Detection of Price Manipulation in Cryptocurrencies

Daniel Sapkota

Abstract

In this study, we use daily OHLC and Margin position data from Bitfinex, and news posted on Reddit from November 10 2017 to November 10 2019 to detect potential price manipulations in Bitcoin (BTC), Ethereum (ETH), Ethereum Classic (ETC), ZCash (ZEC), and Litecoin (LTC). We trained a random forest classifier with a moving training test split starting from a training set of 200 days, and a test set 80 days. We were able to detect potential manipulations at an AUC of 0.57, 0.72, 0.58, 0.58 and 0.54, respectively for BTC, ETH, ETC, ZEC and LTC. Backtesting the test set of 719 days in each coin with the same starting capital and a realistic fee of 0.1% returned 828%, 171%, 169%, 55.9% and -23% when buy and hold would have returned -27%, -86%, 127%, -50% and 37.56% each for ETH, ZEC, BTC, ETC and LTC. These strategies might be promising for potential trading or law enforcement use. We explain the features used, provide examples of the trading logic, and make our hypothesis.

1 Introduction

(Griffin and Shams 2019; Mirtaheri et al. 2019; Xu and Livshits 2019; Gandal et al. 2018; Feder et al. 2018; Li, Shin, and Wang 2019) have previously detected various forms of price manipulations within the cryptocurrency ecosystem. Even the CFTC has warned about cryptocurrency pump and dump schemes (CFTC 2018). In Section 3.1, we detail and provide examples of various types of manipulations in the cryptocurrency ecosystem. Creating an approach that detects all of them is not possible. We create a hypothesis to generalize some price manipulations and attempt to detect them. We hypothesis the following:

- Before a price altering news, Insiders, sometimes trade cryptocurrencies.
- Sometimes, the price goes up due to manipulations without any other significant event. People who manipulate the price, buy it before the pump.
- There are indicators, which can detect these events.

From here, we refer to the events we described in points 1 and 2 as Insider Trading. This is different than Insider Trading in the traditional sense as our definition also includes buying with an attempt to manipulate the price in the future as Insider Trading.

We determine our features based on previous studies and experiments. If our selected feature performs well in an unseen test set in a statistically significant way, price manipulation is likely.

In our manipulation detection algorithm, we created a condition to select eventful days based on news. We use Reddit to track news. Reddit is a discussion forum popular among cryptocurrency enthusiasts. Users submit text and link posts to appropriate subreddits, and other users upvote/downvote it and add comments. Most of the discussion takes place in the comments section. The subreddits for Bitcoin, Ethereum, Litecoin, ZCash, and Ethereum Classic have 1.18M, 0.45M, 0.21M, 15k, and 25k subscribers as of November 10, 2019. The popularity of these subreddits between October 2017 and November 2019, is summarized in Table 1.

Symbol	Avg Posts per Day	Avg Votes per Post	Avg News per Day	Avg Votes per News Post
BTC	238.14	30.45	29.01	21.97
ETH	65.1	14.02	12.73	17.23
LTC	16.52	17.6	2.06	20.25
ETC	7.99	3.69	1.36	5.05
ZEC	3.77	6.68	0.6	8.72

Table 1: Reddit Statistics

Reddit has an upvote and a downvote system. Posts with high votes go to the top, and more people see it. Among the submissions, we filtered out news by selecting 134 domains. The domains were chosen by manually looking at the most used domains in cryptocurrency-based subreddits. We selected traditional news sites, cryptocurrency-based news sites, and the official sites for most cryptocurrencies as news sites. The selected news domains is included in Appendix 9.1. Table 1 shows that these communities are popular enough for our purpose.

Bitcoin and other currencies can be traded in margin with leverage. Bitfinex provides leverage of 3.33 (Bitfinex). These leveraged positions are referred to as longs and shorts. A long position profits when the price goes up, and short

does when the price falls. Bitfinex provides the size of historical long and short positions opened during that time. We use that to calculate most of our features.

2 Contributions of this study

- We used news, chart and margin-based features to detect potential Insider Trading in a statistically signficant way.
- We created a trading algorithm based on this, which showed a huge return.
- We determine indicators from the Margin market, which can predict pumps.

3 Literature Review

We divide this section into two different subsections. First, we cite literature that explains the different types of manipulation that have been observed in cryptocurrencies. Then we explain approaches made at detecting them.

3.1 Manipulations

Since the beginning of financial markets, different types of manipulations have been observed (Markham 2015). (Lin 2016) documented different strategies that have been used to manipulate the financial markets. The strategies include:

- **Cornering and Sqeezing:** In this mechanism, one party obtains a significant portion of a financial commodity and uses that to dictate the price (Markham 2015; Lin 2016). (Lin 2016) writes that this mechanism is less common in liquid markets. In the past, single actors may have strategically made trades to cause massive price changes, cornering, and squeezing the Bitcoin market (Griffin and Shams 2019).
- **Front Running:** In this type of manipulation, a broker performs trade before a market runner after knowing their intention (Hazen 1985; Markham 2015; Lin 2016). To our knowledge, this type of manipulation has not been publicly documented in cryptocurrencies. However, exchanges have had issues. There have been "hacks" which may not be a hack, exchanges have created trades without real money (Gandal et al. 2018) and may still be doing so (Griffin and Shams 2019)
- **Wash Trading:** Wash Trading is a form of manipulation where sham orders are made to create artificial volume and price (Teall 2018). Many cryptocurrency exchanges have meagre fees when trading at high volume. OkEx, a cryptocurrency exchange, has confirmed having a Wash Trading problem (Baker 2019)
- **Spoofing:** Spoofing takes place when order is created without the intention of completing it. We have observed it many times in different cryptocurrencies. Buy and Sell orders bigger than the total weekly volume have been documented. They limit price growth/fall. (unsafecoin 2016; Bacobob 2017)
- **Misinformation:** Fake news and social bots are distorting elections (Bessi and Ferrara 2016). They have been used to manipulate stocks (Renault 2017). (Mirtaheri et al. 2019) found an increase in bot activity in Social Media during price rises too. (Zannettou et al. 2019) found that state-sponsored Russian trolls discussed cryptocurrency in Reddit.
- **Insider Trading:** Company Insiders can have access to more information about the company than an outsider. If they trade based on private information, they will have an advantage and are more likely to make profits/avoid loss. Insider Trading is illegal in most commodities. On July 20, 2016, Ethereum was trading at 11.68$. On July 21, Coinbase, a US-based cryptocurrency exchange, added Ethereum. By July 22, it was trading at 15$. On May 3, 2017, Coinbase added Litecoin. On May 2, Litecoin traded at 15$. By May 4, Litecoin was trading at 25$. Insider trading was widely suspected during these instances. On December 19, 2017, Bitcoin Cash was trading at 2000$. On December 20, 2017, Coinbase added Bitcoin Cash. On the same day, Bitcoin Cash traded at 4000$. Coinbase is facing a class-action lawsuit over Insider Trading in this case. (Zhao 2018)
- **Pump and Dump:** During pump and dumps, a party acquires stock at cheap, artificially pump up the price, and then sell it. (Mirtaheri et al. 2019; Xu and Livshits 2019) examined Telegram Pump and Dump groups. Telegram Pump and Dump groups contain many members. Some groups have a paid structure. A typical Telegram P&D group works like this:
  - The group admin buys a coin at cheap.
  - The admin mentions the time he is going to pump a coin.
  - At that time, the admin posts the name of the currency. In paid groups, paid members see it first. As many people buy, the price goes up. After a while, other members of the group get to see the message too. More people buy.
  - Some people spread misinformation too. The success of the pump depends on the number of people that buy the coin.

  (Xu and Livshits 2019) used ML techniques from features derived from the Telegram data to detect successful and unsuccessful attempts to manipulate large scale and medium scale cryptocurrencies.

It is possible to perform many of these methods in combination. For example, Cornering and Squeezing can be done to dump the price below a certain level. After accumulating an asset at cheap, Wash Trading can be done to create artificial volume. Spoofing orders can be made to create artifical demand. Then, misinformation can be spread. If all of these succeeds, the price goes up, and the pump and dump attempt completes.

3.2 Detection

From October 2013 to December 2013, the price of Bitcoin increased from 139$ to over 1000$. This increase played a massive role in making Bitcoin more mainstream. During this time, Mt Gox was the leading Bitcoin exchange. Since

then, Mt Gox has filed for Bankruptcy. After the bankruptcy filing, (Gandal et al. 2018) studied suspicious trades that took place in Mt Gox during a ten-month period. (Gandal et al. 2018) detected "Willy" and "Markus" bots that operated in Mt Gox during this time. They found that price increased during 80% of the day these bots operated. Price increased by 55% on average when they did not operate. They conclude that these bots probably did not pay real money to perform the trades and were operated by the exchange itself.

From May 2017 to December 2017, the price of Bitcoin increased from 1300$ to over 20000$, creating a cryptocurrency market cap of over 500 Billion $. Although, there was a huge increase in people joining the exchanges, and Google Trends indicating an organic growth pattern, some form of manipulation was suspected. (Griffin and Shams 2019) conclude that a single actor likely drove the price by using a popular form of exchanging mechanism - USDT. They analyzed Bitcoin blockchain, USDT blockchain, and market data and found that less than 1% of hours with heavy tether transaction was associated with 50% of the meteoric rise in Bitcoin. They show that this is not possible at random. Although manipulation alone is not sufficient to create such growth, manipulation played a huge role in the timing of the rise.

In espionage, it is commonly known that strategically altering the key events can cause a huge change in the course of a nation when other factors are ripe for change. (Krafft, Della Penna, and Pentland 2018) performed small scale random trades in cryptocurrencies. They found that small buy actions caused a substantial increase in buy size activity that was hundereds of times the size of the interventions. If it scales up at this rate, given the right circumstance, it explains how manipulations can create a big difference.

Studies have also found small actors manipulating the price of smaller coins. Manipulating small market cap seems easier and more common. (Xu and Livshits 2019) used data from Telegram channels to train a Neural Network that predicted a pump before it happens. By performing hypothetical trade based on market features, they create a model with 0.96 AUC that can return over 60% in 3 months. Before them, (Li, Shin, and Wang 2019) had also conducted a study using data from Telegram pump and dump groups. Recently (Mirtaheri et al. 2019) created a model to detect the success of these pump attempts. They also study Twitter and detect a huge increase in the activity of Twitter Bots during manipulations.

There have been more studies of Insider Trading outside cryptocurrencies. (Donoho 2004) used news and options based features to detect Insider Trading in stocks. (Donoho 2004) created his features from call options and news, and detected Insider Trading.

4 Approach

4.1 Data Collection

First, we selected exchange to collect the data from. Bitfinex was chosen because it allows Margin Trading and has a strong connection with USDT, which has associated with some form of price manipulations. Although Bitfinex is not the biggest exchange and conducts only a fraction of total trade, we assumed that the OHLC data it provided was representative. Five currencies were selected for the sake of convinence. We selected the first five currencies that could be margin traded in Bitfinex.

We used the Bitfinex API to obtain the historical size of Long and Short positions. Then we obtained historic OHLC (Open, High, Low, Close) for the same dates using Bitfinex API.

There are many news sites and scraping data individually from all of them would be time-consuming. Additionally, it would be hard to separate the important from clickbait. So, we used Reddit as Reddit is popular and has a crowd filter mechanism. Reddit API does not provide historic subreddit submission data through an API. So, we used the Pushshift API to download historic submissions in the following subreddits:

- /r/bitcoin for Bitcoin
- /r/ethereum for Ethereum
- /r/zec for ZCash
- /r/litecoin for Litecoin
- /r/ethereumclassic for Ethereum Classic

4.2 Feature Calculation

After getting the required data, we calculated 59 features using the change in long and short positions and news.

First, we filtered out link submission. We counted the number of links in a 24 hour range (starting at midnight) and filtered out news links. Then we used VADER (Valence Aware Dictionary and sEntiment Reasoner) to calculate sentiment on the headlines instead of opening the link and scraping the news post. VADER is a rule-based sentiment classification designed for use in social media data. It returns a number between -1 and 1 from most negative to most positive (Hutto and Gilbert 2014). VADER was chosen instead of SentiStrength, another popular sentiment analysis mechanism, because on 100 random manual examinations, it performed better than SentiStrength.

Then we converted all the Reddit and Bitfinex data to a daily range and calculated features detailed in Table 2, 3, 4.

(Donoho 2004) first used these features for Insider Trading detection in stocks using news and options data. Instead of options data, we used long-short data as the options market for cryptocurrency is relatively small. Features total_news, three_day_count, thirty_day_count, one_thirty_ratio and three_thirty ratio in Table 2 were obtained from (Donoho 2004). The other features in the table were manually tested and calculated. Most of the features on Table 3 were obtained from (Donoho 2004) too.

(Xu and Livshits 2019) had used historic chart based features in their data when creating a model to detect pumps from telegram data. Features 11 to 16 on table 4 were inspired by it. (Xu and Livshits 2019) had used hourly data. However, we obtained higher accuracy with daily data. Features 7 to 10 were inspired by (Donoho 2004)

S.N.	Feature	Range	Meaning
1.	total_news	1D (24H)	Total Number of news posted in Reddit between 12 AM to 12 AM
2.	total_sentiment	1D	Sum of the product of Reddit score and VADER Sentiment
3.	news_dominance	1D	Total Score of news links divided by total score of non news links
4.	normalized_ mean_sentiment	1D	Mean of the score times VADER Sentiment
5.	mean_sentiment	1D	Mean of VADER Sentiment
6.	three_day_count		Number of News posted in the last 3 days
7.	thirty_day_count		Number of News posted in the last 30 days
8.	one_thirty_ratio		Ratio of the number of news in the last day to the number in the last thirty days
9.	three_thirty_ratio		Ratio of the number of news in the last 3 days to the number in the last thirty days
10.	total_ sentiment_ mean	20D	Mean of the daily sentiment for the given time period

Table 2: Reddit Based Features

S.N.	Feature	Range	Meaning
1.	long	1D	Volume of total Longs opened during the end of day
2.	short	1D	Volume of total Shorts opened during the end of day
3.	longshort_ volume	1D	Sum of Total Longs and Total Shorts
4.	longshort_ratio	1D	Total longs divided by total shorts.
5.	mean_volume	20D: Longs and Shorts	Mean daily volume during 20D for long and short
6.	volume_std	20D: Longs and Shorts	Standard Deviation of the daily long short volume during the 20D interval
7.	times_above	20D: Longs and Shorts	How many times above the 20D average the volume was
8.	moving_z_score	20D: Longs and Shorts	How many SD above 20D average the volume was
9.	change	1D: Longs and Shorts	Percentage change in volume compared to the previous day
10.	above_average	Daily: Longs and Shorts	Number of times in the last 5 days that volume exceeded the 20 day Moving Average
11.	change_avg	Daily: Longs and Shorts	Average change (9) in volume
12.	today_avg_ comparision	Daily: Longs and Shorts	Current Percentage change divided by the average change over the last 20 intervals
13.	long_short	Daily	Long Volume divided by the average Short Volume of the 3 biggest days during the last 20 days
14.	short_long	Daily	Short Volume divided by the average Long Volume of the 3 biggest days during the last 20 days

Table 3: Long Short Based Features

S.N.	Feature	Range	Meaning
1.	Open	1D	Price of First Trade during the range
2.	High	1D	Price of biggest Trade during the day
3.	Low	1D	Price of lowest Trade during the day
4.	Close	1D	Price of Last Trade during theday
5.	Volume	1D	Volume traded in USDT
6.	Volume_coin	1D	Volume traded in currency
7.	Volatility	5D	Average of the close values during the last 5 days
8.	volatility_change	5D	Current volatility divided by the 5 day average
9.	volume_change	Daily	Current Volume divided by the 20 day average
10.	two_day_change	Daily	Change in price during the last 2 days
11.	return	1D,2D,3D	Log return of the change in close price during the interval
12.	volume_from	1D,2D,3D	Sum of the volume in USDT in the given interval
13.	volume_to	1D,2D,3D	Sum of the volume in coin in the given interval
14.	return_volatility	1D,2D,3D	Std of the calculated log return during the given timeframe
15.	volume_from_ volatility	1D,2D,3D	Std of the volume in USDT during the given timeframe
16.	volume_to_ volatility	1D,2D,3D	Std of the volume in coin during the given timeframe

Table 4: Price Based Features

4.3 Creating Classification

Random Forest, a Supervised Machine Learning technique, is used to create the ML model. It requires supervised classification of data before training. We name this classification column y. From manual inspection, we created a logic to determine huge rises. To do that, we initially set y to 0. If the predetermined conditions were met, the value of that day and the 14 days before it is set to 1. We number 14 days before a big rise as 1 too because we had hypothesized that buying takes place slowly before a huge rise, (Donoho 2004) had success using it, and we wanted to prevent a heavily imbalanced dataset. The predetermined conditions were:

- The product of total votes and the sentiment of the news post is greater than 10, or the number of news posted that day is greater than or equal to 5 percent of the total posts in the last 30 days. This was done because (Donoho 2004) had used this in a similar system, it was congurent with our logic and, later we found out that using this logic had increased the accuracy, AUC, and profit in our model.
- Tomorrow's price is at least 5 percent greater than today.
- Price 5 days later is at least 5 percent greater than today.
- The percentage change in price over the last 15 days is positive.

This logic classified the major pumps in all five currency fairly well. A figure representing the classified points is included in Appendix 9.2

4.4 Machine Learning Model

In supervised Machine Learning, we divide data into training and test set. The training set is used to train the algorithm. The test set is used to test how it works. There are many approaches to create a training and a test set. We had three options:

- **Train Test Split:**Select the first n% of the data as the training set and use the remaining data as the test set.
- **K-Fold Cross Validation:** Divide the data into k random folds. Train different models with a different test set in each model.
- **Walk Forward Validation:** Start with n days as the training set and y days as the test set. Iterate till the end by increasing the size of the training set by y until the end is reached.

We chose the walk-forward validation as it was robust and provided us with a huge test set that could be tested in different market periods. We started with the first 195 days as training set and 80 days after it, after a buffer of 5 to avoid bias, as the test set. Then, we increased the size of the training set by 80 days i.e., 280 days with the next 80 days as a test set. We repeated this till we reached the end. This way, we had access to a long test set. In each step, we standardized the training and the test set by subtracting it with the mean of the training set and dividing by the standard deviation of the training set. The training set's mean was used to prevent any leakage of future data in the test set.

We trained a different random forest classifier using 100 estimators for each coin.

4.5 Backtesting

After creating the models, the model's prediction on the test set was used to perform backtesting with a realistic 0.1% fee. We started with a capital of 10000$ for each coin. The trading logic for each coin was:

If no position is currently open in that coin:

- Open a position with 95% of the allocated capital for that coin if the prediction is 1.
- Do not do anything if the prediction is 0

If a position is currently open, we first set n to 4. Then:

- Hold for n days and sell all open position if prediction changes to 0 from 1
- Set n to 4 days if the prediction is 1

5 Explanations and Result

The obtained result is summarized in Table 5.

Symbol	AUC	Accuracy	Hold Return (%)	Algorithm Return (%)
ETH	0.72	0.82	-27.38	828.06
BTC	0.57	0.63	127.55	169.17
ZEC	0.58	0.86	-86.75	171.31
ETC	0.58	0.87	-50.86	55.9
LTC	0.44	0.81	37.56	-23.18

Table 5: Result Metrics

Area Under the Curve (AUC) and the Accuracy of the predictions in the test set is provided in Table 5. AUC is considered a better metric in the ML literature than accuracy in an unbalanced dataset like this. High accuracy can be obtained by predicting in the same direction. Our task is price prediction. So although our AUC is not extremely high, it is fairly high for this task.

Symbol	Total Positions	Total Profit	Total Lost	Avg Profit	Avg Loss
ETH	20	14	6	8279.57	-5517.95
ZEC	8	8	0	2141.36	0
BTC	28	17	11	1695.1	-1081.76
ETC	6	4	2	1471.83	-148.7
LTC	14	6	8	1111.42	-1123.34

Table 6: Position Details

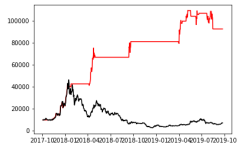
Table 6 shows that the algorithm opened few positions. Nevertheless, the return is enormous. A successful trade returned at least 1000$. The number of successful trades was higher than losing one in all currencies except Litecoin. The average profit during a success was also bigger than the average loss in all of these currencies, excluding Litecoin.

Symbol	one actual	one predicted	zero actual	zero predicted
ETH	178	144	542	576
ZEC	90	44	590	636
BTC	222	229	498	491
ETC	97	31	623	689
LTC	99	68	621	652

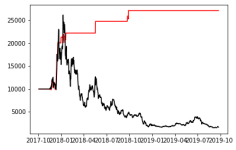
Table 7: Output comparision

The number of predicted and actual ones and zeros is included in Table 7. The number is quite similar.

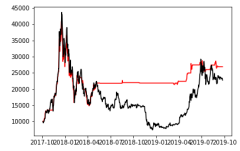
The backtest diagram is included in Appendix 9.4. When a position is opened, there is a green mark. When it is closed, there is a red mark. Our model predicted most of the huge price increases in Ethereum, ZCash, Ethereum Classic and Bitcoin. It did not perform as well for Litecoin.



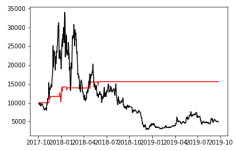
(a) ETH



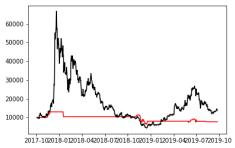
(b) ZEC



(c) BTC



(d) ETC



(e) LTC

Figure 1: Our Algorithm VS holding

We started with 10,000$ in each coin. The movement of the portfolio for each coin is in Figure 1. The black line represents and holding portfolio while the red line shows our algorithm. We can see in the Figures that the portfolio value increases most of the time when the value of the coin increases. However, it does not fall when the price falls. A trading model may be able to increase profit by predicting the falls and performing shorts.

6 Model Examination

ML models are frequently left black box. In this section, we try to understand why our algorithm brought and sold

the way it did. We analyzed the random forest classifier to find out the most significant features. The feature importance function in scikit learn was used for this purpose. The five most important features, contributing about 25% of each coin is in Table 8.

feature	ETH	BTC	ZEC	ETC	LTC
20D_short_ volume_mean	0.11	0.05	0.04	0.02	0.08
20D_long_ volume_mean	0.03	0.06	0.08	0.03	0.04
20D_long_ volume_std	0.04	0.05	0.04	0.03	0.08
short	0.03	0.05	0.07	0.01	0.04
long	0.03	0.04	0.04	0.05	0.03

Table 8: Important Features

All features and their importance in each coin is in Appendix 9.3. These tables show that features derived from long and short were most powerful. They were responsible for nearly 50% of the model in all coins. We analyze some trades our mode primarily made based on these feature this section.

Price change depends on many factors. In this study, we selected some and found that they could predict some price increases correctly. In this section, for intuition, we selected only two features. We only perform a surface examination and do not see how it correlates with other features.

6.1 Ethereum- November 2017 to March 2018

In Appendix 9.4, we can see that the price of ETH was 313$ on November 15, 2017. By January 14 it had gone up to 1257$. Our algorithm had performed a near-perfect trade by buy at 313$ and selling at 1257$.

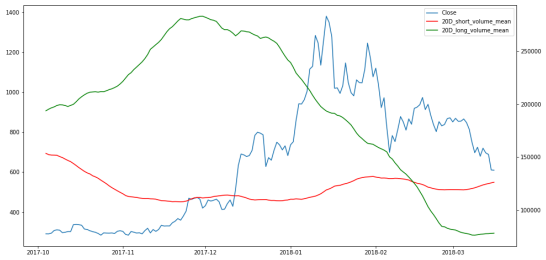


Figure 2: Ethereum around January 2018

Figure 2 shown above can be zoomed for detail. In the figure, the blue line represents the daily close price of Ethereum. Its axis is on the left. On the right, there is another axis for the 20-day short volume mean, and the 20-day long volume mean. The long volume is in green while the short volume is in red. In the figure, we can see the Long volume rises and short volume falls before the price increase in December 2017. Then the long volume remains up for some time. The long volume starts falling in January 2018, before the fall in price. The short volume also increases before the fall. As these are the most powerful feature, and most other features also derived from long and short, we can hypothesize that our algorithm was triggered by the changes in it. The decision turned out to be correct.

6.2 April 2018 - Ethereum

On April 10, our algorithm brought Ethereum at 400$. On April 26, it sold it at 702$.

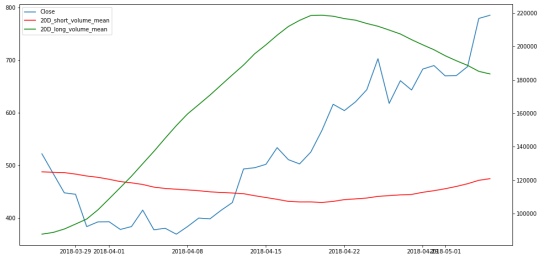


Figure 3: Ethereum around April 2018

We draw a similar figure and have a similar observation in Figure 3. The 20 day long rises. The 20D short average falls. Then the price increases. After April 20, the long average starts falling, and shorts start rising. The price then follows the long average and goes down.

6.3 January 2017 - ZCash

On December 13, 2017, our model brought ZCash at 317$. On December 20, it sold it at 600$.

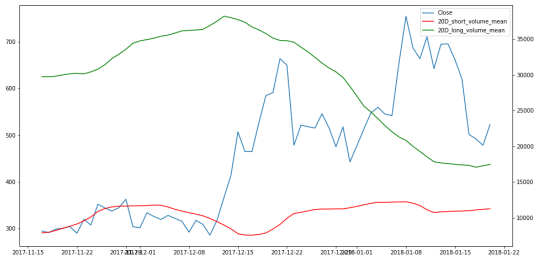


Figure 4: ZCash around January 2017

Figure 4 shows a similar case for ZCash. The 20 day long rises, then the price.

7 Conclusion

We started the study, assuming that Insider Trading takes place based on previous literature. We used features that are possibly indicative of Insider Trading and created an ML model. The model performed much better than holding in the test set.

To learn if our success can happen due to chance, we selected 59 random features between 1 and -1 from a uniform distribution and created a Random Forest classifier, based on the same split mechanism as we did for our algorithm. We ran the simulation for 3000 iterations. None of the 3000 models performed as good as our algorithm. We assumed a starting capital of 10000$ for each coin and added the final value for each to calculate the total return. Using this mechanism, our algorithm had returned 240%. In 3000 iterations, the best random model returned 100%. The average portfolio value at the end was 49034$, slightly less than the starting price. The standard deviation was 7150$. The sum of the 5 AUC our model predicted was 2.99. The best random model had 2.64.

This shows that this behaviour is extremely unlikely on random. Our features were responsible for the correct detection. Thus, Insider Trading probably takes place and can be detected using public data. Although Insider Trading cannot exactly be proved with 100% certainty by this mechanism, we show that it is likely the case.

8 Further Work

In our study, we did not use all possible forms of data available to us. We did not use data from the blockchain. In the future, clustering techniques can be used to cluster addresses and watch the flow of transactions. Features can be created from it.

We used Reddit data in this work. (Mirtaheri et al. 2019) showed that bots are used during pumps. We can use bot detection techniques on Reddit and on Twitter to correlate the bots with pump attempts. (Xu and Livshits 2019; Mirtaheri et al. 2019) had success with Telegram data. A comprehensive model can have a place for them too.

There may be a place for integration of order book data, too, in a sophisticated system. However, integrating all of them may be a very complicated task.

Finally, different groups may have caused different pumps. We may be able to use techniques from cyber forensics and unsupervised machine learning to find types of groups and find a signature. Research in this direction will be helpful for Law Enforcement too. *

References

Bacobob. 2017. Maidsafecoin (maid) - price & trading topic.

Baker, P. 2019. Okex comes clean on wash trading problem.

Bessi, A., and Ferrara, E. 2016. Social bots distort the 2016 us presidential election online discussion. *First Monday* 21(11-7).

Bitfinex. Intro to margin trading.

CFTC. 2018. Customer advisory: Beware virtual currency pump-and-dump schemes.

Donoho, S. 2004. Early detection of insider trading in option markets. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 420–429. ACM.

Feder, A.; Gandal, N.; Hamrick, J.; and Moore, T. 2018. The impact of ddos and other security shocks on bitcoin currency exchanges: Evidence from mt. gox. *Journal of Cybersecurity* 3(2):137–144.

Gandal, N.; Hamrick, J.; Moore, T.; and Oberman, T. 2018. Price manipulation in the bitcoin ecosystem. *Journal of Monetary Economics* 95:86–96.

Griffin, J. M., and Shams, A. 2019. Is bitcoin really untethered? *Available at SSRN 3195066*.

Hazen, T. L. 1985. *The law of securities regulation*. West Publishing Company.

Hutto, C. J., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Krafft, P. M.; Della Penna, N.; and Pentland, A. S. 2018. An experimental study of cryptocurrency market dynamics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 605. ACM.

Li, T.; Shin, D.; and Wang, B. 2019. Cryptocurrency pump-and-dump schemes. *Available at SSRN 3267041*.

Lin, T. C. 2016. The new market manipulation. *Emory LJ* 66:1253.

Markham, J. 2015. *Law enforcement and the history of financial market manipulation*. Routledge.

Mirtaheri, M.; Abu-El-Haija, S.; Morstatter, F.; Steeg, G. V.; and Galstyan, A. 2019. Identifying and analyzing cryptocurrency manipulations in social media. *arXiv preprint arXiv:1902.03110*.

Renault, T. 2017. Market manipulation and suspicious stock recommendations on social media.

Teall, J. L. 2018. *Financial trading and investing*. Academic Press.

unsafecoin. 2016. A theory about the 500 btc buy wall.

Xu, J., and Livshits, B. 2019. The anatomy of a cryptocurrency pump-and-dump scheme. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 1609–1625.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019. Who let the trolls out?: Towards understanding state-sponsored trolls. In *Proceedings of the 10th ACM Conference on Web Science*, 353–362. ACM.

Zhao, W. 2018. Coinbase hit by lawsuit over alleged insider trading.

9 Appendix

9.1 News Sites

coindesk.com	marketwatch.com	0xproject.com	allcryptocurrencies.news	tezos.foundation
medium.com	theguardian.com	discord.gg	crypto-lines.com	tezosfoundation.ch
cointelegraph.com	coinbase.com	coinwhalenews.com	iotahispano.com	
bitcoin.com	wired.com	cryptotown.io	iota-news.com	
cryptocoinsnews.com	bbc.co.uk	investinblockchain.com	ethereumworldnews.com	
newsbtc.com	cryptobrowser.site	publish0x.com	oracletimes.com	
bloomberg.com	vice.com	cointopper.com	litecoin.com	
cnbc.com	insidebitcoins.com	etherscan.io	monerobase.com	
bitcoinmagazine.com	ft.com	coingeek.com	moneroblocks.info	
forbes.com	themerkle.com	craigwright.net	ripple.com	
bitcoinist.com	fortune.com	coinblockdesk.com	ripplenews.tech	
bitcoinist.net	tumblr.com	coingecko.com	omisego.network	
github.com	coinspeaker.com	yours.org	omise.co	
zerohedge.com	coinjournal.net	cryptodaily.co.uk	coinspot.com.au	
ambcrypto.com	beincrypto.com	coinidol.com	coincodex.com	
businessinsider.com	seekingalpha.com	businessdigit.com	smartlands.io	
bitcoinfeeds.com	arstechnica.com	cryptolinenews.com	todaysgazette.com	
steemit.com	rt.com	blockchain.info	samcrypto.com	
wsj.com	wikipedia.org	coin.dance	linkedin.com	
newsforyou.today	theverge.com	linuxfoundation.org	discussions.app	
bitcoinvox.com	btcfeed.net	dashforcenews.com	theeoswriter.io	
nytimes.com	tradingview.com	dashnews.org	eoswriter.io	
yahoo.com	bbc.com	dash.org	theaccountingblockchain.io	
reuters.com	ibtimes.co.uk	dashcentral.org	eosauthority.com	
coinfox.info	bitguru.co.uk	dashpaymagazine.com	neonbeginner.com	
toshitimes.com	the-blockchain-journal.com	cryptobriefing.com	neonewstoday.com	
bravenewcoin.com	qz.com	financemagnates.com	neo.org	
techcrunch.com	washingtonpost.com	thedashtimes.com	z.cash	
ccn.com	hackernoon.com	ethereum.org	cash.foundation	
cnn.com	telegraph.co.uk	google.com	xrpnewsonline.com	
cntldr.com	thenextweb.com	stackexchange.com	thecoinshark.net	
btcmanager.com	nasdaq.com	thecoinrepublic.com	koinalert.com	
8btc.com	binaryoptionevolution.com	iota.org	santiment.net	

Table 9: News Sites

9.2 Classification y

(a) Bitcoin

(b) Ethereum

(c) Litecoin

(d) ZCash

(e) Ethereum Classic

9.3 Feature Strength

feature	ETH	BTC	ZEC	ETC	LTC
20D_short_volume_mean	0.11	0.047	0.035	0.025	0.081
last_price	0.041	0.032	0.018	0.037	0.032
High	0.04	0.032	0.015	0.026	0.037
20D_long_volume_std	0.039	0.049	0.044	0.034	0.084
Open	0.039	0.023	0.016	0.024	0.022
longshort_ratio	0.038	0.028	0.032	0.026	0.03
Low	0.036	0.024	0.015	0.038	0.03
long_20D_short	0.035	0.021	0.041	0.035	0.038
thirty_day_count	0.034	0.023	0.031	0.05	0.019
20D_long_volume_mean	0.034	0.058	0.075	0.031	0.041
long	0.034	0.036	0.035	0.048	0.03
short_20D_long	0.032	0.052	0.033	0.026	0.033
20D_short_volume_std	0.031	0.021	0.022	0.034	0.03
longshort_volume	0.03	0.027	0.029	0.032	0.019
short	0.03	0.047	0.069	0.01	0.042
20D_total_sentiment_mean	0.03	0.022	0.032	0.059	0.032
Close	0.028	0.025	0.013	0.025	0.035
20D_long_change_avg	0.023	0.02	0.023	0.034	0.021
20D_short_times_above	0.019	0.014	0.014	0.022	0.015
volume_from_1_day	0.016	0.009	0.018	0.009	0.009
Volume	0.015	0.008	0.017	0.01	0.009
20D_long_times_above	0.015	0.024	0.037	0.027	0.015
volume_from_3_day	0.015	0.016	0.032	0.007	0.025
20D_short_change_avg	0.015	0.019	0.024	0.053	0.039
volume_from_2_day	0.014	0.014	0.016	0.007	0.017
20D_short_above_avg	0.013	0.006	0.016	0.005	0.004
5d_volatility	0.011	0.012	0.015	0.009	0.017
20D_long_moving_z_score	0.011	0.013	0.022	0.012	0.006
volume_to_3_day	0.011	0.021	0.025	0.009	0.021
5d_20d_volatility_change	0.01	0.015	0.008	0.021	0.005
two_day_change	0.01	0.017	0.005	0.012	0.006
20D_short_moving_z_score	0.01	0.013	0.015	0.025	0.012
vol_from_vol_3_day	0.01	0.007	0.012	0.012	0.008
volume_to_2_day	0.009	0.014	0.02	0.005	0.014
20D_short_today_avg_comparision	0.007	0.01	0.006	0.007	0.004
20D_long_above_avg	0.007	0.007	0.007	0.003	0.003
vol_from_vol_2_day	0.007	0.005	0.006	0.005	0.006
volume_change	0.006	0.012	0.006	0.014	0.005
return_vol_3_day	0.006	0.01	0.011	0.015	0.01
three_day_count	0.005	0.012	0.007	0.005	0.004
return_3_day	0.005	0.007	0.006	0.009	0.005
one_thirty_ratio	0.005	0.007	0.003	0.003	0.004
vol_to_vol_3_day	0.005	0.009	0.008	0.011	0.006
normalized_mean_sentiment	0.005	0.006	0.001	0.004	0.003
return_1_day	0.005	0.006	0.004	0.009	0.004
20D_short_change	0.004	0.009	0.004	0.008	0.005
total_news	0.004	0.007	0.001	0.003	0.002
return_2_day	0.004	0.006	0.004	0.008	0.003
volume_to_1_day	0.004	0.009	0.007	0.007	0.009
return_vol_2_day	0.004	0.007	0.007	0.006	0.004
total_sentiment	0.004	0.006	0.002	0.006	0.003
mean_sentiment	0.004	0.006	0.001	0.004	0.003
vol_in_coin	0.004	0.01	0.012	0.004	0.008
vol_to_vol_2_day	0.004	0.007	0.005	0.005	0.007
20D_long_today_avg_comparision	0.003	0.009	0.005	0.01	0.01
20D_long_change	0.003	0.008	0.007	0.006	0.004
three_thirty_ratio	0.003	0.011	0.006	0.009	0.01
news_dominance	0.003	0.008	0.0	0.003	0.003

9.4 Backtest Chart

(a) Bitcoin

(b) Ethereum

(c) ZCash

(d) Litecoin

(e) Ethereum Classic