THE JBBA

PEER REVIEWED RESEARCH

OPEN ACCESS
ISSN Online: 2516-3957
ISSN Print: 2516-3949
https://doi.org/10.31585/jbba-2-2-(2)2019

# Cryptocurrency Investing Examined

Jim Kyung-Soo Liew[1], Richard Ziyuan Li[2] , Tamás Budavári[2], Avinash Sharma[3]
Johns Hopkins University, USA
[1]Carey Business School
[2]Department of Applied Mathematics and Statistics
[3]Bioengineering and Biomedical Engineering
**Correspondence:** kliew1@jhu.edu

**Received**: 17 January 2019 **Accepted**: 28 March 2019 **Published**: 28 May 2019

**Abstract**

In this work we examine the largest 100 cryptocurrency returns ranging from 2015 to early 2018. We concentrate our analysis on daily returns and find several interesting stylized facts. First, principal components analysis reveals a complex daily return generating process. As we examine data in the most recent year, we find that surprisingly more than one principal component appears to explain the cross-sectional variation. Second, similar to hedge fund returns, cryptocurrency returns suffer from the "beta-in-the-tails" hidden risk. Third, we find that predicting cryptocurrency movements with machine learning and artificial intelligence algorithms is marginally attractive with variation in predictability power per crypto-currency. Fourth, lower volatile cryptocurrencies are slightly more predictable than more volatile ones. Fifth, evidence exists that efficacy of distinct information sets varies across machine learning algorithms, showing that predictability may be much more complex given a set of machine learning algorithms. Finally, short-term predictability is very tenuous, which suggests that near-term cryptocurrency markets are semi-strong form efficient and therefore, day trading cryptocurrencies may be very challenging.

**Keywords:** *AI, Bitcoin, Cryptocurrencies, Machine Learning, PCA, Beta-in-the-Tails*

**JEL Classifications:** *G12, G14, G17, G40, G*

**Competing Interests:**
*None declared.*

**Ethical approval:**
*Not applicable.*

**Author's contribution:**
*JL[1], RL[2], TB[2], and AS[3] designed and coordinated this research and prepared the manuscript in entirety.*

**Funding:**
*None declared.*

**Acknowledgements:**
*None declared.*

## 1. Introduction

Cryptocurrency is a digital asset designed to work as a store of value and a medium of exchange[1]. As of February 28th, 2018, the total market capitalization of the cryptocurrency market stood at $448 billion and consists of 1,524 types of currencies. Amongst the many controversies surrounding cryptocurrencies, a popular topic of debate is whether it should be classified as a commodity, investment, property, currency or digital currency. Bitcoin puts cryptocurrencies center stage in the popular press and with the recent painful pull back in early 2018, the interest in Bitcoins in particular continues to hold. Bitcoins started 2017 at $998.33 and grew 14x to finish the year at $14,156.40, as is shown in Fig. 1. As of February 28th, the price was $10,559.20.

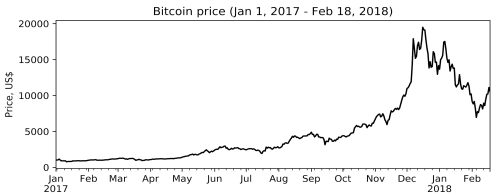

Figure 1: Bitcoin price from Jan 1, 2017 to Feb 18, 2018

Bitcoin, the first successful cryptocurrency, was created in January 2009, in the aftermath of the financial crisis of 2008, by an unknown person or a group of people under the Japanese name of Satoshi Nakamoto. Bitcoin utilizes a technology called blockchain, which is a combination of cryptography, consensus algorithms, economic incentives and distributed ledger to secure its transactions. While the technical discussion of blockchain is beyond the scope of this work, this technology has endowed Bitcoin with many important characteristics, such as;

- Decentralization,
- Trusted network built upon potentially untrustworthy nodes,
- Transparency, and
- Immutability history, etc.

Many cryptocurrencies were invented after Bitcoin, but Bitcoin continues to be the most popular, as evidenced by it having the largest market capitalization and trading volume, shown in Table 1 below. Subsequently, our investigation primarily focuses on Bitcoin prices in this research.

| Index | Name | Price | Market Cap ($Billion) | Volume (24 hrs $Billion) |
|---|---|---|---|---|
| 1 | Bitcoin | $10,559.20 | $178.4 | $6.9 |
| 2 | Ethereum | $869.63 | $85.1 | $2.0 |
| 3 | Ripple | $0.921 | $36.0 | $0.33 |
| 4 | Bitcoin Cash | $1,223.85 | $20.8 | $0.38 |
| 5 | Litecoin | $208.43 | $11.6 | $0.78 |
| 6 | NEO | $135.27 | $8.8 | $0.33 |
| 7 | Cardanol | $0.317 | $8.7 | $0.12 |
| 8 | Stellar | $0.346 | $8.2 | $0.037 |
| 9 | EOS | $8.64 | $6.0 | $0.38 |
| 10 | IOTA | $1.89 | $5.3 | $0.044 |

Table 1: Top Ten Cryptocurrencies

(Source: CoinMarketCap.com, data as of February 28th, 2018.)

While participants of the Bitcoin blockchain can transfer Bitcoins with each other directly, most investors have to go to cryptocurrency exchanges if they want to purchase Bitcoins with U.S. dollars or other traditional currencies. While the quoted prices from different exchanges can vary largely, arbitrage was very difficult due to the lack of easy access to short Bitcoins, until CBOE and CME introduced Bitcoin futures in December 2017.

### 1.2 Artificial Intelligence (AI)

Similar to cryptocurrency, AI is another increasingly intriguing technological development. AI represents a broad range of techniques including machine learning, deep learning, natural language processing, etc. Its application is rapidly penetrating every aspect of human society - e-commerce, autonomous vehicles, image recognition, to name a few. A detailed discussion of AI techniques and their application, unfortunately, is beyond the scope of this paper.

Financial institutions are increasingly testing and deploying AI techniques to obtain an edge in their business, such as in trading. Money managers have been employing thousands of quantitative experts to develop sophisticated AI models for predicting prices, identifying signals, monitoring sentiment, etc. While the efficacy of these efforts is still debatable, AI models and strategies are prevailing in every market (equity, commodity, FX, etc.). It is, therefore, only a matter of time before practitioners and academic researchers begin using AI techniques to analyze cryptocurrency markets. We hope our findings herein will serve as an important contribution to this growing field.

### 1.3 Our Research Results

In this paper, we first analyze the top 100 cryptocurrencies using correlation analysis and principal component analysis (PCA). Daily returns reveal that in some period there exists a single dominant component however, in the most recent prior year there appears to be two components that help explain the variation of the cryptocurrency returns. Next, we compare cryptocurrencies with traditional assets. We also perform Liew [2013]'s beta-in-the tail analysis to examine potential hidden risks. We find some evidence that similar to hedge funds, cryptocurrencies may suffer from this hidden risk.

Finally, we conduct rolling prediction analysis on 57 cryptocurrencies with 11 AI algorithms. Our results show that predictability may be difficult and there are many heterogeneous effects here. Some information sets perform better with some family of algorithms, and larger cryptocurrencies with lower volatility maybe more predictable than smaller cryptocurrency with higher volatility.

The remainder of this paper is organized as follows: Section 2 reviews prior literature, Section 3 presents our data and preliminary analysis, Section 4 describes the methodology, Section 5 provides the results and Section 6 summarizes and concludes.

## 2. Literature Review

While there are many cases and projects about Bitcoin price predictions online, scarce academic research presently exists regarding Bitcoin price predictability. We review the most important prior research in this subject by aggregating them into three different groups.

The first group attempts to predict Bitcoin prices with information about the Bitcoin blockchain network. For example, Madan et al. [1] from Stanford use three machine learning algorithms to predict the sign of daily price change of Bitcoin based on data about the Bitcoin blockchain network, including average confirmation time, block size, hash rate, etc. They report a highest accuracy of 98.7%. Another group of Stanford researcher, Greaves et al. [2] perform similar analysis, getting a classification (sign of hourly price change) accuracy of 55%. In addition to information about the blockchain network, McNally [3] adds daily open, high, low, and close prices as explanatory variables, reporting a classification (signs of daily price changes) accuracy of 52%. El-Abdelouarti Alouaret [4] moves further by including the S&P 500 index and EUR/USD rate, as well as a variable named bitcoins days destroyed. Similar to sentiment analysis, it also includes a variable representing daily page view on the Wikipedia item "Bitcoin". It also uses

  

vector autoregression and recurrent neural network to conduct price prediction instead of classification.

The second group of studies focus on the relationship between social media data and Bitcoin performance. For instance, Mai et al. [5] analyze Bitcoin-related user posts from a forum and Twitter and demonstrate that more bullish posts are associated with higher future Bitcoin returns. They also conclude that the social media effects on Bitcoin performance are driven by the "silent majority", and the impact of forum posts is larger than that of tweets. Stenqvist et al. [6] try to predict Bitcoin price (up/down) using sentiment analysis on Twitter, and report that the sentiment change over a 30-minute period is useful for predicting price movement of 2 hours later, resulting in an accuracy of 79%. Instead of performing sentiment analysis on all social media content posted, Kim et al. [7] extract the hottest topics on a Bitcoin-related forum and define a time series score to represent the "strength" of each topic. While these scores are not significant in Granger causality tests, a deep learning model with these scores as inputs leads to prediction (for price and transaction volume) accuracies ranging from 50%+ to 80%+. Interestingly, Kaminskt [8], by analyzing Twitter posts, claims that social media sentiments mirror the Bitcoin market activity, rather than being predictive.

Instead of Bitcoin blockchain network data and social media data, some papers examine the performance of Bitcoin in other ways. Chu et al. [9] fits log returns in fifteen popular parametric distributions in finance and find that the generalized hyperbolic distribution is the most appropriate. Balcilar et al. [10] perform causality-in-quantiles tests and point out that Bitcoin trading volume can predict price returns but fail to predict volatility. Indera et al. [11] use Multi-layer Perceptron (MLP) to predict Bitcoin price based on historical open, high, low, and close, as well as the moving average technical indicators, reporting significant results (in mean mean-squared error ).

The third group of research comprises of researchers attempting to use every factor to predict Bitcoin price. Georgoula et al. [12] and Garcia et al. [13] contribute their work in this way. As they provide many conclusions, we are not summarizing here.

## 3. Data and Preliminary Analysis

### 3.1 Cryptocurrency

As we mentioned above, there are 1,524 different cryptocurrencies as of February 28, 2018, and they are traded at many different exchanges (markets). Fortunately, CoinMarketCap.com collects transaction data of these cryptocurrencies from various exchanges and publishes both up-to-date and historical data for free, which can be obtained through their API. Taking advantage of this resource, we scrap the historical data of the top 100 cryptocurrencies, in terms of market capitalizations as of February 18, 2018. Before selecting the top 100, we remove those with relatively short history[1]. Therefore, all selected cryptocurrencies date back to at least January 1, 2017, and Fig. 2 shows the number of cryptocurrencies under analysis over time. The data includes close price, trading volume, and market capitalization during the period of January 1, 2015 to December 31, 2017.

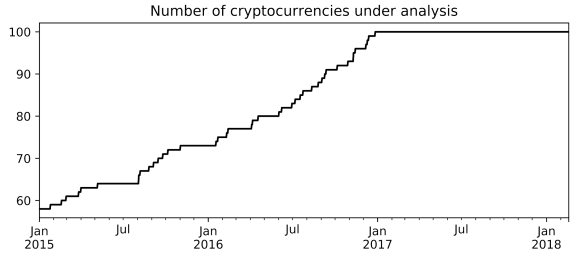


Figure 2: Number of cryptocurrencies under analysis (Jan 2015 - Feb 2018)

#### 3.1.1 Price returns

We calculate daily, weekly, and monthly returns for each cryptocurrency as (holding period returns):

$$R_t = \frac{P_t}{P_{t-1}} - 1$$

We conduct normality tests on all returns series and find that during Jan 1, 2015 to Feb 18, 2018, none of the daily price returns of any cryptocurrency is normal at the significance level of 95%. For weekly returns, two cryptocurrencies yield normal returns. And ten of them have normal monthly returns. Therefore, we think it is more appropriate to use holding period returns rather than log returns.

Table 2, Table 3, and Table 4 provide statistical summary of price returns of Bitcoin (BTC), Ethereum (ETH), and Ripple (XRP), respectively, which are the top 3 cryptocurrencies in terms of market capitalization, as of February 18, 2018. All the three have an average daily return of less than 1% as well as single-digit weekly returns.

Table 2: Statistics summary for price returns of Bitcoin (Jan 2015 - Feb 2018)

| | Count | Mean | Standard deviation | Minimum | Median | Maximum |
|---|---|---|---|---|---|---|
| Daily | 1144 | 0.0039 | 0.0403 | -0.2115 | 0.0026 | 0.2525 |
| Weekly | 163 | 0.0268 | 0.1053 | -0.2834 | 0.0187 | 0.5097 |

Notes: the "Count" means the number of daily returns and etc. This note applies to the next three tables.

Table 3: Statistics summary for price returns of Ether (Aug 2015 - Feb 2018)[i]

| | Count | Mean | Standard deviation | Minimum | Median | Maximum |
|---|---|---|---|---|---|---|
| Daily | 926 | 0.0097 | 0.0798 | -0.7280 | -0.0002 | 0.5103 |
| Weekly | 132 | 0.0682 | 0.2514 | -0.3394 | 0.0098 | 1.4227 |

Table 4: Statistics summary for price returns of Ripple (Jan 2015 - Feb 2018)

| | Count | Mean | Standard deviation | Minimum | Median | Maximum |
|---|---|---|---|---|---|---|
| Daily | 1144 | 0.0065 | 0.0914 | -0.4600 | -0.0035 | 1.7937 |
| Weekly | 163 | 0.0494 | 0.2808 | -0.3311 | -0.0169 | 1.9992 |

Table 5 presents the average statistics summary for the top 100 cryptocurrencies. On average, these cryptocurrencies have an average history of 30 months[ii]. Due to some volatile cryptocurrencies, the average returns and average standard deviations are larger than those for the top 3 shown above.

Table 5: Average statistics summary for price returns of the Top 100 cryptocurrencies (Jan 2015 - Feb 2018)

| | Count | Mean | Standard deviation | Minimum | Median | Maximum |
|---|---|---|---|---|---|---|
| Daily | 962 | 0.0452 | 0.4701 | -0.5580 | -0.0009 | 9.0874 |
| Weekly | 137 | 0.1636 | 0.9940 | -0.5356 | 0.0064 | 9.2084 |

Notes:

1. First, we calculate the statistics summary for each cryptocurrency, including count, mean, standard deviation, minimum, median, and maximum. Then, we calculate the averages of these statistics of all cryptocurrencies.
2. Not all cryptocurrencies have history back to January 2015. The missing values are dropped before calculating the statistics.

### 3.1.2 Correlations

To reveal the relationship between various cryptocurrencies, we calculate the correlations of price returns between the top 100 of them. Fig. 3 present the heatmaps of the correlations of daily returns. And Table 6 provides statistics summary for the correlations across all top 100. Obviously, most of the cryptocurrencies are positively correlated and correlations are getting higher when the time frame becomes larger. Another interesting finding is that correlations between large market-cap cryptocurrencies are higher than correlations between smaller market-caps.[iii] Therefore, we can conclude that most cryptocurrencies are moving in herds with lower double-digit correlations, and this phenomenon is stronger between large market-caps.

Finally, to find out how correlations among cryptocurrencies develop over time, we perform a rolling analysis as is shown in Fig. 4. On each day, we calculate the correlations based on daily returns of the preceding 60 (180) days, and then we use the arithmetic mean as the average correlation for that day. That said, the statistic represents the level of correlation of the overall cryptocurrency market during the past 60 (180) days. Obviously, an interesting finding is the spike of market correlation in the second half of 2017, which was exactly accompanied with the rising hotness of cryptocurrencies.

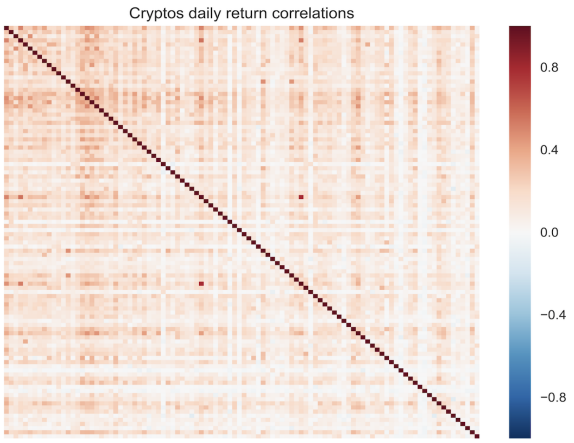


Figure 3: Correlations of daily price returns between top 100 cryptocurrencies (Jan 2015 - Feb 2018)

Table 6: Statistical summary for correlations of returns between top 100 cryptocurrencies (Jan 2015 - Feb 2018)

| | Mean | Standard deviation | Minimum | Median | Maximum |
|---|---|---|---|---|---|
| Daily | 0.1210 | 0.0522 | 0.0052 | 0.1290 | 0.2289 |
| Weekly | 0.1569 | 0.0659 | 0.0036 | 0.1729 | 0.2855 |

Notes:
First, for each cryptocurrency, we calculate the mean of its correlations with other cryptocurrencies. Then, we calculate these statistics of the means of correlations.

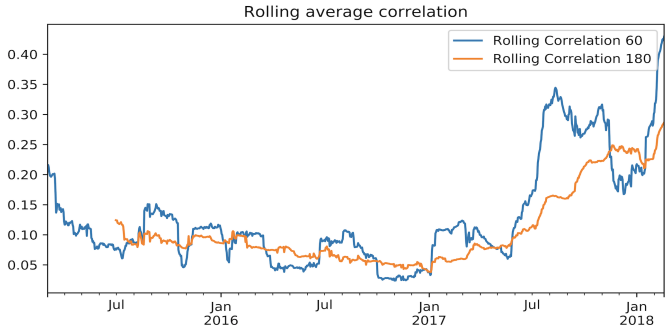


Figure 4: Rolling average correlation (60-days and 180-days, Jan 2015 - Feb 2018)

To have a closer look at Bitcoin, we summarize the statistics of its correlations of price returns with other cryptocurrencies in Table 7. On average, Bitcoin has a correlation of price returns (daily, weekly) of about 0.20 with other cryptocurrencies. In addition, Table 8 lists the most and least correlated cryptocurrencies with Bitcoins. One interesting cryptocurrency stood out upon a quick inspection - Litecoin (LTC) is highly positively correlated with Bitcoin in both time frames.

We also examine the autocorrelation of Bitcoin, as is shown in Fig. 5. The autocorrelations for daily returns fall between -0.05 and 0.05, implying a low autocorrelation nature.

Table 7: Statistics summary for correlations of between Bitcoins and other cryptocurrencies (Jan 2015 - Feb 2018)

| | Mean | Standard deviation | Minimum | Median | Maximum |
|---|---|---|---|---|---|
| Daily | 0.2211 | 0.1158 | -0.0140 | 0.2225 | 0.5035 |
| Weekly | 0.1897 | 0.1382 | -0.1135 | 0.1962 | 0.4976 |

Notes: These statistics are calculated based on the correlations of price returns between Bitcoins and the other 99 cryptocurrencies.

Table 8: Most and least correlated cryptocurrencies with Bitcoins (Jan 2015 - Feb 2018)

| | Daily returns | | Weekly returns | |
|---|---|---|---|---|
| | Symbol | Correlation | Symbol | Correlation |
| Most correlated | PPC | 0.5035 | SBD | 0.4976 |
| | LTC | 0.5006 | LTC | 0.4706 |
| | DOGE | 0.4740 | GOLOS | 0.4463 |
| | NMC | 0.4678 | EMC2 | 0.4315 |
| | WAVES | 0.4401 | NMC | 0.4281 |
| Least correlated | PASC | -0.0140 | ZOI | -0.1135 |
| | PURA | 0.0029 | GAME | -0.0991 |
| | NYC | 0.0244 | PIVX | -0.0915 |
| | MOON | 0.0248 | EMC | -0.0829 |
| | EXP | 0.0306 | CRW | -0.0681 |

Notes: the ranks are based on magnitudes of correlations.

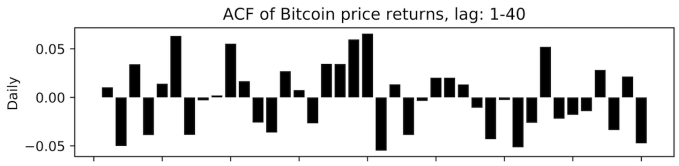


Figure 5: Autocorrelation function of Bitcoin daily price returns (Jan 2015 - Feb 2018)

Notes: the lags range from 1 to 40.

### 3.1.3 Principal Component Analysis (PCA)

To uncover the common drivers of price returns, we employ a popular dimensionality reduction technique - PCA. The starting time of each cryptocurrency varies, thus, to avoid artificially creating biasedness by filling backward on the missing leading values, we select three subsets of time for our PCA analysis and only employ overlapping series. First, we select the 59 cryptocurrencies which have full history back to January 1, 2015. Second, we select the 74 cryptocurrencies with full history back to January 1, 2016. Finally, we select the 100 cryptocurrencies which have returns back to January 1, 2017. We perform PCA for our three periods employing daily price returns.

Figure 6, Figure 7, and Figure 8 present the results for 2015 to Jan 2018, 2016 to Jan 2018, and 2017 to Jan 2018, respectively. In the first and second case, the first principal component captures the majority of the variance, with less variation explained by the other four principal components. In the third case, the period from 2017 to February 2018 the daily returns appear to differ in their structure. Figure 8 displays that the variation explained by the second principal component gains significantly as the first principal component fall to less than 60%.

Clearly, 2017 was a banner year for cryptocurrency and the addition of more retail investors could be one of the explanations of why this period may have a different underlying structure in the return generating process compared to the two other periods. Retail investors became more heavily involved purchasing cryptocurrencies as evidenced by CoinBase having more accounts than Charles Schwab in November 27, 2017[iv]. This changing investor base could possibly bring in more of a herding and momentum behavior if these retail investors are susceptible to known biases similar to those affecting stock retail investors.

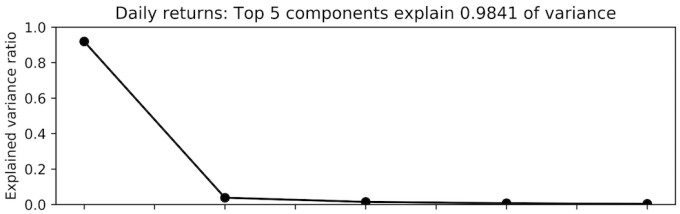


Figure 6: Explained variance ratios for PCA components (58 cryptocurrencies, Jan 2015 - Feb 2018)

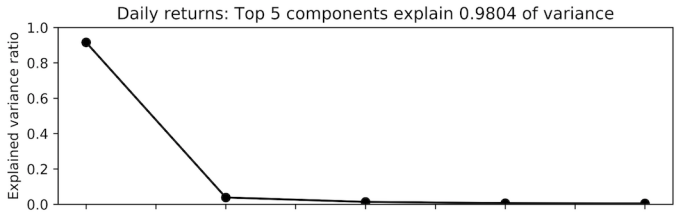


Figure 7: Explained variance ratios for PCA components (73 cryptocurrencies, Jan 2016 - Feb 2018)

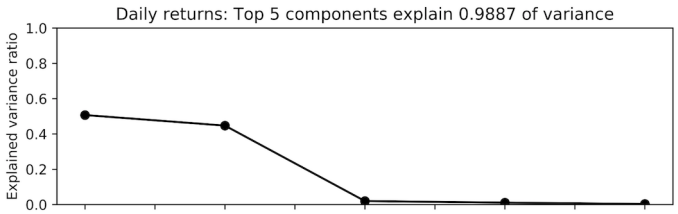


Figure 8: Explained variance ratios for PCA components (100 cryptocurrencies, Jan 2017 - Feb 2018)

### 3.2 Traditional assets

Recent literature [14] shows that Bitcoin provides diversification to portfolio comprised of traditional assets. We dig in and investigate the cross-market relationship between the

top 100 cryptocurrencies and traditional assets. Daily prices of following assets are downloaded from Bloomberg Terminal:

- S&P 500 index (SPX Index): It is a capitalization-weighted index of 500 stocks trading in the U.S. stock market.
- MSCI World Index (MXWO Index): It is a free-float weighted equity index covering stocks trading in developed markets.
- MSCI Emerging Markets Index (MXEF Index): It is a free-float weighted equity index covering large and mid-cap stocks trading in emerging markets.
- US Dollar Index: a measure of the value of the U.S. dollar relative to the value of a basket of currencies of the majority of the U.S.'s most significant trading partners.
- Gold spot price (in US$)
- Bloomberg Commodity Index (BCOM Index): It is an index reflecting commodity futures price movement.
- VIX Index: The measure of volatility implied by S&P 500 index options, calculated and published by CBOE.

Table 9 presents the correlations between Bitcoin, other cryptocurrencies, and traditional assets, calculated in terms of daily returns. Obviously, Bitcoin is barely correlated to any traditional assets at the daily level (absolute correlations < 0.1). It exhibits a slightly positive correlation to S&P 500, MSCI, USD, Gold, and Commo, while demonstrating a negative correlation to Emg and VIX. Not surprisingly Bitcoin is positively associated with the first PCA component and very highly correlated to the market capitalization weighted cryptocurrency returns.

Table 9: Correlations between daily returns of cryptocurrencies and traditional assets (Jan 2015 - Feb 2018)

| | BTC | VW | SP500 | MSCI | Emg | USD | Gold | Commo | VIX |
|---|---|---|---|---|---|---|---|---|---|
| BTC | 1.0000 | 0.9416 | 0.0441 | 0.0232 | -0.0212 | 0.0134 | 0.0419 | 0.0351 | -0.0921 |
| VW | 0.9416 | 1.0000 | 0.0538 | 0.0316 | -0.0204 | -0.0049 | 0.0526 | 0.0359 | -0.0975 |
| SP500 | 0.0441 | 0.0538 | 1.0000 | 0.9093 | 0.4480 | 0.0831 | -0.1674 | 0.2967 | -0.7880 |
| MSCI | 0.0232 | 0.0316 | 0.9093 | 1.0000 | 0.6587 | -0.0413 | -0.1262 | 0.3836 | -0.7283 |
| Emg | -0.0212 | -0.0204 | 0.4480 | 0.6587 | 1.0000 | -0.0426 | -0.0053 | 0.3641 | -0.3848 |
| USD | 0.0134 | -0.0049 | 0.0831 | -0.0413 | -0.0426 | 1.0000 | -0.4070 | -0.2427 | -0.0828 |
| Gold | 0.0419 | 0.0526 | -0.1674 | -0.1262 | -0.0053 | -0.4070 | 1.0000 | 0.2441 | 0.1365 |
| Commo | 0.0351 | 0.0359 | 0.2967 | 0.3836 | 0.3641 | -0.2427 | 0.2441 | 1.0000 | -0.2224 |
| VIX | -0.0921 | -0.0975 | -0.7880 | -0.7283 | -0.3848 | -0.0828 | 0.1365 | -0.2224 | 1.0000 |

Notes: "VW" is the market cap weighted price returns. "MSCI" is the MSCI developed market index. "Emg" is the MSCI emerging market index. "Commo" is the Bloomberg Commodity Index.

### 3.3 Beta-in-the-Tails Analysis (BTA)

In this section we estimate the potential hidden risks in the cryptocurrency markets. In particular, we examine the stability of their betas for Bitcoin and the VW index with respect to the market, which we employ the S&P 500 as a proxy. Edwards and Caglayan [15] document changes in hedge fund correlation in bull and bear markets. Liew [16] introduces the beta-in-the-tail analysis for hedge funds and documents the vanishing diversification benefits as a hidden risk for hedge fund investors. In down periods the beta associated to hedge fund increases and thus decreasing the perceived diversification benefits. Similarly, we find such an occurrence for cryptocurrencies and warn potential investors to be vigilant with regards to the beta-in-the-tail risk.

Upon visual inspection we document the increasing betas in down S&P 500 daily return periods. We argue that beta-in-the-tail is a significant hidden risk for cryptocurrency investors when employing daily returns.

The methodology for daily beta-in-the-tail analysis follows: First, order all the daily returns on the S&P 500 from least to greatest. Associated to each S&P 500-day period we link both the Bitcoin return and MarketCap Weighted Index return for that day. Next, we anchor the worst daily returns for the S&P 500 and use thirty days of returns to run our regressions. That is, we estimate the beta associated with the worst thirty days in our sample period. At this point, it is important to note that the time dimension has been compromised with this sorting of the daily returns.

The regression is the crypto-returns regressed on the S&P 500 returns. Assuming that the risk-free daily returns are zero yields the CAPM's beta of Sharpe [17] and Litner [18] for the given cryptocurrency index. By anchoring the worst return day for the S&P 500 and expanding the window of daily returns we plot the slope coefficients with inclusion of another daily return. When the window has been expanded to include all the daily returns then the final regression corresponds to the beta for the whole period.

The Betas are reported in the left y-axis and the average daily returns for the window period is reported in the right y-axis on the black dashed line. Standard deviation bands surround the beta estimates. Notice that as more observations are included the standard deviation of the beta estimates reduces. The beta-in-the-tails based on daily returns reach above 1.0 compared this to the whole period beta of close to zero for Bitcoin and VW Index, respectively, as seen on the furthest left bottom corner of Fig. 9.

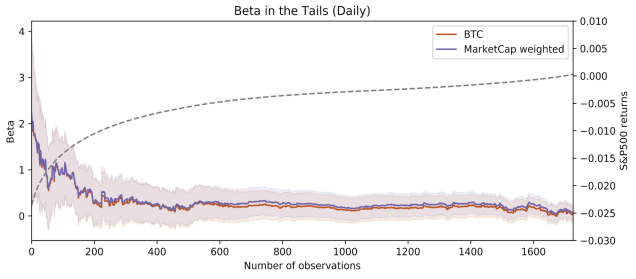


Figure 9: Beta in the Tails (daily)

Notes: Calculated based on daily returns from April 2013 to Feb 2018.

Given that cryptocurrencies trade seven days a week and twenty-four hours a day in contrast to stocks which typically trade only five days a week and six and a half hours a day, we repeat the analysis excluding the weekend in Fig. 10, Beta in the Tail Excluding the Weekends. We arrive at a similar pattern with an increase in the beta in down S&P 500 days. Beta-in-the-tails appears robust to non-trading weekdays.

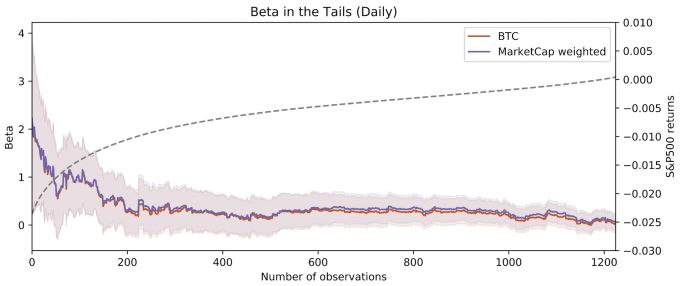


Figure 10: Beta in the Tails (daily, excluding weekends)

Notes: Calculated based on daily returns from April 2013 to Feb 2018.

## 4. Methodology - Rolling Prediction Analysis

In this section, we firstly give a brief introduction to the 11 machine learning algorithms we tested. Next, we describe the way we roll the prediction analysis. Finally, we present our data.

### 4.1 Algorithms

In this subsection, we introduced the 11 machine learning algorithms. Our problem can be easily described with linear models – we have a set of variables (x, a matrix with each column being a variable and each row being value for the corresponding day) such as historical returns, volatility and etc., and a target variable (y, a column vector); and we want to train a model that predicts y with out of sample input x.

There are three strands of algorithms in our analysis: 1) linear models, including LASSO, ElasticNet, Stochastic Gradient Descent, and Bayesian Regression; 2) tree-based models, including Decision Tree, Extra Tree Random Forest, AdaBoost, and Gradient Tree Boosting; 3) other models, including KNN, Support Vector Machine, and Multi-layer perceptron. We briefly introduced each of the algorithms as below.

A typical objective function of linear models is as below:

$$\min_{\omega} \frac{1}{n}\sum_{i=1}^{n} L\left(y_i - f(x_i)\right) + \alpha * R(\omega) \quad (1)$$

where $L$ is loss function, $R$ is regularization term, $f$ is the fitted function.

Least Absolute Shrinkage and Selection Operator (LASSO):

LASSO [19] is a linear model that performs both variable selection and regularization. In contrast to simple linear regression, its objective function is as below. We use the scikit-learn default parameters: squared loss function and L2 regularization with $\alpha = 1.0$.

$$\min_{\omega} \frac{1}{2*n}\|X_\omega - y\|_2^2 + \alpha * \|\omega\|_1 \quad (2)$$

**ElasticNet (EN):**

EN [19] is a linear model that performs regression with both L1 and L2 regularization. This gives it the property of both LASSO and ridge regression, and the objective function is as below. We use the scikit-learn default selection of $\alpha = 1.0$.

$$\min_{\omega} \frac{1}{2*n}\|X_\omega - y\|_2^2 + \alpha * \rho * \|\omega\|_1 + \frac{\alpha*(1-\rho)}{2}\|\omega\|_2 \quad (3)$$

**Stochastic Gradient Descent (SGD):**

SGD [19] is an efficiency method to fit linear models. It searches for minima or maxima through iterations. We use the scikit-learn default parameters: squared loss function and L2 regularization with $\alpha = 0.0001$.

$$\min_{\omega} \frac{1}{n}\|X_\omega - y\|_2^2 + \alpha * \|\omega\|_2 \quad (4)$$

**Bayesian Regression (BR):**

BR [19] provides another way of performing linear regression, where linear model can be written as below:

$$y_i = \alpha + \beta * x_i \; with \; y_i \sim N(\mu_i, \sigma) \quad (5)$$

That is, y follows a normal distribution with mean μ and σ, while μ is a linear function with parameters α and β. In this way, the model can be estimated using maximum likelihood function instead of minimizing squared errors:

$$\max_{\alpha,\beta,\sigma} \prod_{i=1}^{n} N(y_i; \alpha + \beta * x_i, \sigma) \quad (6)$$

**Decision Tree (DT):**

DT [19] is a non-parametric method that can be used for both classification and regression. The tree is built for classifying or predicting test points based on several rules. For classification problems, the leafs of the tree are the classification labels, and for regression problems, the leafs are continuous values. We use the default parameters provided by scikit-learn: using mean square error as splitting criterion, and without max depth of trees.

**Extra Tree Random Forest (ETRF):**

Random forest [19] is an ensemble method built on many trees, and each tree is built through training on a sample of the entire train set with replacement. In addition, when splitting a node during the construction of trees, the best split is measured among a random subset of features rather than all features. This randomness leads to lower variance and larger bias. On the other hand, ETRF moves even further regarding randomness in splitting the nodes – splitting thresholds are randomly assigned instead of searching for the most discriminative thresholds. We use the default parameters provided by scikit-learn: 10 trees without max depth of trees and using mean square error as splitting criterion.

**Adaptive Boosting (AdaBoost):**

AdaBoost [19] is an ensemble algorithm that fits a sequence of relatively weak models with repeatedly modified data. More specifically, it firstly trains on the original train set and assesses the errors. Then it modifies the train set by assigning more weights to poorly modeled points. The processes are repeated for multiple times. Decision Tree is usually used as the base model in AdaBoost. We use the default parameters provided by scikit-learn: 50 Decision Tree models as base estimators.

**Gradient Tree Boosting (GTB):**

Gradient Boosting [19] is another ensemble algorithm that also fits a sequence of relatively weak models with repeatedly modified data. More specifically, it firstly trains on the train set and the original predicted targets. Then it modifies the predicted targets to be certain type of residuals between the true values and the predicted (trained) values. The processes are repeated for multiple times. GTB is the combination of Decision Tree and Gradient Boosting. We use the default parameters provided by scikit-learn: 100 Decision Tree models as base estimators and without max depth.

**K-nearest Neighbor (KNN):**

Typically, KNN [19] method is designed for classification, where discrete labels are determined by the majority of certain amount of nearest data points. However, KNN can also be used for regression where the labels are continuous. The label assigned to a test point is determined based on the mean of the labels of its nearest data points. Scikit-learn provides three methods of searching for nearest neighbors: 1) brute force – compare distances of all pairs of data points; 2) K-D tree – use tree-based structures to reduce the calculations of distances; and 3) ball tree – partition data in a series of nesting hyper-spheres when constructing trees. As scikit-learn supports auto method selection based on input data, we use this option. Also, we use the default parameters provided by scikit-learn: 5 nearest neighbors and uniform weights.

**Support Vector Machine (SVM):**

For regression, SVM [19] finds the classifiers represented by hyperplanes that separate the different groups as wide a margin as possible. The hyperplanes are represented by the normal vector v and the bias b, which can be found by solving a constrained optimization problem:

$$\min_{\omega} \|\omega\| \; A = \pi r^2 \qquad (7)$$

$$s.t. \, y_i * (\omega' X_i - \beta) \geq 1, i = 1, \ldots, n$$

SVM can also be used for regression, where similar kernel method is applied.

**Multi-layer Perceptron (MLP):**

Given a set of features and a target y, MLP [19] can learn a non-linear function estimator for either classification or regression. It trains using backpropagation with no activation function in the output layer, which can also be seen as using the identity function as activation function. Therefore, it uses the square error as the loss function, and the output is a set of continuous values. We use the default parameters of scikit-learn: one hidden layer with 100 hidden units and "relu" as activation function.

### 4.2 Rolling Methodology

We perform rolling prediction analysis. That is, we train our models based on prior historical data and predict future returns. The procedure then rolls forward by expanding the train set by one day and then repeating the training and prediction procedure. A detailed description is as below.

Suppose we stand on day $D_t$, and we want to predict the n-day (n>=1) price returns ahead. To allow the prediction to take place at any time of day $D_t$, we only refer to information up to the previous day $D_{t-1}$. There are two important considerations:

Our predicted variable (y) is calculated as: $R_t = \frac{P_t}{P_{t-1}} - 1$

and our explanatory variables (X), we can only use variables up to day $D_{t-1}$ .For example, the m-day historical return on $D_t$: $HR_{t-m,t-1} = \frac{P_{t-1}}{P_{t-m}} - 1$.

Table 10 provides an example of our data structure.

Table 10: An example of data structure of rolling prediction

| Date | Predicted variable (y) | Explanatory variables (X) | |
|---|---|---|---|
| | n-day returns | Historical m-day returns | Historical k-day moving averages |
| $D_t$ | $P_{t+n}$ / $P_t$ -1 | $P_{t-1}$ / $P_{t-1-m}$ -1 | SUM($P_{t-k}$, …, $P_{t-1}$)/k |
| $D_{t+1}$ | $P_{t+1+n}$ / $P+1_{t+1}$ -1 | $P_t$ / $P_{t-m}$ -1 | SUM($P_{t-k+1}$, …, $P_t$)/k |

Another problem concerning time series rolling analysis is time series leakage. More specifically, standing on day $D_t$, though we have access to historical information (X) up to the previous day ($D_{t-1}$), but we do not have the predicted variable (y), whose calculation involves the close price on day (t+n). That said, standing on day $D_t$, if we want to train a model and predict the n-day returns ahead, the train set can only be constructed based on data from day $D_0$ to $D_{t-n}$ (the predicted variable for $D_{t-n}$ is $R_{t-n} = \frac{P_{t-1}}{P_{t-1-n}} - 1$)

Finally, we repeat our rolling method with a specific example. Suppose we have constructed a time series data set of 1,000 days: the y is a series of 30-day returns and X is a matrix of size 1,000 by 20 (20 explanatory variables). We want to experiment a rolling prediction of 30-day returns. We set the minimum train set size as 100. First, we train a model based on the data from $D_0$ to $D_{99}$ (the predicted variable for $D_{99}$ is $R_{99} = \frac{P_{128}}{P_{98}} - 1$; then we use the trained model to predict the $R_{130} = \frac{P_{159}}{P_{129}} - 1$ based on $X_{130}$ (a 1 by 20 row vector) which contains information up to day $D_{129}$. Next, we expand the train set to include data from $D_0$ to $D_{100}$ and repeat the training and prediction. The analysis is rolled until we get $R_{1000}$.

### 4.3 Explanatory variables

Table 11 shows the explanatory variables in our rolling prediction analysis (predicting 30-day returns for Bitcoin). Based on the preliminary analysis above, we decide to exclude USD index, gold, and VIX, due to their relatively low correlations with Bitcoin. The variables are constructed in the abovementioned rolling way and standardized using StandardScaler in scikit-learn, which centers the data with sample mean and the scales them into unit variance.

In addition, we categorize these variables into eleven "information sets". In the later sections, we will examine the relative importance of each information set for Bitcoin, in terms of their contribution to the performance of our machine learning algorithms.

## 5. Model Results

### 5.1 Rolling prediction analysis (30-days) for Bitcoin

We recalculate predicted prices based on predicted 30-day returns, as is shown in Figure 11. As the ill-performance of Multi-layer perceptron during the second half of 2017 leads to poor readability, we present results of the top 3 algorithms (in terms of accuracy) from Jan 2017 to Jan 2018 in Figure 12. Obviously, none of them successfully forecasted the big price crash in Jan 2018. On the other hand, Figure 13 and Figure 14 show the accuracy and RMSE, respectively, both of which are calculated in a cumulative way (expanding the data by one prediction for each time). As the number of predictions increases, accuracy of all algorithms stabilizes in the range of 50 to 65 percent.

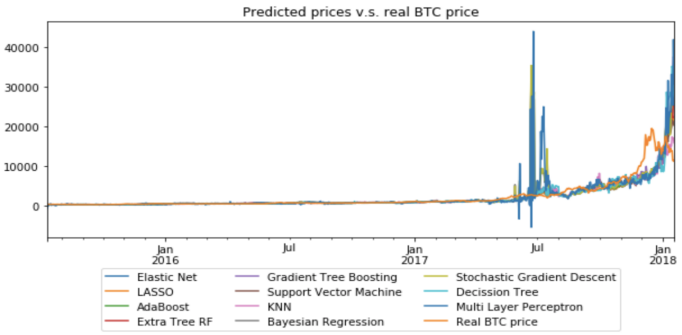


Figure 11: Predicted price vs. Real BTC price (predicting 30-day returns)

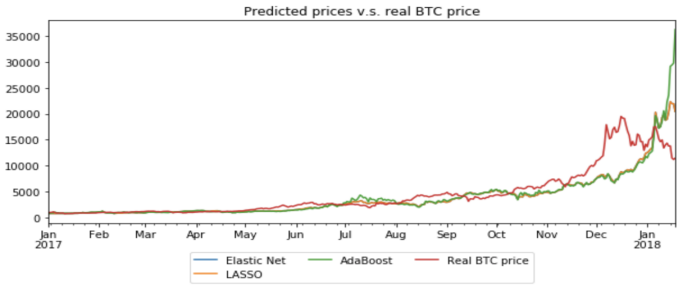


Figure 12: Predicted price vs. Real BTC price (predicting 30-day returns)
Notes: This figure shows results from Jan 2017 to Feb 2018 for the top 3 algorithms (in terms of accuracy).

### 5.2 Important information sets for Bitcoin

As stated above, to reveal the potentially useful information sources in predicting Bitcoin prices, we categorize all variables into 10 information sets: 1) price returns, 2) price momentum, 3) rolling volatility, 4) volume, 5) S&P 500, 6) Developed equity market, 7) Emerging equity market, 8) commodity, 9) market capitalization weighted returns of cryptocurrencies (crypto VW), and 10) the 30-day rolling correlation of the overall cryptocurrency market (rolling volatility).

We first run the rolling prediction analysis with all information sets as input, and next, we repeat the analysis for 10 times by removing one information set each time. The "relative importance" of each information set is measured as the difference between the accuracies with and without the corresponding information set as input. That is, a positive difference indicates positive contribution of the information set and negative difference implies the opposite.

Figure 15 shows the heatmap presenting the relative importance of each information set for each algorithm. Overall speaking, none of the information sets has significant impact on any algorithms, as the relative importance fall in the range between -0.05 and 0.05. However, a closer inspection would reveal that, on average, rolling volatility (past 15 days and 30 days) and correlation among cryptocurrency market (past 30 days) are useful information for most algorithms, while the market

Table 11: Explanatory variables

| | Variable name | Definition | Information set |
|---|---|---|---|
| 1 | Price_ret10 | Historical 10-day price returns | Historical price returns |
| 2 | Price_ret30 | Historical 30-day price returns | |
| 3 | Price_momentum_MA10 | The ratio of price to 10-day moving average minus 1 | Price momentum |
| 4 | Price_momentum_MA30 | The ratio of price to 30-day moving average minus 1 | |
| 5 | Volume_momentum_MA10 | The ratio of trade volume to 10-day moving average minus 1 | Volume Momentum |
| 6 | Volume_momentum_MA30 | The ratio of trade volume to 30-day moving average minus 1 | |
| 7 | Price_volatility15 | The standard deviation of the daily price returns over the past 15 days | Rolling volatility |
| 8 | Price_volatility30 | The standard deviation of the daily price returns over the past 30 days | |
| 9 | SP500_ret15 | S&P500 historical 15-day price returns | S&P 500 |
| 10 | SP500_momentum_MA15 | The ratio of price to 15-day moving average of S&P500 minus 1 | |
| 11 | Developed_ret15 | MSCI developed equity market historical 15-day price returns | Developed equity market |
| 12 | Developed_momentum_MA15 | The ratio of price to 15-day moving average of MSCI developed equity market minus 1 | |
| 13 | Emerging_ret15 | MSCI developing equity market historical 15-day price returns | Emerging equity market |
| 14 | Emerging_momentum_MA15 | The ratio of price to 15-day moving average of MSCI developing equity market minus 1 | |
| 15 | Commodity_ret15 | Bloomberg Commodity Index historical 15-day price returns | Commodity |
| 16 | Commodity_momentum_MA15 | The ratio of price to 15-day moving average of Bloomberg Commodity Index minus 1 | |
| 17 | VW_returns10 | 10-day market-cap weighted returns 57 cryptocurrencies * | Market capitalization weighted returns of cryptocurrencies |
| 18 | VW_returns30 | 30-day market-cap weighted returns 57 cryptocurrencies * | |
| 19 | PC1 ** | The first principal component of PCA on x-day returns of 57 cryptocurrencies * | Principal components of cryptocurrencies |
| 20 | PC2 ** | The second principal component of PCA on x-day returns of 57 cryptocurrencies * | |
| 21 | Crypto_corr30 | The average correlation between the predicted coin and other cryptocurrencies over the past 30 days | Rolling correlation of the overall cryptocurrency market |

Notes:

1. * All the "57 cryptocurrencies" above means the 57 cryptocurrencies which have full data back to January 1, 2015.

** The PCA is conducted in a rolling base.

capitalization weighted historical returns (15-day and 30-day) and emerging equity market are the least beneficial.

### 5.3 Rolling prediction analysis for other Cryptocurrencies

We also examine the analysis for the 57 cryptocurrencies with available data back to January 1, 2015. Many cryptocurrencies are slightly predictable if the algorithms with the highest accuracies are chosen. Bitcoin yields the highest best accuracy as displayed in Fig. 14 below. Another finding is that higher

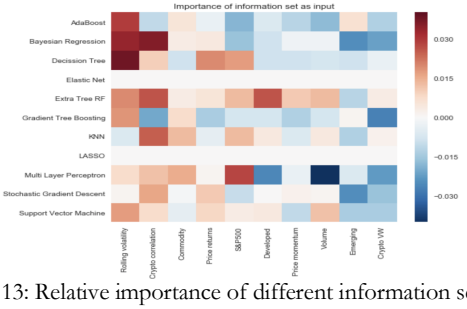


Figure 13: Relative importance of different information sets on predicting 30-day Bitcoin returns

prediction accuracy is associated with larger market capitalization and lower volatility. But we also see that higher predictability is accompanied by larger dispersion among different algorithms.

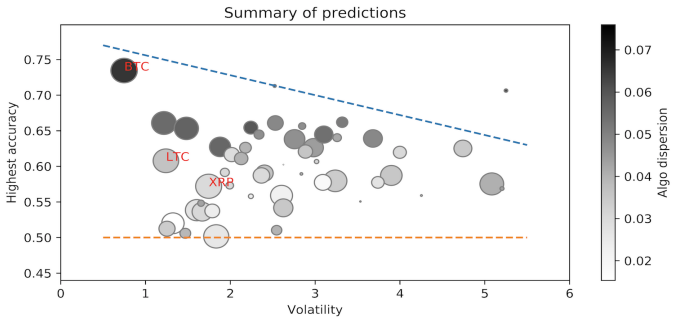


Figure 14: Summary of rolling prediction results (predicting 30-day returns)

Notes:

1. The volatility is calculated by annualizing the daily volatility over the sample period (Jan 1, 2015 - Feb 18, 2018). We limit the range of x-axis to be [0, 6] for the purpose of readability, and as result 8 cryptocurrencies are removed from the figure.
2. The highest accuracy: we run 11 algorithms for each cryptocurrency and pick the one with highest accuracy.
3. The size of dots is based on the market capitalization of each cryptocurrency, i.e., Bitcoin is the largest.
4. The color of dots is based on the standard deviations of accuracies generated by 12 algorithms (algo dispersion).

Fig. 15 presents a performance summary of the 12 algorithms. LASSO dominates in predicting the 30-day returns of cryptocurrencies. And one average, all algorithms generate accuracies in the range of 50 to 60 percent, which is above random guess but still far from accurate prediction.

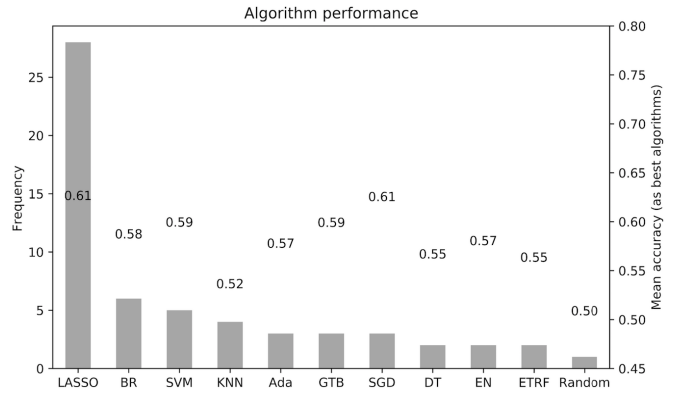


Figure 15: Summary of algorithm performance (predicting 30-day returns)

Notes:

1. The frequency is the times an algorithm performs the best among the 11 algorithms plus random guess.
2. The mean accuracy is calculated by averaging the accuracies when the corresponding algorithm performs the best.

## 6. Conclusion

Cryptocurrencies have captured the attention of many investors across the spectrum from retail to institutional - see Liew and Hewlett [14]. In this work we extend our understanding of the behavior of cryptocurrencies. We document several interesting findings. First off, we find that PCA reveals that the return generating process is much more complex than that for stock returns. Generally speaking, the financial community agrees that the "market" is the first dominant PCA in stock returns. However, for cryptocurrencies daily returns reveals that in some period there exists a single dominant component however, in the most recent prior year there appears to be two components that help explain the variation of the cryptocurrency returns. Next, we document a strong beta-in-the-tails hidden risk associated with Bitcoin daily returns. Similar to hedge fund cryptocurrencies may have some unstable tail behaviors.

Our analysis of machine learning algorithms applied to the data from cryptocurrencies hints that predictability may be difficult and there are many heterogeneous effects here. Some information sets perform better with some family of algorithms, and larger cryptocurrencies with lower volatility maybe more predictable than smaller cryptocurrency with higher volatility. Some care should be taken given the many moving parts across the cryptocurrency industry. The complexity will lead to possible risks of overfitting machine learning algorithms.

## References:

[1] I. Madan, S. Saluja, and A. Zhao, "Automated bitcoin trading via machine learning algorithms," URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan, vol. 20, 2015.

[2] A. Greaves and B. Au, "Using the bitcoin transaction graph to predict the price of bitcoin," No Data, 2015.

[3] S. McNally, J. Roche, and S. Caton, "Predicting the price of Bitcoin using Machine Learning," in 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), 2018, pp. 339–343.

[4] Z. El-Abdelouarti Alouaret, "Comparative study of vector autoregression and recurrent neural network applied to bitcoin forecasting," PhD Thesis, ETSI_Informatica, 2017.

[5] F. Mai, Q. Bai, J. Shan, X. S. Wang, and R. H. Chiang, "The impacts of social media on Bitcoin performance," 2015.

[6] E. Stenqvist and J. Lönnö, Predicting Bitcoin price fluctuation with Twitter sentiment analysis. 2017.

[7] Y. B. Kim, J. Lee, N. Park, J. Choo, J.-H. Kim, and C. H. Kim, "When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation," PloS one, vol. 12, no. 5, p. e0177630, 2017.

[8] J. Kaminski, "Nowcasting the bitcoin market with twitter signals," arXiv preprint arXiv:1406.7577, 2014.

[9] J. Chu, S. Nadarajah, and S. Chan, "Statistical analysis of the exchange rate of bitcoin," PloS one, vol. 10, no. 7, p. e0133678, 2015.

[10] M. Balcilar, E. Bouri, R. Gupta, and D. Roubaud, "Can volume predict Bitcoin returns and volatility? A quantiles-based approach," Economic Modelling, vol. 64, pp. 74–81, 2017.

[11] N. Indera, I. Yassin, A. Zabidi, and Z. Rizman, "Non-linear autoregressive with exogeneous input (NARX) Bitcoin price prediction model using PSO-optimized parameters and moving average technical indicators," Journal of Fundamental and Applied Sciences, vol. 9, no. 3S, pp. 791–808, 2017.

[12] I. Georgoula, D. Pournarakis, C. Bilanakos, D. Sotiropoulos, and G. M. Giaglis, "Using time-series and sentiment analysis to detect the determinants of bitcoin prices," Available at SSRN 2607167, 2015.

[13] D. Garcia and F. Schweitzer, "Social signals and algorithmic trading of Bitcoin," Royal Society open science, vol. 2, no. 9, p. 150288, 2015.

[14] J. K.-S. Liew and L. Hewlett, "The case for Bitcoin for institutional investors: Bubble investing or fundamentally sound?," Available at SSRN 3082808, 2017.

[15] F. R. Edwards and M. O. Caglayan, "Hedge fund performance and manager skill," Journal of Futures Markets: Futures, Options, and Other Derivative Products, vol. 21, no. 11, pp. 1003–1028, 2001.

  

[16] J. Liew, "Hedge fund index investing examined," Journal of Portfolio Management, vol. 29, no. 2, p. 113, 2003.
[17] W. F. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," The journal of finance, vol. 19, no. 3, pp. 425–442, 1964.
[18] J. Lintner, "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets: A reply," The review of economics and statistics, pp. 222–224, 1969.
[19] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[i] The data of Ethereum provided by coinmarketcap.com starts on Aug 7, 2015.

[ii] The average history is calculated using the data for only 2015 to 2017, thus it is not the exact length of average history. But as most of the top 100 cryptocurrencies came into being after 2015, this calculation approximates the real length of average history.

[iii] For horizontal axis, cryptocurrencies are ranked by market capitalizations from the right (large) to the left (small). For vertical axis, they are ranked by market capitalizations from the top (large) to the bottom (small).

[iv] Accessed on Mar 14, 2018: https://www.cnbc.com/2017/11/27/bitcoin-exchange-coinbase-has-more-users-than-stock-brokerage-schwab.html