Bitcoin Daily Directional Price Prediction

Project Report

Dianna Chen
ESSEC Business School & CentraleSupélec
dianna.chen@essec.edu

Quentin Boutoille-Blois
ESSEC Business School & CentraleSupélec
quentin.boutoilleblois@essec.edu

Victor Terras-Dober
ESSEC Business School & CentraleSupélec
victor.terrasdober@essec.edu

Adel Nehme
ESSEC Business School & CentraleSupélec
adel.nehme@essec.edu

ABSTRACT

In recent years, the cryptocurrency markets have become rife with high-frequency traders largely preoccupied by technical strategy and order flow. Marshalled by the relative reliability of such approaches, traders and bots dance around each other in collective pursuit of minute gains. As Bitcoin and other digital assets gain further momentum, the intensity of this high-frequency rat-race will only amplify. The challenge arises - is there bluer ocean beyond these shark-infested water. Our approach entailed leveraging data on the Bitcoin blockchain to develop a classification model that predicts the direction of Bitcoin prices. Using a rolling basis differenced time series cross validation lasso logistic regression, we were able to predict with 80% accuracy the direction of Bitcoin prices. Furthermore, taking inspiration from existing literature, we investigated whether social signals can potentially have an impact on the direction of Bitcoin prices, by incorporating Reddit subscriber data in our approach. We conclude with a brief discussion on the future of Bitcoin and potential predictive variables that can influence our model.

CCS CONCEPTS

• **Computing Methodologies** → **Machine Learning**; *Artificial Intelligence*; • **Machine Learning Approaches** → Classification and Regression Trees;

KEYWORDS

Bitcoin, Cryptocurrency, Long Short-Term Memory Networks, Logistic Regression, Support Vector Machine, Directional Prediction, Time Series, Cross Validation

1 INTRODUCTION

1.1 What is Bitcoin

The ostensible purpose of our exploration is to predict daily directional movement of Bitcoin prices. Bitcoin is a digitally native asset which in recent months has accumulated evident symbolic recognition among governments and financial institutions alike. [3] Trends in recent weeks would suggest the beginning of parabolic growth in interest amongst retail investors and popular media, who are largely drawn by historic, unprecedented opportunities for speculative action. The asset was introduced in late 2008, and unleashed onto the world through a seminal whitepaper by a group of programmers under the pseudonym Satoshi Nakamoto. Bitcoin is a digital asset derived from mathematical cryptography and conceived as an alternative to government-backed fiat currencies. [3] Over the last 8 years, hundreds of alternative cryptocurrencies (otherwise termed altcoins) have emerged, each seeking to establish a decentralized economic universe of its own. Indeed, beyond the basis of technological superiority as it relates to characteristics of currency itself (ie privacy, scalability, transaction speed, etc), altcoins are driving the shift towards cryptocurrencies as a displacement of more secure, more efficient, and more trustworthy economic structures. Thus, while it was originally conceived as a method of payment, Bitcoin's symbolic status amongst retail investors has turned it into speculative target of historic proportions, superseding the coin's original transactional intent. [10] In this paper we seek to better understand the behaviour and speculative dynamics of bitcoin in relation to elements of its underlying microstructure.

1.2 Motivation

The natural question is, why? Why construct a model aimed at directional prediction for an asset class yet unregulated, yet immature, and yet tiny in the greater landscape of traditional market capitalizations? The reason lies in the significance of bitcoin on both a behavioural and historic level. Before we proceed, it is important to acknowledge that the immaturity of cryptocurrencies in themselves, paired with the speculative nature of bitcoin, engenders little fundamental consensus on bitcoin value formation nor behaviour. [4] Everything observed thus far, and everything the market will continue to observe over the coming years, is novel. Thus, the reader must be aware that a plethora of opinions exist, more than one of which is likely significant.

Observing price action from a behaviour standpoint, bitcoin's path since inception can be considered a compelling social experiment. How do people react to the "foreign" notion of digital currency over time? How does the perception of risk interact with unprecedented greed? More, to what extent do market dynamics depart from changes in bitcoin microstructure? By observing bitcoin's directional predictability, we indirectly observe all of these dynamics unfold.

1.3 Applications

The practical value of our model is concentrated in two areas. Primarily, such a predictive model may be used for automated swing trading. Swing trading encapsulates short-term trading strategies for positions which typically last from 2 to 6 days, and which may last as long as 2 weeks.

Secondly, in a broader sense, we can observe the performance of our algorithm as a proxy for how sensitive bitcoin price is to underlying microstructure fundamentals (as much as such fundamentals can be defined at this stage of market immaturity). As such, we interpret the performance of our model as a measurement of bitcoin's evolution from a store of value to an asset with parabolic speculative properties.

The second point is especially significant because of trends which have emerged over the course of recent weeks. Namely, bitcoin's year-to-date gain has been over 1350%, as it continues to shatter all expectations of price ceiling at the time of writing. New capital inflows from institutions and retail investors, heavy reallocation of South Korean capital stemming from the country's ban on ICOs, and the increasing magnitude of Bitcoin's investment opportunity cost have contributed to the accelerating price bubble observed thus far. In terms of emotion, greed, and fear. The market is at an all time high. Thus, present-day conditions offer a rich context in which to conduct our experiment.

2 PROBLEM DEFINITION

Our task is to classify the daily price direction of Bitcoin Y given a set of features X pertaining to the Bitcoin blockchain as well as the daily increase in the Bitcoin Subreddit [19] as a means to capture sentiment. Our initial assumptions going into the data is that Bitcoin's price appreciation over the past few years is directly measurable through structural changes in the Bitcoin microstructure and through evolving characteri.
Our classification's performance is measurable through a number of metrics. Since this problem is a time series binary classification one, and that state of the art performance on this specific issue has been calculated using these metrics, we will measuring performance using the following:

(1) True Positives (TP): Number of instances where the model predicted positive price direction where the price actually went up.
(2) True Negatives (TN) = Number of instances where the model predicted negative price direction where the price actually went down.
(3) False Positives (FP) = Number of instances where the model predicted positive price direction where the price actually went down.
(4) False Negatives (FN) = Number of instances where the model predicted negative price direction where the price actually went up.
(5) Sensitivity or Recall or True Positive Rate $TPR = \frac{TP}{TP+FN}$
(6) Specificity or True Negative Rate $TNR = \frac{TN}{TP+FN}$
(7) Precision or Positive Predictive Value $PPV = \frac{TP}{TP+FP}$
(8) False Discovery Rate $FDR = \frac{FP}{FP+TP}$
(9) Accuracy:

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}$$

(10) F1-score or Harmonic mean of precision and recall

$$F1 = \frac{2TP}{2TP + FP + FN}$$

3 RELATED WORK

Given that Bitcoin and cryptocurrencies in general are a relatively new phenomenon, research for Bitcoin price direction is widely underdeveloped. Yet, a different number of papers have discussed various methodologies for Bitcoin price prediction, or direction movement. These approaches different in two fundamental aspects; choice of data, and choice of model. The choice of data can be clustered into two broad categories; sentiment-based data and non sentiment-based data. Sentiment-based data pertains to public sentiment regarding Bitcoin price, whereas non sentiment-based data pertains to directly observed quantitative characteristics such as price and characteristics of Bitcoin microstructure. Shah et al. [21] developed an automated Bitcoin trading client using Bayesion regression. Bayesion regression refers to utilizing empirical data as a proxy to perform Bayesian inference. They were able to reach an 89% return on investment after 50 days of trading, however their success was largely rooted in having two-second interval pricing data, with predictions done over specific hand-picked clusters capturing high price variation and derived from the kâĹŠmeans algorithm.

Furthermore, Pournarakis et al. [15] developed a sentiment-based approach to assess the correlation between public sentiment indicators and BitcoinâĂŹs price. Their methodology entailed developing a vector error-correction model to capture a long-term positive correlation between daily Bitcoin price and the frequency of daily Bitcoin Wikipedia views and Bitcoin Twitter sentiment derived from utilizing support vector machines on raw Twitter Data. They found both Wikipedia views, their Twitter sentiment ratio and other structural features of the Bitcoin Blockchain had positive effects on the Bitcoin price. Using a sentiment based approach as well, Colianni et al. [5] developed a purely sentiment driven approach using logistic regression, naive bayes and support vector machines purely on twitter and twitter sentiment data pertaining to Bitcoin. They managed to acheive impressive performance at 95% accuracy using naive bayes when using tweets as feature vectors using Porter stemmings, but however they did not perform cross validation as well, leaving their results in the same sphere of skepticism as Madan et al. [12]

More interestingly, Kim et al. [11] developed an interesting method by crawling through dedicated forums for cryptocurrencies, and managed to perform sentiment analysis on the user posts and comments in order to capture the sentiment of knowledgeable retail investors on the price of Bitcoin. Their method entailed performing a 10 fold cross validation with a 6 day lag using a time series model incorporating two linear models that they designed and gave an admirable score of 71%. However, a drawback in their method is that it may not be sustainable as Bitcoin is continuously evolving over time and the role of public sentiment in driving the prediction may wane in the long term. Another drawback of using a sentiment based approach is discussed by McNally [14], he posits that a model designed around sentiment can be subject to mass manipulation via misinformation on social media platforms.

Greaves et al. [9] developed a networked a based approach to extract features from the Bitcoin Blockchain. Their data consisted of a network of over 37 million Bitcoin transactions with the key objective of extracting features of the Bitcoin network topology. Their research goal was to pedict Bitcoin price's an hour in advance, and were able to marginally outperform the market scoring 55.1% using ANN. They also utilized Logisitc Regression and scored 54.3%, which was their baseline method of choice. Their main takeaway from their study is that their methodology has exhausted all the predictive value in the Bitcoin network and the transaction behavior accompanying it. Additionally, they posit that true predictive solutions for Bitcoin price direction lie outside the Bitcoin blockchain and the network of transactions attached to it.

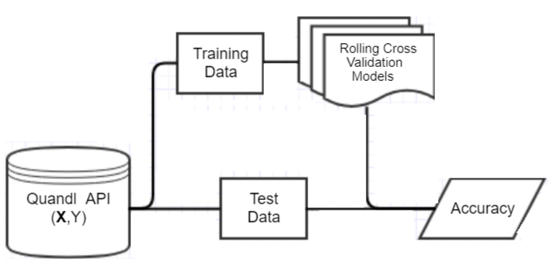
What can be considered "state of the art" performance is Madan's et al. [12] paper on Bitcoin price direction movement. Their data exclusively consisted of features related to the Bitcoin blockchain, price and payment network. Their methodology can be a characterized in a two phase modeling process where the first phase consists of utilizing Logisitic Regression to predict the price direction of Bitcoin using five years worth of daily Bitcoin blockchain data, providing them an accuracy score of 98.7%. The second phase consists of utilizing the same models but on ten minutes and ten second intervals, yielding them a performance of 50-55%. Our objective was to emulate the first phase of their methodology, however we were not able to recreaete their results due to two primary factors. One of the lesser important factors, is that Bitcoin is fundementally changing over time at a rapid pace. Increased speculation has made it more volatile and Madan et al. captured patterns in

the Bitcoin blockhain at a fundementally different time in it's maturity. More importantly, Madan et al. failed to cross validate his results. Thus, risk of overfitting is considerably high. We have encountered similar overfitting problems when trying to emulate their approach, scoring 90% on our test set using a ridge logistic regression, but scored abysmally on other test sets.
McNally [14] made the same observation in his research on predicting the price of Bitcoin using long short term memory networks and recurrent neural networks. His choice of data were mainly Bitcoin price data, specifically opening and closing price of Bitcoin in USD. His approach using Long Short Term Memory yielded him an accuracy score of 52.78%, barely outperforming the market baseline ARIMA prediction by 2.72%.

4 METHODOLOGY

4.1 Overview

Our methodology entails a simple yet statistically correct pipeline which splits our data between test and training set, performs a rolling basis cross validated time series forecasting and we derives performance on the test set.



4.2 Data

4.2.1 Data Source. All of our blockchain data was sourced from the Quandl API [18]. Reddit subscribers were scraped from the Reddit website [19].

4.2.2 Features Used.

(1) Number of Bitcoin transactions/day: Number of Bitcoin transactions/day
(2) Number of Bitcoin users: Number of Bitcoin users
(3) Volume of Bitcoin transactions: Volume of all transactions
(4) Bitcoin Average Block Size: Average Blockchain size
(5) Total Bitcoin Block Size: Total Blockchain size
(6) Bitcoin Transaction Confirmation Time: Amount of time needed to confirm a transaction
(7) Bitcoin Miner Revenue: Revenue made by miners
(8) Bitcoin Hash Rate: Computational resources needed to mine a bitcoin
(9) Bitcoin Cost per Transaction: Cost per Bitcoin transaction
(10) Bitcoin cost percentage on transaction volume: Proportion of Cost relative to total transaction volume
(11) Bitcoin USD trade volume: Total Volume in USD
(12) Bitcoin Estimated Transaction Volume: Estimated transaction volume
(13) Bitcoin Total Output Volume: Bitcoin Total Volume
(14) Bitcoin Transactions/Block: Number of transactions per block
(15) Bitcoin Number of Adresses Used: Number of unique devices on the network
(16) Bitcoin Non Popular Adresses: Non popular adresses
(17) Bitcoin Total Transactions: Total transactions on the chain in USD
(18) Bitcoin Number of Transactions: Number of transactions on the chain
(19) Bitcoin Transaction fees: Transaction Fees
(20) Bitcoin Market Cap: Bitcoin Market Capitalization
(21) Bitcoin Mining Operating Margin: Operating margin of all miners
(22) Reddit Subcribers: Bitcoin Subreddit subcribers

It is essential to note that feature selection techniques involving variance threshold or manual domain knowledge feature selection led to a decrease in performance throughout the course of our modeling.

4.3 Stationarity

Stationarity constitutes an important part of time series forecasting, a stationary process has the property that the mean, variance and autocorrelation structure do not change over time. In the case of non-stationary processes where statistic properties change over time it becomes very difficult to predict movements in the time series using conventional techniques. In this condition, it is always best to study stationary processes or to

render the process stationary. Most economic time series are far from stationary even after some seasonality adjustments and obvious pre-processing. So, it has become common to look for long term trends and reverse it using mathematical transformations. Bitcoin price is no exception to this, it even follows an exponential trend, as scholars argue whether it is a bubble or not [4], the price has risen 13 times in the course of 2017 and it only took 90 days to go from $4,000 to $8,000.
A common process to verify if a time series is stationary is to do a unit root test, which means try and find if the time series variable has a unit root, if there are none then the process is stationary else it is not. One of the most well-known and used test is the Augmented Dickey fuller test which tests the correlation between two successive observations in a time series and tries to determine the presence of a unit root in the autoregressive equation.
Thus, to resolve the issue of stationarity, several solutions were at our disposal. Using price as an example, we first looked at the returns instead of price but the variance was still not stationary. Finally, we found 2 effective ways to make the series stationary: one is to study the difference of the logarithm of the price and a second is to leverage the price direction. We used this technique on all our data set, utilizing the direction of the difference for all our features.

4.4 Cross Validation

4.4.1 Cross Validation in Related Work. What can be considered state of the art in the problem of Bitcoin price direction prediction was Madan's et al. [12] paper where they achieved 98.7% accuracy using Logistic Regression. Initially, this paper was the basis of our undertaking. However, we learned that their method did not employ any cross validation and have probably suffered from overfitting. An attempt to recreate their approach led us to solidify this conclusion, as replication of their method led us to poor results.

4.4.2 Rolling Basis Cross Validation. Our time series cross validation technique consisted of a rolling basis cross validation. We used a three hundred day training set with a 20 day cross validation window, with final scores being predicted on a 10 day test set (*the choice of training set was due to the structural changes Bitcoin has endured in the past year, whereas the choice of test set was shrunk in order to control for accuracy in a time series forecast.*). In this procedure, the training set consists of observations that occured prior to the cross validation set, the cross validation accuracy is computed by averaging over all forecasts, and then the final prediction accuracy is computer on the test set. An illustration of this technique can be found below

Figure 1: Augmented Dickey Fuller Test on Price Data

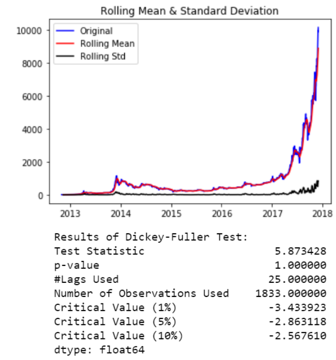


Figure 2: Augmented Dickey Fuller Test on Price Direction Data

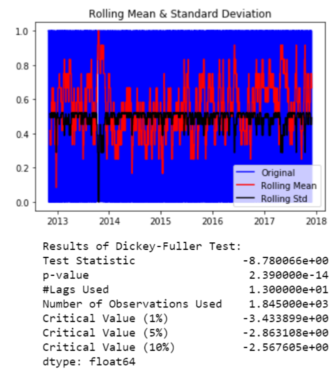
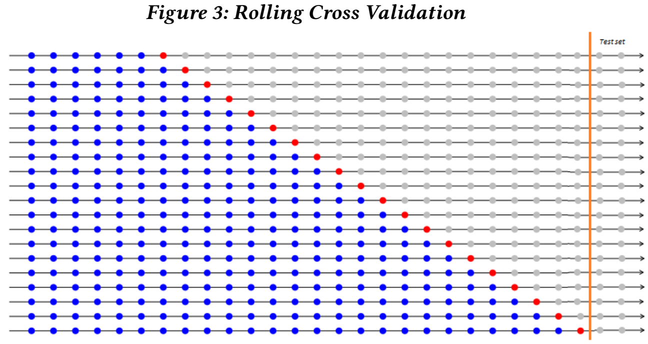


Figure 3: Rolling Cross Validation



4.5 Models Used

4.5.1 Logistic Regression. Logistic Regression is one of the most widely used machine learning models in binary classification. Logistic Regression describes the relationship between a binary dependent variable y (usually given as either $+1$ or -1), and a set of independent variables (in this case our n-dimensional feature vector $x = (X_1, X_2, ..., X_n)$. The goal of Logistic Regression is to examine the training data, and determine the best separation between data points belonging to each class of the target variable y. At the heart of Logistic Regression is estimating the log odds of an observation belonging to a class y, in the case of an n-dimensional feature vector, the function is defined by

$$\log(\frac{p(y=1)}{1-p(y=1)}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

The choice of parameters β is calculated by minimizing the negative log likelihood, the quantity one minimizes, given m training samples is thus:

$$L(\beta) = p(y|X;\beta) = \prod_{i=1}^{m} p(y^i|x^i;\beta)$$

4.5.2 Lasso Logistic Regression. Lasso logistic regression has the similar problem formulation as logistic regression, with the sole difference being an addition to the minimization function that adds a penalty term to the log likelihood. The penalty term is designed to minimize the out of sample error.

$$L + \lambda \sum_{i=1}^{n} |\beta_i|$$

Where λ is a free parameter which controls the degree of penalty used.

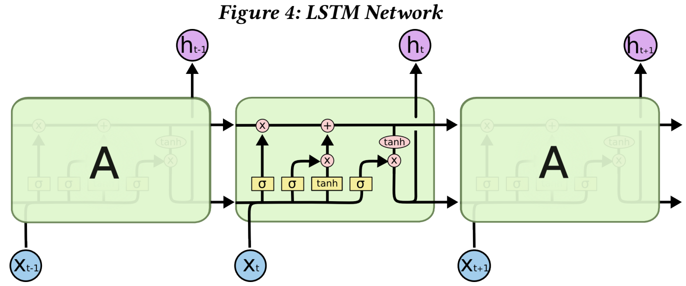
4.5.3 Ridge Logistic Regression. Ridge logistic regression has the similar problem formulation as lasso logistic regression, with the sole difference being a slight to change to the penalty term added to the log likelihood. The penalty term is designed to shrink the coefficients that contribute most to the error.

$$L + \lambda \sum_{i=1}^{n} \beta_i^2$$

Where λ is a free parameter which controls the degree of penalty used.

4.5.4 Long Short Term Memory Network. Long short term memory networks are a unique kind of Recurrent Neural Networks that are optimized to learn long term dependencies. They are particular suited to learning behavior and retaining information for a long period of time and it is inherent in their behavior. A brief overview of LSTM networks can be found below

Figure 4: LSTM Network



4.5.5 Random Forest. Random forest is an ensemble of binary decision trees which avoids overfitting. Generally deep decision trees will have low bias and a high variance because of the way they are constructed. Random forest is a way of averaging multiple deep decision trees, with the goal of reducing the variance with the trade-off of increasing the bias and reduced interpretability. Random forest works in two steps. The first one is bootstrapping to reduce the variance of the classification, we pick randomly with replacements m elements of the training set and we do this b times. This step is very important because it reduces the variance without augmenting the bias. The second step is called feature bagging, it means that instead of running the same tree decision algorithm on each of the b bags, the features selected are going to change randomly for each of the b bootstraps. Finally, the output of the algorithm will be constituted of the mode of the classes, that is to say the classes that appear the most.

5 MODELS PERFORMANCE

Our modeling consisted of three phases, first was an attempt to emulate Madan's et al. [12] approach using logistic regression to attain a 90% range accuracy. We were successful at reaching a test score of 90% without rolling cross validation and differencing, however we clearly understood that our model was not able to generalize well on other test sets. This was clearly due to overfitting in their approach.

The second and third phases were an attempt to perform a sets of models, one on 300 days worth of differenced training data with a 20 day cross validation window, with accuracy derived from a rolling forecast on a 10 day horizon test set. The other on 70 days worth of differenced training data also with a 20 day cross validation window, with accuracy derived from a rolling forecast on a 10 day horizon test set. The highest performing model was lasso logistic regression on the 300 day training data, scoring us a test score of 80%, with a poorer performance derived from the 70 day training data. Ridge logistic regression scored marginally lower, at 70%. Both LSTM and random forest scored considerably lower, showcasing an accuracy of 40% and 70 % respectively with the 300 days training set and even lower with the 70 days training set. The rationale behind the third phase of the project was to ascertain whether structural changes in the Bitcoin price direction has any effects on our predictive ability. However, a golden rule emerges for all of our models: More data means better accuracy.

Table 1: Test set performance with 300 days training set

	Accuracy	Precision	Recall	F1 Score
LSTM	40%	40%	40%	40%
Random Forest	70%	49%	70%	58%
Ridge Logistic Regresison	70%	49%	70%	58%
Lasso Logistic Regression	80%	88%	80%	81%

6 TOOLS USED

(1) Pandas [16]: An open source library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.
(2) Numpy [7]: Open source package for scientific computing with Python. Primarily used for its powerful N-dimensional array objects.
(3) Matplotlib [6]: A Python 2D plotting library which produces publication quality figures.
(4) Searborn [17]: A visualization library based on the Matplotlib library.
(5) Scikit-Learn [20]: One the most popular packages for Machine Learning libraries on Python, fully compatible with Numpy and Pandas, it offers accessible efficient tools for Machine Learning and Data Analysis.
(6) Keras [1]: One of the most popular packages for developing deep neural networks.

7 FINAL DISCUSSION

The dynamics of this parabolic state of affairs are intrinsically and deeply embedded with emotion. Thus, while we have taken a micro-structure approach, we believe that incorporating measures of market and media sentiment can significantly improve our model's performance. [13]
Popular characterizations of Bitcoin price as an asset price bubble peripherally on the cusp of collapse are dramatizations of a notion widely acknowledged in institutional and retail spaces alike. However, while speculative behaviour is currently the primary driver of Bitcoin price, it may not continue to be so in the coming years, and a sentiment-driven approach will thus likely be relevant only until the bubble reaches full collapse. Based on a generalized conception of asset price bubbles and their structure, as the bitcoin price reaches its top, it may begin a jagged descent, prompting market participants to take profit and move elsewhere. At some distant point in the future, Bitcoin's symbolic value may be pierced, and as it's core development team works to improve its technical foundation, the coin may eventually be re-branded towards a transactional currency. Regardless, throughout such a cycle, sentiment is the key driver. [8] Indeed, bitcoin microstructure pre-disposes the coin towards being highly exposed to sentiment. Such factors have been widely explored in the literature. For future reference, we summarize them below:

(1) Heavy participation from retail investors: Additionally, a broad trend within the literature confirms the dominance of speculative intent in the cryptocurrency markets. Heavy retail participation in combination with overwhelming speculative intent naturally predisposes bitcoin prices to be unusually reactive to news and market sentiment. [22]
(2) Short-sale constraints and limited market depth: Even as the largest cryptocurrency market by capitalization, bitcoin has (amongst other asset classes) limited depth and severe short-sale constraints. Where speculators harbour divergent opinions about the degree of directional movement of a risky asset, pessimistic investors without the ability to go short are forced out of the market, which leads prices to over-represent optimistic sentiment. Thus, limited market depth and short-sale constraints cause bitcoin prices to be

more sensitive to speculative activity and cycles of attention-driven buying. [10]

(3) No prevailing consensus on fundamental value formation [2]
(4) Deterministic Supply [3]

REFERENCES

[1] [n. d.]. Keras. ([n. d.]). FranÃğoisChollet

[2] Jamal Bouoiyour and Refk Selmi. 2017. The Bitcoin price formation: Beyond the fundamental sources. (07 2017).

[3] Burniske and white. [n. d.]. Bitcoin: Ringing the Bell for a New Asset Class. ([n. d.]). https://www.research.ark-invest.com/hubfs/1_Download_Files_ARK-Invest/White_Papers/Bitcoin-Ringing-The-Bell-For-A-New-Asset-Class.pdf

[4] Eng-Tuck Cheah and John Fry. 2015. Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin. *Economics Letters* 130, Supplement C (2015), 32 – 36. https://doi.org/10.1016/j.econlet.2015.02.029

[5] Stuart Colianni, Stephanie Rosales, and Michael Signorotti. 2015. Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis.

[6] Num Focus. [n. d.]. Matplotlib Documentation. ([n. d.]). http://matplotlib.org

[7] Num Focus. [n. d.]. Numpy Documentation. ([n. d.]). http://www.numpy.org/

[8] David Garcia and Frank Schweitzer. 2015. Social signals and algorithmic trading of Bitcoin. *Open Science* 2, 9 (2015). https://doi.org/10.1098/rsos.150288 arXiv:http://rsos.royalsocietypublishing.org/content/2/9/150288.full.pdf

[9] Alex S. Greaves and Benjamin Au. 2015. Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin.

[10] Rishab Guha. 2015. Understanding Bitcoin as a Speculative Asset.

[11] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. 2016. Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies. *PLOS ONE* 11, 8 (08 2016), 1–17. https://doi.org/10.1371/journal.pone.0161197

[12] Isaac Madan and Shaurya Saluja. 2014. Automated Bitcoin Trading via Machine Learning Algorithms.

[13] Martina Matta, Ilaria Lunesu, and Michele Marchesi. 2015. The Predictor Impact of Web Search Media On Bitcoin Trading Volumes. (11 2015).

[14] Sean McNally. 2016. *Predicting the price of Bitcoin using Machine Learning*. Master's thesis. Dublin, National College of Ireland. http://trap.ncirl.ie/2496/

[15] Demitrios E. Pournarakis, Dionisios N. Sotiropoulos, and George M. Giaglis. 2017. A computational model for mining consumer perceptions in social media. *Decision Support Systems* 93, Supplement C (2017), 98 – 110. https://doi.org/10.1016/j.dss.2016.09.018

[16] PyData. [n. d.]. Pandas Documentation. ([n. d.]). https://pandas.pydata.org/

[17] PyData. [n. d.]. Seaborn Documentation. ([n. d.]). https://seaborn.pydata.org/index.html

[18] Quandl. [n. d.]. Quandl API. ([n. d.]). https://www.quandl.com/

[19] Reddit. [n. d.]. Bitcoin Subreddit. ([n. d.]). https://www.reddit.com/r/Bitcoin/

[20] Scikit. [n. d.]. Scikit Learn Documentation. ([n. d.]). http://scikit-learn.org/stable/index.html

[21] Devavrat Shah and Kang Zhang. 2014. Bayesian regression and Bitcoin. *CoRR* abs/1410.1231 (2014). arXiv:1410.1231 http://arxiv.org/abs/1410.1231

[22] Paul C. Tetlock. 2011. All the News That's Fit to Reprint: Do Investors React to Stale Information? *The Review of Financial Studies* 24, 5 (2011), 1481–1512. https://doi.org/10.1093/rfs/hhq141