UNDERSTANDING HOW TWEET VOLUMES AND SENTIMENT AFFECT THE PRICE OF BITCOIN

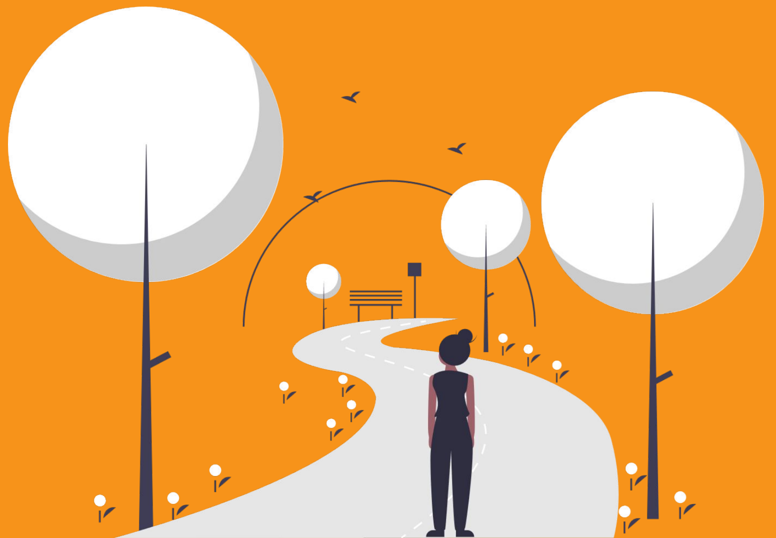
A STUDY ON PRICES FROM JANUARY 2021 - MARCH 2021

FINE460: FINANCIAL ANALYTICS
April 1st, 2022

David Perez, Youcef Sahnoune. Jonathan Steinberg

AGENDA

1. *The Context & Question*
2. *Our Process*
    a. *Data Collection & Analysis*
    b. *Variable Selection*
    c. *Multiple Regression*
    d. *Classification*
3. *Main Results*
4. *Potential Caveats*
5. *Appendix*



How we reached our ultimate research question

What factors generally impact cryptocurrency? What impact could social media “hype” have? How do investigate this?

How do **Tweet volumes** and **sentiment** impact the **price movement of Bitcoin?**

We took an iterative approach to understanding our most relevant variables

DATA ANALYSIS Explaining Bitcoin's price using Tweets	PREDICTIVE ANALYSIS Predicting Bitcoin's price using Tweets
PCA	MULTIPLE LINEAR REGRESSION
LASSO/RIDGE	REGRESSION TREES & RANDOM FOREST
CLUSTERING	

DATA COLLECTION

Data Collection Process

TWEETS DATA	SENTIMENT DATA	BTC PRICING DATA
snScrape	**Vader**	**FTX**
Python package which gathers an array of attributes on all tweets that contain a specific text search query (in this case, “BTC” or “Bitcoin”) between specified start and end dates Collected 6,915,458 tweets between for the first three months of 2021	Sentiment analysis package for social media texts. Vader compute a positive, negative, and neutral polarity score for each tweet, as well as a compound score that normalizes the aforementioned three scores Compound score ranges from -1 to 1, -1 being entirely negative and 1 being entirely positive	Bitcoin price can be gathered through FTX’s REST API, which allows us to gather candlestick data on the BTC/USD market for different levels of granularity between specified start and end dates

We then merged all data into hourly intervals to generate the following variables…

The starting set of variables we used in our analysis

Variable Name	*Description*
DateTime	*The date and hour of any given number of tweets*
Price	*The price of Bitcoin (BTC) at the end of the next hour*
Log_price	*The logarithmically transformed price of BTC for stationarity purposes, and our dependent variable*
Tweets	*# of BTC-related tweets in any given hour*
Likes	*# of BTC-related tweet likes in any given hour*
Replies	*# of replies to BTC-related tweets in any given hour*
Retweets	*# of retweets of BTC-related tweets in any given hour*
Quotes	*# of retweets of BTC-related tweets in any given hour that also have additional commentary*
CompoundMean	*The average sentiment "score" of the tweets across the hour (a positive and higher value indicates a more positive sentiment while a negative and lower value indicates more negative sentiment)*
VerifiedX	*X being tweets, replies, retweets, quotes, and likes, verified dictates whether or not the activity is done by a "verified" user on Twitter*
PositiveX	*X being tweets, replies, retweets, quotes, and likes, Positive dictates whether or not the activity has been attributed a positive sentiment (compound score > 0.05)*
NeutralX	*X being tweets, replies, retweets, quotes, and likes, Positive dictates whether or not the activity has been attributed a neither negative nor positive sentiment (-0.05 < compound score < 0.05)*
NegativeX	*X being tweets, replies, retweets, quotes, and likes, Negative dictates whether or not the activity has been attributed a negative sentiment (compound score < -0.05)*
Overall Activity	*# of BTC-related replies, likes, retweets, and quotes combined for any given hour*
X lag1	*A variable created for all the above variables (excluding DateTime) that was lagged by one hour*

In total, we ended up looking at 76 independent variables, and 2,160 price observations of Bitcoin

Vader is pretty accurate....

TraderKøz
@TraderKoz

I love $BTC
I love $ETH
I love $LTC
I love $DOT
I love $SNX
I love $AAVE
I love $COMP
I love $YFI
I love $SUSHI
I love $SOL
I love $LINK
I love $ALGO
I love $ATOM
I love $FTT

I LOVE CRYPTO

Compound Score: 0.9969

BitcoinExchangeRate
@bitcoinrate247

Average Bitcoin market price is: USD 28,982.56, EUR 23,673.86

Compound Score: 0.0000

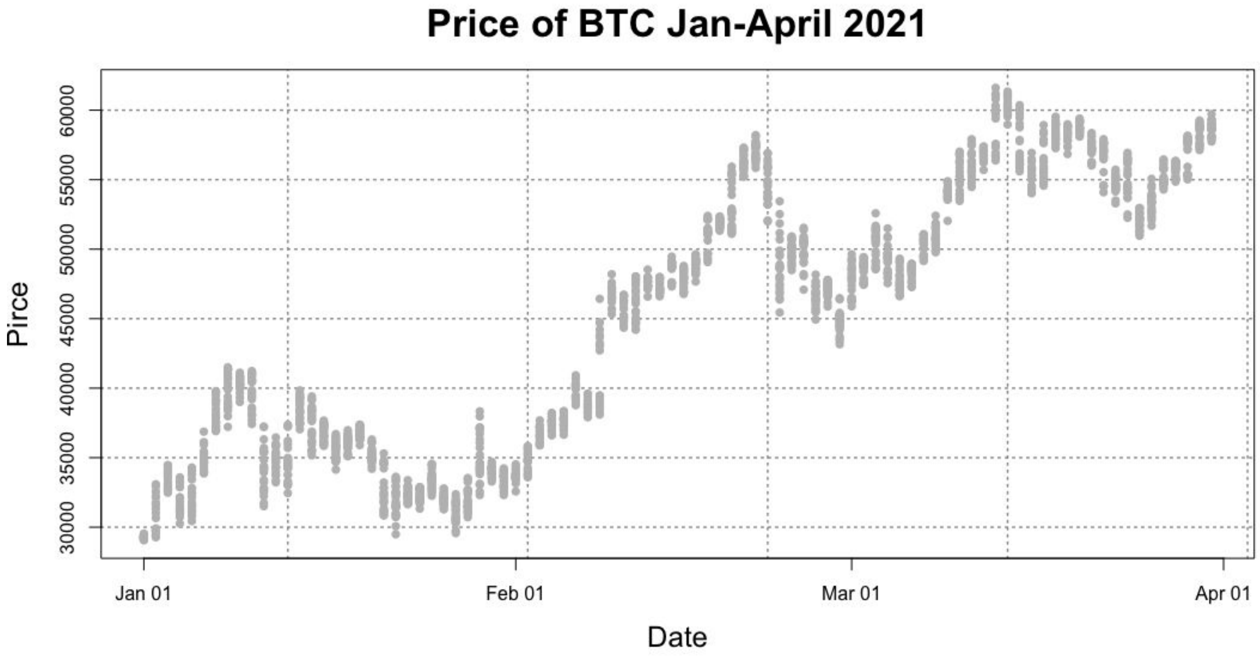
ShitcoinSurfer
@Shitcoin_Surfer

Replying to @UltraXBT

Fuck this scam. Fuck fuck fuck fuck fuck scam Bitcoin ponzi scam fuck this fuck scam scam fuck this scam fuck fuck fuck dog.

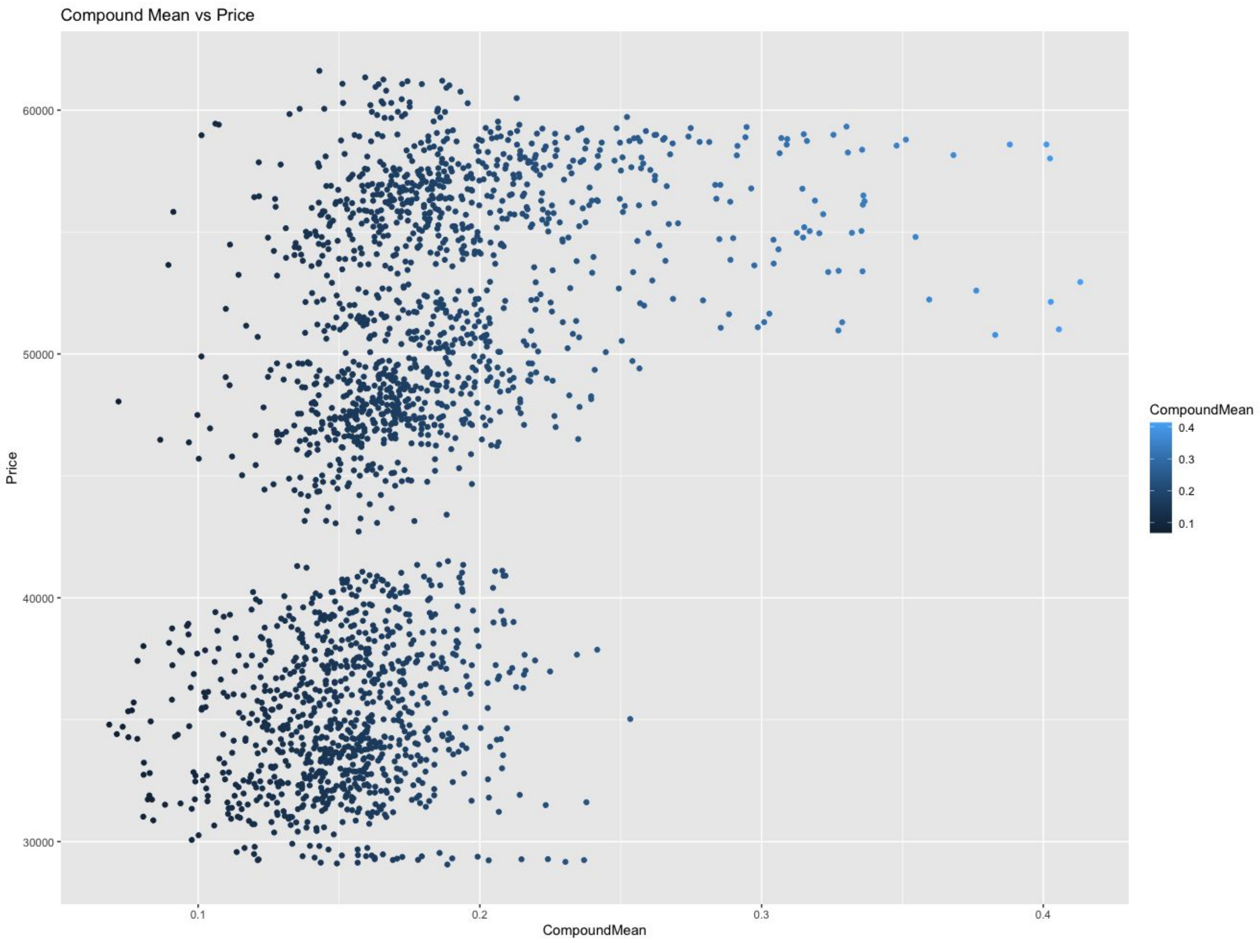
Compound Score: -0.9970

DATA ANALYSIS

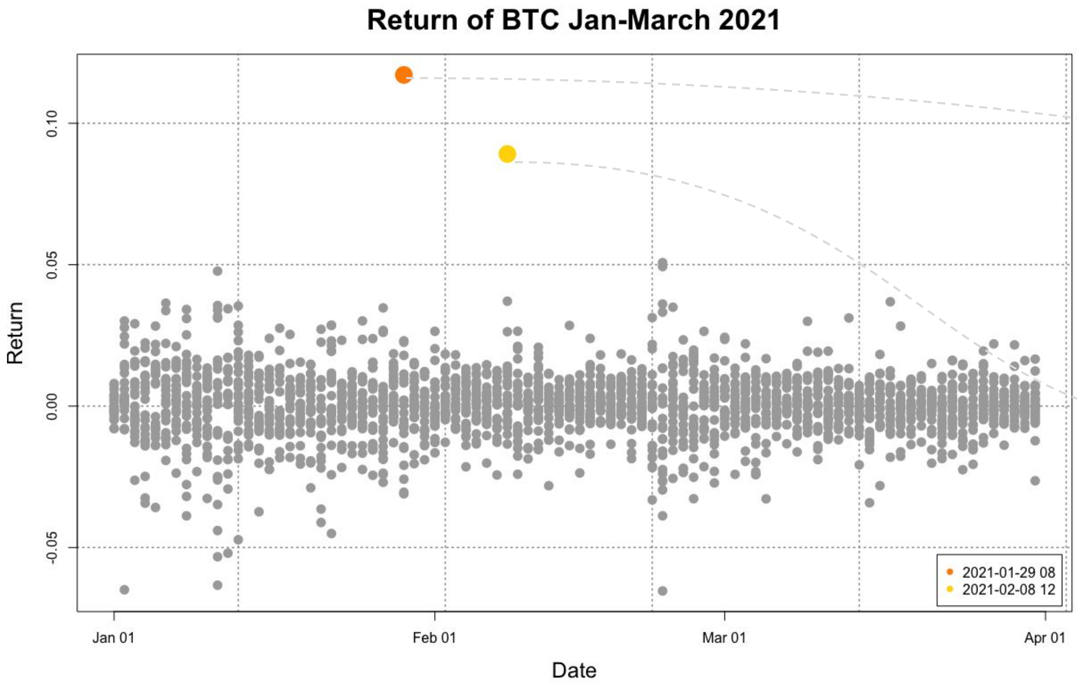


From January-April 2021, the price of bitcoin more than doubled

The relationship between Price and Sentiment



In our analysis, we noticed two return outliers



Two return outliers

Return Outliers

January 29th (3:00am)

Bitcoin spikes 20% after Elon Musk adds #bitcoin to his Twitter bio

PUBLISHED FRI, JAN 29 2021·5:20 AM EST | UPDATED FRI, JAN 29 2021·8:03 AM EST

Elon Musk
@elonmusk

In retrospect, it was inevitable

3:22 AM · Jan 29, 2021 · Twitter for iPhone

53.6K Retweets **7,480** Quote Tweets **578.1K** Likes

February 8th (8:00am)

Tesla buys $1.5 billion in bitcoin, plans to accept it as payment

PUBLISHED MON, FEB 8 2021·7:48 AM EST | UPDATED MON, FEB 8 2021·1:43 PM EST

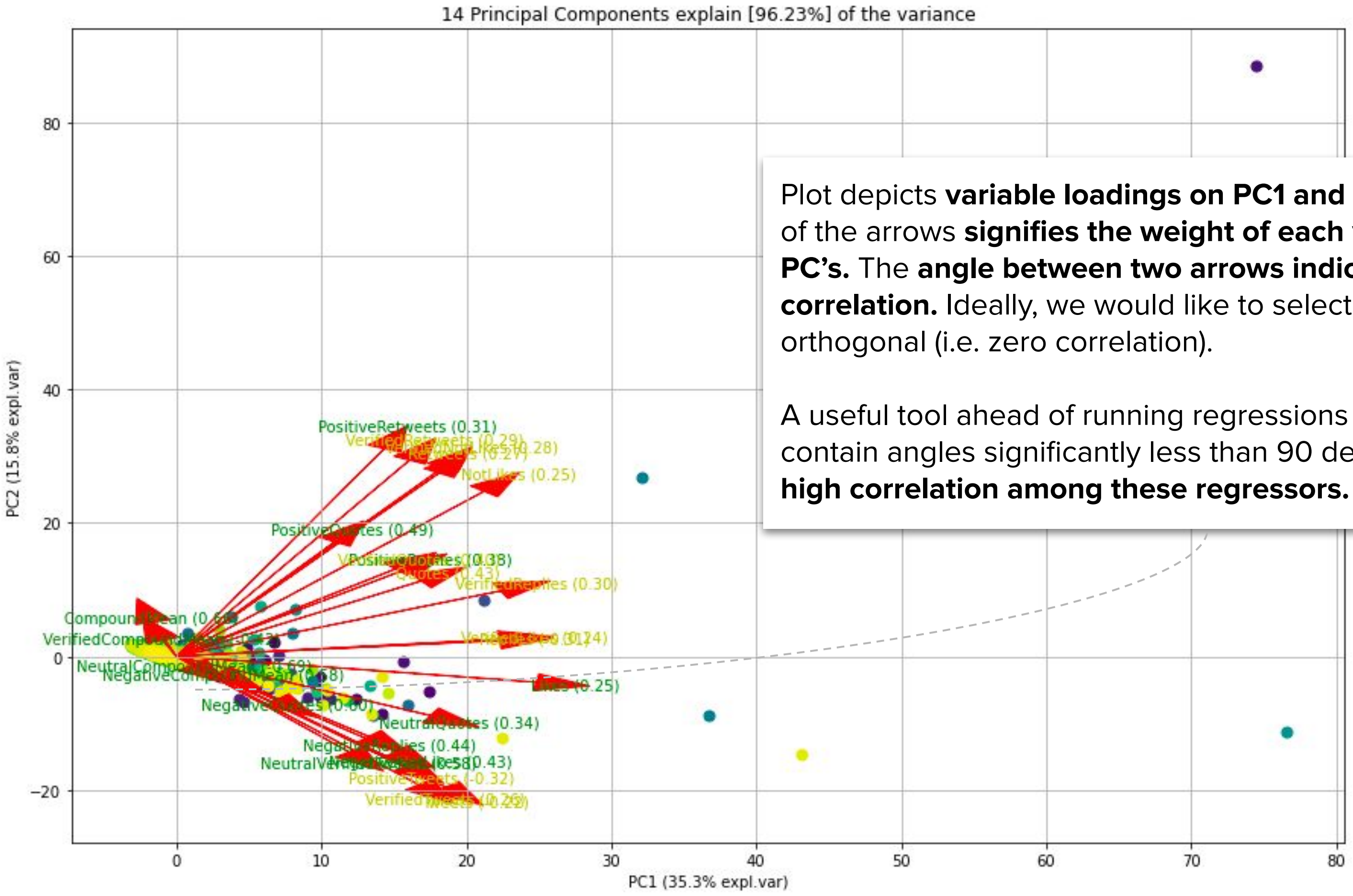
KEY POINTS

- Tesla announced in an SEC filing Monday that it has bought $1.5 billion worth of bitcoin.
- The company also said it would start accepting bitcoin as a payment method for its products.
- CEO Elon Musk has been credited for raising the prices of cryptocurrencies, including bitcoin, through his messages on Twitter.

Key finding: There is evidence that tweets about Bitcoin affect its price

VARIABLE SELECTION

PCA Plot of PC1 and PC2



Plot depicts **variable loadings on PC1 and PC2**. The magnitude of the arrows **signifies the weight of each variable on the PC's.** The **angle between two arrows indicates their correlation.** Ideally, we would like to select variables that are orthogonal (i.e. zero correlation).

A useful tool ahead of running regressions as many arrow pairs contain angles significantly less than 90 degrees, **suggesting high correlation among these regressors.**

Running LASSO and Ridge on all variables gave us an initial starting point

76 variables, many of which are highly correlated or insignificant

RIDGE

Illustrative example of optimal Ridge model output (non-exhaustive)

```
Tweets                    .
Likes                     .
Replies                   5.291363e-07
Retweets                  .
Quotes                    .
NotLikes                  5.233692e-08
CompoundMean              1.916068e+00
VerifiedTweets            .
VerifiedLikes             .
VerifiedReplies           .
```

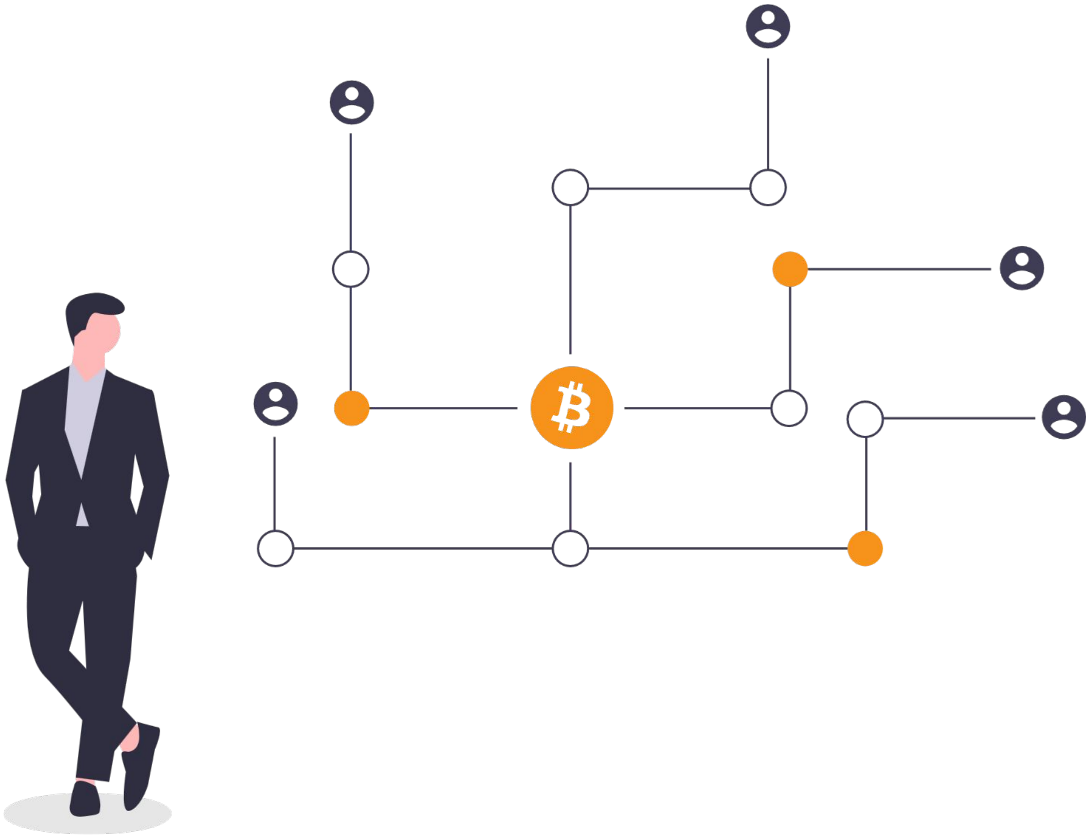
LASSO

Illustrative example of optimal LASSO model output (non-exhaustive)

```
VerifiedNotLikes             .
VerifiedCompoundMean         .
PositiveTweets               .
PositiveLikes                .
PositiveReplies              .
PositiveRetweets             .
PositiveQuotes               1.569049e-06
PositiveNotLikes             1.784591e-08
PositiveVerifiedTweets      -6.447730e-04
PositiveCompoundMean         1.041679e+00
```

17 variables related mostly to sentiment, activity, and engagement

PREDICTIVE ANALYSIS



We then used the variables that we selected to run a multiple regression

STEP 1: FOCUSING ON PREDICTIVE POWER

SIMPLE LINEAR REGRESSIONS → EVALUATE ON ADJ. R^2 & SIGNIFICANCE → SELECT BEST PREDICTORS

STEP 2: FOCUSING ON OUT-OF-SAMPLE PERFORMANCE

CREATE A MODEL ON TRAINING DATA → EVALUATE MSE ON TEST DATA → TUNE MODEL TO MINIMIZE MSE

Adjusting for linear regression

PROBLEMS	WHAT WE DID
LACK OF STATIONARITY	Took the log transformation of price, and examined historical data of BTC to understand the mean-reversion of BTC
HIGHLY CORRELATED VARIABLES	Examined the overall correlation matrix, and only selected variables with higher individual predictive powers (in relation to log_price)
TOO MANY VARIABLES	Ran another series of LASSO and Ridge analyses on the initial multiple regression to ensure we were only pulling out significant and important variables, and consolidated variables where we could

Understanding the Multiple Regression Results

```
Call:
lm(formula = lead(log_price, n = 1) ~ CompoundMean + NeutralTweets +
    NeutralCompoundMean + PositiveCompoundMean + NegativeVerifiedTweets +
    Compoundmeanlag1 + Overall_activity, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53019 -0.15294  0.02446  0.14146  0.46638

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            9.673e+00  7.228e-02 133.826  < 2e-16 ***
CompoundMean           9.965e-01  2.124e-01   4.691 2.97e-06 ***
NeutralTweets          5.957e-05  1.164e-05   5.117 3.51e-07 ***
NeutralCompoundMean    4.708e+01  1.896e+01   2.484  0.01311 *
PositiveCompoundMean   1.119e+00  1.521e-01   7.355 3.13e-13 ***
NegativeVerifiedTweets -3.275e-03  9.959e-04  -3.288  0.00103 **
Compoundmeanlag1       1.168e+00  1.988e-01   5.875 5.20e-09 ***
Overall_activity       1.519e-07  7.985e-08   1.902  0.05733 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.182 on 1502 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.3007,    Adjusted R-squared:  0.2974
F-statistic: 92.26 on 7 and 1502 DF,  p-value: < 2.2e-16
```

ADJUSTED R^2

29.74%

(not horrible for tweets)

MSE

0.035

(+/- **$7,500** deviation)

While our out-of-sample predictions weren't necessarily accurate, we still came to interesting findings on the relationship between BTC and Tweets

GENERAL SENTIMENT

Represented by the high coefficients and the high number of significant variables related to the "core" Compound Score variable

Interestingly, PositiveCompoundMean seemed to have the highest impact coefficient-wise, indicating that hours where tweets were overwhelmingly positive strongly contribute to price increase of Bitcoin

This supports our initial "hype" hypothesis, in spite of the out-of-sample inaccuracy of the model

POPULARITY

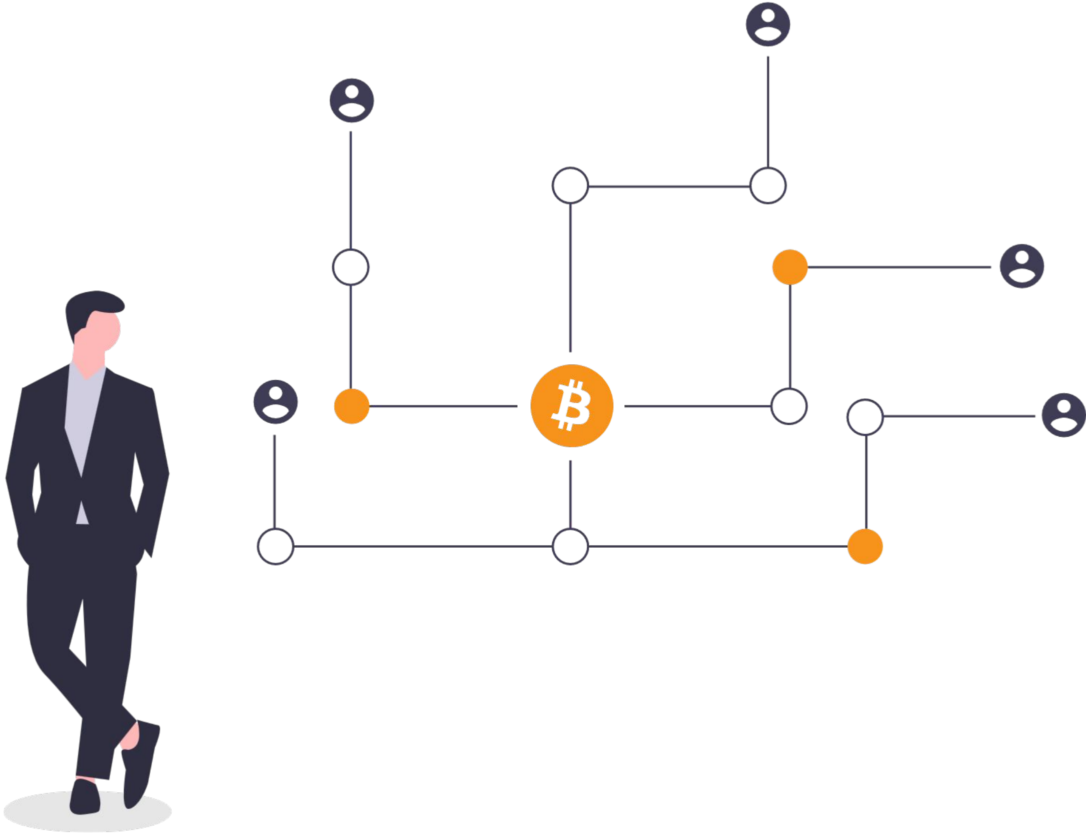
The importance of the variables tied to activity and whether or not a tweet is verified was also apparent in our regression, representing how famous tweeters/trending tweets can have a measurable impact on price

NegativeVerifiedTweets here had a strong negative impact on price (as the number of tweets increases), indicating that more famous people (or verified people) will have a higher impact on the price of Bitcoin

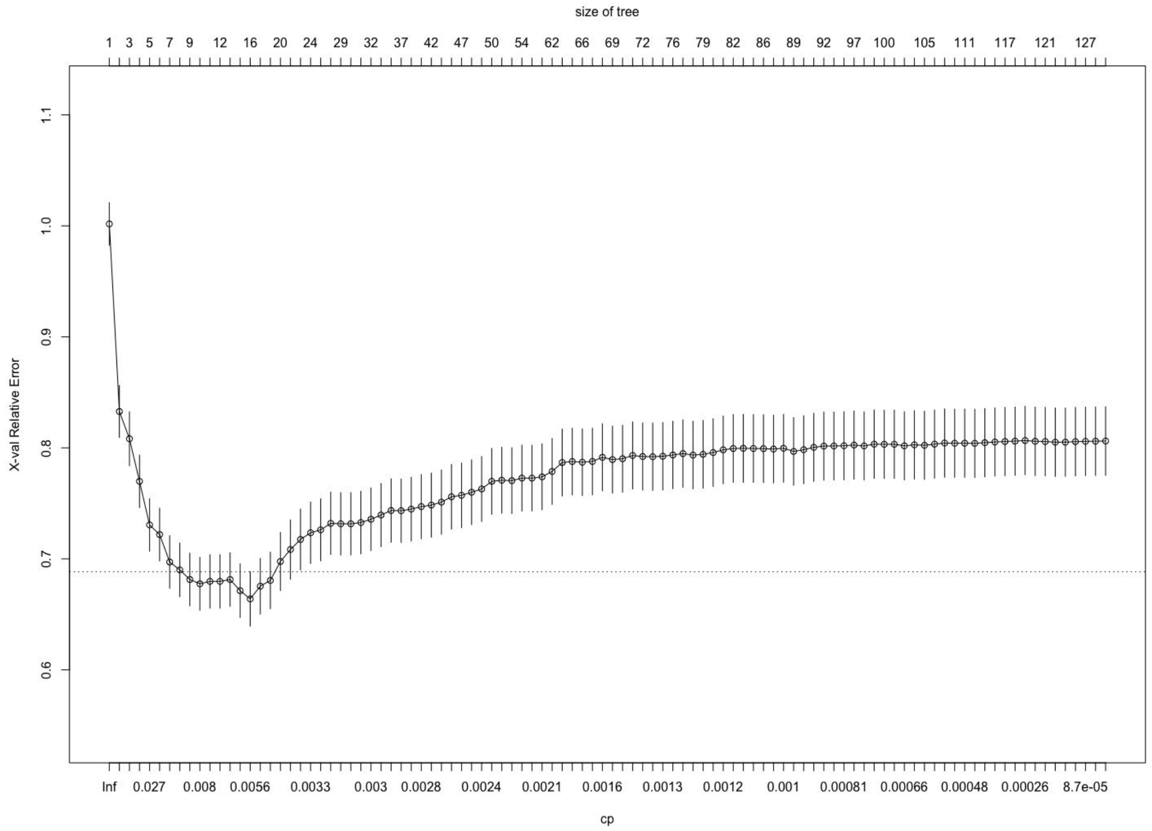
This also supports our initial "hype" hypothesis, as more popular players can have a higher magnitude impact

That said, there was more analysis to be done…

REGRESSION TREES & RANDOM FOREST



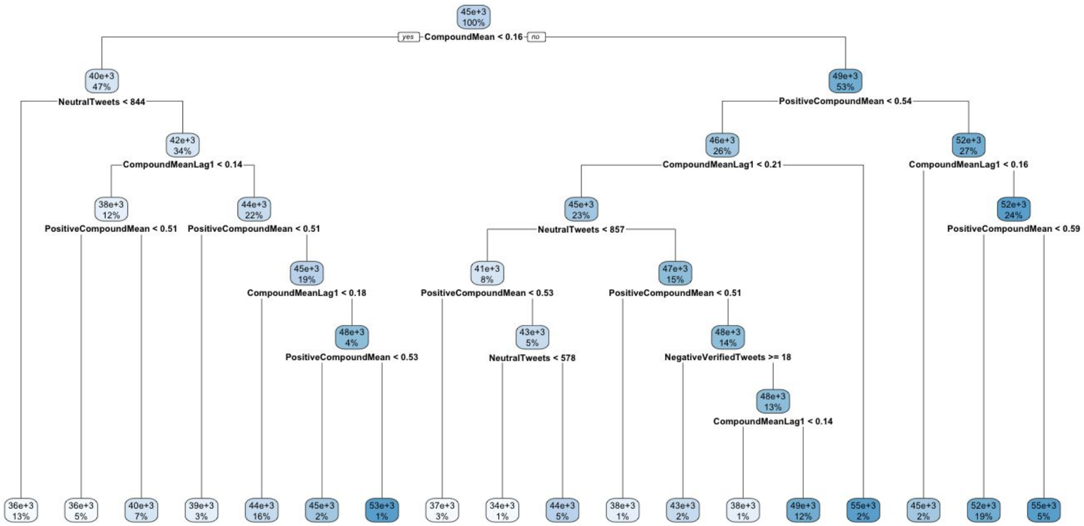
Continuous Regression Tree: Finding the Optimal CP



OPTIMAL CP

0.005218272

Continuous Regression Tree: Optimal Regression Tree



OPTIMAL REGRESSION TREE MSE

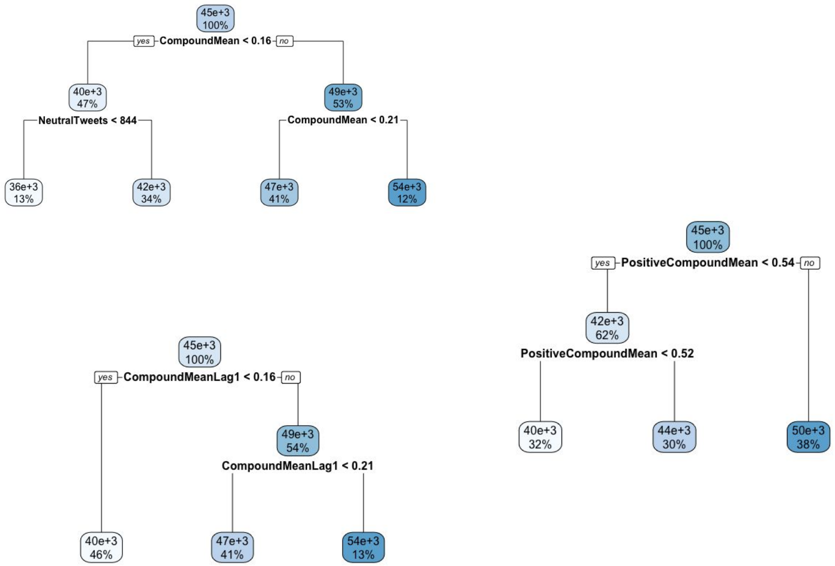
58,391,173

(+/- **$7,641.41** deviation)

Continuous Random Forest

Random forest constructs many regression trees and combines them for a more accurate prediction

For each tree, a random bootstrapped sample of sqrt(n) predictors is taken, which eliminates biases due to possible multicollinearity



MSE was found to be **4,908,6974** (+/- **$7,006.21** deviation)

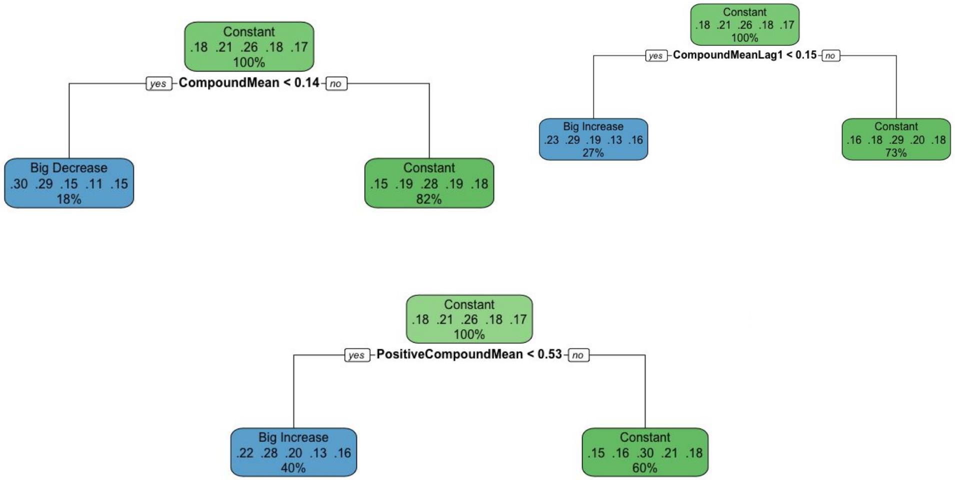
Classification Random Forest

We wanted to make our model more intuitive for investors by creating three classes…

Decrease: Return < -0.005

Constant: -0.005 < Return < 0.005

Increase: Return > 0.005



Accuracy was found to be **40.28**%

Classification Random Forest (Cont'd)

Confusion Matrix

	Constant	Decrease	Increase
Constant	534	64	104
Decrease	261	40	85
Increase	270	57	95

Sensitivity: 0.8056

Specificity: 0.2753

Sensitivity: 0.1585

Specificity: 0.9011

Sensitivity: 0.1362

Specificity: 0.8759

Our approach to data analysis did contain a number of caveats

FINAL TAKEAWAY

As seen with Elon Musk, there is evidence that tweets about bitcoin affected its price... However, our models only did an adequate job at best at encapsulating this relationship/predicting price based off twitter data. Clearly, there is more than effects that price of bitcoin than tweets alone.

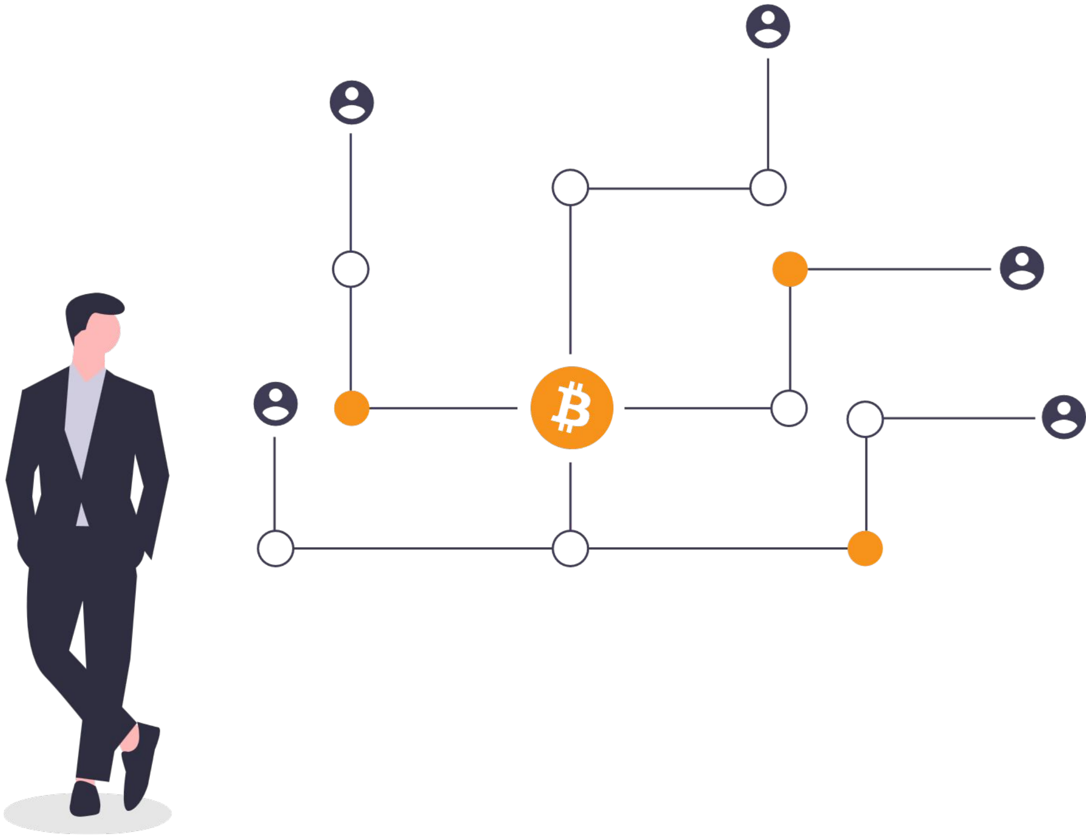
CAVEATS

Only pulling 3 months of data might be limiting and may result in a biased result

We assumed Bitcoin had a historical mean reverting property, thus allowing us to use log(price) for the purposes of regression and stationarity

Tweets about Bitcoin
vs.
Tweets that mentioned Bitcoin

APPENDIX



Across the last decade, Bitcoin has quickly risen in popularity among retail, institutional, and enthusiast investors, dictating cryptocurrency financial markets

12,700%

Price increase in Bitcoin from 2015 - today

$29.7bn

Trading volume in the last 24h in the US

57x

How much more popular Bitcoin is than “US dollar” on Google search terms

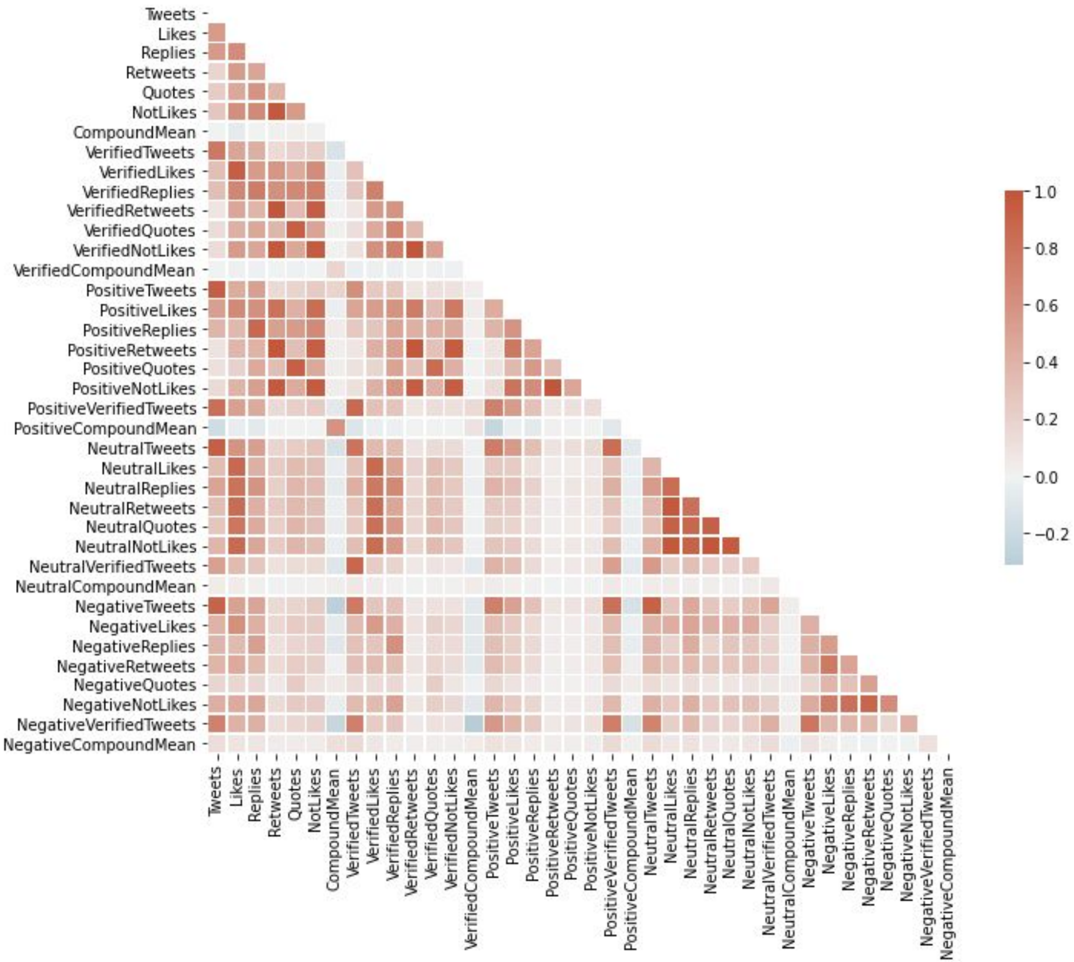
Summary Statistics

	COUNT	MIN	MEAN	MAX
PRICE	2,160	29,057	45,053	61,611
TWEETS	2,160	1,107	3,202	23,811
COMPOUND SCORE MEAN	2,160	0.06845	0.17165	0.41310
ACTIVITY	2,160	5,255	37,587	1,503,654

Indicates generally positive sentiment towards BTC. This makes sense given Bitcoins price double from Jan 1 - April 1

More in-depth Summary Statistics

Price	Tweets	Likes	Replies	Retweets	Quotes	NotLikes	Activity	CompoundMean
Min. :29057	Min. : 1107	Min. : 3793	Min. : 504	Min. : 544	Min. : 56.0	Min. : 1222	Min. : 5255	Min. :0.06845
1st Qu.:35705	1st Qu.: 2209	1st Qu.: 13071	1st Qu.: 1621	1st Qu.: 1966	1st Qu.: 242.0	1st Qu.: 3974	1st Qu.: 17301	1st Qu.:0.14680
Median :47088	Median : 2799	Median : 21110	Median : 2452	Median : 3156	Median : 394.0	Median : 6160	Median : 27478	Median :0.16569
Mean :45053	Mean : 3202	Mean : 28845	Mean : 3270	Mean : 4778	Mean : 694.3	Mean : 8742	Mean : 37587	Mean :0.17165
3rd Qu.:54227	3rd Qu.: 3690	3rd Qu.: 33899	3rd Qu.: 3674	3rd Qu.: 5093	3rd Qu.: 656.0	3rd Qu.: 9574	3rd Qu.: 43832	3rd Qu.:0.18793
Max. :61611	Max. :23811	Max. :1277685	Max. :88761	Max. :781699	Max. :77568.0	Max. :860265	Max. :1503654	Max. :0.41310



Correlation Heat Map

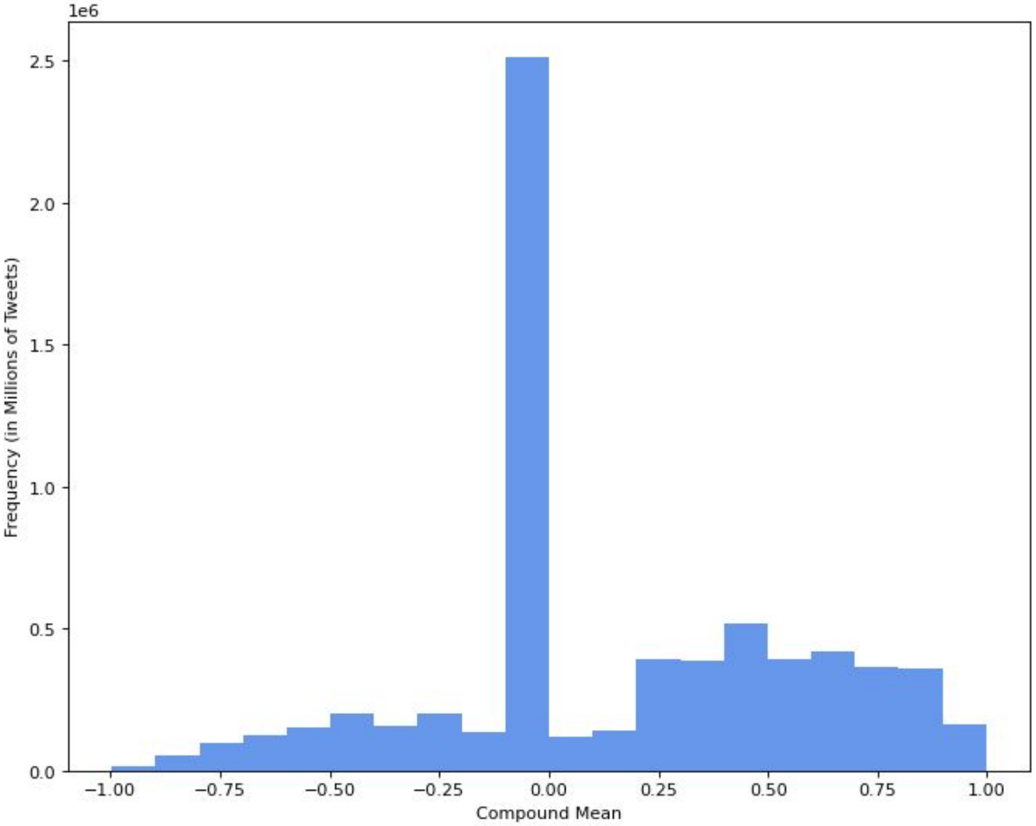
The matrix here suggests some significant multicollinearity among regressors

We used this as a tool to verify the results of the multiple linear regression

Some key mentions:

- Very high correlation between Tweets and Neutral Tweets
- Very high correlation between Replies and Verified Activity

Histogram of Compound Mean



In this histogram we can see that more tweets were positive as opposed to negative, consistent with our previous findings. However, the vast majority of tweets were found to be neutral in our dataset.

K-MEANS CLUSTERING

K-Means: We used the silhouette method to choose the optimal number of clusters

$$Silhouette_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

The Silhouette Method is used to gauge **quality of cluster assignment** for each observation

Attempts to **minimize cohesion** and **maximize separation**

COHESION

How similar observations are within the same cluster

- a_i: average distance between observation i and all other observations in same cluster

SEPARATION

How different clusters are from each other

- b_i: average distance between observation i and all other

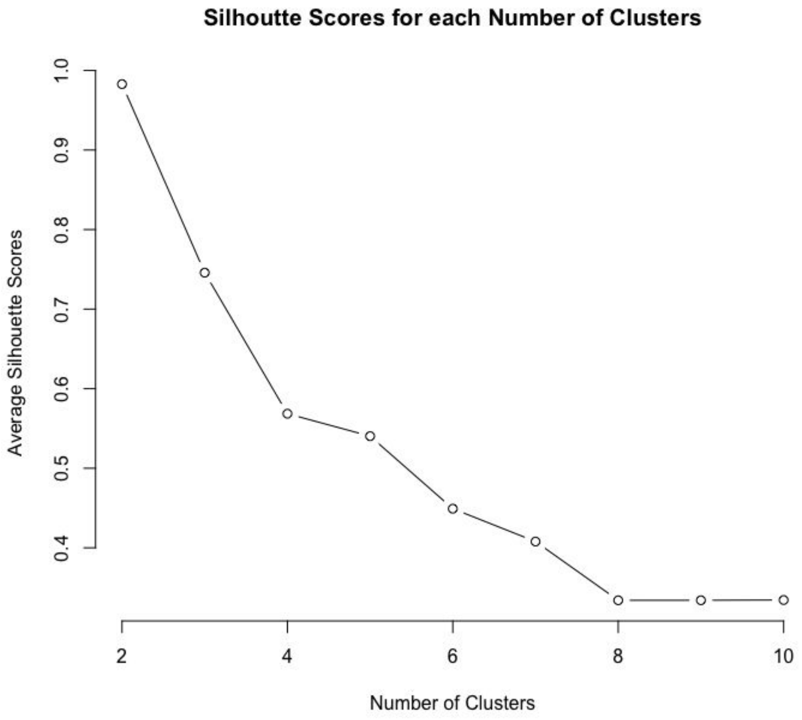
POSITIVE VALUE

A positive silhouette score indicates that a particular object is well matched to its own cluster, and not well to another

HIGHER VALUE

When using silhouette, higher values are better in validating clusters

K-Means: We used the silhouette method to choose the optimal number of clusters



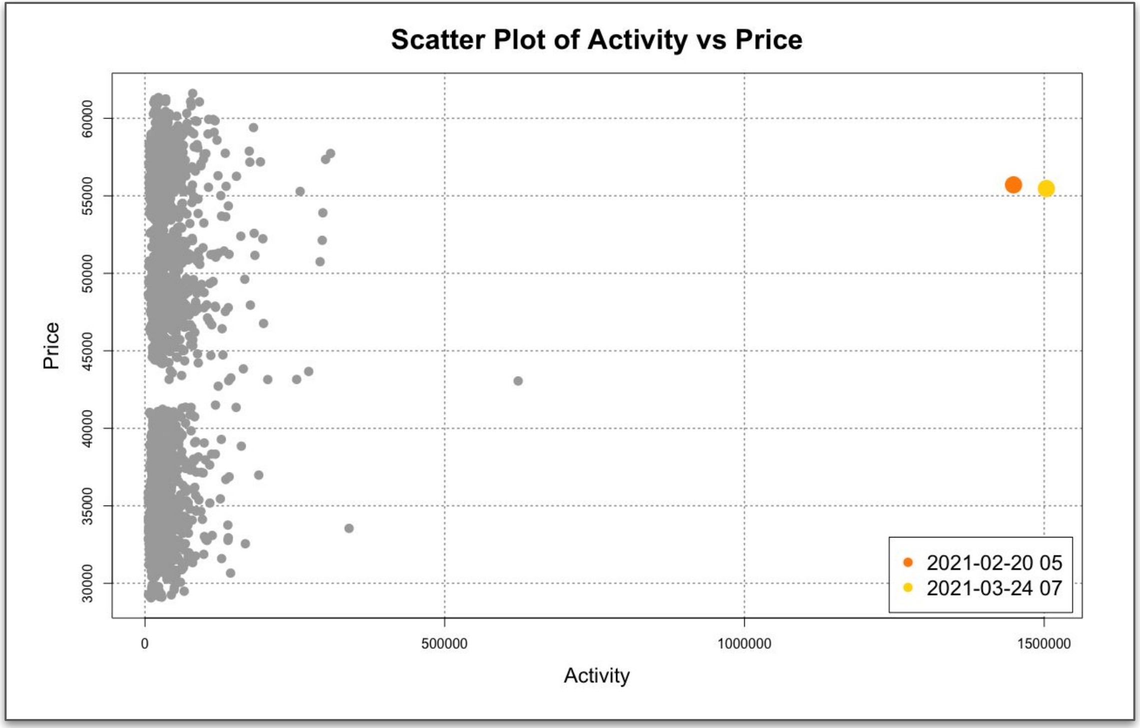
K-means clustering was performed for all variables in our dataset for each value of K between 2 - 10

The optimal amount of clusters was found to be 2

Cluster #1: Tweets on 2021-02-20 05 & 2021-03-24 08 (UTC)

Cluster #2: Remaining observations

K-Means: Analysis of Cluster #1



Observations in cluster #1 significantly more activity than any other hour

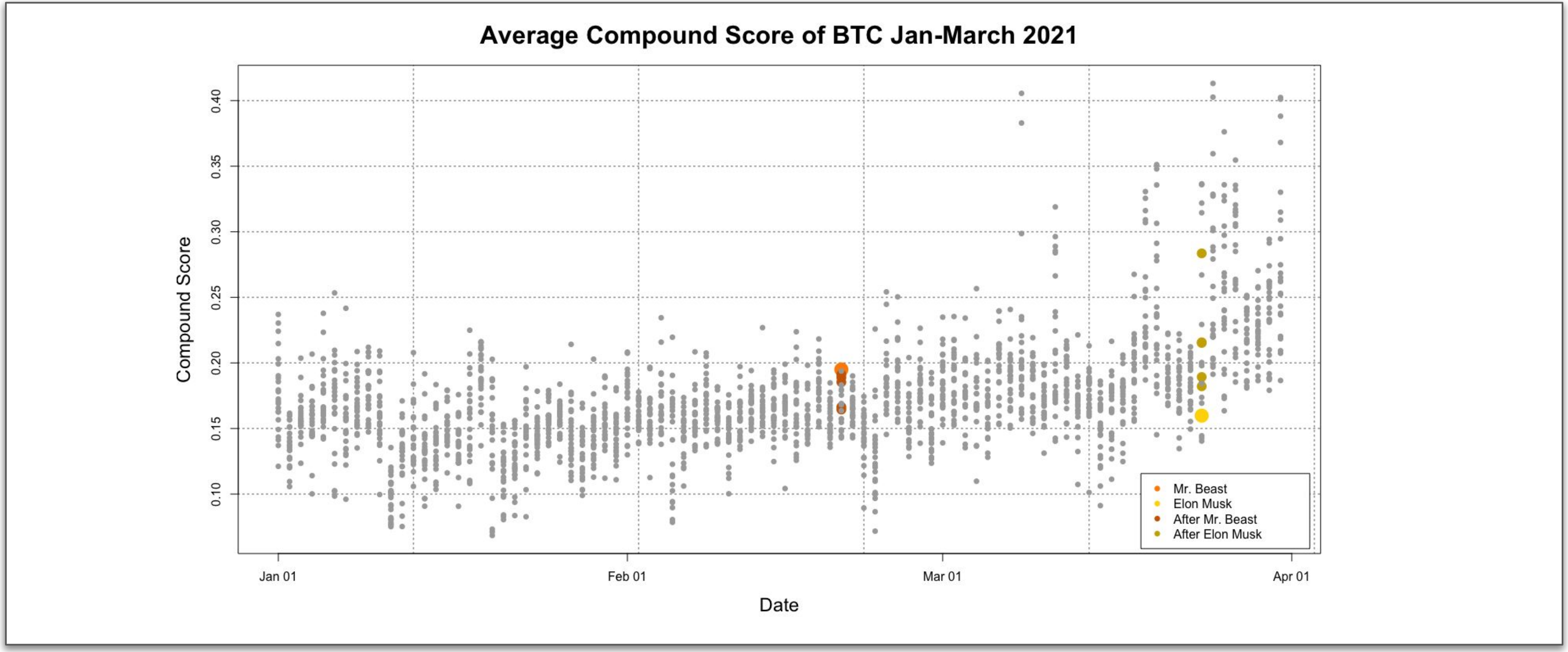
K-Means: Analysis of Cluster #1

Tweet with most activity: February 20th (12:00am)

Tweets with 2nd, 3rd, 7th most activity: *March 24th (3:00am)*

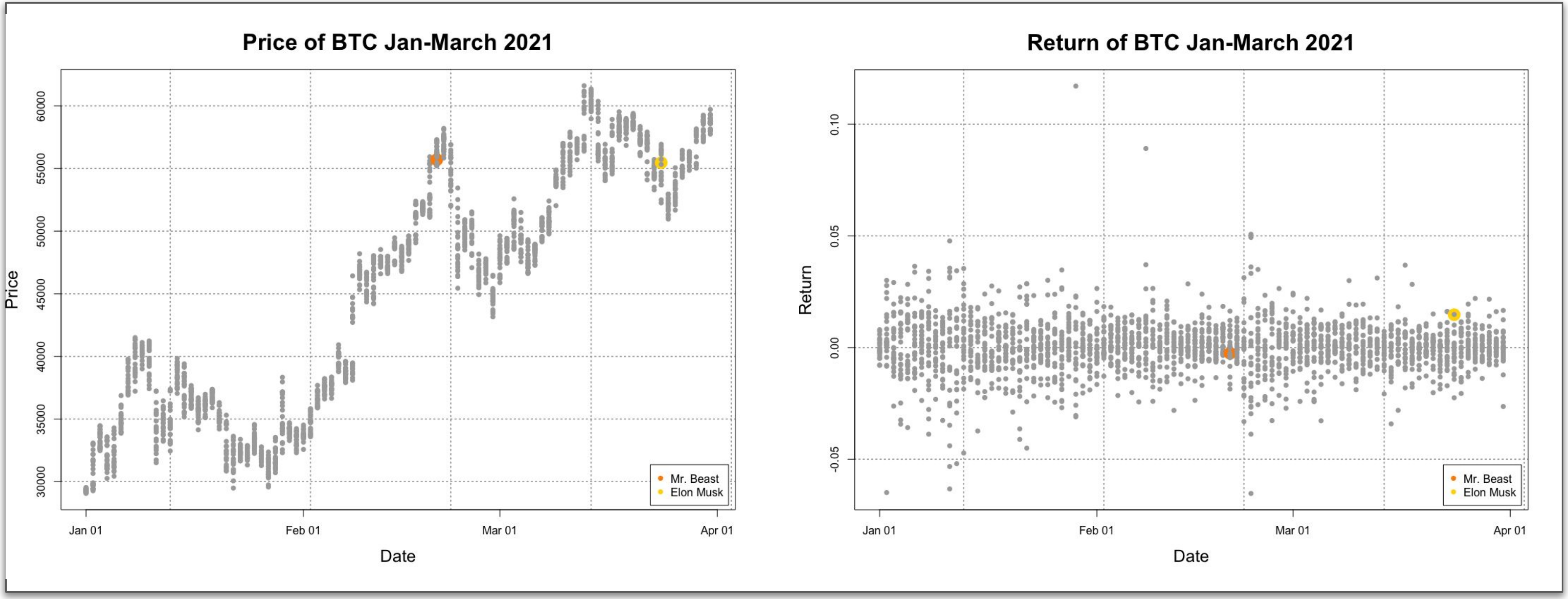
Our cluster found the most popular tweets about bitcoin, but did these tweets affect its price?

K-Means: Analysis of Cluster #1



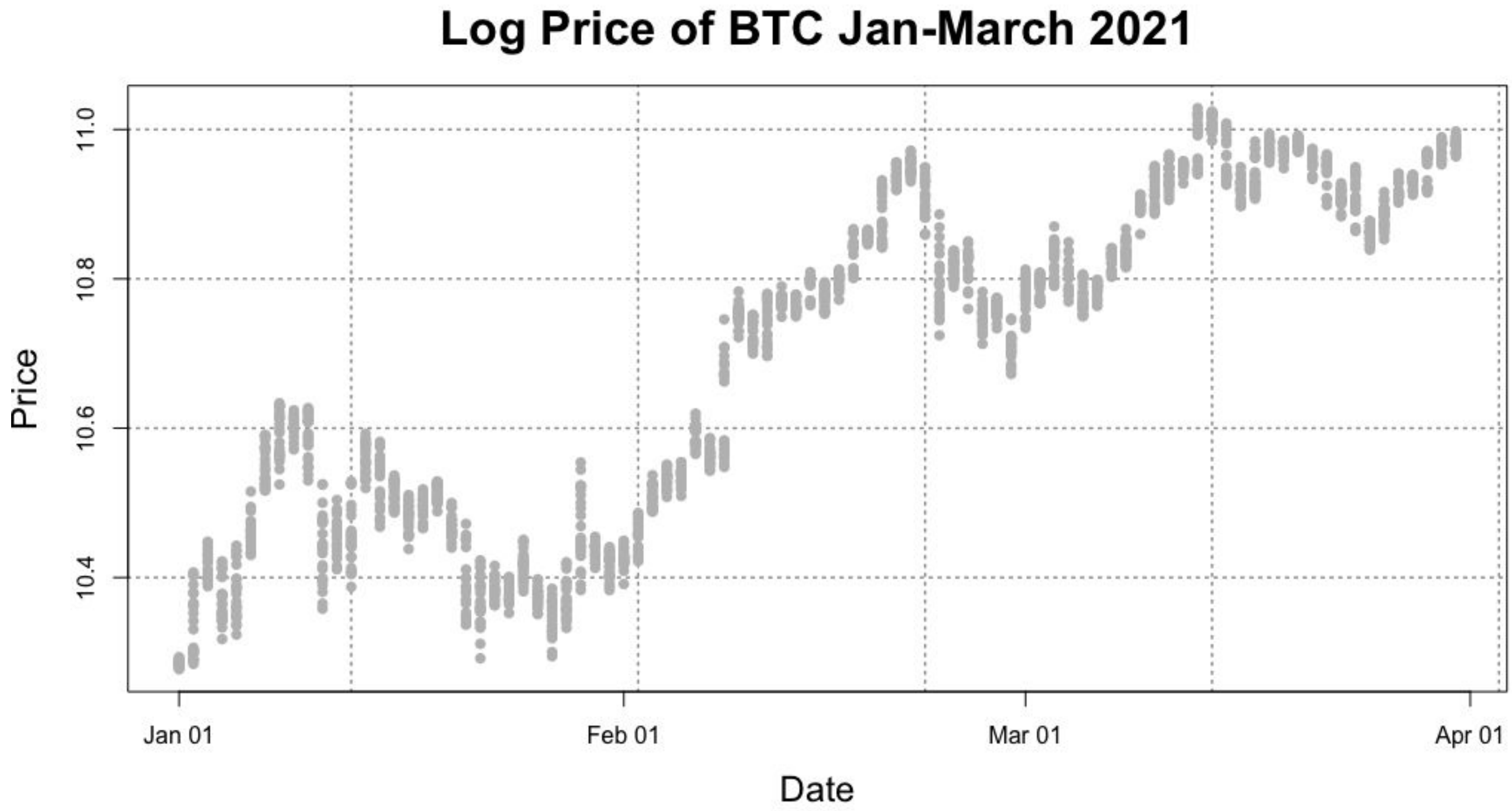
Tweets were more positive after Elon Musk and more negative after Mr. Beast

K-Means: Analysis of Cluster #1



Observations in cluster #1 do not exhibit any effect on BTC Price or Return in the next hour

STATIONARITY

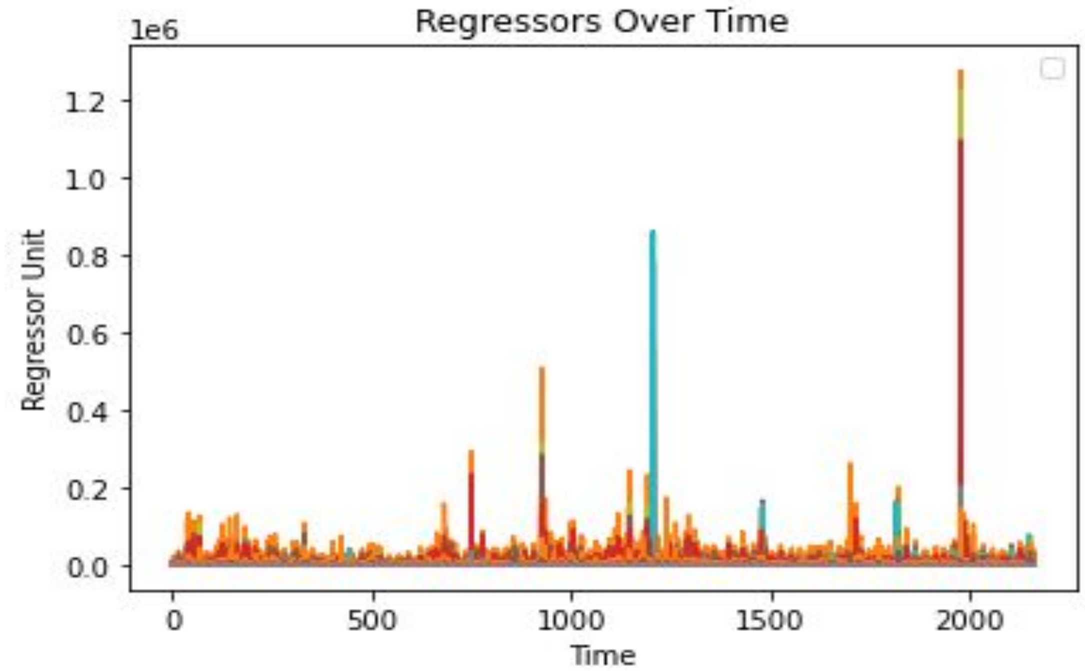

Log Price of BTC Jan-March 2021
Price
11.0
10.8
10.6
10.4
Jan 01
Feb 01
Mar 01
Apr 01
Date

Argument for Stationarity for Bitcoin Prices

If the trend reverts to a common mean, we can argue that bitcoin prices are stationary

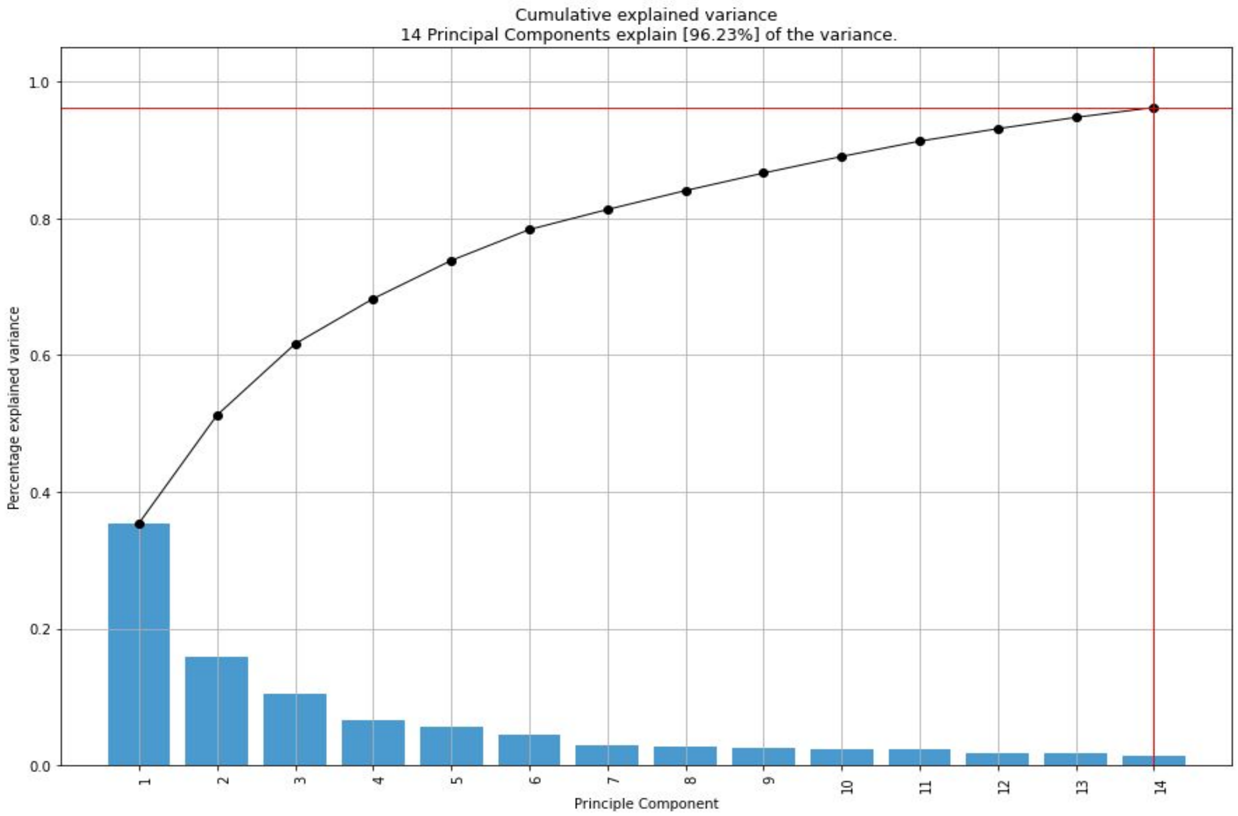
Stationarity Among Predictor Variables

Reversion to the mean. Omission of any transformation to preserve interpretation of regressors. Superimposed image of all regressors as a time series. Volatility clusters that revert to the mean.



PCA

Explained Variance by Principal Component



Principal components significantly reduce our initial regressor count

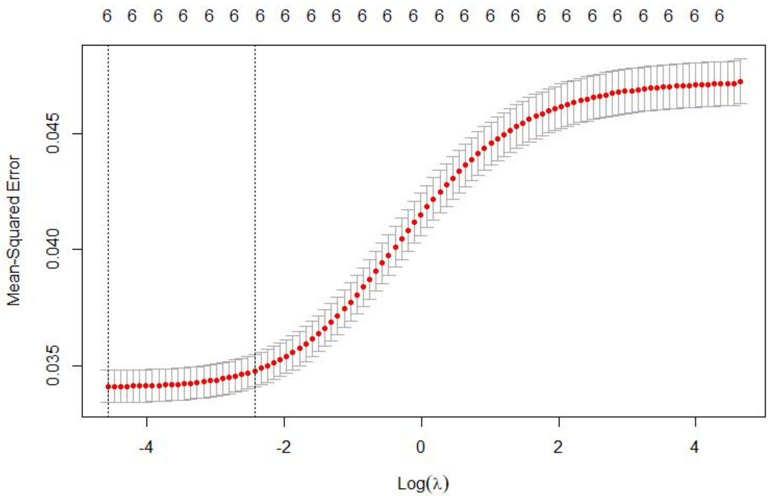
Factor loadings were not used as regressors because we favoured interpretability of our model rather than performance

Loadings on different PC's have less economic significance by construction

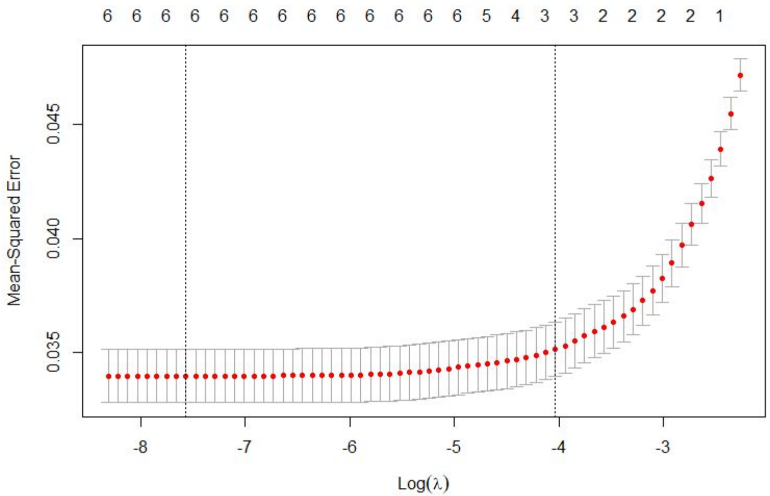
LASSO/RIDGE

Ridge& LASSO Plots

Plotting the ridge analysis, which outlines the minimum MSE at around 0.035



Plotting the LASSO analysis, which outlines the minimum MSE at around 0.035



Lasting Ridge and LASSO tables in developing the final models

RIDGE

```
> coef(best_model_ridge)
8 x 1 sparse Matrix of class "dgCMatrix"
                                   s0
(Intercept)             9.709278e+00
CompoundMean            1.914379e+00
NeutralTweets           6.397695e-05
NeutralCompoundMean     4.399023e+01
PositiveCompoundMean    1.128919e+00
NegativeVerifiedTweets -3.682043e-03
Compoundmeanlag1        .
Overall_activity        1.334898e-07
```

LASSO

```
> coef(best_lambda_lasso)
8 x 1 sparse Matrix of class "dgCMatrix"
                                   s0
(Intercept)             9.735573e+00
CompoundMean            1.917278e+00
NeutralTweets           5.364101e-05
NeutralCompoundMean     3.477853e+01
PositiveCompoundMean    1.087085e+00
NegativeVerifiedTweets -2.609585e-03
Compoundmeanlag1        .
Overall_activity        1.067810e-07
```

While RIDGE and LASSO both suggested removing Compoundmeanlag1, we found that including this variable actually contributed to both a higher R^2, and a lower MSE

SIMPLE LINEAR REGRESSIONS

Illustrative example of the simple linear regression process

Running a regression on each column

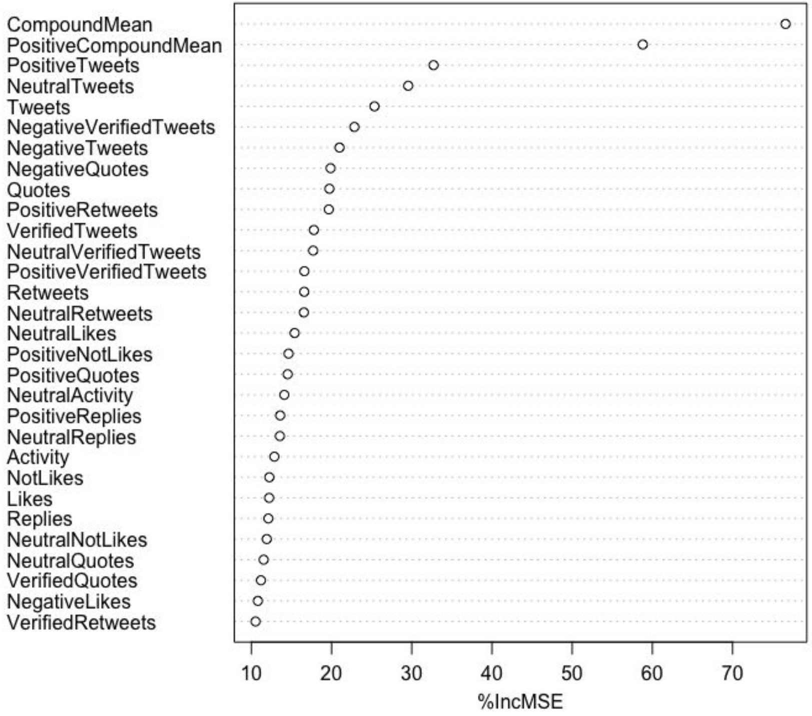
```
for (i in 1:76)#length(colnames(df_lag1))
{
  x<-(df_lag1[,i])
  y<-(lead(df_lag1$log_price,n=1))

  Reg_results[[i]] <- lm(y~x)
  w = lm(y~x)
  R_results <- c(R_results, summary(w)$r.squared)
}
```

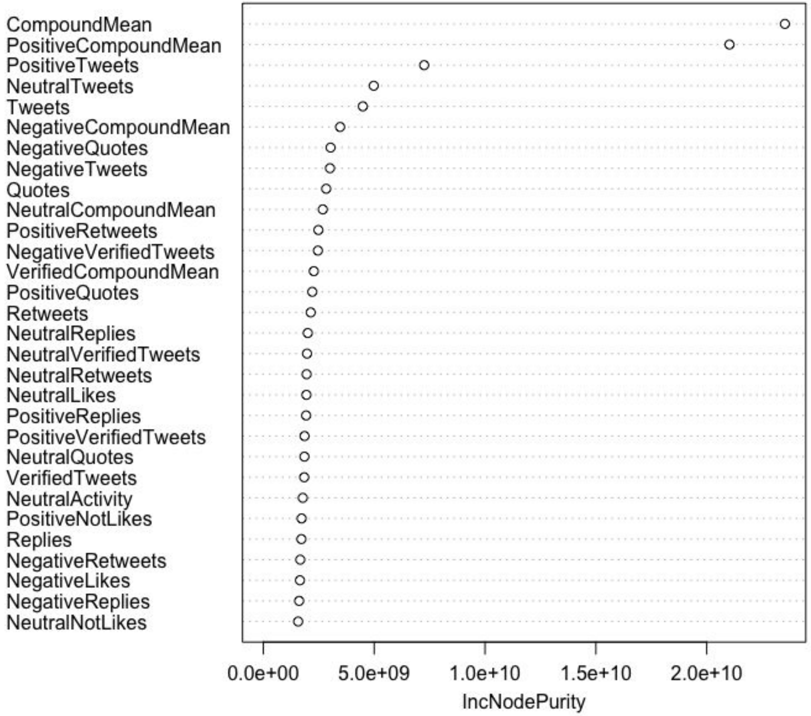
Looking at the coefficients, intercepts, p_values, and R^2 of each individual relationship

unlist.var_	unlist.Coe	unlist.inte	unlist.p_v	unlist.R_results.
Tweets	1.22E-05	10.65359	4.11E-05	0.00777
Likes	4.50E-07	10.67957	0.000122	0.006821
Replies	4.62E-06	10.67744	9.74E-05	0.007017
Retweets	7.30E-07	10.68908	0.006125	0.003478
Quotes	6.25E-06	10.68823	0.003254	0.004007
NotLikes	7.50E-07	10.68601	0.000807	0.005192

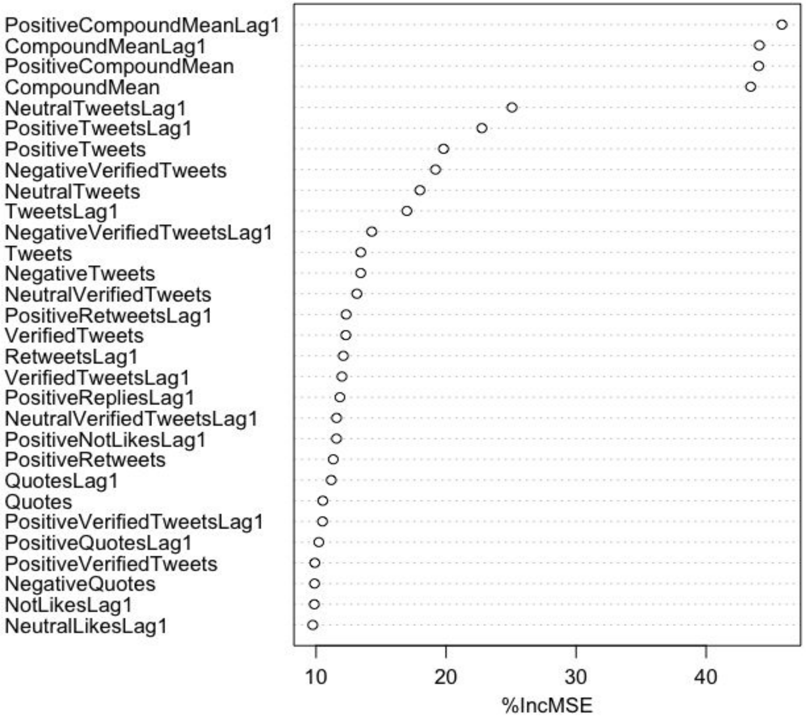
VARIABLE IMPORTANCE

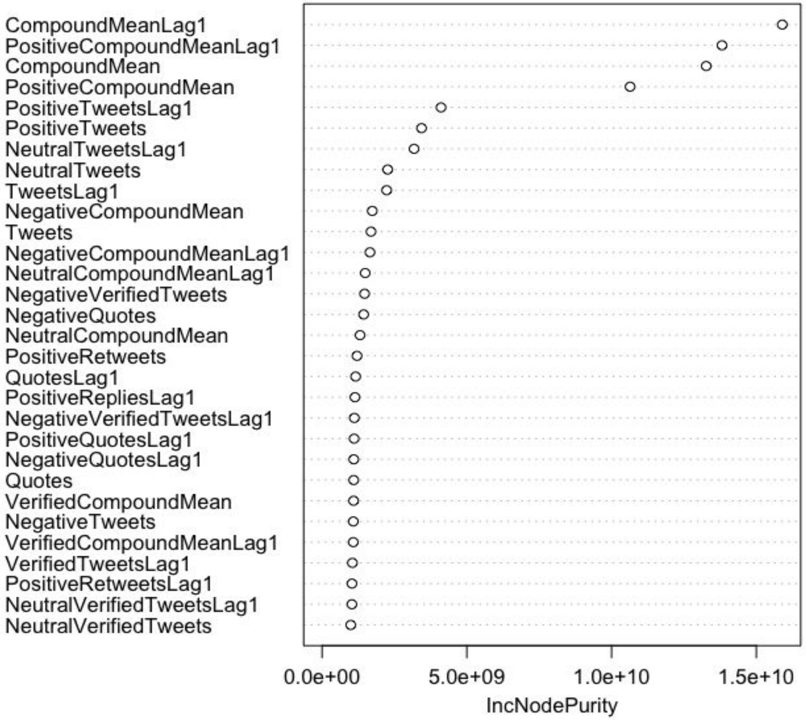
Continuous Variable Importance: Without Lags



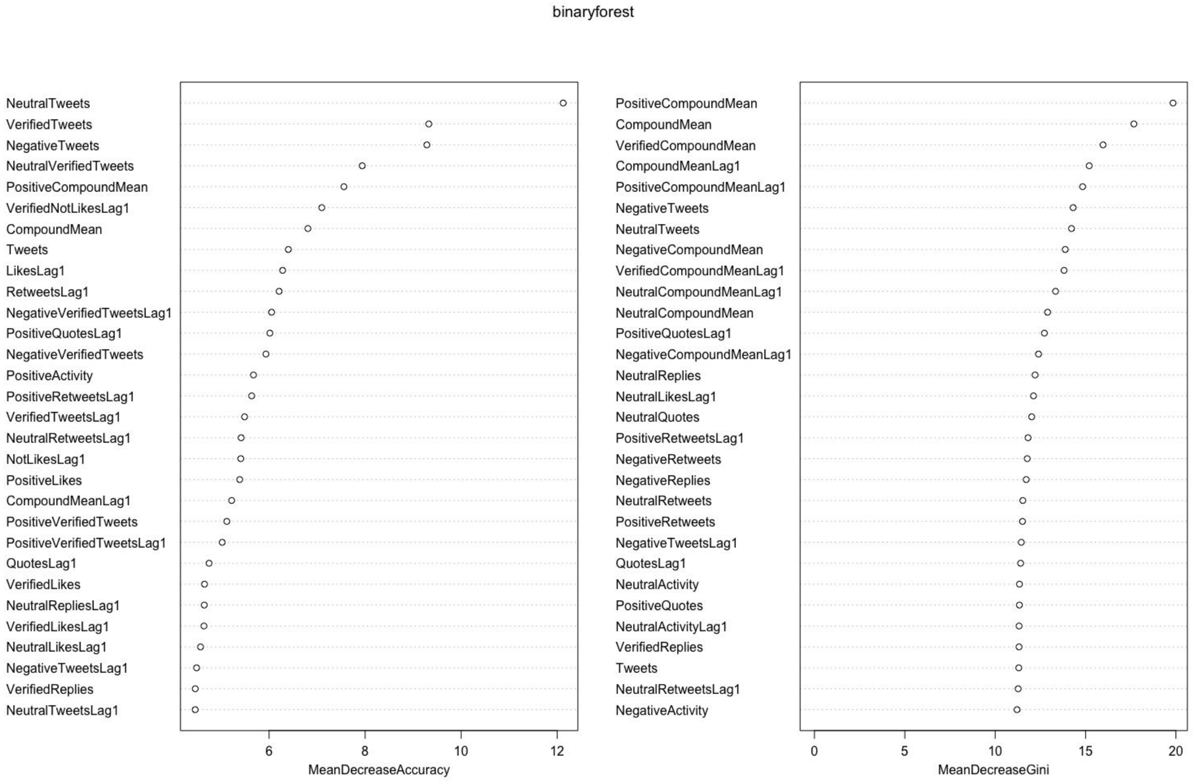


Continuous Variable Importance: With Lags



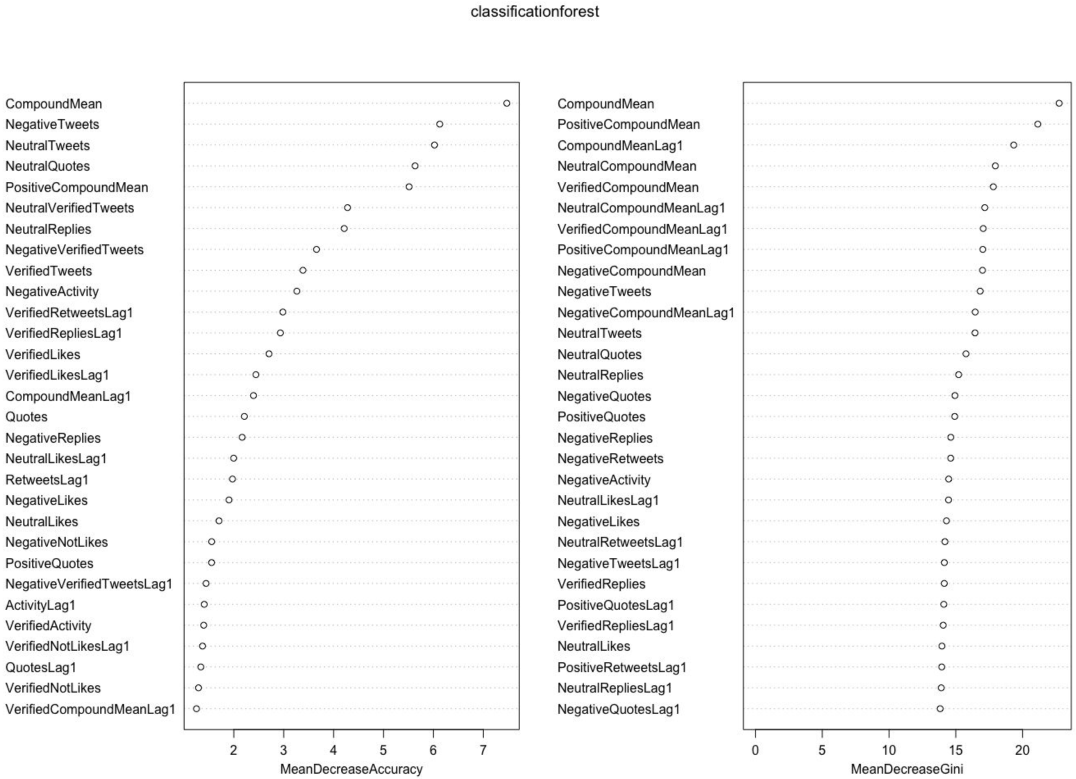


Classification Random Forest: Variable Importance



Interestingly enough, neutral tweets played a massive part in the analysis here, alongside variables that represent the general sentiment of the tweets

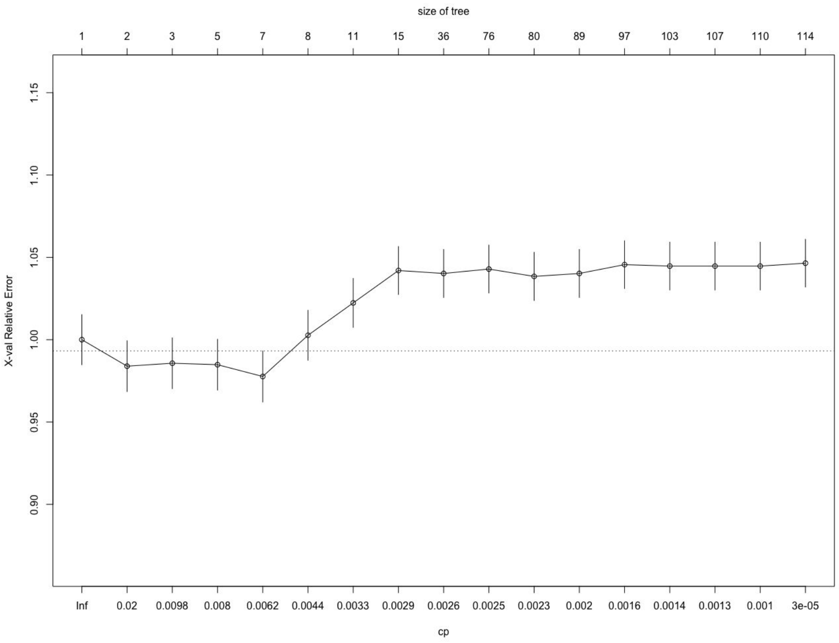
Multi-Classification Random Forest: Variable Importance



Common take away from all variable importance plots; compound mean and compound mean of lag 1 were found to be the most important variables, suggesting we should be included them in our models

CLASSIFICATION REGRESSION TREE

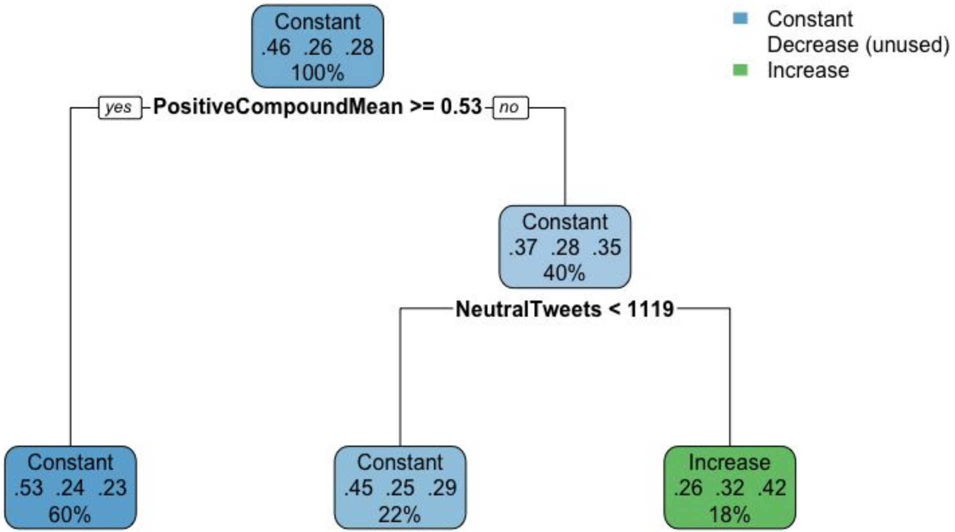
Classification Regression Tree: Finding the Optimal CP



OPTIMAL CP

0.00536193

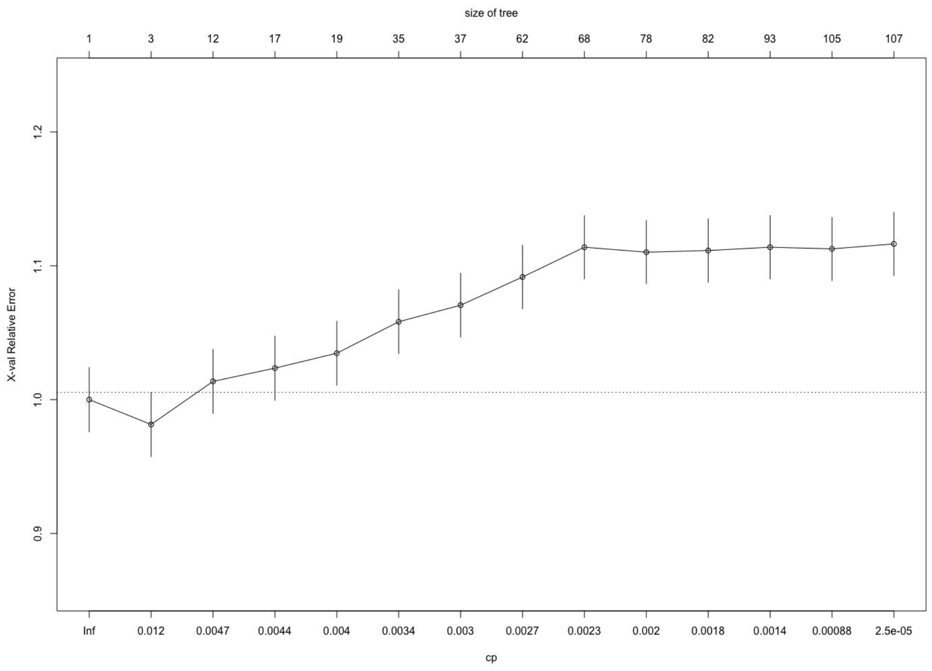
Classification Regression Tree: Optimal Regression Tree



Regression tree is easy to read and intuitive for people trading on this strategy in real time

MULTI-CLASSIFICATION MODEL

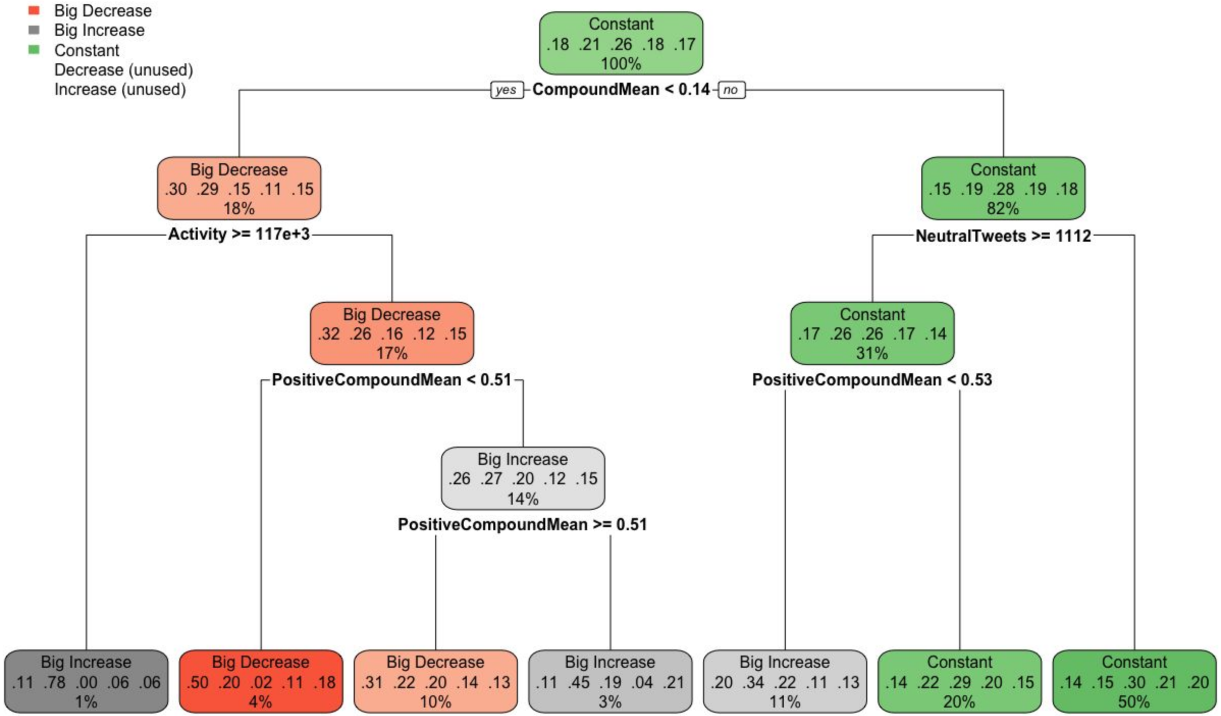
Multi-Classification Regression Tree: Finding the Optimal CP



OPTIMAL CP

0.004950495

Multi-Classification Regression Tree: Optimal Regression Tree



Interestingly, 70% of the observations were classified as constant, recommending a hold position most of the time

Multi-Classification Random Forest

What if we increase the classes...

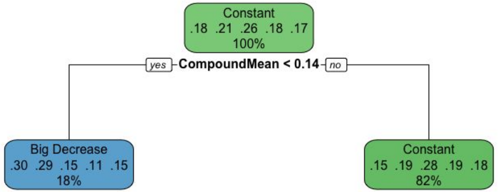
Big Decrease: Return < -0.0075

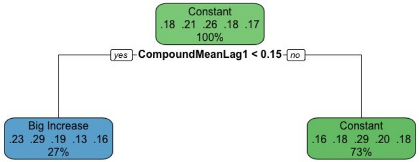
Decrease: -0.0075 < Return < 0.0025

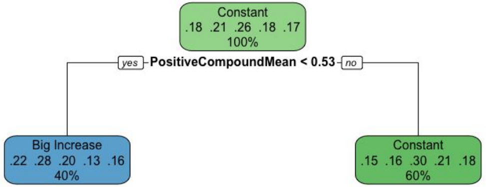
Constant: -0.0025 < Return < 0.0025

Increase: 0.0025 < Return < 0.0075

Big Increase: Return > 0.0075







Accuracy was found to be **24.85**%

Multi-Classification Random Forest (Cont'd)

Confusion Matrix

	Big Decrease	Big Increase	Constant	Decrease	Increase	
Big Decrease	46	82	83	31	28	Sensitivity: 0.18852 Specificity: 0.9006
Big Increase	52	98	116	32	20	Sensitivity: 0.2933 Specificity: 0.7871
Constant	39	85	169	54	44	Sensitivity: 0.4453 Specificity: 0.5832
Decrease	33	48	130	34	26	Sensitivity: 0.1702 Specificity: 0.8628
Increase	27	46	124	39	24	Sensitivity: 0.1136 Specificity: 0.9205