Price Forecasting for Bitcoin: Linear Regression and SVM approaches*

Corazza Marco[1][0000−0003−3376−3752] and Giovanni Fasano[2][0000−0003−4721−8114]

[1] Ca' Foscari University, Dep. of Economics, Cannaregio 873, 30121 Venice (Italy)
[2] Ca' Foscari University, Dep. of Management, Cannaregio 873, 30121 Venice (Italy)
{corazza,fasano}@unive.it

Abstract. A long term analysis for the most renowned crypto asset (namely Bitcoin) is considered. Bitcoin has currently the largest market capitalization among the crypto assets, and in the last years has steadily attracted the attention of both private and institutional investors. Since 2008 Bitcoin price has always experienced high volatility, too, which earned it the title of risky asset in the literature of FinTech. Hence, practitioners have stressed the key role of robust models to reliably predicting its value, not to mention the relevance of a comparative analysis with safe haven assets as silver and gold. This paper focuses on a couple of approaches to predict long term Bitcoin price. Basically the first one relies on more standard regression and linear models. Our second proposal is conversely based on applying a Support Vector Machine (SVM), in the class of Machine Learning (ML) methods, which do not require any of the assumptions typically needed by solvers for standard regression problems. We remark that both the above proposals are inherently data–driven.

Keywords: Bitcoin · Long Term Price Forecast · Linear Regression Models · Support Vector Machines.

1 Introduction

In this paper, the authors tackle the challenging task of long term price forecasting for crypto assets, with a specific focus on Bitcoin. Bitcoin [7] is considered one of the most famous and largest market capitalization crypto assets. It was created in 2008 by an anonymous researcher or team known as Satoshi Nakamoto. It is also characterized by its decentralized nature, unlike fiat currencies such as the Dollar, Euro, Pound or Yen [13].

Bitcoin transactions are finalized through special actors named *miners*, who validate them after official exchanges have collected them into *blocks*. In addition, peer-to-peer movements of tokens on the Bitcoin network can also be completed by miners without intermediaries (i.e. the exchanges). Transactions are validated by network nodes through complex cryptographic computations

* Supported by GNCS of Istituto Nazionale di Alta Matematica (INδAM), Italy.

and are recorded in a public distributed ledger called *blockchain*. Miners, special nodes in the network, are rewarded with newly minted bitcoins[1] for solving these cryptographic problems. The rewarding policy of Bitcoin miners changes every four years, leading to events known as halvenings, which halve the reward associated with each mined block of token transactions.
Price prediction for Bitcoin has garnered increasing interest in the last decade. However, the high volatility of BTC price poses a significant challenge for accurate forecasting. Speculative trading, leveraged transactions, and the relatively recent emergence of Bitcoin as an asset contribute to its price fluctuations. Existing literature has proposed various approaches, including Machine Learning (ML)-based models that use intrinsic mode functions (IMFs) coupled with support vector machines (SVMs), see e.g. [1], as well as the optimization method LASSO for ML, see e.g. [10].
In this paper, the authors present a twofold approach to Bitcoin long term price forecasting. First, they use linear and regression models applied for BTC price prediction. Then, they study the role of the so called Stock-to-Flow (SF) ratio in regression problems related to Bitcoin. They propose a novel ML-based technique that combines mathematical programming and SVMs to estimate Bitcoin price in the long term. This approach does not rely on statistical assumptions typically associated with regression models.
The paper also hints some methodological concerns in the literature of Bitcoin price prediction, and aims to provide a foundation for more reliable analyses in [4], where the use of linear least squares schemes, variable scaling, and the computation of the SF ratio are better investigated. In particular, theoretical contributions using SVMs and the bootstrap method are also included in [4]. Hence, the current paper mainly addresses methodological concerns and provides a possible foundation for analysis related to the growing body of research in this field.

The remainder of the paper is organized as follows: in Sections 2 and 3 we give indications about the first proposed approach, based on a linear and regression techniques; then, Section 4 better details the computation of the SF ratio for BTC. Section 5 gives some hints about the authors' second approach based on SVMs. Finally, Section 6 will include some final remarks.

2 Regression problems and least squares optimization

Linear regression problems typically assume that we are given the $p+1$ random variables $\{X_1, \ldots, X_p, Y\}$, with

$$Y = \sum_{i=1}^{p} \beta_i X_i + u_i, \qquad \beta_i \in \mathbb{R}, \quad i = 1, \ldots, p, \tag{1}$$

being $\{u_i\}$ statistical errors such that $\mathbb{E}(u_i | X_1, \ldots, X_p) = 0$, for any i. Then, if $X_1, \ldots, X_p, Y$ are *independent and identically distributed* (i.i.d.) we can approach

[1] Observe that the name *Bitcoin* or *BTC* (indifferently used in the present paper) indicates the crypto asset while *bitcoin* is used to identify its token.

the solution of the linear regression problem through solving the Linear Least Squares problem

$$\min_{m\in\mathbb{R}^p,\ q\in\mathbb{R}} \sum_{\ell=1}^{N}\left[\left(q+\sum_{i=1}^{p} m_i X_i^{(\ell)}\right) - Y^{(\ell)}\right]^2, \tag{2}$$

being N the number of samples for the random variables $(X_1,\ldots,X_p,Y)$. However, the solution of (2) relies on theoretical assumptions on the quantities $X_1,\ldots,X_p,Y,u_1,\ldots,u_p$ which may not be fulfilled (e.g., when the samples do not follow a normal distribution). Hence, a suitable reliability test on the solutions of (2) is needed, typically involving the p–value and/or the R^2 indicators.

This paper reports two approaches to estimate the price of Bitcoin (i.e. Y) *vs.* time *or vs.* its SF (i.e. X), being $Y = aX + b + u$, and $a, b \in \mathbb{R}$. Since *the error u may not be normally distributed*, due to the complex dynamics of Bitcoin price, we show that the solution of the linear least squares problem (2) might be possibly poor with respect to a SVM-based approach (see also [6]).

3 Linear and Least Squares models for Bitcoin price

This section provides an overview on BTC price, in order to spot some light on the *high volatility asset* fame that it has always gained in its history. In particular, we focus on data from the last decade (see `https://finance.yahoo.com`). Figure 1 gives evidence of the most recent dynamics of BTC price with a couple of scatter plots. Both the plots report 3288 price (USD) samples for BTC from September 30th, 2014 to September 30th, 2023, and the scale on the ordinate axis (i.e. 10^4) immediately reveals the large volatility that Bitcoin price has experienced in its history. Moreover, if y_i is BTC price at date x_i (with $x_i \in \{1,\ldots,3288\}$) then the two parallel lines in both the subplots solve the problem

$$\begin{aligned} \min_{m,q_1,q_2\in\mathbb{R}} \quad & q_2 - q_1 \\ & y_i \geq m x_i + q_1, \qquad i = 1,...,3288 \\ & y_i \leq m x_i + q_2, \qquad i = 1,...,3288 \end{aligned}$$

so that the *narrowest* stripe (delimited by the continuous lines) containing all the points can be identified. In addition, the dashed line provides a trendline for BTC price and we highlight that its computation does not rely on the standard assumptions required by regression analysis, as recalled in Section 2. Note that the two subplots essentially differ on the nature of the data used for their computation. In the upper subplot daily *close* prices are used, while in the lower subplot daily *open* prices are used, with very similar results on the long term analysis (i.e. very similar values for the parameters m, q_1 and q_2).

As an additional introductory analysis, before treating the problem of long term BTC price forecast through a ML approach, we also carried on a complete numerical experience based on least squares minimization. The main outcomes

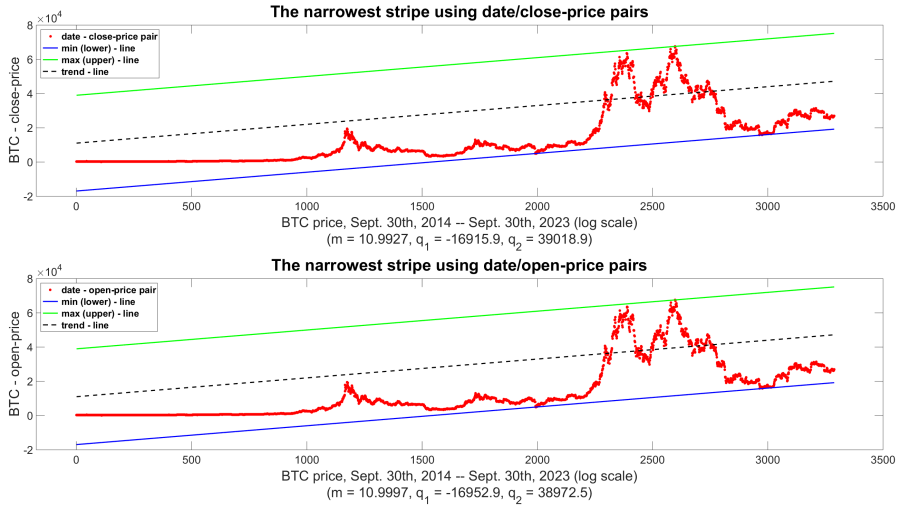

Fig. 1. Linear models for BTC price, in order to predict the narrowest region where BTC price is confined. On the abscissa axis we have 3288 ticks, each corresponding to the a day between September 30th, 2014 and September 30th, 2023.

can be found in Figure 2, where now the *upper* and *lower* subplots solve respectively the problems:

$$\min_{m,q\in\mathbb{R}} \sum_{i=1}^{3288} [(y_i - mx_i) - q]^2$$

and (each $\xi_i \in \mathbb{R}$ represents itself an unknown of the problem)

$$\min_{m,q,\xi_i\in\mathbb{R}} \sum_{i=1}^{3288} [(\xi_i - mx_i) - q]^2$$
$$y_i^{open} \leq \xi_i \leq y_i^{close}, \qquad i = 1, ..., 3288.$$

In practice, in the upper subplot of Figure 2 the dashed line solves a standard least squares (LS) problem as in (2). Conversely, the dashed line in the lower subplot of Figure 2 represents a constrained least squares problem, where we duly consider that the daily prices $\{y_i\}$ range between the values $\{y_i^{open}\}$ and $\{y_i^{close}\}$, i.e. the daily open/close prices. As a natural consequence, the solution in the upper plot can partially preserve the properties of a linear regression problem (as specified in Section 2), while the solution in the lower subplot is not endowed with statistical properties, though allowing a more robust analysis[2] to possibly better encompass also daily volatility for BTC. The results in Figures 1 and 2 represent a starting point for a more thorough numerical and theoretical

[2] In particular, in Figure 2 the final objective function value is $\approx 3.6652e+11$ (upper subplot) and $\approx 3.3196e+11$ (lower subplot), respectively.

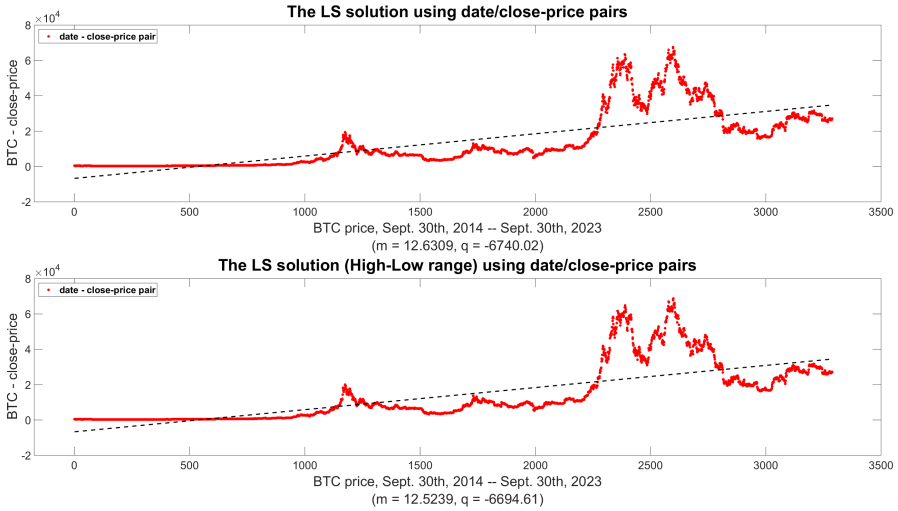

Fig. 2. Least Squares models for BTC price, in order to predict long term BTC price. On the abscissa axis we have 3288 ticks, each corresponding to the a day between September 30th, 2014 and September 30th, 2023.

investigation that these authors are carrying on (partial intermediate results are summarized in [4]). In particular, future work will have to issue and give a more sound foundation to common claims and questions raised by practitioners, in the literature related to Bitcoin (see also [3, 8]). On one hand these results have to comply with a more general approach adopted following the classical Box–Cox transformation [2]; on the other hand, they also have to retain a meaningful insight in view of practical Bitcoin price forecast.

4 The Stock-to-Flow (SF) ratio for BTC price analysis

Here we intend to introduce the SF ratio for Bitcoin, and possibly suggest the importance of this parameter for long term BTC price forecast. By definition the SF ratio is given by

$$SF = \frac{\text{Total amount of BTCs generated}}{\text{Amount of BTCs generated in the period } \Delta T},$$

and in case $\Delta T = 365$ *days* it basically represents the number of years, at the current annual production rate, which are necessary to obtain the overall BTC current stock. This also highlights that the SF may be conceived as a proxy for the scarcity of an asset. As an example[3] the formula for the computation of the

[3] Data for Bitcoin prices can also be retrieved from `www.coinmarketcap.com`

SF ratio associated with Bitcoin and referred to the end of September 2021, is given by

$$SF_{Sept\_2021} \approx \frac{18,700,000}{\frac{24\cdot 60}{10}\cdot 463\cdot 6.25} \approx 45. \tag{3}$$

In the last relation the computation was carried on using the next information:

- about every 10 minutes one newly minted block is added to Bitcoin blockchain, so that about $24 \cdot 60/10$ newly mined blocks are added every day;
- exactly 6.25 bitcoins are minted and rewarded to miners for each newly added block to the blockchain of Bitcoin. The value 6.25 will be halved in 2024;
- the time window ΔT for computing the flow in (3) is different with respect to 365 days (being indeed *463 days*, as suggested in [3] and [8]).

5 An SVM-based approach for long term Bitcoin price prediction

Here we briefly sketch a second approach for the analysis of long term Bitcoin price. More explicitly, we combine a Multiobjective Technique (MT) with a SVM framework (see also [5, 11, 12] for some references). We obtain similar outcomes using this second model, with respect to those from the first approach, in terms of Bitcoin price prediction. However, the next two improvements are obtained:

- we do not require any assumption typically needed by regression;
- through a bootstrap analysis we are able to obtain significantly better statistical results with our second approach based on SVM, with respect to the first one. In particular, we can show that our prediction of Bitcoin price is more robust (details can be found in [4]).

On this purpose, in order to combine MTs with SVMs, once more let be given the pairs $(\bar{x}_i, \bar{y}_i)$, $i = 1, \ldots, N$, of Bitcoin SF and Bitcoin price. Then, we preliminarily identify the two subsets $L_{\max}$ and $L_{\min}$ of $\{(\bar{x}_i, \bar{y}_i)\}$, associated with different weak Pareto fronts, being in particular:

- $L_{\max} \equiv L_{West-North}$: the weak Pareto front associated with (at once) the minimization of the SF ratio and the maximization of Bitcoin price;
- $L_{\min} \equiv L_{East-South}$: the weak Pareto front associated with (at once) the maximization of the SF and the minimization of Bitcoin price.

In other words, $L_{\max}$ includes those (*desirable*) points with high performance of Bitcoin price vs. its SF. Conversely, $L_{\min}$ contains those (*undesirable*) points with poor Bitcoin price performance vs. its SF. Hence, $L_{\max}$ and $L_{\min}$ include points which may be associated with extreme opposite performances of Bitcoin price vs. its SF. Now, we detail how to use the last subsets in a SVM-based framework, in order to determine the maximally distant separating hyperplane from them (see also [9] for details).

In the procedure we propose we set $A_0 = L_{\max}$ and $B_0 = L_{\min}$; then, we generate the sequences of sets $\{A_k\}$ and $\{B_k\}$, with $k = 0, 1, 2, \ldots$, by iteratively following the next steps:

- solve an SVM classification problem that yields the line $H(\beta^{(k)}, \beta_0^{(k)}; x, y) = 0$ and linearly separates the sets A_k and B_k. We remark that this line is maximally distant (i.e. it has a maximum *margin*) from the points in the sets A_k and B_k;
- we identify a point $(\bar{x}_{\max}^{(k)}, \bar{y}_{\max}^{(k)})$ in $\{(\bar{x}_i, \bar{y}_i),\ i = 1, \ldots, N\} \setminus \{A_k \cup B_k\}$ and assign the label $z_k \in \{-1, +1\}$ to it. If $z_k = +1$ we generate the novel sets $A_{k+1} = A_k \cup \{(\bar{x}_{\max}^{(k)}, \bar{y}_{\max}^{(k)})\}$ and $B_{k+1} = B_k$, otherwise we set $A_{k+1} = A_k$ and $B_{k+1} = B_k \cup \{(\bar{x}_{\max}^{(k)}, \bar{y}_{\max}^{(k)})\}$. The point $(\bar{x}_{\max}^{(k)}, \bar{y}_{\max}^{(k)})$ is such that among the points in $\{(\bar{x}_i, \bar{y}_i),\ i = 1, \ldots, N\} \setminus \{A_k \cup B_k\}$ it maximizes the distance from the line $H(\beta^{(k)}, \beta_0^{(k)}; x, y) = 0$. Furthermore, if $H(\beta^{(k)}, \beta_0^{(k)}; \bar{x}_{\max}^{(k)}, \bar{y}_{\max}^{(k)}) > 0$ then $z_k = +1$, otherwise $z_k = -1$.

As a further comment, observe that according with the last procedure, at step k we iteratively consider the pair of sets (A_k, B_k), so that we first compute the maximally distant separating hyperplane between them. Then, we identify one point among the pairs $\{(\bar{x}_i, \bar{y}_i),\ i = 1, \ldots, N\} \setminus \{A_k \cup B_k\}$ which is maximally distant from the last hyperplane. This point will be assigned to either A_{k+1} or B_{k+1}, so that at the next step the pair (A_{k+1}, B_{k+1}) will be used to compute the next separating hyperplane. Finally, when at step m the condition

$$\{(\bar{x}_i, \bar{y}_i),\ i = 1, \ldots, N\} \setminus \{A_m \cup B_m\} = \emptyset$$

is fulfilled, then the SVM-based algorithm stops and the last separating hyperplane is used as a trend–line for Bitcoin price prediction.

Several theoretical results can be proved starting from this MT-SVM-based approach. Reporting all them is far beyond the purposes of the present paper; however, the interested reader may refer to [4] for complete materials. As an additional result, observe that we used Bootstrap to reinforce the statistical outcomes associated with the MT-SVM-based approach. Note that Bootstrap is a widely used technique to infer statistics on a population (of results). In its basic version, it is essentially based on performing re–sampling with replacement of the original dataset associated with the population. Then, invoking the Central Limit Theorem (CLT) simple statistics (i.e. the *mean value* and the *standard deviation*) are sought. Lastly, we recall that according with the CLT, when i.i.d. random variables are summed up (or averaged), then their properly normalized sum approaches a normal distribution, regardless of the original distribution of the random variables.

6 Conclusions and future work

In the past decade, Bitcoin has gained popularity as a digital asset for potential investments from both private individuals and institutional investors. This paper

presents models for long term BTC price forecast. These authors believe that various factors beyond the SF ratio significantly influence Bitcoin long term price, and this paper confirms that the use of the SF on Bitcoin price analysis can be of impact, as also suggested in [3]. Further advances are likely from future extensions of the present work, by incorporating a Machine Learning-based approach where multiple market indicators (based on the SF) are possibly combined.

Acknowledgements Marco Corazza and Giovanni Fasano wish to thank Istituto Nazionale di Alta Matematica (INδAM), Giovanni Fasano wishes to thank *Consiglio Nazionale delle Ricerche – Istituto di Ingegneria del Mare (CNR–INM)*, for the support they received.

References

1. Aggarwal, D., Chandrasekaran, S., Annamalai, B.: A complete empirical ensemble mode decomposition and support vector machine-based approach to predict Bitcoin prices. Journal of Behavioral and Experimental Finance **27**(100335) (2020)
2. Box, G., Cox, D.: An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological) **26**(2), 211–252 (1964)
3. Buy Bitcoin Worlwide: Bitcoin stock to flow model live chart. www.buybitcoinworldwide.com
4. Caliciotti, A., Corazza, M., Fasano, G.: From regression models to Machine Learning approaches for long term Bitcoin price forecast. (accepted for publication) Annals of Operations Research (2023)
5. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press (2000)
6. Graybill, F., Iyer, H.: Regression Analysis: Concepts and Applications. Duxbury Press, Belmont, CA (1994)
7. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system. http://www.bitcoin.org/bitcoin.pdf (2008)
8. PlanB: Modeling Bitcoin's Value with Scarcity. https://medium.com/@100trillionUSD/modeling-bitcoins-value-with-scarcity-91fa0fc03e25
9. Pontiggia, A., Fasano, G.: Data Analytics and Machine Learning paradigm to gauge performances combining classification, ranking and sorting for system analysis. Technical Report 05-21, Department of Management, University Ca' Foscari of Venice, Italy (2021)
10. Reddy, L.S., Sriramya, D.: A research on Bitcoin price prediction using Machine Learning algorithms. International Journal of Scientific & Technology Research **9**(4), 1600–1604 (2020)
11. Vapnik, V.: The Nature of the Statistical Learning Theory. Springer Verlag, New York (1995)
12. Vapnik, V.: Statistical Learning Theory. Wiley (1998)
13. Vigna, P., Casey, M.: The Age of Cryptocurrency: How Bitcoin and Digital Money Are Challenging the Global Economic Order - First edition. St. Martin's Press (2015)