FIN-525

Financial Big Data

Project

Students:
Linas Syvis
Alexander Rusnak

Professor:
Damien Challet

Lausanne, January 14, 2021

1 Introduction

In the currently available academic literature on high-frequency trading it is considered that in most cases high-frequency traders improve market liquidity [1], [2] and efficiency [3]. It is also known that in some cases the high-frequency traders cause or magnify huge losses for the market. Such cases are known as "flash crashes". There is no rigorous definition of what can be called a flash crash, however, literature agrees that it is an event in financial markets wherein the selling or purchasing of financial assets rapidly amplifies price movement. The result appears to be a rapid sell-off or buy-out of securities that can happen over a few minutes, resulting in dramatic declines or increases. High-frequency trading companies are said to be highly responsible for flash crashes recently [4]. Indeed, they consume liquidity during flash crashes, which is instead provided by "traditional" slow traders, and generate a transitory price impact which is unrelated to the permanent impact. In order to be able to foresee and prevent such events it is largely complicated. There may be too many variables in the financial markets, for machine learning in its current state of development to deal with [5].

However, the data that needs to be considered when it comes to trading cryptocurrencies is of a smaller magnitude. Data in cryptocurrencies might be more speculative and co-dependent, which may result in far more predictable patterns than in the financial markets [5].

The cryptocurrency markets operate without breaks and cryptocurrency trading is often done via automated trading programs. High-frequency trading, on the other hand, is not possible on any of the larger cryptocurrency exchanges in order not to break the exchange itself. However, cryptocurrency markets are still prone to the inaccurate feedback loops that define automated market behavior. In other words, trading bots detect increased volume which activates new trading programs which detect increased volume and so on and so forth.

In this paper, we analyze such market behaviour in the cryptocurrency markets. More specifically, we define more rigorously what a flash crash is, analyze them, and develop a model detecting them for the largest cryptocurrencies.

2 Data analysis

In this paper we use data of the largest cryptocurrencies by their Market Cap. according to [6]. The data was downloaded from Kaggle website [7]. It contains the historical trading data (OHLC) of the cryptocurrency/USD pairs at 1 minute resolution reaching back until the year 2013. It was originally collected from the Bitfinex exchange using their API. There are no timestamps for time periods in which the exchange was down. Also, if there were time periods without any activity or trades there will be no timestamp as well.

2.1 Exploratory data analysis

We first begin our analysis by taking a global view of the structure of the data, which is visualized in Figure 10. The data originally consisted of 6 columns: *time*, *open*, *close*, *high*, *low*, *volume*. Since timestamps were encoded in a 13-digit number, in order to better understand them, we transformed them into a date and time format. There were no missing values, which, when modelling, need to be replaced. However, it seems that some missing data might be filled in by the exchange itself (row 2) and may not be too reliable. Also, some cryptocurrencies, like XRP or LTC, had large time gaps of no data in 2013 and 2014. After a more in depth analysis, we observed that such problems occur mostly with older data.

	time	open	close	high	low	volume	ticker
2013-04-01 00:07:00	1364774820000	93.25	93.30	93.30	93.25	93.300000	BTCUSD
2013-04-01 00:08:00	1364774880000	100.00	100.00	100.00	100.00	93.300000	BTCUSD
2013-04-01 00:09:00	1364774940000	93.30	93.30	93.30	93.30	33.676862	BTCUSD
2013-04-01 00:11:00	1364775060000	93.35	93.47	93.47	93.35	20.000000	BTCUSD
2013-04-01 00:12:00	1364775120000	93.47	93.47	93.47	93.47	2.021627	BTCUSD

Figure 1: Data structure.

We further look into the data more in detail and pick a few of the largest cryptocurrencies we used in the analysis. In Figure 2 we can see the exchange rate of BTC/USD evolution over time as the blue line. We notice that the coin has grown exponentially: since 2017 it *exploded* and is currently around an all time high.

This *explosion* is a reason we cannot have other cryptocurrencies in the same graph as BTC - their lines are barely visible. ETH - the second largest cryptocurrency looks much smaller when judging just according to price. However, its' circulating supply (which is around 6x larger) makes the coin's market capitalization much closer to the one from Bitcoin. We show the evolution of market capitalization of the same coins in Figure 12 in the Appendix. If we visualize ETH/USD and LTC/USD in another graph, they show a similar pattern as BTC/USD (especially ETH/USD). There might be many reasons for such similarity: high correlation between the cryptocurrencies; the effect of strengthening/weakening USD, performance of stocks in financial markets, etc.

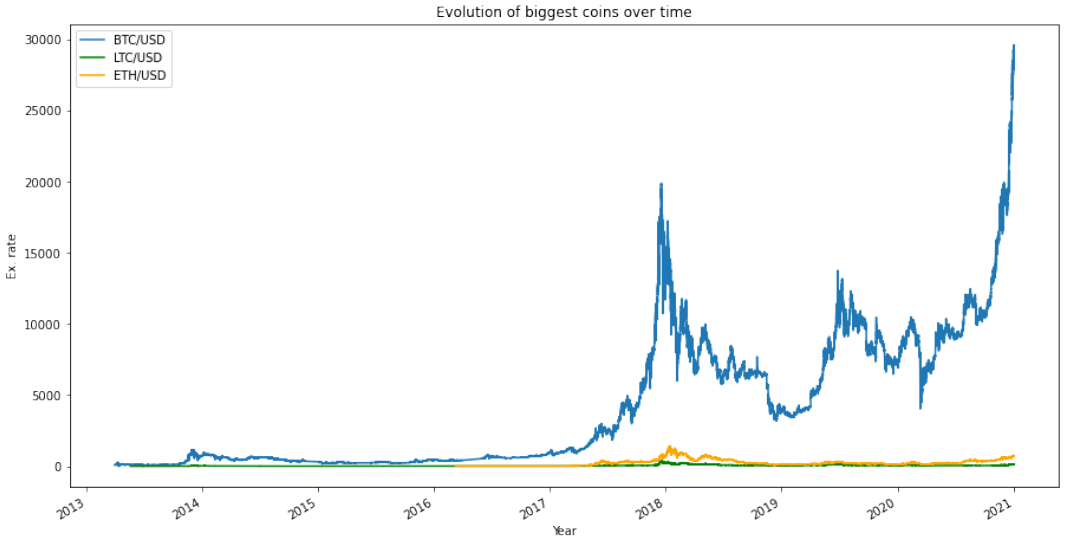

Figure 2: BTC/USD evolution over time.

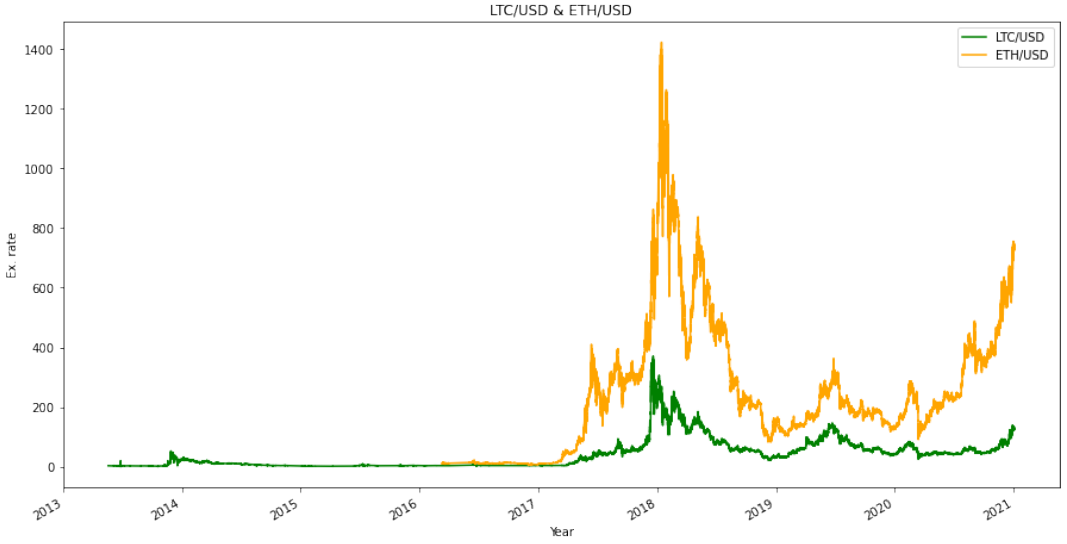

Figure 3: ETH/USD and LTC/USD evolution over time.

Next, we take a look at one of the most recent crashes of Bitcoin, which happened in August, 2020. When looking at the Figure 4, we observe a sharp decline of around 12 percent in a span of 15 minutes. Such drop was then followed by a small recovery. However, the effect of the decrease stayed for the longer term. The reason why this Bitcoin flash crash happened was not immediately clear. A most common explanation found was that it was caused by so-called *whales* who control large amounts of Bitcoin and other cryptocurrencies moving the market. The market is more easily pushed around by whales when trading volumes are lower, such as early on that Sunday morning. Indeed, when looking at the data, the volumes dramatically increased during the flash crash. The Figure 11 in Appendix displays

this phenomenon.

One might be able to exploit such rate changes and generate profit, by having a *flash crash* detection system. To take advantage of a sharp movement, it is essential to study the underlying reasons, the environment just before and after the crash and the aftermath.

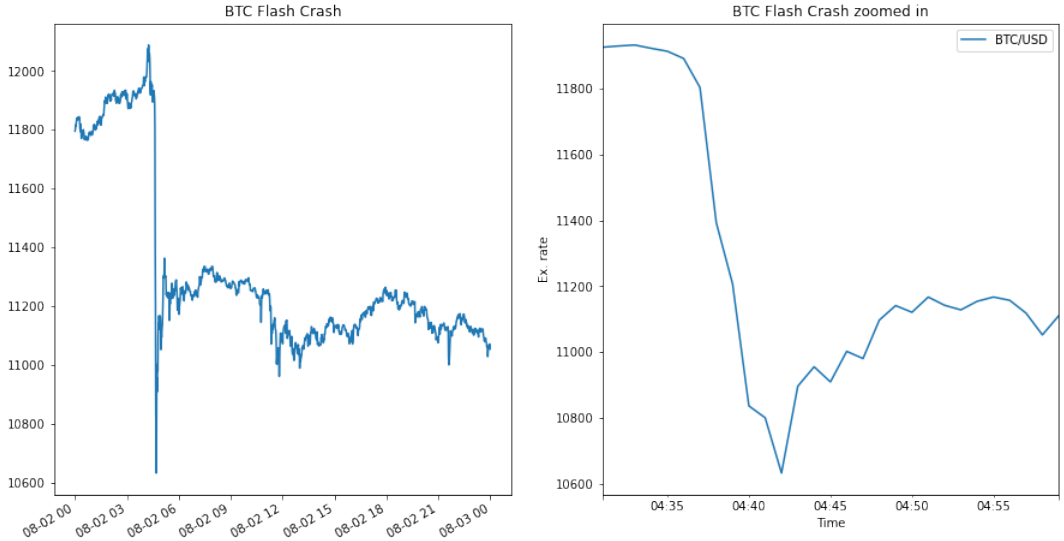

Figure 4: BTC/USD flash crash.

2.2 Definition of a flash crash

As mentioned in the introduction, in the scientific literature, there is no specific definition of a flash crash. Intuitively, it is some abnormal decline that happens in a very short period of time. Now, what one may call 'abnormal' is abstract and 'very short period of time' depends on the situation. So, there are at least two parameters that one has to specify: period of time and magnitude, when defining what a flash crash is. In this project, we decided to call 'flash crashes' movements of 2% or more in the period of 15 minutes. One may say that a change of 2% is relatively small for cryptocurrencies, however, that happening in a 15 minutes timespan was not that common. For example, Bitcoin had around 1200 such crashes since 2013, out of over 3 million timestamps. This makes for a good amount of data points when modelling. It was important not to restrict the flash crash definition too much, because then we would only have a few data points to learn from. It was also important not to have too loose of a definition for a flash crash in order not to label too many points as flash crashes. In the end, it should still be a relatively rare phenomenon.

3 Modelling

3.1 Volume-synchronized probability of informed trading

As an additional metric to describe our data and provide more structured information to our classifier, we implemented the VPIN flow toxicity metric, which stands for volume-synchronized probability of informed trading. Such metric was introduced by Easley et al. [8].

Probability of informed trading has a well documented history of being used to estimate flow toxicity in high frequency data, and thus is useful for estimating flash crashes as it can provide an approximation of how informed trading events are. It has been demonstrated that VPIN has a correlation with high short-term toxicity induced volatility, and can predict large price moves from this volatility. If the trades are extremely toxic (uniformed) then it is likely they are entering feedback loops or stochastic processes that could lead to flash crashes. Traditional PIN approximation rates model daily buys and sells of stocks, but are based on a time-variant approach of the arrival rate of particular traders. Since the flow of trades does not always distribute equally between time frames, it is important to equalize this distribution when determining the probability of informed trading. Thus, the VPIN metric recalibrates the data to have equal volume of trades in each time interval, which overcomes variance in volume flow in highly active or evolving markets.

In the figure below, we compare the VPIN metric across time for 8 coins in our dataset. It demonstrates clearly that the bulk of toxic flow occurs in a more recent time span for smaller / less stable coins, which have demonstrably higher volatility than larger and more established coins like bitcoin.

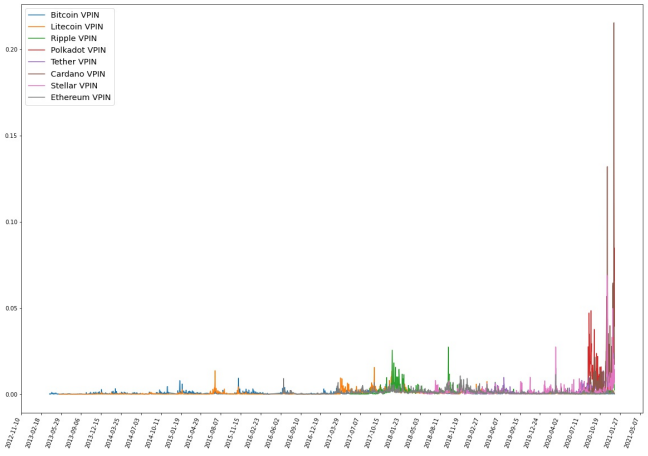

Figure 5: VPIN Metric

Coin Name	Mean Volume	Coef. of Variation	Mean VPIN
Bitcoin	17.97	0.946444	0.0007116
Ethereum	142.64	0.744645	0.0012860
Tether	13521.55	0.009645	0.0035474
Ripple	33354.14	0.786306	0.0016066
Cardano	7122.49	0.227688	0.0241297
Litecoin	168.14	0.857291	0.0011752
Polkadot	416.19	0.187772	0.0190554
Stellar	9526.83	0.567248	0.0047683
Tops Coins	8033.75	0.540879	0.0070351
All USD Paired Coins	17006.08	2.090281	N/A

We also found a relationship between the crashes themselves and the VPIN metric, which is a result we were looking for. The correlation between the VPIN metric and the flash crash binary variable was 0.153934. Future analysis could be done to correlate the directionality and percent change of the crashes to the VPIN metric.

3.2 Gradient Boosted Tree Classifier

In order to classify which subsequent frames might contain a crash event, we decided to use a gradient boosted tree classifier fed with the preprocessed data of the 8 top coins on which we focused our analysis. We staged our problem as a binary classification problem, with 0 representing no crash in the following frame, and 1 representing a crash in the following frame. Gradient boosted trees are based on a standard random forest algorithm, which itself is based on an ensemble of decision trees. The decision trees create delineating partitions between different values of the data that it uses to sort them into classes. Since random forests use many trees, they can learn more complex combinations of data that point to class differentiation. The gradient boosted aspect refers to the way in which the trees are initialized, the model incorporates the gradient with respect to the error from the last created tree, and uses it to generate progressively more accurate trees as it trains. This approach has been demonstrated to be substantially more computationally efficient and accurate than a standard random forest. In particular, we use the XGBoost library for our implementation. At each time step, we feed the model the open, close, high, and low values of that step. We also give the price change, raw volume, adjusted volume, and the VPIN of the preceding day. As previously mentioned, the target of the model is a binary class label representing either a crash event in the following time step, or the lack of a crash event.

3.3 Acceleration

Training our model on such large amounts of data is a computationally expensive task, and one that consumes a lot of time on our relatively small devices. To that end, we accelerated training the XGBoost algorithm by distributing computations

to the GPU. To test the speed up capabilities, we trained using a subset of our data (just the training set of LiteCoin, or 1,455,199 rows) on both the CPU and GPU of one of our local machines. In total, this smaller dataset took 166.89 seconds to train on the CPU and only 143.74 seconds when distributed to the GPU, a 13.86 percent reduction in compute time. When scaled up to our larger top coins dataset, this is a substantial saving in time of approximately 2 minutes.

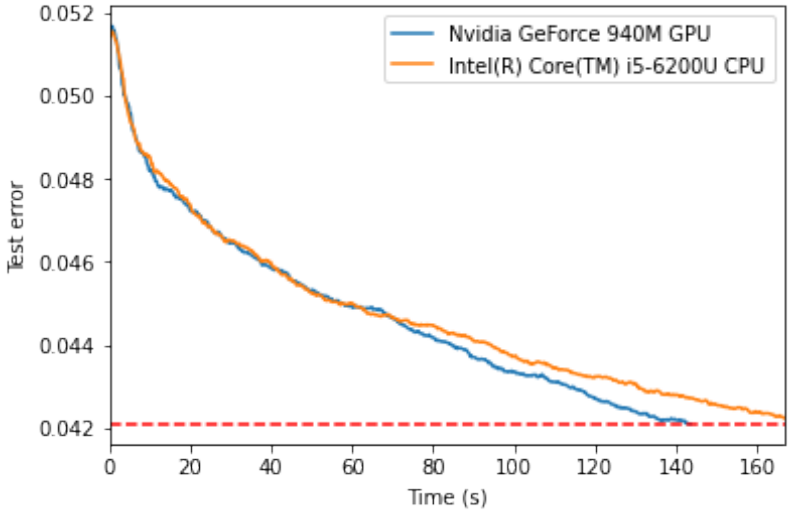

Figure 6: Hardware Comparison on LiteCoin Data

In addition to using the GPU for our model, we also used a Dmatrix for our train and test sets instead of a standard pandas dataframe or numpy matrix. A DMatrix is an optimized data structure that is part of the XGBoost ecosystem, and generally increases memory efficiency and speed. We utilized dask as well wherever appropriate in our code to speed up computations on the CPU by optimizing their execution. We also saved with the parquet file type in any function where the repetitive access of files was needed, and thus could be sped up by quicker read / write speeds.

3.4 Dask implementation

As these days the datasets and computations scale faster than CPUs and RAM, we need to find ways to scale our computations too. For this reason, in addition to our two limited laptops with i5 and i7 processors and 8GB of RAM, we used Dask. Dask is a dynamic task scheduling open-source library, which splits up large computations and routes parts of them efficiently onto distributed hardware.

To avoid excess memory use, Dask is good at figuring out how to evaluate computations in a low-memory environment by pulling in chunks of data from disk, doing the necessary processing, and throwing away temporary values rapidly. Dask enabled us to efficiently parallel computations on our laptops by using and abusing their multi-core CPUs. As a result it saved significant amount of time when labeling the data into "flash crash" / "not flash crash" and modelling multiple currencies.

3.5 Data Pipeline

We used multiple base datasets of varying sizes, but generally used the same preprocessing pipeline. After taking the raw cryptocurrency data and labeling it with crashes and calculating the VPIN, we changed the datetime to a singular float number representing the 24 hour clock time of the event. Our input to our model was then: float time, high, low, open, close, raw volume, change in price over frame, and vpin. We then used a standard scaler on all the features, and split them into train and test datasets for the model. The main datasets we ended up using were: the top coins dataset consisting of 9,386,046 rows, the LTC coin set consisting of aforementioned 1,455,199 rows, and a final USD paired dataset of all the coins we could process in time / handle in the memory of our computer when training the model, of 12,317,928 rows. We also used a multi coin (IOT, EOS, ETH) JPY paired dataset of 436,924 rows and a multi coin (IOT, EOS, ETP) ETH paired dataset of 792,569 rows for cross pair compatibility testing. Our largest computation hurdle in terms of time was processing the labels for the data and calculating the VPIN, while our biggest challenge in terms of memory allocation was handling the largest datasets while training the model.

4 Results

We found that we were accurately able to train our XGBoost model on our dataset to predict which frames preceded flash crashes. Our main observations in regards to model performance were: quantity of data vs specificity of data, feature importance, and applicability across currency pairings.

4.1 Quantity vs Specificity

We observed through our experiments with different models that not all coins have the same patterns, but that in general having a larger dataset for training is particularly useful even for more specific, particular coins. We observed that while training a model on one coin with a sufficiently large dataset performed slightly better on that particular coin then a more general model, it was horrible in terms of generalization as it overfit to the specific coin. That is, as the singular coin model became better at targeting that particular coin, it became much worse at targeting other coins which were out of sample. This does not bode well for the robustness or generalizabilty of the single coin models across the cryptocurrency trading space or across time. Furthermore, when trained and tested on the entire dataset, the top coins model achieved superior loss values than the single coin model did when tested on the single coin test set. This could be due to class imbalances in the larger dataset, but we find this unlikely considering our test coin (LiteCoin) did not have substantial volatility relative to other coins in our set.

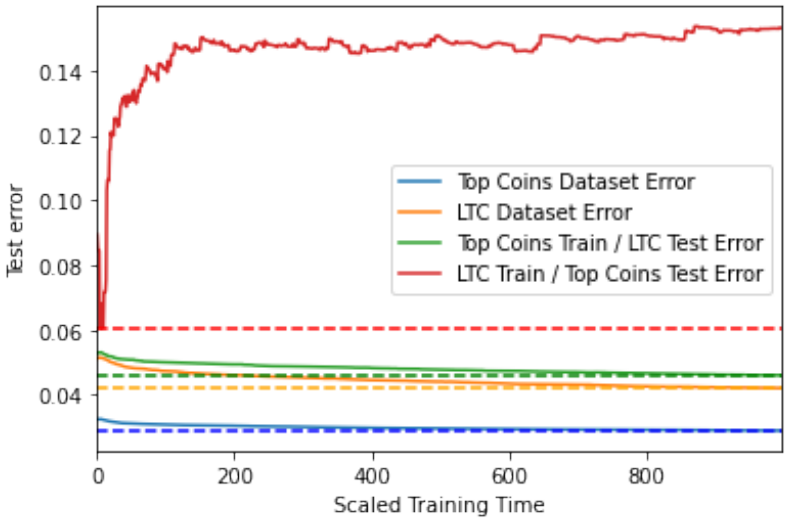

Figure 7: Comparative Loss on Different Train and Test sets

4.2 Feature Importance

We observed that the feature importance of different models varied based on training runs due to the stochastic nature of the model, but primarily based on the dataset.

When using the smaller LTC set, the primary feature of importance was actually the float time, but when looking at the larger datasets, the VPIN metric was the most important predictor. This is indicative of the importance of the VPIN metric to predicting highly volatile markets and flash crashes.

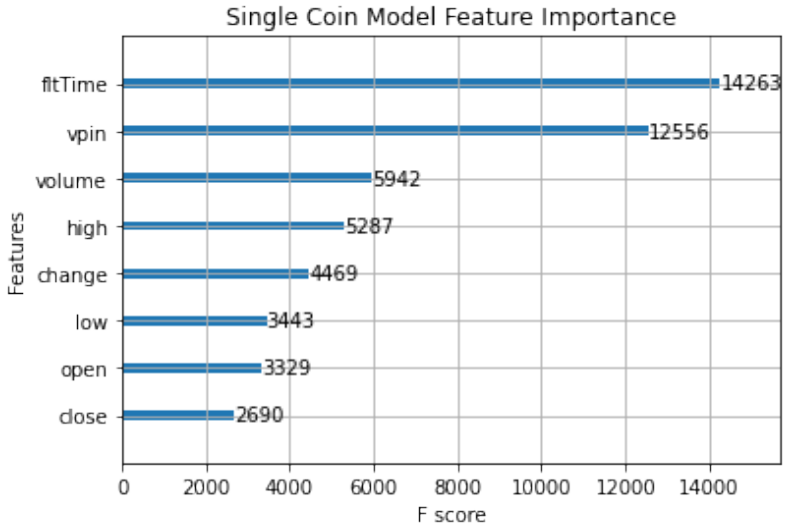

Figure 8: LTC Dataset Feature Importance

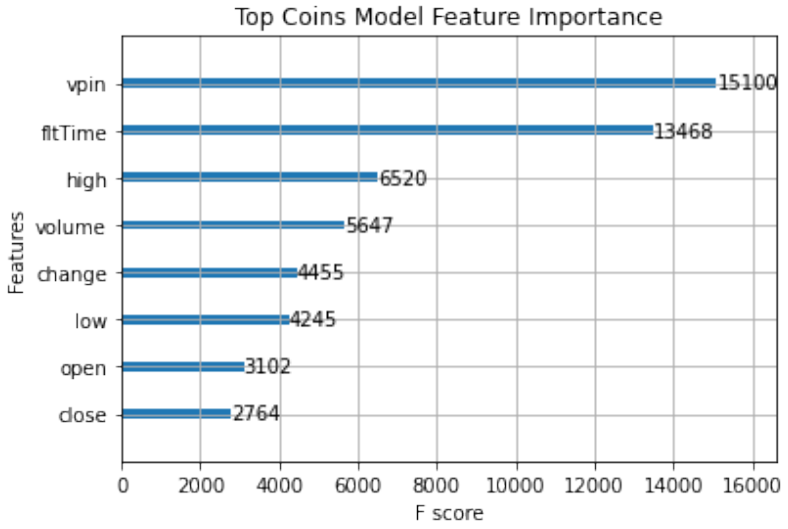

Figure 9: Top Coins Dataset Feature Importance

4.3 Applicability Across Currency Pairings

In addition to examining which models performed best on more broad and more specific datasets, we also explored the efficacy of models trained on Crypto-USD paired data to other pairings. In particular we looked at Crypto-JPY pairings, and

Crypto-Crypto pairings (with ETH as the common base pairing for the Crypto-Crypto set). We also used our Max Coins USD dataset for training in this analysis, as we learned in our earlier analysis that more training data led to better out of sample results/ more generalizable results. We observed that there was divergence between the JPY pairings and USD pairings, and between the ETH pairings and USD Pairings (i.e. as the model increased its performance on the USD pairing training set, it's performance on the other pairings decreased). However, the performances were not equal. With the JPY pairing, the error remained relatively low in relation to the USD pairing error. In contrast, the ETH pairings had the worst error rate we observed throughout all of our testing, indicating there is a substantial divergence in the movements of the market for the crypto-crypto pairings. This makes logical sense as the JPY and USD share many characteristics because they are both highly traded fiat currencies from large westernized economies, but we did not expect the difference in performance to be so stark. Further examination of this disparate performance could be a path for future research.

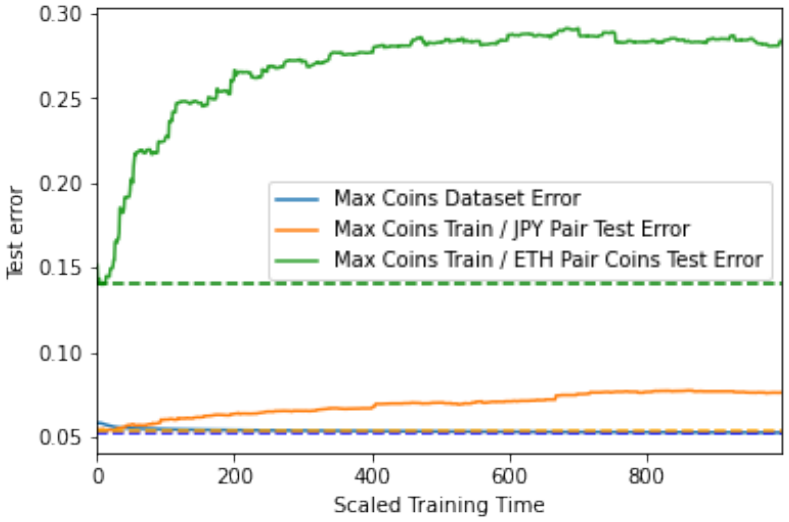

Figure 10: Max Coins Model performance on Max Coins, JPY Pairing, and ETH Pairing Test Sets

5 Conclusion

While high frequency traders and algorithmic traders add substantial liquidity to markets, they also can be responsible for high volatility and large scale price moves known as 'flash crashes.' Though much research has been done on this topic in regards to equities, traditional currencies, and other securities, the cryptocurrency space remains less examined. It is particularly interesting to analyze because of the unique restrictions on trading in regards to coin size and mining time that define the moves of the market. We analyzed copious amounts of cryptocurrency data using the volume-synchronized probability of informed trading metric to determine flow toxicity, and found that it does have a correlated relationship to and predictive effect for flash crashes. We then utilized an XGBoost gradient boosted tree model to further test the efficacy of the VPIN metric, which we confirmed on larger datasets as the most important variable for prediction. We also tested the predictive power of our tree model across more specific coin datasets and other currency pairings, finding that while more data does lead to a less fragile model, there are substantial differences between currency pairing types, particularly cryptocurrency-cryptocurrency pairs. In summation, we effectively created a predictive classifier, using large amounts of data, for flash crashes in the cryptocurrency space and confirmed the importance of the VPIN metric for predicting these crashes.

6 Appendix

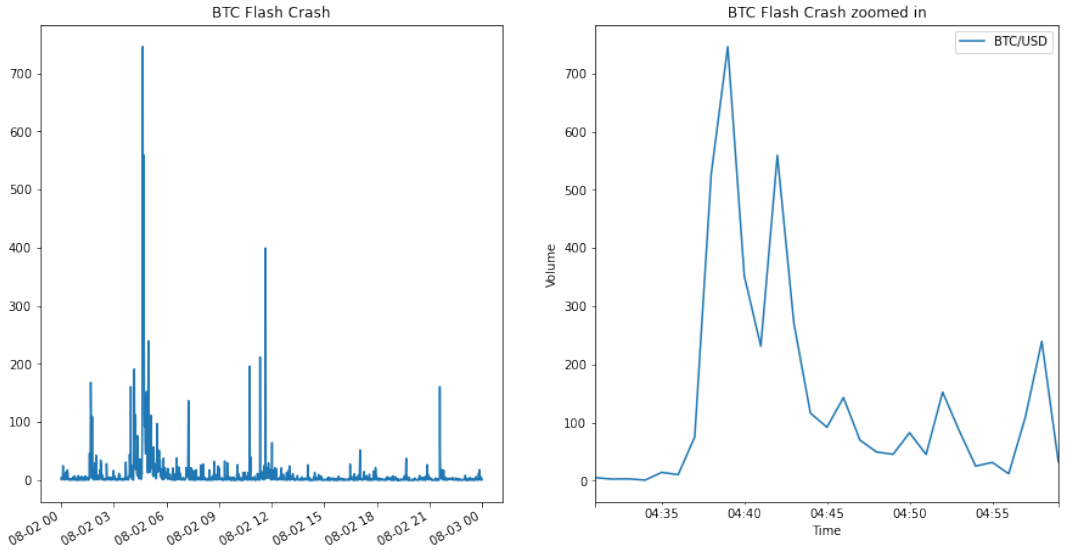

Figure 11: BTC/USD flash crash Volume.

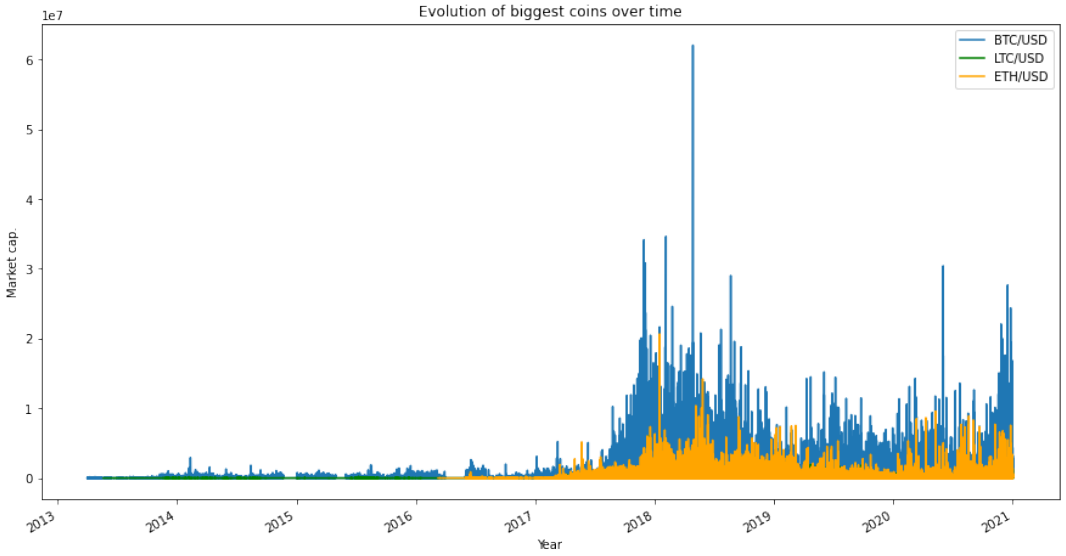

Figure 12: Cryptocurrencies market cap over time.

References

[1] J. Hendershott and Menkveld, “Does algorithmic trading improve liquidity?” 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2010.01624.x

[2] C. M. Jones, “What do we know about high-frequency trading?” [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2236201

[3] B. C. E. H. Chaboud, A. P. and C. Vega. (2014) Rise of the machines: Algorithmic trading in the foreign exchange market. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12186

[4] W. Kenton. (2019) Flash crash. [Online]. Available: https://www.investopedia.com/terms/f/flash-crash.asp

[5] P. Tan. (2019) Ai struggles to beat financial markets, but could it beat cryptocurrency markets? [Online]. Available: https://medium.com/swlh/ai-struggles-to-beat-financial-markets-but-could-it-beat-cryptocurrency-markets-7ddd7cb6f6a1

[6] Coinmarketcap. Cryptocurrency prices by market cap. [Online]. Available: https://coinmarketcap.com/

[7] Kaggle. 400+ crypto currency pairs at 1-minute resolution. [Online]. Available: https://www.kaggle.com/tencars/392-crypto-currency-pairs-at-minute-resolution

[8] M. O. David Easley, Marcos M. Lopez de Prado, “Flow toxicity and liquidity in a high-frequencyworld,” 2012.

[9] Dask, “Why Dask?” 2020. [Online]. Available: https://docs.dask.org/en/latest/why.html

[10] Jinzhi Jiang , “Volume-Synchronized Probability of Informed Trading” 2015.

[11] Forbes, “A Massive Bitcoin Flash Crash Just Created $1 Billion Of Crypto Chaos” 2020.

[12] Patrick Tan, "AI Struggles to Beat Financial Markets, But Could It Beat Cryptocurrency Markets?" (2020), [Online]. Available: https://medium.com/swlh/ai-struggles-to-beat-financial-markets-but-could-it-beat-cryptocurrency-markets-7ddd7cb6f6a1

[13] Torben Andersen, Oleg Bondarenko, "The Trouble with VPIN" (2012), [Online]. Available: https://insight.kellogg.northwestern.edu/article/the$_t rouble_w ith_v pin$

[14] Bellia, Mario; Christensen, Kim; Kolokolov, Aleksey; Pelizzon, Loriana; Renò, Roberto, "High-frequency trading during flash crashes: Walk of fame or hall of shame?", [Online]. Available: https://www.econstor.eu/bitstream/10419/215430/1/1693370115.pdf

[15] Avanidhar Subrahmanyam, "Algorithmic trading, the Flash Crash, and coordinated circuit breakers", https://www.sciencedirect.com/science/article/pii/S2214845013000082

[16] Fry, John and Serbera, Jean-Philippe, "Modelling and mitigation of Flash Crashes", [Online]. Available: https://mpra.ub.uni-muenchen.de/82457/1/MPRA$_paper_8$2457.*pdf*

[17] CFA Institute, "Flash crashes", [Online]. Available: https://www.cfainstitute.org/en/advocacy/is crashes

[18] Ian Poirier, "High-Frequency Trading and the Flash Crash: Structural Weaknesses in the Securities Markets and Proposed Regulatory Responses", https://repository.uchastings.e

[19] Anton Golub, John Keane, and Ser-Huang Poon, "High Frequency Trading and Mini Flash Crashes", https://arxiv.org/pdf/1211.6667.pdf