

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN FINANCIAL SERVICE-FRAUD DETECTION IN CRYPTOCURRENCY EXCHANGES REPORTED VOLUME

Yusi Li

08/2019

1 1. Problem Background

Bitcoin is a global fungible commodity with low transaction cost, near-zero transportation costs and low-to zero storage cost. As a result, we would expect the bitcoin market to be uniquely orderly and efficient, with tight spreads and nearly perfect arbitrage. Unfortunately, Public perception and data suggest exactly the opposite. Public perception holds that the bitcoin market is in fact uniquely disorderly and inefficient. This is a rational response to the information most people have at their disposal. For example, leading data aggregators show prices on different exchanges separated by hundreds of dollars. The real market for bitcoin is significantly smaller, more orderly, and more regulated than is commonly understood. Report volume adds to roughly \$6 billion/day, but under the hood the exchanges that report the highest volumes are unrecognizable. The vast majority of this reported volume is fake and or non-economic wash trading.

The data comes from CoinMarketCap.com, the most widely cited source for bitcoin volume. It is used by every major media outlet in the world. Despite its widespread use, the CoinMarketCap.com data is wrong. It includes a large amount of fake and non-economic trading volume, thereby giving a fundamentally mistaken impression of the true size and nature of the bitcoin market. And approximately 95% of this volume is fake or non-economic in nature.

According to Bitwise(<https://www.bitcointradevolume.com>), there are only ten real exchange as list in figure 1 and figure 2.

95% of the reported bitcoin volume is fake. Below is the real volume. It's smaller, but more orderly and regulated than is widely understood.

[Read The Full Analysis →](#)

LATEST VOLUME DATA

\$1,636,624,104

Bitcoin Spot 24 hr Volume
Updated hourly. Data as of: 08/14/19 10pm ET.

Figure 1: 10 real exchange of bitcoin

The real exchanges for Bitcoin have some similar characteristics compared with fake exchanges. For example, the bid and ask are random distribution in fake exchanges, while the mix of buying











 Spot Exchange · Malta	\$625,270,854 24h Volume
 Spot Exchange · United States	\$302,142,633 24h Volume
 Spot Exchange · United States	\$229,493,226 24h Volume
 Spot Exchange · Taiwan	\$163,960,685 24h Volume
 Spot Exchange · United Kingdom	\$143,845,757 24h Volume
 Spot Exchange · Japan	\$75,115,574 24h Volume
 Spot Exchange · United States	\$39,579,743 24h Volume
 Spot Exchange · United States	\$21,450,293 24h Volume
 Spot Exchange · United States	\$18,406,864 24h Volume
 Spot Exchange · United States	\$17,358,477 24h Volume

Figure 2: 10 real exchange of bitcoin

and selling activities are unequal and streaky. The spread in the real exchanges is much more smaller than that in the fake exchanges. And so on.

Machine learning and artificial intelligence are being rapidly adopted for a range of applications in the financial services industry, including using by market regulators for surveillance and fraud detection. The US Securities and Exchange Commission staff leverages "big data" to develop text analytics and machine learning algorithms to detect possible fraud and misconduct. For instance, the SEC staff uses machine learning to identify patterns in the text of SEC filings. With supervised learning, these patterns can be compared to past examination outcomes to find risks in investment manager filings.

In the rest of the report, I try to use the machine learning method to find some index or method to detect whether a Bitcoin exchange is real or fake in the volume.

2 2. Solution 1

I get the data from Blockchain Transparency Institute(<https://www.bti.live>), including the orderbook, quote and trade daily data in several real and fake exchanges during June 2019. The raw data is like figure 3. And then I clean the data and calculate several variables as in figure 4. The variables include whether the platform is fraud or not(yes is 1, no is 2), the average spread during one day, the standard deviation of spread during one day, the average ask size during one day, the average bid size during one day, the average trading size during one day, the standard deviation of trading size during one day, the minimum and maximum size during one day.

	price	size	symbol_id	taker_side	time_coinap	time_exchar	uuid
0	8216.44	3.00E-06	BINANCE_SF	SELL	2019-06-14	00:00.1	18746e56-95d8-4acf-90ac-40f5ab37411a
1	8216.43	0.014235	BINANCE_SF	SELL	2019-06-14	00:00.1	60f3c166-ac14-47b4-bd32-1e82d053fb04
2	8216.43	0.016146	BINANCE_SF	SELL	2019-06-14	00:00.2	56c65d1d-9894-4248-af9a-42b2c2004bd3
3	8216.43	0.001907	BINANCE_SF	SELL	2019-06-14	00:00.3	f9419817-5bc3-4615-b4df-205703853053
4	8216.43	0.022562	BINANCE_SF	SELL	2019-06-14	00:00.4	9d7ddc43-3026-46d6-9232-38a2448b0944
5	8217.12	0.065926	BINANCE_SF	SELL	2019-06-14	00:00.6	d9d1c3ee-0f72-4ac5-8060-e5bef5854f43
6	8217.14	0.022051	BINANCE_SF	SELL	2019-06-14	00:00.9	75de882e-b738-4f5e-857b-044746956df1
7	8217.56	0.147975	BINANCE_SF	BUY	2019-06-14	00:01.4	7e091469-cb3d-4fc5-98ca-7ab7ac13abab
8	8217.14	0.000216	BINANCE_SF	SELL	2019-06-14	00:01.8	463755aa-a564-49ce-8b8a-b3d66144b084
9	8217.12	4.10E-05	BINANCE_SF	SELL	2019-06-14	00:01.8	0ea2ecac-9fc5-42f3-b464-39915d125130
10	8216.43	0.030181	BINANCE_SF	SELL	2019-06-14	00:01.8	ed6dab9c-3dd9-4445-bcc0-a4ac2a8327a3
11	8216.43	0.005519	BINANCE_SF	SELL	2019-06-14	00:01.9	9293163b-27cf-412e-a767-89f9ebdf8c2d
12	8216.43	0.009452	BINANCE_SF	SELL	2019-06-14	00:01.9	065dbaaa-9fae-48bd-8a2a-95453dcd8d3f
13	8215.18	0.103491	BINANCE_SF	SELL	2019-06-14	00:02.1	1c788aa6-886b-499b-af42-d6c7e5b5d60a
14	8215.05	0.187	BINANCE_SF	SELL	2019-06-14	00:02.1	b586cc20-1759-4507-8fa4-60b35a8a622f
15	8215	0.088955	BINANCE_SF	SELL	2019-06-14	00:02.1	9041e5e2-3987-4ea3-ab81-7c4b72853b30
16	8215	0.012172	BINANCE_SF	SELL	2019-06-14	00:02.7	84b20259-4a4d-4538-913e-4c986ce1785a
17	8216.44	0.013952	BINANCE_SF	BUY	2019-06-14	00:02.7	b7eec559-5e51-4644-90b6-5a23f222476a
18	8215	0.048873	BINANCE_SF	SELL	2019-06-14	00:03.1	3b4223fa-76a9-4e79-bd11-fb9a59dcf821
19	8215	0.011373	BINANCE_SF	SELL	2019-06-14	00:03.1	c235694c-978b-4678-8412-79f8f0aab0a7
20	8216.44	0.022246	BINANCE_SF	BUY	2019-06-14	00:03.3	e469b70b-9630-42b9-86df-2bf502a7e7dd
21	8218.19	0.116304	BINANCE_SF	BUY	2019-06-14	00:03.3	e3e0de21-df1e-4467-a86d-987fb35aeae7
22	8218.2	0.099849	BINANCE_SF	BUY	2019-06-14	00:03.3	bdae27f1-8e0c-4251-bdca-f2cae58778b0
23	8218.21	0.130577	BINANCE_SF	BUY	2019-06-14	00:03.3	f302228b-d8d3-4721-bfa7-0ba24c9caa95
24	8218.36	0.007765	BINANCE_SF	BUY	2019-06-14	00:03.3	14d23b24-f604-4d2d-b1ac-e8eaf9ec21bc
25	8218.38	0.022518	BINANCE_SF	BUY	2019-06-14	00:03.3	cf865546-a2e0-4770-af5e-0145dcf0eeaa

Figure 3: raw quote data

First I use the Linear Discriminant Analysis Method, I model the distribution of the predictor whether fake or not separately in each of the response classes, and then use Bayes' theorem to flip these around into estimate. When these distributions are assumed to be normal, it turns out that the model is very similar in form to logistic regression. The probability that the prediction is correct is around 95.45%. The coefficients of linear discriminants of the variables spread, std spread, ask size, bid size, size, std size, min size and max size are $1.793393e-01$, $1.301611e-01$, $-1.218621e-01$, $3.811368e-04$, $-7.178249e-01$, $6.360914e-01$, $4.744763e+03$, $-1.393099e-02$.

The second method is Tree-based Model. These involve stratifying or segmenting the predictor

platform	fraud	spread	std_spread	ask_size	bid_size	size	std_size	min_size	max_size
BBX	1	11.11135	19.28896	4517.037	4434.669	10180.09	5435.097	1	24910
BBX	1	15.98962	26.42525	5414.121	4801.848	10165.16	5604.921	1	25726
BINANCE	0	2.245064	1.404633	0.6528636	0.9545803	0.1100969	0.6224294	0	169
BINANCE	0	2.037066	1.170097	0.4296061	0.4834115	0.0985885	0.3715447	0	47.6107
BINANCE	0	1.87641	1.056414	0.3840389	0.3062354	0.09798546	0.4804092	0	56.32691
BINANCE	0	1.658106	1.082466	0.7354835	0.5439611	0.09493304	0.5273991	0	187.73096
BINANCE	0	1.483091	0.9783503	0.6905761	0.5383296	0.09815308	0.58849	0	138.5763
BINANCE	0	1.669762	1.028691	1.167021	0.9049837	0.1149154	0.7592252	0	189.09446
BINANCE	0	1.559781	0.9732691	0.4820439	0.4613288	0.08928409	0.3512463	0	47.6252
BINANCE	0	1.574475	1.030128	0.4939306	0.5343147	0.1028166	0.3976796	0	50.625
BINANCE	0	1.466574	0.9760306	0.6092738	0.9270577	0.09466304	0.3907431	0	49.96856
BINANCE	0	1.887665	1.348812	1.31361	0.5111044	0.1049886	0.5450671	0	70.51352
BITFOREX	1	6.66107	2.784476	0.09499465	0.06268447	0.8946646	1.232194	0.0001	8.9835
BITFOREX	1	7.153139	3.1546	0.08165707	0.1165245	0.8979372	1.348739	0.0001	8.9881
BITMART	1	9.964873	5.669485	1.195404	1.047522	17597.81	21407.43	0.26	159197.46
BITMART	1	11.60516	21.74203	1.192798	1.155098	14101.83	18568.36	0.08	148568.23
COINBENE	1	11.98274	4.906747	0.1105343	0.7166192	2.117978	2.068078	0.0001	10.7574
COINBENE	1	11.77869	5.043045	0.1308487	0.5323746	2.666291	1.681842	0.0001	12.7086
HUOBIPRO	1	1.345242	1.098397	0.7416319	8530.508	0.09357676	0.3745814	0	38.8472
HUOBIPRO	1	1.264915	1.172109	0.7816433	0.9499117	0.08344753	0.4102846	0	56.6074
OKEX	1	-6.979279	48.02706	0.4979486	0.4244039	0.3310836	1.404843	0	86.06085
OKEX	1	-181.2996	274.7711	0.5566752	0.4831103	0.3647935	1.531081	0	64.31649

Figure 4: calculate the variables using the raw data

space into a number of simple regions. In order to make a prediction for a given observation, we typically use the mean or the mode of the training observations in the region to which it belongs. The result of decision tree model is in figure 5. Random forests provide an improvement over bagged trees by way of a random forest small tweak that decorrelates the trees. In building a random forest, at each split in the tree, the algorithm is not even allowed to consider a majority of the available predictors. The importance of the predictors from high to low are in figure 6.

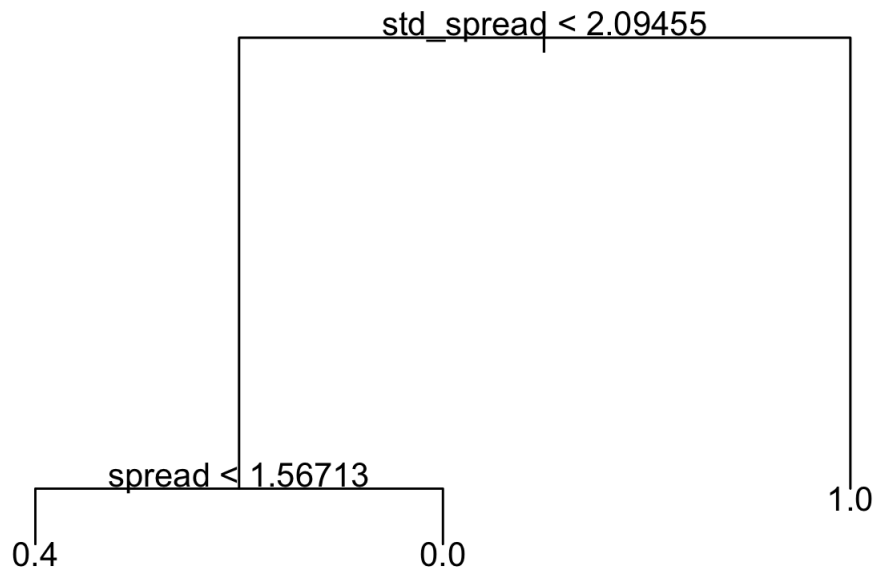


Figure 5: Result of Decision tree model

3 3. Conclusion 1

Artificial intelligence and machine learning can be used widely in financial service, especially in surveillance and fraud detection. In March 2019, the news of fraud trade volume in cryptocurrency exchange has made a big attention. According to the LDA Model and Decision Tree Model, the trading size, bid ask spread, the standard deviation of bid ask spread and trading size are the most important index to detect the fraud of Bitcoin exchange. The Bitwise Asset Management even called for the issuers to field applications for bitcoin or bitcoin future ETFs to avoid the fraud of volume and better detection. The Bitcoin ETF or the index created by the variables listed below in my machine learning method can intend to provide direct exposure to bitcoin, priced off the equivalent of a crypto consolidated tape, while custodianship assets at a regulated, insured, third-party custodian.

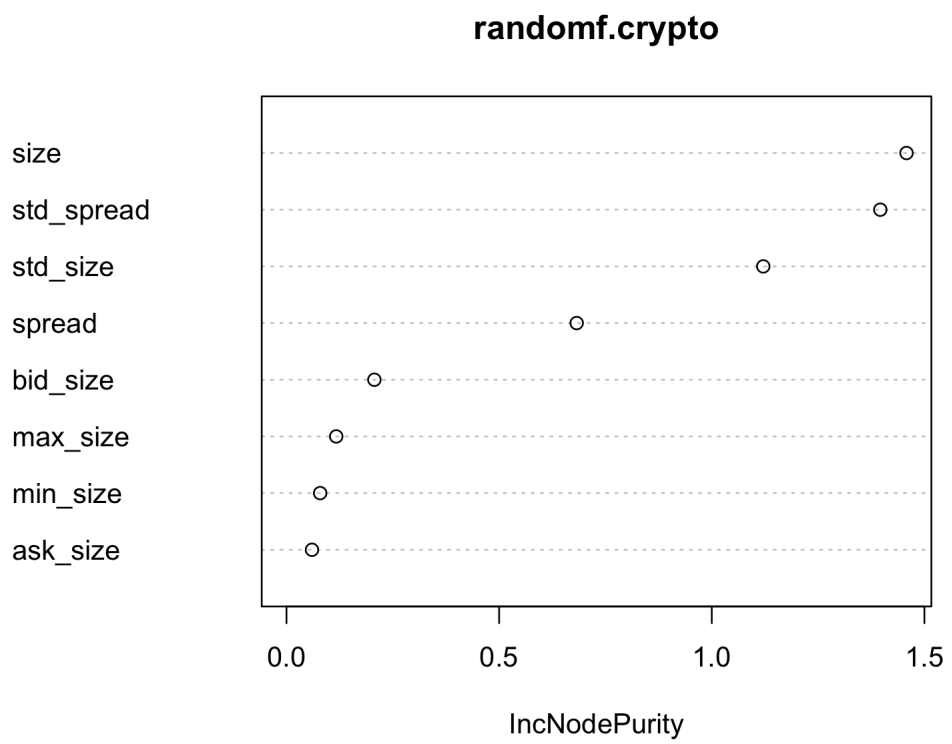


Figure 6: calculate the variables using the raw data