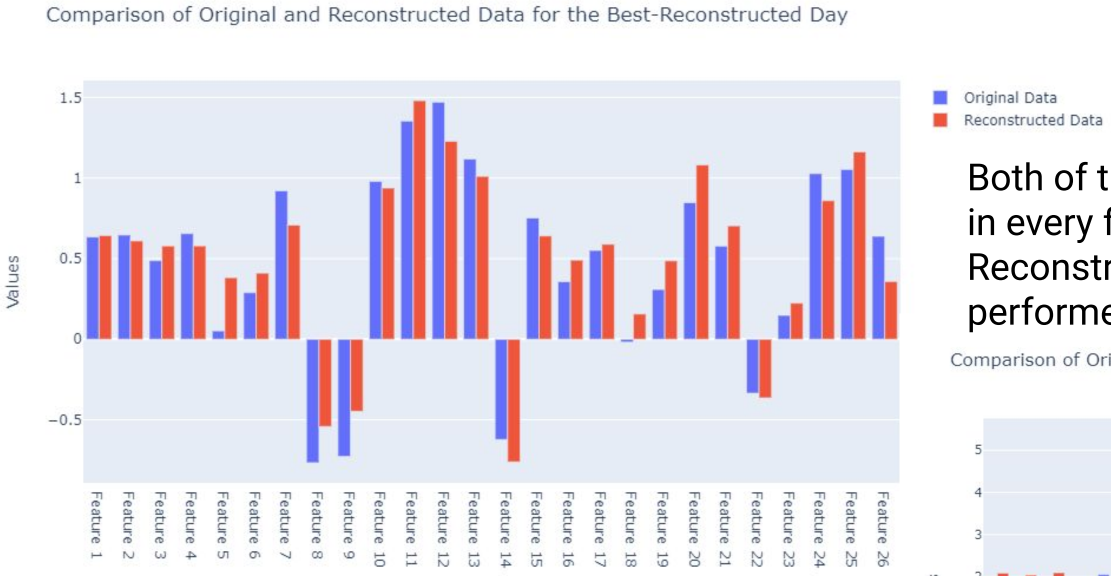
	Min	Max	Mean	Median	STD Dev	Variance	CAGR (%)
Bitcoin	15782	67550	32097	24191	14260	203350642	-35
SPY	3520	4805	4175	4108	324	104822	-5
CL	66	125	90	87	13	177	-2
Corn	512	818	671	671	72	5212	9
Gold	1620	2054	1810	1806	84	7051	2
Copper	3.15	4.91	4.08	4.20	0.43	0.19	-3
SGDUSD	0.70	0.77	0.73	0.73	0.01	0.00	0
EURUSD	0.96	1.17	1.07	1.07	0.05	0.00	-4
ETH	994	4730	2276	1732	1043	1088551	-49
LTC	43	278	99	89	48	2328	-46
DOGE	0.05	0.30	0.12	0.09	0.06	0.00	-47

Amount of variance explained by PCA components

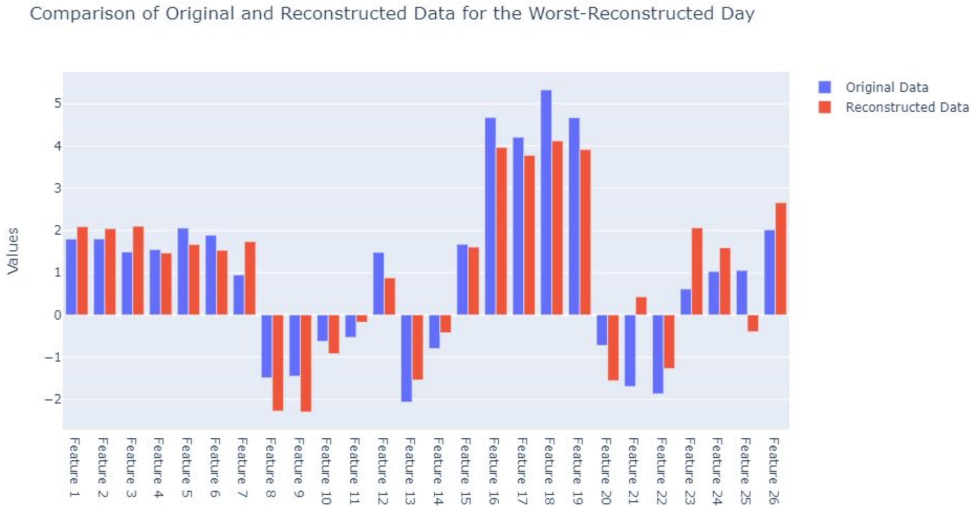


Apply PCA with 6 components, we can see that the first 6 PC can explain 80% of variance.

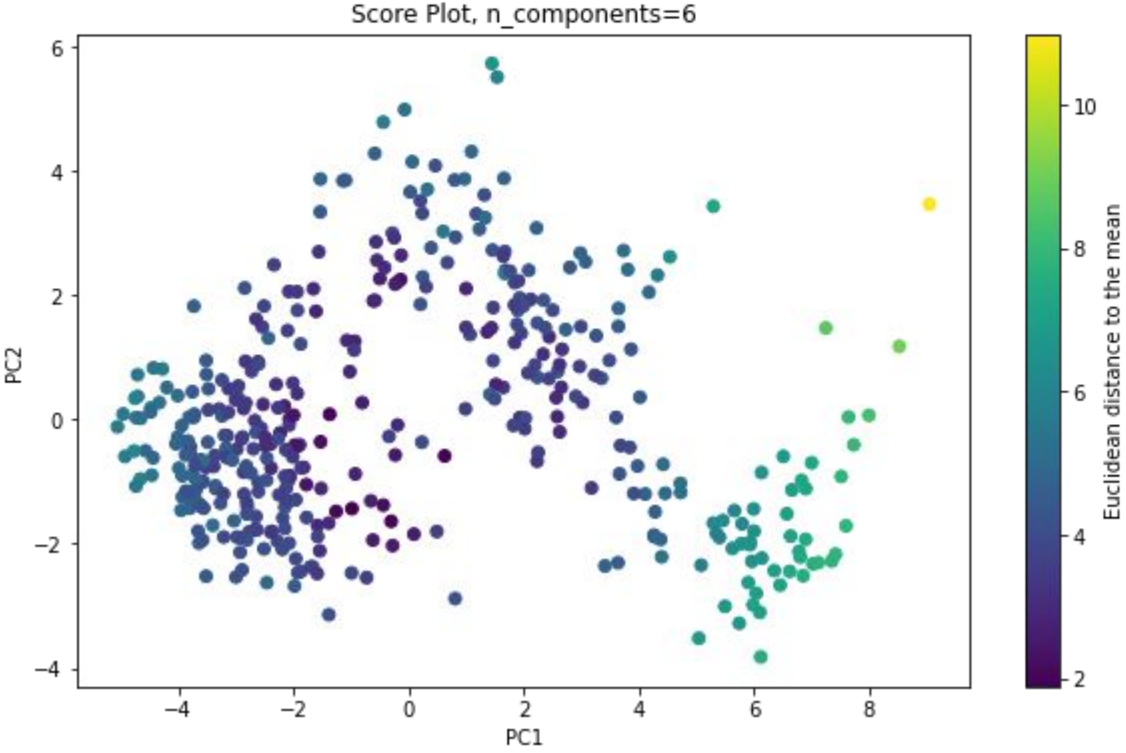
The best/worst reconstructed day using the first 6 PCs



Both of the 2 plots do not shows large difference in every features between original data and Reconstructed data, which means the PCA we performed could be reliable.

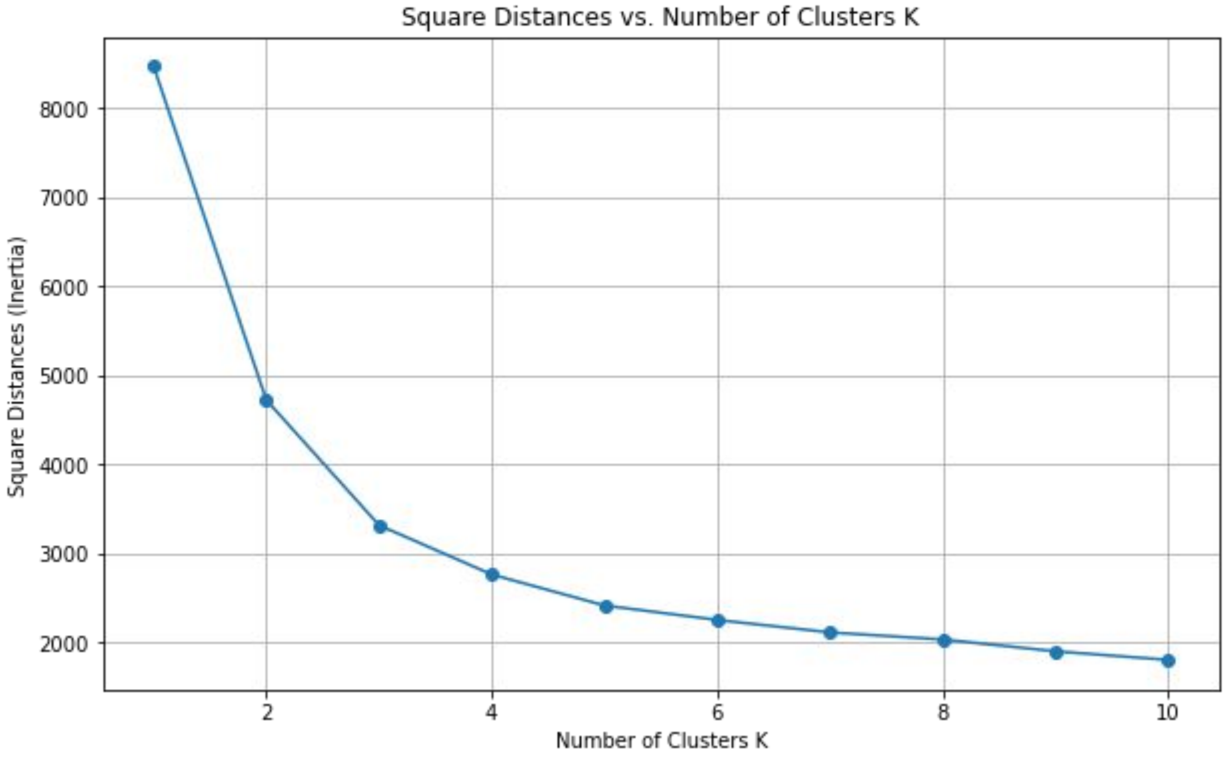


Score Plot



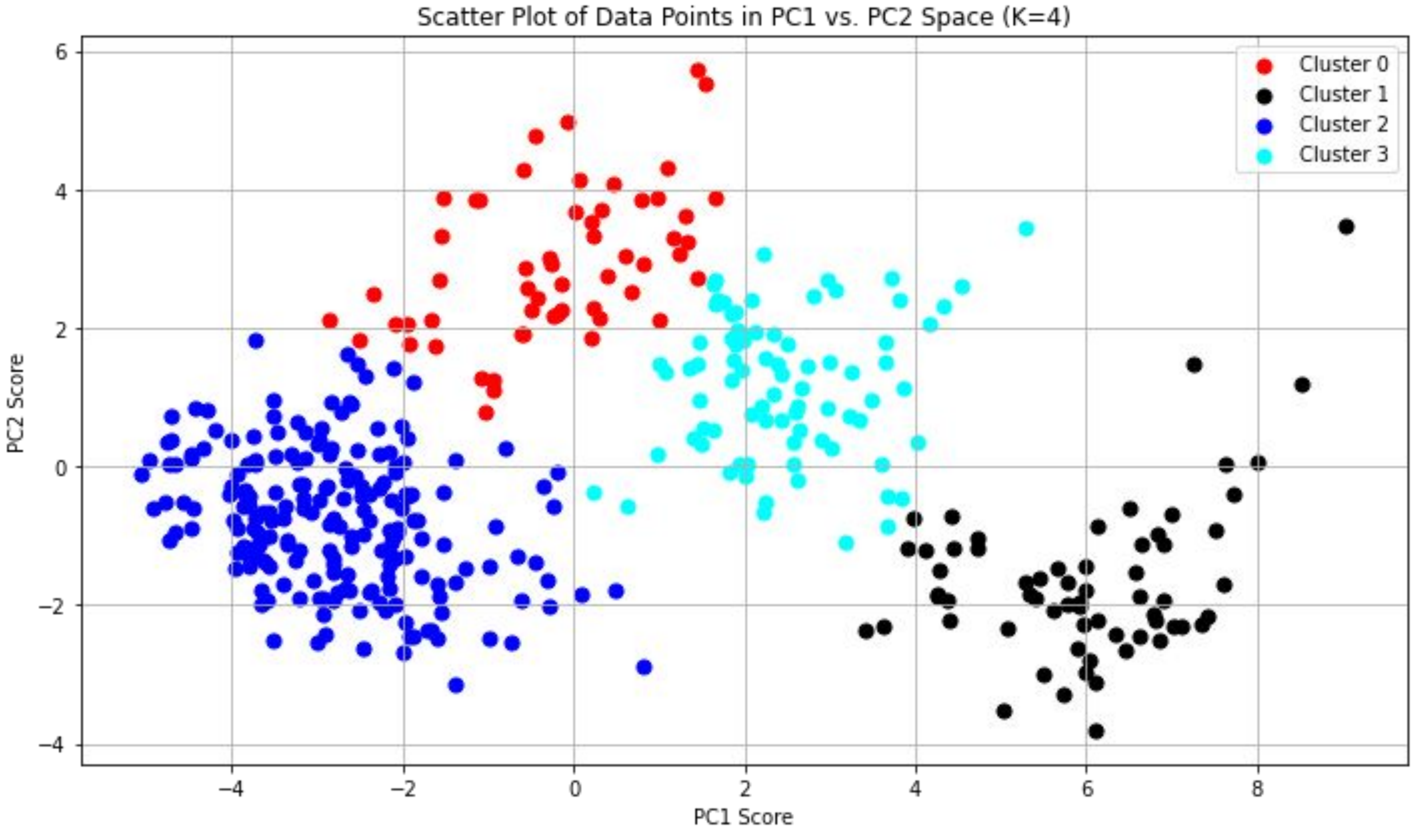
Score Plot, n_components=6
PC1
PC2
Euclidean distance to the mean
−4
−2
0
2
4
6
8
10

Elbow method



By using elbow method, we can see that choosing k=4 can reduce lots of distance in the same time avoid having too many clusters.

Fit K-Means model with K=4



Loadings for the First Two PCs

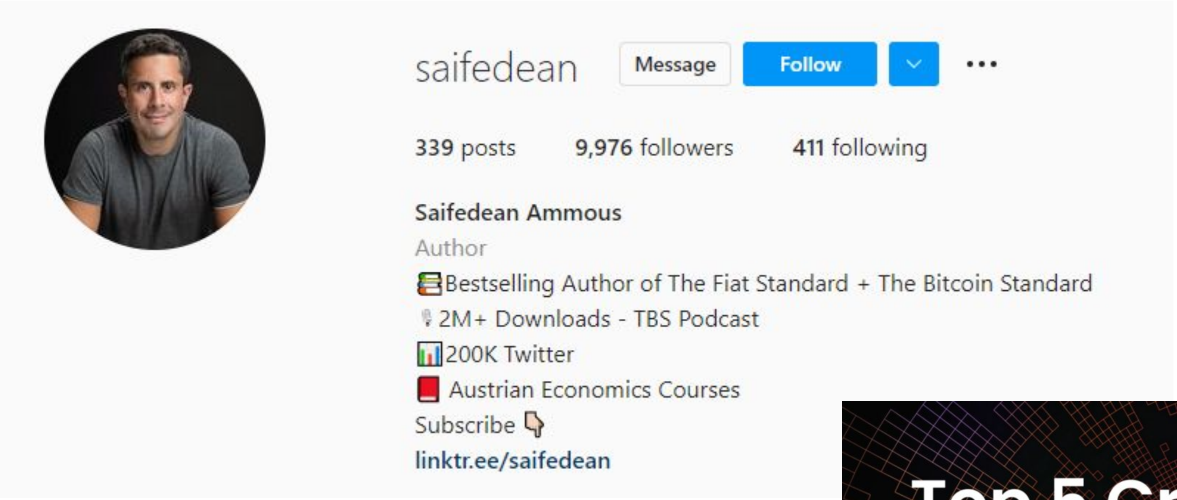
- Low PC1 value might mean that this day have low cryptocurrency price and Google Trend Popularity rating and related tweets' number, but high Average twitter sentiment score and yield of US treasury bill.
- Low PC2 value might mean that his day have low Crude oil price, Corn price,USD to SGD conversion rate, related tweets and news articles' number, but high other cryptocurrencies's price, USD to EURO conversation rate, Average twitter sentiment score and Google Trend Popularity rating.

Explanations

- By analysing the loadings of the first two PC, which explains nearly 60% of the variance of our data set, We can learn that most of our features are useful in future modeling, as most of our features have high positive or negative correlations with the principal components.

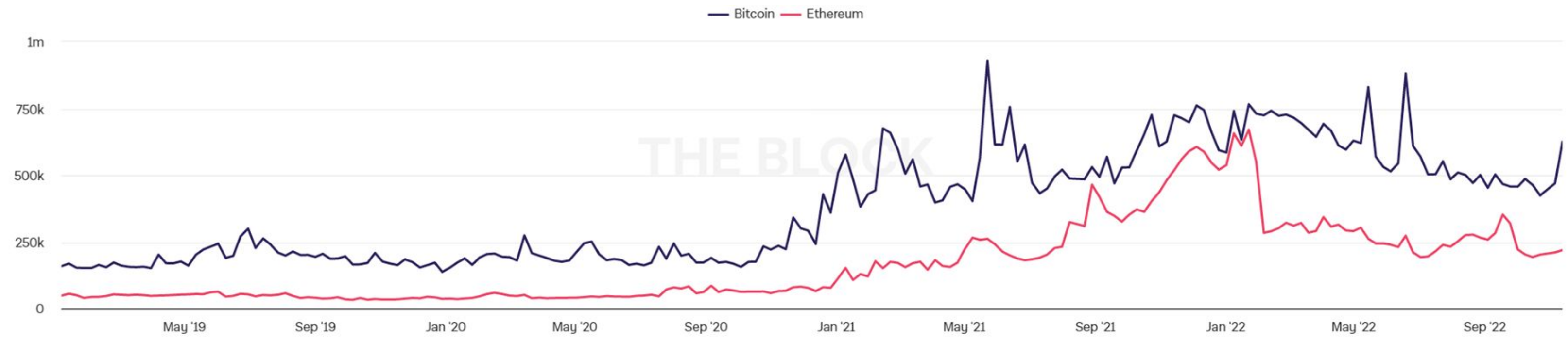
- From PC1, we can learn that users on Twitter may have a more positive sentiment towards cryptocurrencies on days with lower prices and popularity, and lower yields on US Treasury bills.

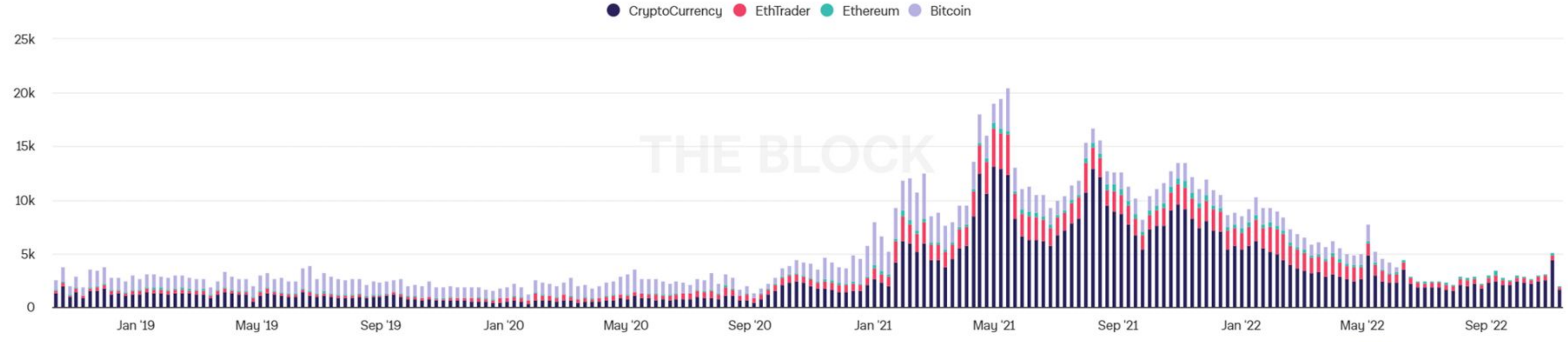
- From PC2, we can learn that on days with lower crude oil and corn prices and a lower USD to SGD conversion rate, users may be more interested in and invest more in other cryptocurrencies, and the overall sentiment towards cryptocurrencies may be more positive.



4. Influencer Marketing Hub: 39 Instagram Crypto Influencers To Follow - Accessed February 2022
5. Investing.com - Is Ether or Elon Pulling Crypto Up - Accessed February 2022
6. Medium - Top 5 Crypto To Buy in November 2021 - Accessed February 2022

Activity on Twitter is in line with the price movement for both, Bitcoin and Ethereum





SOURCE: SUBREDDITSTATS
UPDATED: NOV 16, 2022

7. The Block - Data - Web Traffic Twitter - Accessed November 2022
8. The Block - Data - Web Traffic - Reddit - Accessed November 2022

Data processing

Date	BitcoinVolumel	BitcoinDat	BitcoinClo	Bitcoin1D	Bitcoin1W	ETHDate	DOGEDat	LTCDate	SPYDate	UST1YDat	UST10YDa	CLDate	GoldDate	CopperDa	CornDate	sgdusdDa	EuroDate	All_Twitte	Positive_T	Negative_	Neutral_T	Tweet_Av	News_Art	News_Ser	Crypto_Gl
11-10-2021	48730828378	54734.1	57484.8	54952.8	48208.9	4636.17	0.25571	261.263	4350.65	0.001	0.0159	79.92	1759.3	4.328	5.33	1.34736	1.15955	1844	691	220	933	0.07514	12	0.05852	56
12-10-2021	30966005122	57526.8	56041.1	54734.1	49175	3908.5	0.16442	148.598	4350.65	0.0011	0.0156	79.92	1759.3	4.283	5.225	1.3648	1.12933	1311	492	162	657	0.0774	9	0.11616	56
13-10-2021	41684252783	56038.3	57401.1	57526.8	51486.7	3606.2	0.23257	177.628	4363.8	0.001	0.0152	79.82	1794.7	4.5205	5.1225	1.35111	1.15996	1302	486	119	697	0.085	7	0.09293	56
14-10-2021	36615791366	57372.8	57321.5	56038.3	55338.6	3786.01	0.23227	180.118	4438.26	0.0012	0.0159	80.77	1797.9	4.6315	5.1675	1.34868	1.15915	1369	478	146	745	0.0803	4	0.04646	56
15-10-2021	51780081801	57345.9	61593.9	57372.8	53802.1	3862.63	0.23378	188.814	4471.37	0.0012	0.0159	81.73	1768.3	4.734	5.2575	1.34844	1.15929	3092	954	392	1748	0.06424	1	0	56
16-10-2021	34250964237	61609.5	60892.2	57345.9	53929.8	3830.38	0.23729	186.189	4476.4	0.0011	0.0159	81.73	1767.43	4.73283	5.2575	1.34795	1.15956	1305	486	149	671	0.08456	1	0.04545	56
17-10-2021	29032367511	60887.7	61553.6	61609.5	54952.8	3847.1	0.2379	183.731	4481.43	0.0011	0.0159	81.73	1766.57	4.73167	5.2575	1.3477	1.1597	1283	496	126	663	0.08772	1	0.09091	56
18-10-2021	38055562075	61548.8	62026.1	60887.7	54734.1	3748.76	0.24728	185.557	4486.46	0.0011	0.0165	81.69	1765.7	4.7305	5.3275	1.34896	1.16139	1669	649	190	830	0.0821	1	0.13636	58.9167
19-10-2021	40471196346	62043.2	64262	61548.8	57526.8	3877.65	0.24579	188.94	4519.63	0.001	0.0165	82.44	1770.5	4.714	5.3025	1.3449	1.1632	2567	936	244	1388	0.07867	1	0.1	64.75
20-10-2021	40788955582	64284.6	65992.8	62043.2	56038.3	4155.99	0.25409	206.871	4536.19	0.0012	0.0168	83.42	1784.9	4.756	5.3925	1.3427	1.1655	3190	1205	279	1706	0.08254	1	0.09231	70.5833
21-10-2021	45908121370	66002.2	62210.2	64284.6	57372.8	4054.32	0.24239	196.869	4549.78	0.0013	0.0166	82.5	1781.9	4.5865	5.3225	1.34681	1.16253	1752	659	202	892	0.08668	1	0.07692	76.4167
22-10-2021	38434082775	62237.9	60692.3	66002.2	57345.9	3970.18	0.24422	190.822	4544.9	0.0013	0.0165	83.76	1796.3	4.526	5.38	1.34725	1.1629	1378	514	152	712	0.08073	1	0.06154	82.25
23-10-2021	26882546034	60694.6	61393.6																		423	0.07583	1	0.04615	88.0833
24-10-2021	27316183882	61368.3	60930.8																		471	0.08227	1	0.03846	91
25-10-2021	31064911614	60893.9	63039.8																		697	0.08514	1	0.03077	91.75
26-10-2021	34878965587	63032.8	60363.8																		737	0.08311	1	0.01538	93.25
27-10-2021	43657076893	60352	58482.4																		717	0.09231	1	0	94.75
28-10-2021	45257083247	58470.7	60622.1																		703	0.08793	1	0.06667	96.25
29-10-2021	36856881767	60624.9	62228																		540	0.08931	1	0.2	97.75
30-10-2021	32157938616	62239.4	61888.8																		430	0.10237	1	0.19276	99.25
31-10-2021	32241199927	61850.5	61319																		616	0.10351	2	0.18551	100
01-11-2021	32852737090	61320.4	61004.4																		631	0.09859	2	0.17827	99.4167
02-11-2021	34075811416	60963.3	63226.4																		759	0.08758	3	0.16378	98.25
03-11-2021	35298885742	63254.3	62970																		662	0.07514	4	0.14929	97.0833
04-11-2021	36521960068	62941.8	61452.2																		617	0.0984	5	0.13481	95.9167
05-11-2021	37745034394	61460.1	61125.7																		569	0.09029	6	0.12032	94.75
06-11-2021	38968108720	61068.9	61527.5																		507	0.08249	7	0.10583	93.5833
07-11-2021	39579645883	61554.9	63327																		619	0.06968	7	0.09859	93
08-11-2021	40191183046	63344.1	67566.8																		1174	0.08245	7	0.09134	92.4167
09-11-2021	41414257372	67549.7	66971.8																		1004	0.07476	8	0.07685	91.25
10-11-2021	42637331698	66953.3	64995.2																		971	0.07408	9	0.06237	90.0833
11-11-2021	35880633236	64978.9	64950																		630	0.08854	10	0.04788	88.9167
12-11-2021	25775869261	64864	64155.9	64978.9	61460.1	4084.45	0.16902	158.042	4682.85	0.0017	0.016	79.69	1868.5	4.405	5.85	1.35244	1.14771	1453	500	196	758	0.07231	1	0.08571	87.75
13-11-2021	30474228777	64158.1	64469.5	64864	61068.9	4651.46	0.26171	258.093	4682.83	0.0018	0.0161	79.69	1867.87	4.406	5.85	1.35223	1.14594	1185	441	133	611	0.07672	1	0	86.5833
14-11-2021	25122092191	64455.4	65466.8	64158.1	61554.9	4626.36	0.26291	278.008	4682.82	0.0018	0.0163	79.69	1867.23	4.4065	5.85	1.35212	1.14505	1390	495	140	755	0.08448	1	0.29167	86
15-11-2021	30558763548	65521.3	63557.9	64455.4	63344.1	4557.5	0.25653	262.763	4682.8	0.0017	0.0163	79.75	1866.6	4.407	5.84	1.35202	1.14416	1537	557	146	835	0.08885	10	0.0611	86.5
16-11-2021	46844335592	63721.2	60161.2	65521.3	67549.7	4216.37	0.23725	230.599	4700.9	0.0018	0.016	79.74	1854.1	4.3585	5.775	1.35389	1.13665	1757	585	190	982	0.07549	7	0.04141	87.5
17-11-2021	39178392930	60139.6	60368	63721.2	66953.3	4287.59	0.23749	229.504	4688.67	0.0018	0.0159	77.55	1870.2	4.273	5.815	1.35639	1.13212	827	301	92	434	0.08086	11	0.12727	88.5
18-11-2021	41388338699	60360.1	56942.1	60139.6	64978.9	4000.65	0.22138	204.423	4704.54	0.0018	0.0154	78.41	1861.4	4.31	5.7925	1.35553	1.13212	1573	532	186	856	0.07838	11	0.10644	89.5
19-11-2021	38702407772	56896.1	58119.6	60360.1	64864	4298.31	0.23311	218.139	4697.96	0.0019	0.0157	75.94	1851.6	4.4085	5.77	1.35647	1.13684	1558	501	185	876	0.07403	10	0.15704	90.5
20-11-2021	30624264863	58115.1	59697.2	56896.1	64158.1	4409.93	0.23302	227.08	4692.95	0.0019	0.016	75.94	1836.5	4.40417	5.77	1.35838	1.13375	1179	378	113	689	0.0795	1	0.0875	91.5
21-11-2021	26123447605	59730.5	58730.5	58115.1	64455.4	4269.73	0.22633	221.772	4687.95	0.002	0.0163	75.94	1821.4	4.39983	5.77	1.36028	1.13066	1646	517	136	994	0.06786	5	0.08047	92
22-11-2021	35036121783	58706.8	56289.3	59730.5	65521.3	4088.46	0.21968	209.158	4682.94	0.0021	0.0167	76.75	1806.3	4.3955	5.8425	1.36219	1.12757	1516	511	155	850	0.07614	9	0.07345	91.0909
23-11-2021	37485803899	56304.6	57569.1	58706.8	63721.2	4340.76	0.22651	216.389	4690.7	0.0024	0.0164	78.5	1783.8	4.4235	5.8825	1.3643	1.12421	1434	542	138	757	0.08589	9	0.07593	89.2727
24-11-2021	36635566789	57565.9	56280.4	56304.6	60139.6	4239.98	0.22183	209.807	4701.46	0.0022	0.0156	78.39	1784.3	4.458	5.855	1.36575	1.12449	1195	438	144	614	0.07766	10	0.16958	87.4545

Dataset 1 was created by merging all the raw CSV files obtained from various sources and compiled into one singular data-frame.

For NaN values, we used “interpolate”

This Dataset contained 507 rows and 25 columns

A total of 12,765 data points

Using Dataset 1, we built a preliminary model to evaluate its performance.

Model performance with dataset 1 will be referred to as ModelOG

Model Type	**ModelOG-R2**	**ModelOG-OSR2**	**ModelOG-Accuracy**
Linear Model	0.993	0.994	0.994
Decision Tree Regressor	NA	-0.128	0.595
Decision Tree Classifier	NA	-0.417	0.490

- Model had extremely poor performance on the CART models.
- Based on the R2 and OSR2 values, we also suspected overfitting of the models

- Realized that Dataset 1 is not good enough for the model to make better predictions that the baseline

Using the variables that were significant in the Linear Regression Model, we decided to build a new model with fewer variables to reduce noise and overfitting.

Model performance with dataset 1 with variables removed will be referred to as ModelOG2

Model Type	**ModelOG2-R2**	**ModelOG2-OSR2**	**ModelOG2-Accuracy**
Linear Model	0.960	0.938	0.938
Decision Tree Regressor	NA	-0.417	0.595
Decision Tree Classifier	NA	-0.830	0.475

Despite the change, the CART model continued to demonstrate poor performance.

- Identified the two main reasons the OSR2 continued to be negative:
    1. Limited data to train the model
    2. Noise in the data as a result of excessive models

- Realized that this can be detrimental to other models such as Random Forest and Gradient Boosting Regressor

Using the significant variables from dataset 1, we decided to focus on building a two new dataset.

Dataset 2 will have the same time frame as Dataset 1

- Dataset 2 will include additional information about each of the financial market indicators
- This includes the price of each asset for a time period ranging from 1D ago to 1Y ago (total of 7 points for each)
- Adds on top of Dataset 1

Dataset 3 will an extended time frame

- Dataset 3 was generated using the Yahoo Finance API
- Yahoo Finance had data for all financial market indicators from 2000
- Yahoo Finance had data for Bitcoin price and volume from 2014
- Dataset has a time period starting September 17 2015
- Dropped all Twitter sentiment data since information prior to 2020 was not available
- Dropped all Bitcoin News Sentiment Data since information prior to 2021 was not available
- Included the News Sentiment Data from the Federal Reserve of San Francisco
- Fairly different from Dataset 1

Feature Engineering And Advanced Data Collection

Dataset 2
- Start: September 17 2015
- End: April 21 2023
- Variables: 89
- Rows: 1898
- Total number of data-points: 169,000

Dataset 3
- Start: October 11 2021
- End: March 4 2023
- Variables: 25
- Rows: 507
- Total number of data-points: 12,700

Dataset 23
- Start: October 11 2021
- End: March 4 2023
- Variables: 103
- Rows: 507
- Total number of data-points: 52,200

For all significant financial indicators

1. Open Price
2. Close Price
3. Price 1 Day Ago
4. Price 1 Week Ago
5. Price 1 Month Ago
6. Price 3 Month(s) Ago
7. Price 6 Month(s) Ago
8. Price 9 Month(s) Ago
9. Price 12 Month(s) Ago

Also considered the Log value of all prices to normalize the prices

Additional Features for Bitcoin Price

1. Trade Volume ($)
2. Price Movement 1D
3. Price Movement 1W
4. Price Movement 1M
5. Price Movement 3M
6. Price Movement 6M
7. Price Movement 9M
8. Price Movement 12M

Each of the price movement were converted to categorical variables.

1 signifies upward movement
0 signifies downward movement

Prior Work

Major findings

- Some of the papers we reviewed were written by academics **(Phillipas 2019)** to derive and build theories linked with Bitcoin and media. Some of the others we reviewed were written by students seeking industry applications **(Tandon, Revankar 2021)**
    - A lot of the papers are published by students as part of their graduate or PhD these

- In addition, a vast number of blogs covering similar topics and approaches
    - The blogs posted are usually by hobbyists in the space **(Yang 2019; Riveroll 2019)**

- 90% of references in our literature review were published from 2018 - 2022, indicating a clear interest in Bitcoin prediction and inquiry after the value skyrocketed

- Each of these journals or articles provided either a unique approach, or adapted their models to address diverse problems.

  (Mittal; Dhiman 2019) in their research focused on short-term price fluctuations whereas **(Kaman 2020)** focused on simple sentiment analysis.

Bibliography

Sources

A. Barr, K. (2022, November 15). *Vast majority of people who invest in bitcoin inevitably lose money, study shows*. Gizmodo. Retrieved February 15, 2023, from https://gizmodo.com/bitcoin-crypto-bank-of-international-settlements-1849784466

B. Coulter, Kelly Ann. “The Impact of News Media on Bitcoin Prices: Modelling Data Driven Discourses in the Crypto-Economy with Natural Language Processing.” Royal Society Open Science 9, no. 4 (April 20, 2022). https://doi.org/10.1098/rsos.220276.

C. Gurrib, Ikhlaas, Firuz Kamalov, and Linda Smail. “Bitcoin Price Forecasting: Linear Discriminant Analysis with Sentiment Evaluation.” *ArabWIC 2021: The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research, Sharjah, UAE*, 2021. https://doi.org/10.1145/3485557.3485561.

D. Kaastra, Iebeling, and Milton Boyd. “Designing a Neural Network for Forecasting Financial and Economic Time Series.” *Neurocomputing* 10, no. 3 (April 1996): 215–36. https://doi.org/10.1016/0925-2312(95)00039-9.

E. Kaman, Sweta. “News Sentiment Analysis by Using Deep Learning Framework.” *News Sentiment Analysis By Using Deep Learning Framework*, 2020. https://doi.org/10.14293/s2199-1006.1.sor-.ppcv5ia.v1

F. *Mai, Feng, Zhe Shan, Qing Bai, Xin (Shane) Wang, and Roger H.L. Chiang. “How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis.” Journal of Management Information Systems 35, no. 1 (2018): 19–52*

G. Mittal, Aditi, Vipasha Dhiman, Ashi Singh, and Chandra Prakash. “Short-Term Bitcoin Price Fluctuation Prediction Using Social Media and ...” IEEE Xplore, 2019. https://ieeexplore.ieee.org/document/8844899.

H. Philippas, Dionisis, Hatem Rjiba, Khaled Guesmi, and Stéphane Goutte. “Media Attention and Bitcoin Prices.” Finance Research Letters 30 (September 2019): 37–43. https://doi.org/10.1016/j.frl.2019.03.031.

I. Riveroll, F. (2020, March 14). *Predicting bitcoin price with business news (python)*. Medium. Retrieved February 15, 2023, from https://medium.com/swlh/predicting-bitcoin-price-with-business-news-python-f3bcf60f5818

J. Sattarov, Otabek, Heung Seok Jeon, Ryumduck Oh, and Jun Dong Lee. “Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis.” *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, 2020. https://doi.org/10.1109/icisct50599.2020.9351527.

K. Steinert, Lars, and Christian Herff. “Predicting Altcoin Returns Using Social Media.” PLOS ONE 13, no. 12 (2018).

Sources

L. *Tandon, Chahat, Sanjana Revankar, Hemant Palivela, and Sidharth Singh Parihar. "How Can We Predict the Impact of Social Media Messages on the Value of Cryptocurrency? Insights from Big Data Analytics." International Journal of Information Management Data Insights 1, no. 2 (2021): 100035.*

M. Throuvalas, A. (2022, December 20). *Portion of bitcoin supply held by retail investors reaches all-time high: Glassnode*. Decrypt. Retrieved February 15, 2023, from https://decrypt.co/117685/portion-bitcoin-supply-held-retail-reaches-all-time-high-glassnode

N. Yang, C. (2019, February 3). *How to use news articles to predict BTC price changes*. Medium. Retrieved February 15, 2023, from https://towardsdatascience.com/how-to-use-news-articles-to-predict-btc-price-changes-c37e305a97f