Predicting Cryptocurrencies Market Phases through On-Chain Data Long-Term Forecasting

1st Bruno Casella*
Computer Science Department
University of Turin
Turin, Italy
0000-0002-9513-6087

2nd Lorenzo Paletto*
Computer Science Department
University of Turin
Turin, Italy
0000-0001-7225-9378

2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC) | 979-8-3503-1019-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICBC56567.2023.10174989

***Abstract*—Blockchain, the underlying technology of Bitcoin and several other cryptocurrencies, like Ethereum, produces a massive amount of open-access data that can be analyzed, providing important information about the network's activity and its respective token. The on-chain data have extensively been used as input to Machine Learning algorithms for predicting cryptocurrencies' future prices; however, there is a lack of study in predicting the future behaviour of on-chain data. This study aims to show how on-chain data can be used to detect cryptocurrency market regimes, like minimum and maximum, bear and bull market phases, and how forecasting these data can provide an optimal asset allocation for long-term investors.**

***Index Terms*—Blockchain, Bitcoin, cryptocurrencies, on-chain data, artificial intelligence, prediction, asset allocation, trading rules, markets**

I. Introduction

Blockchain is a subset of technologies belonging to the family of Distributed Ledgers that has emerged as a disruptive technology advancement in both private and public sectors. Blockchain is a distributed ledger that facilitates transaction registration processes and monitoring assets. Blockchain is a decentralized technology addressing some problems of the centralized architectures, such as the single point of failure and the privacy issues when transferring information to it. To modify the blockchain without a central authority, all the network members need to reach a consensus. Blockchain is increasingly being adopted across various industries since it is decentralized, distributed and transparent.

The main blockchain application is for cryptocurrencies. The first blockchain application is Bitcoin (BTC) [1], proposed by Satoshi Nakamoto (pseudonym indicating a person or a group of people) in 2008.

After Bitcoin, many other cryptocurrencies based on blockchain have been proposed, such as ETH, the coin of the Ethereum [2] blockchain, an ecosystem for decentralized applications (Dapps). Most of the cryptos can be traded on dedicated centralized or decentralized exchanges. Although the main purpose of BTC was to be an alternative payment method, nowadays, it is considered a speculative asset due to its high volatility. Moreover, BTC's price works as a pulling power for all the other assets, affecting in this way all the cryptos of the market. Most investors, regardless of the sectors in which they operate, during the analysis of an asset use different approaches, for example, the fundamental, the technical, or the quantitative analysis.

Cryptocurrency investors have another available type of analysis, the on-chain analysis. Indeed, blockchain produces a vast amount of open-access data that can be analyzed, providing information about the network's activity and the price of its respective token. We refer to these data as on-chain data. Due to their intrinsic nature, on-chain data are recorded as time series. The on-chain analysis is the analysis of these data.

Most works [3], [4] focus on training AI models for predicting the price of BTC or ETH using on-chain data as input for the algorithm. However, the predictions won't be helpful due to the price's highly fluctuating nature. Moreover, the dataset should not be made of only past prices because the price depends on many factors like news, market trends for related/unrelated assets, and insider trading news. A high-quality dataset containing prices and all these external data sources may not exist. For this reason, it might be helpful to decrease the task's difficulty and to predict only the price trend rather than the price itself. However, only a few financial works [5] try to predict market phases such as up-trend or down-trends. This work aims to show how the on-chain data can be used as indicators to understand the market regimes of cryptocurrencies and how predicting these data can be useful to provide an optimal asset allocation for long-term investments. The main contributions of this work are:

- we extensively collected and pre-processed on-chain data from Glassnode [7].
- we show how on-chain data can be used as indicators to assess market phases.
- we provide results of experiments on on-chain data using stochastic processes and Deep Learning models to forecast future values.

The rest of the paper is organized as follows. In Section

* Equal contribution

Bruno Casella and Lorenzo Paletto are with the Computer Science Dpt. of the University of Turin, Corso Svizzera 185, Turin, Italy (emails: [bruno.casella, lorenzo.paletto]@unito.it)

979-8-3503-1019-1/23/$31.00 ©2023 IEEE

Section II, we discuss the recent related works. Section III shows how it is possible to assess market trends through on-chain data. Section IV shows and discusses experimental results. Finally, Section V concludes the paper.

II. Related Work

Literature is rich in works on price prediction using stochastic processes such as SARIMA models and AI models. Some studies [8] have stated that Machine Learning (ML) is the best technology capable of learning price patterns with respect to traditional statistical methods that may require unrealistic assumptions. Other works [9], [10] have tested ML models and technical indicators in predicting bitcoin values.

Deep Learning (DL), a subfield of ML concerned with Neural Networks (NN), has been used to predict the price in a lot of markets, from forex to cryptos. A recent work [6] developed a complete trading system trained on historical forex data from 2010 to 2021, using two different DL architectures: the first was a ResNet-50, a convolutional neural network (CNN) typically used for image classification tasks; the second was a Transformer, an attention-based mechanism. The leitmotif of the work was to understand if deep learning can improve technical analysis of forex data to predict future price movements. Unsurprisingly, the authors had a negative answer because NNs are evaluated according to their prediction error. However, they do not take into account the global impact on a broader trading system.

[3] proposes a novel framework that predicts the price of Bitcoin employing the change point detection technique for stable prediction in unseen price range and giving in input to the NN on-chain data. In particular, the authors used self-attention-based multiple long short-term memory, where the modules of LSTM [11] are used for on-chain variable groups, while the attention mechanism for the prediction model.

Another recent work [4] adopted an LSTM using a subset of on-chain data as input. The authors proposed three self-adaptive techniques to select optimal hyperparameters for the NN so that the training is quick and effective, each of which converges on a set of optimal parameters to predict the price of Ethereum. DL and blockchain have also been used to create novel secure data-sharing frameworks for softwarized UAV environments [13] and to protect confidential information in the context of the Industrial Internet of Things [14].

However, since AI models do not take into account the global scenario, including social sentiment, unpredictable market movements, news, and market trends of related/unrelated assets, they are still not fully reliable. For this reason, lowering the task target from price prediction to the market regime can be a good strategy to obtain more reliable results. However, only a few studies try to predict market regimes, such as market maximum or minimum and bear or bull markets. [5] predicts U.S. bear and bull stock markets with dynamic binary time series models to obtain optimal asset allocation decisions.

To our knowledge, there are no studies on predicting bear and bull crypto markets and predicting on-chain data. This work aims at filling this lack of studies by showing how on-chain data can be used as indicators for long-term investments.

III. On-chain data

On-chain data are constantly generated from their reference blockchain and are recorded as time series due to their intrinsic nature. Examples of on-chain data can be the size of the blockchain, the number of blocks, transactions, wallet balances or the fees paid to miners. These metrics inform about the state of the blockchain network. On-chain analytics refers to analyzing on-chain metrics to extrapolate insights and to determine market trends and sentiment, two crucial factors in deciding whether or not an investment is worth it.

All the metrics we used in this work have been downloaded from Glassnode. We selected six on-chain time series. For each of them, we will give a brief definition and description of how it can be used to determine market cycles:

- **New addresses**: the number of unique addresses that appeared for the first time in a transaction of the native coin in the network. It gives us people's interest in the asset. It can be seen (Fig. 2a) that new addresses peaks coincide with BTC price peaks and that this metric anticipates BTC market trends.
- **Active addresses**: the number of unique addresses that were active in the network either as a sender or receiver. Only addresses that were active in successful transactions are counted. It gives us people's interest in the asset. It can be seen (Fig. 2b) that active addresses peaks coincide with BTC price peaks and that this metric anticipates BTC market trends.
- **Block Height**: the block height, i.e. the total number of blocks ever created and included in the main blockchain. It tells us if the asset is overpriced (price over block) or if the asset is underpriced (price below block). This is clearly shown in Fig. 2c
- **Fees**: the total amount of fees paid to miners. Issued (minted) coins are not included. Fig. 2d shows that cyclic price peaks coincide with fee peaks.
- **Hash rate**: the average estimated number of hashes per second produced by the miners in the network. It can be used like the block height metric: if BTC is overpriced, the price will be over the hash rate; otherwise, if BTC is underpriced, the price will be below the hash rate (Fig. 2e).
- **Spent Output Profit Ratio**: the Spent Output Profit Ratio (SOPR) is computed by dividing the realized value (in USD) divided by the value at creation (USD) of a spent output. Or simply: price sold / price paid. SOPR appears to oscillate around the number 1. During a bull market, values of SOPR below one are rejected, while during a bear market, values of SOPR above 1 are rejected (Fig.2f).

From these descriptions, it is clear that having a good forecast of these metrics is a critical point for long-term asset allocation. In the next section, we will show how these data can be forecasted using stochastic processes such as SARIMAX or Deep Learning methods such as recurrent neural networks (RNNs) or CNNs. Figure 1 shows the flow of the proposed market phase prediction method.

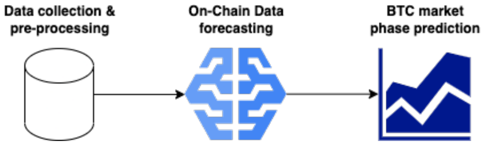


Fig. 1: Flow of the proposed method

IV. Experiments

To conduct our experiments, we adopted Python 3 as a programming language and Keras as DL framework. Each experiment is run on an Intel(R) Core(TM) i5-5257U CPU (2 cores per CPU).

Dataset: We compared stochastic processes and DL models, on the six time series presented in section III: new addresses, active addresses, block height, fees, hash rate and SOPR. The details of the datasets are summarized in TableI.

TABLE I: Details of the time series

Metric	Time steps	Weekly steps	Min	Max	Mean
NA	4.380	627	94	800.180	210.618
AA	////	626	2,4e+01	6,8e+06	2,3e+06
BH	4.383	627	33.114	716.598	386.457
Fees	4.116	605	0	1.495	58,04
HR	1.826	262	7,7e+06	2,0e+20	3,1e+19
SOPR	3.285	470	0,81	1,24	1,01

Preprocessing: Some of the time series had multiple values for single days. For those days, we averaged the values to have one value for each day. We also removed the first entries of the time series if they were equal to zero because those parts did not bring information. We performed a differentiation of order one to make the time series stationary. Stationarity was tested with the Dickey-Fuller test. We lastly split the series into train, validation and test set, which were all [0-1]-normalized with respect to the train set. All the data were daily except for Active Addresses, which was weekly. Weekly data were created from daily data by computing the mean values over the weeks and applying the preprocessing steps described above. Validation and test set corresponded to the first and last six months of 2021 for both daily and weekly data.

Model: we compared three different models on each of the time series:

- **SARIMA**: We used a SARIMA(p,d,q)(P,D,Q,s) model independently for each time series. Seasonality was set to 7 for daily data and 5 for weekly data. All the other parameters were chosen by fitting models within ranges defined for each parameter, typically [0,3], and comparing the Akaike and Bayesian Information Criterion. This process was automatically conducted using the function *auto_arima* contained in the statistical library *pmdarima*.
- **LSTM** [11]: we used a NN made of two LSTM layers of 100 units, followed by a Dense layer. The best learning rate was chosen according to the *LearningRateScheduler* function of Keras; the optimizer was NAdam, i.e. Adam with Nesterov momentum; the loss function used was Huber.
- **CNN**: as architecture, we adopted WaveNet [12], a 1-D CNN first proposed by Google DeepMind in 2016. Hyperparameters have been chosen with the same methods as LSTM.

Prediction technique: for SARIMA, we tested three different prediction techniques:

- **Out-of-sample** forecasting by iteratively predicting the next point of the test set and appending the prediction to the history.
- **In-sample** forecasting by iteratively forecasting the first point of the test set, adding the true value to the history, and repeating the process for all the next data points.
- **Multi-step** forecasting by using *forecast(step=n)*.

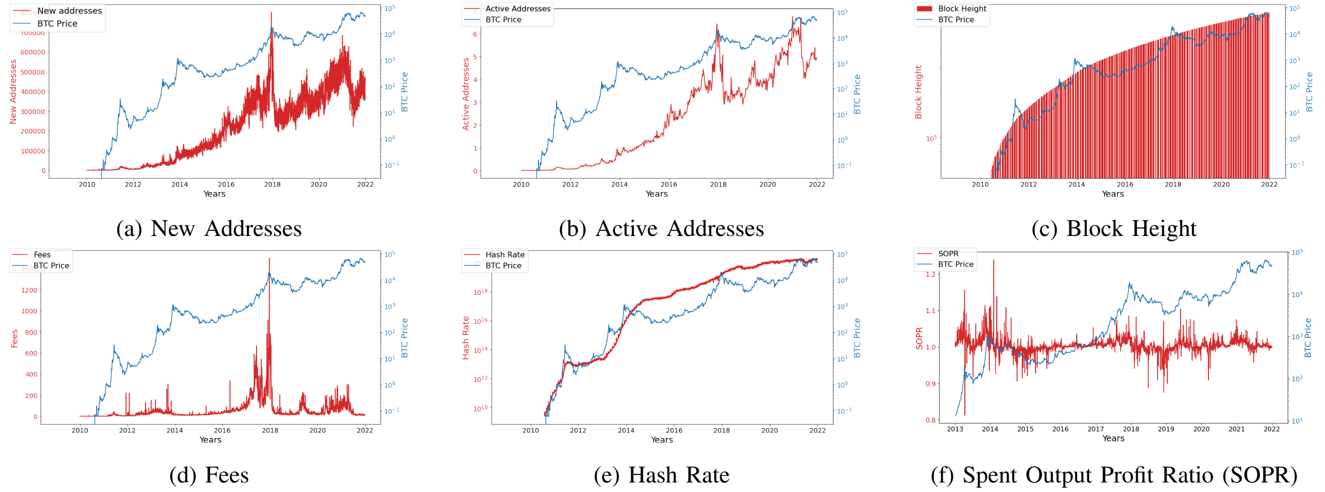
For DL models, we tested only the out-of-sample and the multi-step techniques. The algorithms have been compared with the standard metric of mean absolute error (MAE). Results are reported in tables II to IV.

TABLE II: Models' MAE with out-of-sample prediction.

Time series	SARIMA	LSTM	CNN
Daily data			
New Addresses	115.684	**100.814**	**100.814**
Active Addresses	////	////	////
Block Height	63.212	**4.586**	**4.586**
Fees	**86,71**	92,58	92,58
Hash Rate	**2,31e+19**	2,38e+19	2,38e+19
SOPR	0,0286	**0,0085**	**0,0085**
Weekly data			
New Addresses	121.632	**24.215**	**24.215**
Active Addresses	**625.173**	2.146.443	2.146.443
Block Height	75.683	**31.690**	**31.690**
Fees	95.84	**48.02**	**48.02**
Hash Rate	2,15e+19	**2,04e+19**	**2,04e+19**
SOPR	0,022	**0,0107**	**0,0107**

TABLE III: Models' MAE with multi-step prediction.

Time series	SARIMA	LSTM	CNN
Daily data			
New Addresses	115.684	31.588	**28.912**
Active Addresses	////	////	////
Block Height	63.212	**499**	5.873
Fees	86,72	19,66	**4,34**
Hash Rate	2,31e+19	1,28e+19	**8,31e+18**
SOPR	0,0272	0,035	**0,0026**
Weekly data			
New Addresses	121.633	**14.646**	16.320
Active Addresses	625.172	761.671	**267.371**
Block Height	75.647	**8.658**	17.778
Fees	95.84	11,58	**3,88**
Hash Rate	2,13e+19	**1,07e+19**	1,29e+19
SOPR	0,0264	0,0056	**0,0028**



(a) New Addresses (b) Active Addresses (c) Block Height

(d) Fees (e) Hash Rate (f) Spent Output Profit Ratio (SOPR)

Fig. 2: Relationship between on-chain data and BTC price

TABLE IV: Models' MAE with in-sample prediction (SARIMA only).

	NA	AA	BH	Fees	HR	SOPR
Daily	33.937	////	641	4,25	1e+19	0,0043
Weekly	19.766	225.045	3.509	5,31	9,36e+18	0,0039

Tables II to IV show that Multi-step prediction outperforms both Out-of-sample and In-sample forecasting. The best performances are achieved by DL models (in particular by the Wavenet CNN) on almost all the metrics and with all the different forecasting techniques. Unsurprisingly, the out-of-sample forecasting results are the worst because the error accumulates iteratively; this shows that this technique can be only used for short-term forecasting, which prevents error accumulation. Graphs of the predictions are omitted for space constraints.

V. Conclusions

In this paper, we tested stochastic processes and DL models on six on-chain metrics that, until now, have been used only as input for ML algorithms. To the best of our knowledge, this is the first work forecasting on-chain data and showing how they can be used for statistical hedging. For this work, we tested AI models on six on-chain time series; however, our method can be used with every relevant on-chain metric. In future work, we aim to test state-of-the-art DL models and stochastic processes to improve our predictions and to collect further datasets to provide a more comprehensive picture of the market phase. All the code is available at this link.

Acknowledgements

This work has been supported in part by the Spoke ”FutureHPC & BigData” of the ICSC – Centro Nazionale di Ricerca in ”High Performance Computing, Big Data and Quantum Computing”, funded by European Union – NextGenerationEU, and in part by the European Union within the Horizon 2020 Research and Innovation Programme “European Processor Initiative—Specific Grant Agreement 1” under Grant 826647, EPI.

References

[1] S.Nakamoto, *”Bitcoin: A Peer-to-Peer Electronic Cash System”*. [Online]. https://bitcoin.org/bitcoin.pdf, 2009

[2] V. Buterin, *”Ethereum White Paper: A Next Generation Smart Contract & Decentralized Application Platform”*. [Online]. Available: https://ethereum.org/en/whitepaper, 2013

[3] G. Kim et al., *”A Deep Learning-Based Cryptocurrency Price Prediction Model That Uses On-Chain Data”*, IEEE Access, 2022

[4] N. Jagannath et al., *”An On-Chain Analysis-Based Approach to Predict Ethereum Prices”*, IEEE Access, 2021

[5] H. Nyberg, *”Predicting bear and bull stock markets with dynamic binary time series models”*, Journal of Banking & Finance 37 (2013) 3351–3363.

[6] M. Fisichella, F. Garolla, *”Can Deep Learning Improve Technical Analysis of Forex Data to Predict Future Price Movements?”*, IEEE Access, 2021.

[7] G. Studio. Glassnode Studio-on-Chain Market Intelligence. Glassnode Studio. Aug. 15, 2021. [Online]. Available: https://studio.glassnode.com

[8] A. M. Khedr et al., *”Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey”*, Intell. Syst. Accounting, Finance Manage., vol. 28, no. 1, pp. 3–34, 2021.

[9] S. McNally et al., *”Predicting the price of bitcoin using machine learning”*, in Proc. 26th Euromicro Int. Conf. Parallel, Distrib. Netw.-Based Process. (PDP), Mar. 2018, pp. 339–343.

[10] J. Z. Huang et al., *”Predicting bitcoin returns using high-dimensional technical indicators”*, J. Finance Data Sci., vol. 5, no. 3, pp. 140–155, Sep. 2019.

[11] K. Greff et al., *”LSTM: A Search Space Odyssey”*, IEEE Trans. Neural Networks Learn. Syst. vol.28, no. 10, pp. 2222-2232, 2017.

[12] A. va den Oord et al., *”WaveNet: A Generative Model for Raw Audio”*, CoRR abs/1609.03499, 2016, http://arxiv.org/abs/1609.03499.

[13] P. Kumar et al., *Blockchain and Deep Learning Empowered Secure Data Sharing Framework*, IEEE International Conference on Communications Workshops, 2022.

[14] R. Kumar et al., *Blockchain and Deep Learning for Cyber Threat-Hunting in Software-Defined Industrial IoT*, IEEE Internation Conference on Communications Workshops, 2022.