This is the accepted version of the following article

Petr Hajek, Lubica Hikkerova, Jean-Michel Sahut (2023). How well do investor sentiment and ensemble learning predict Bitcoin prices?. *Research in International Business and Finance.* Volume 64, January 2023, 101836. DOI: 10.1016/j.ribaf.2022.101836

This version is licenced under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0.International

Publisher's version is available from: https://www.sciencedirect.com/science/article/pii/S0275531922002227

How well do investor sentiment and ensemble learning predict Bitcoin prices?

Petr Hajek[a], Lubica Hikkerova[b,*], Jean-Michel Sahut[c]

[a] Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Studentska 84, Pardubice, Czech Republic, e-mail: petr.hajek@upce.cz

[b] IPAG Business School, Paris, France, e-mail: lubicahikkerova@gmail.com

[c] IDRAC Business School, Lyon, France

* corresponding author, e-mail: lubicahikkerova@gmail.com

Abstract

Investor sentiment is widely recognized as the major determinant of cryptocurrency prices. Although earlier research has revealed the influence of investor sentiment on cryptocurrency prices, it has not yet generated cohesive empirical findings on an important question: How effective is investor sentiment in predicting cryptocurrency prices? To address this gap, we propose a novel prediction model based on the Bitcoin Misery Index, using trading data for cryptocurrency rather than judgments from individuals who are not Bitcoin investors, as well as bagged support vector regression (BSVR), to forecast Bitcoin prices. The empirical analysis is performed for the period between March 2018 and May 2022 and had a total of 1537 observations. The results of this study suggest that the addition of the sentiment index enhances the predictive performance of BSVR significantly. Moreover, the proposed prediction system, enhanced with an automatic feature selection component, outperforms state-of-the-art methods for predicting cryptocurrency for the next 30 days.

Keywords: Bitcoin, price, cryptocurrency, sentiment, prediction, feature selection, support vector regression

Highlights

- Investor sentiment based on Bitcoin Misery Index (BMI) considered for cryptocurrency forecasting.
- Automated feature selection component used to identify relevant predictors of Bitcoin prices.
- BMI-based investor index substantially improves the performance of bagged support vector regression.
- Predicting Bitcoin prices for the period 2018–2022 using different forecasting horizons (1 to 30 days).
- Outperforms existing univariate and multivariate models in terms of prediction errors.

1. Introduction

Due to the fact that cryptocurrencies are not backed by a government, their value is not decided by economic fundamentals; it is more likely to be determined by the market supply and demand and will have relatively significant volatility (Balcilar et al., 2017; Koki et al., 2022; Jia et al., 2022). Thus, trust among Bitcoin users is likely to be revealed in their attitudes toward and emotions regarding Bitcoin (Kaabia et al., 2020; White et al., 2020; Karaa et al., 2021). Moreover, especially in times of crises and stresses, the dynamics of cryptocurrency markets change and investors may overreact to new information (Akhtaruzzaman et al., 2022a, 2022b; Goodell and Goutte, 2021).

Recent studies on the drivers of cryptocurrency prices have increasingly emphasized the role of investor sentiment, as has previous literature on traditional assets and currencies (Ahn and Kim, 2020). Indeed, Kristoufek (2013) was the first to point out the association between the number of searches on Google and Bitcoin prices. However, Kaminski and Gloor (2014) suggested that the emotions expressed on Twitter were not reliable predictors of Bitcoin prices. This stream of literature has since been extended to the study of other assets and proxies by several authors (Karalevicius et al., 2018; Eom et al., 2019; Rognone et al., 2020, Chen et al., 2020; Oad Rajputet al., 2020; Guégan and Renault, 2021; Gaies et al., 2021; López-Cabarcos et al., 2021).

While these publications demonstrate a consensus in the literature about the influence of investor sentiment on cryptocurrency prices, the empirical findings indicate that one fundamental concern remains unanswered: What is the predictive power of investor sentiment on cryptocurrency prices?

Existing prediction models used for cryptocurrency forecasting either rely on univariate models (Lahmiri and Bekiros, 2019; Livieris et al., 2020) or use multiple cryptocurrency determinants, such as commodity prices, stock market prices (Sun et al., 2020), and indices of the cryptocurrency market (Liu et al., 2021; Guo et al., 2021). While the performance of the multivariate forecasting models has been encouraging in terms of cryptocurrency forecasting, their accuracy may have been adversely affected by the inclusion of irrelevant factors. Moreover, most studies on cryptocurrency forecasting have neglected the effect of investor sentiment. Hence, the previous studies used have failed to provide a transparent prediction model that would allow for the evaluation of the contribution of investor sentiment to cryptocurrency price predictions.

Consequently, the purpose of this paper is to propose a prediction model that would integrate: (1) investor sentiment and other features (e.g., previous prices of Bitcoin, technical indicators, cryptocurrency market features, and cryptocurrency metrics), (2) an automated feature selection component to identify relevant predictors, and (3) a bagged support vector regression (BSVR) model for forecasting Bitcoin prices up to 30 days in advance.

The contributions of this paper are fourfold. First, it highlights the fact that the prediction model using previous prices of Bitcoin, as considered in previous univariate models, is not effective enough to accurately predict the Bitcoin price in the next 30 days. The findings of this study suggest that other predictors are needed and that investor sentiment might further improve the accuracy of predictions. Second, this paper demonstrates how the proposed prediction system outperforms current prediction models of cryptocurrency prices. This can be attributed to the fact that the proposed model, unlike the models in previous studies, effectively uses not only investor sentiment but also a number of Bitcoin metrics. Third, this paper reveals the value of our sentiment variable, the Bitcoin Misery Index (BMI), which is based on actual cryptocurrency trading data and not on information conveyed by individuals who may not even be Bitcoin investors, as is the case for existing sentiment indices based on Google Trends (Li et al., 2021). Fourth, this paper shows that the selection of certain features substantially improves the predictive performance of our model based on BSVR. This particular area of predicting cryptocurrency prices has been overlooked in previous studies, and this lack has not only limited the accuracy of the machine learning-based models, but also reduces the transparency of the existing prediction models.

This research paper is structured as follows. Section 2 provides an overview of prior research on forecasting methods applied to cryptocurrencies. It then demonstrates the shortcomings of these studies and validates our methodological approach. Section 3 describes the data and our methodology, as well as the benefits of BSVR over alternative forecasting methodologies. The fourth section analyzes the empirical outcomes of our various models. Section 5 offers conclusions about the research.

2. Related Literature

In this section, we provide an overview of previous work on forecasting cryptocurrencies and the theoretical rationale for incorporating investor sentiment into a cryptocurrency forecasting model. Over the past five years, researchers have shown an increased interest in cryptocurrency forecasting using both parametric and nonparametric time series models.

As shown in Table 1, univariate forecasting models applied to time series of different cryptocurrencies have so far been preferred in the literature. More recent studies have used multivariate approaches to utilize larger sets of cryptocurrency predictors that would consider the potentially important effects of other commodity and financial assets such as precious metals, foreign exchange rates, and stock markets (Sun et al., 2020; Liu et al., 2021). In addition, features associated with cryptocurrency markets have been included such as trading volumes and commission charges (Guo et al., 2021).

Regarding the methods used for forecasting, neural networks have proven to be effective models for capturing intricate nonlinear patterns in cryptocurrency time series data (Khaldi et al., 2019; Lahmir and Bekiros, 2021). However, complex neural network–based forecasting models may have a poor power of generalization, while at the same time being typically more computationally intensive than conventional machine learning models. To capture both local trends and high-level temporal patterns in cryptocurrency time series, convolutional neural networks (CNNs) were combined with long short-term memory (LSTM) neural networks (Alonso-Monsalve et al., 2020; Livieris et al. 2020, 2021b). The disadvantage of this approach, however, was the problem related to LSTM of a vanishing gradient resulting from the dynamics of nonlinear time series. Moreover, previous studies have not taken advantage of the ability of such models to process high-frequency multivariate data. In fact, the above-mentioned multivariate time-series models have been shown to be powerful in conjunction with support vector machine (SVM) prediction methods, owing to the efficiency of SVMs in dealing with high-dimensional data (Akyildirim et al., 2021). However, models based on SVMs, and this also applies to SVR, are not well suited for large and noisy data, such as cryptocurrency data (Lahmiri and Bekiros, 2019, 2020). To address this issue, boosting-based ensemble models have been considered an effective solution, owing to their robustness to overfitting and capacity to handle noisy data (Sun et al., 2020). Motivated by these considerations, we here propose an SVR-based model enhanced with bagging ensemble learning. It is generally considered more effective when learning a cryptocurrency forecasting model that only relevant predictors be used in the multivariate model (Mudassir et al., 2020). However, existing models only use the feature selection embedded in forecasting methods such as Random Forest (Mudassir et al., 2020), which lacks transparency and has a predictive performance that depends on the forecasting method used. To overcome this problem, here we perform the feature selection process independently of the forecasting method used by estimating the relevance of features based on their predictive capability while also considering their redundancy.

Concerning the predictive performance of the models for Bitcoin prices, which is the cryptocurrency used in the present study, the time period used to train and test the model turned out to be crucial because of the high volatility in certain periods. Therefore, comparing the performance of each model with the performance of each of the other models was problematic. In general, however, non-parametric models based on machine learning outperformed their parametric competitors in terms of the accuracy of predictions in various studies.

Insert Table 1 about here

While information gathered from cryptocurrency markets is an important aid for investment decisions, it is widely acknowledged that such information is inadequate to accurately predict the evolution of cryptocurrencies. Indeed, most information influencing investors' decisions takes a linguistic, rather than a numerical, form. Trading based on investor sentiment is nothing new in studies in this field, and it allegedly beat trading strategies that were based strictly on data from the cryptocurrency market (López-Cabarcos et al., 2021; Guégan and Renault, 2021). In the context of the cryptocurrency market, several approaches have been introduced to assess investor sentiment.

The first group of approaches to estimate cryptocurrency-related sentiment used market-based measures. For example, Bukovina and Marticek (2016) used a decomposition of Bitcoin into two parts, a rational component and a less rational component. The average number of transactions per block represented the rational part by reflecting the fundamentals, that is, the relationship between supply and demand. The second component represented the revenue of the miner per transaction. The latter captured less rational factors, such as speculative investments induced by sentiment. Consequently, this component was recommended as a plausible measure of sentiment. Although this kind of measure is easy to calculate from the financial data available, it nevertheless represents an indirect measure of investor sentiment, lacking in interpretability and requiring a strong theoretical justification. Moreover, it was reported that such a sentiment indicator could only explain a small portion of the total volatility of Bitcoin (Bukovina and Marticek, 2016).

A second group used a direct survey-based measure of investor sentiment (Anamika et al., 2021). This type of measure is based on investor opinions regarding market expectations. However, such surveys can only be conducted infrequently and require a representative sample of respondents. This in turn often results in measurement errors.

A third approach was based on Google Trends to reveal periods of cryptocurrency sentiment and used statistical tests to explain a large portion of the price clustering in Bitcoin by using Google Trends (Baig et al., 2019). In particular, Chen et al. (2020) developed a sentiment index to assess the sentiment surrounding the fear of the coronavirus by using the number of Google search requests for coronavirus-related terms. Their results showed that market volatility was amplified by fear sentiment, and that this fear sentiment could be used to explain the high trading volumes and negative returns of Bitcoin. The search volume from Google was also used to construct FEARS (Financial and Economic Attitudes Revealed by Search), an investor sentiment index proposed by Burggraf et al. (2021) to show that the Bitcoin market was not efficient according to the efficient market hypothesis. Similar findings were provided by several other studies (Oad Rajput et al., 2020; Kapar and Olmo, 2021). On the one hand, using search-based sentiment measures was beneficial because the underlying data were available at a high frequency compared with the data from survey-based sentiment indices. On the other hand, recent evidence in behavioral finance suggested that more relevant sentiment information was covered by social media-based sentiment indices (Brochado, 2020).

A fourth approach was to extract sentiment from social media or newspapers. Georgoula et al. (2015) extracted sentiment values from Twitter by categorizing daily posts related to cryptocurrencies as having positive sentiment or negative sentiment. This Twitter sentiment ratio was reportedly positively associated with Bitcoin prices in the short term. Similarly, news articles and blog posts were collected by Karalevicius et al. (2018) to calculate the frequencies of the positive and negative words used. Statistical tests confirmed interactions between the media-based sentiment score and the Bitcoin price. Notably, a tendency among investors to overreact to media content was observed in the short term. The Bitcointalk database was used by Ahn and Kim (2019) to collect comments posted on bulletin boards. Their study employed three sentiment dictionaries to obtain sentiment scores and found that sentiment disagreement implied a high volatility in the cryptocurrency price. López-Cabarcos et al. (2021) extracted social media sentiment from the StockTwits social network, and showed how investor sentiment significantly affected Bitcoin volatility in stable (less volatile) periods. Guégan and Renault (2021) also used StockTwits sentiment based on the number of bullish (positive) and bearish (negative) user postings. Their study revealed a strong interaction between investor sentiment and Bitcoin returns in very short frequencies (only up to 15 minutes). Some serious weaknesses of social media–based sentiment are that (1) they are posted by noise and often poorly informed investors vulnerable to certain opinions and emotions, and (2) it is difficult to

extract an accurate sentiment score due to the ambiguity in the meaning of investors' postings, which are usually written in less formal language. As an alternative to social media, other authors used newspapers to examine the effect of news sentiment on the price of Bitcoin (Rognone et al., 2020). Surprisingly, the price was positively affected by news with both positive sentiment and negative sentiment, indicating investor enthusiasm for cryptocurrencies. The key problem of news–based sentiment was that the news was concerned with what had happened, as opposed to what is likely to happen in the future (Kearney and Liu, 2014; Flori, 2019). Another bias resulted from the effects of periods when cryptocurrencies received increased attention in the media (Gaies et al., 2021). To summarize the above, while the idea of using investor sentiment to predict future Bitcoin prices was appealing, none of these approaches have proved to be really convincing so far.

3. Methodology

3.1 Data and Features

To determine the extent to which investor sentiment could be used to predict Bitcoin prices, daily data were collected from Refinitiv[1] for the period from March 2018 to May 2022. Bitcoin prices (i.e., daily closing prices) are denominated in US dollars.

The innovative BMI, established in 2018 by Tom Lee from Fundstrat Global Advisors, was used in this research paper as a measure of investor sentiment. This index helps traders to gauge investor attitudes toward Bitcoin prices and, more broadly, toward the cryptocurrency sector. The BMI scale ranges from 0 to 100, with 0 representing the highest degree of misery and 100 representing the highest degree of pleasure; therefore, the BMI signifies how miserable or pleased Bitcoin holders are. Due to the significant volatility of Bitcoin, the BMI is an indicator used by traders to determine the ideal periods in which to produce gains. Hence, it provides investors with a tool for applying the well-known "buy the dip" investment strategy to Bitcoin trading, that is, to use the high volatility, which is often considered a disadvantage of Bitcoin, to their advantage. The index is calculated using the successful percentage of total trade deals and their volatility. The BMI is computed using the winning trade ratio's *z*-score (a measure of the number of standard deviations from the mean of a single data point) and the upside/downside volatility *z*-score.

[1] For descriptive statistics for all variables used in the empirical analysis, see Table A in the Appendix. Source: https://www.refinitiv.com/en

Because the BMI is thought to be a critical component in predicting whether an investor should buy or sell Bitcoin, we study how the BMI can forecast Bitcoin prices. To our knowledge, only one study has examined the effect of investor sentiment on Bitcoin returns using this metric (Gaies et al., 2021); it revealed a significant but asymmetric impact of the BMI on the Bitcoin return in the short term and long term. The positive effect of an optimistic shock was larger than the negative effect of a pessimistic shock in the short run. Conversely, in the long run, Bitcoin returns were more responsive to pessimistic shocks than to optimistic ones (Gaies et al., 2021). No study has otherwise yet used the BMI to forecast Bitcoin prices, so the present use of the BMI is a significant innovation. The majority of prior studies have used search mines associated with the term "Bitcoin" on the Internet as an estimate of user moods. They derived their sentiment variables primarily from search engines and social media data, such as Google searches, Twitter, Wikipedia, Facebook, blogs, financial and business magazines, newspaper articles, positive and negative news, and forums (Rognone et al., 2020).

Without a doubt, the profiles of a sizable fraction of the writers and disclosers of this data are very different from the profiles of genuine users (Kaabia et al., 2020). Indeed, the writers and disclosers may not possess Bitcoins or be members of the cryptocurrency community, and this fact may create a bias in the effectiveness of the sentiment index constructed from these data. Another source of bias might be the difference in influence between individual users, which cannot be examined with these kinds of data.

There were two other sources of bias in previous studies. First, the sentiment variable was constructed using word classification dictionaries. This implied a significant subjectivity bias in attributing the sentiment of each word, as well as the possibility that the aggregate sentiment score for a word arrangement may have differed from the actual hidden sentiment derived from the entire sentence. Second, media-based sentiment variables were expected to be impacted by times of high media attention for Bitcoin, as well as periods of low media interest. Without accounting for this unreported impact, the emotion variable may have been subject to endogeneity bias. Because the BMI is calculated as the ratio of successful trades to total transactions and their volatility, it avoids all of these issues and is therefore more suited for our research. As shown by Gaies et al. (2021), the BMI is a market sentiment index that avoids individual interpretation bias, because it is built on real crypto-currency trading data and not on information conveyed on Google or Twitter by individuals, most of whom are not Bitcoin investors, as with existing sentiment indices based on Google Trends (Li et al., 2021).

Along with the BMI, and consistent with current research on the predictors of Bitcoin prices (Balcilar et al., 2017; Huang et al., 2018), we included features from the cryptocurrency market, oil prices, and technical indicators used by traders that are available on Refinitiv.com and Reuters.com (two main real-time market data providers of the financial community):

Cryptocurrency market:

- Network usage, such as block counts and sizes, because blockchain fundamentals were found to determine the prices of cryptocurrencies (Liebi, 2022).
- Transactions, as inter-exchange transaction features improved the performance of Bitcoin price prediction in previous studies (Guo et al., 2021).
- Mining features, such Bitcoin's hash rate and difficulty of finding a block, because the cost of production is reportedly important in explaining Bitcoin price dynamics (Hayes, 2019).
- Market forces: demand-based determinants (the number of active addresses, counts and values of transactions), and supply-based determinants (issuance, market capitalization, current and expected supply of amounts of units in circulation), representing the interaction between cryptocurrency supply and demand (Bouri et al., 2019; Ciaian et al., 2016).
- Market data for Bitcoin and other cryptocurrencies, including the price and volume of Ethereum, Bitcoin Cash, Ripple, and Litecoin, and the volatility and return on investment for Bitcoin (Balcinar et al., 2017; Liu et al., 2021).

Oil price

- The WTI (West Texas Intermediate) crude oil price is a macroeconomic determinant included on the grounds that cryptocurrency markets are vulnerable to price fluctuations in oil markets (Jin et al., 2019).

Technical indicators:

- Chaikin's volatility indicator which compares the spread between high and low Bitcoin prices. The results of previous empirical studies suggested that the use of technical indices could help predict Bitcoin returns that are difficult to capture with fundamentals (Akyildirim et al., 2021; Alonso-Monsalve et al., 2020). Chaikin's volatility indicator is a popular one also used in previous related research (Huang et al., 2019).
- Psychological line: technical analysis indicator of Reuters.com. This is the ratio of the number of rising periods over the total number of periods for Bitcoin. This represents

the buyer's purchasing power in proportion to the seller's selling power. This indicator was used to predict the price of Bitcoin in previous studies (Mallqui and Fernandes, 2019; Mallqui et al., 2021) because it provided information about the direction and strength of the cryptocurrency market trend.

- Accumulation/distribution (A/D) indicator is a cumulative indicator that uses the Bitcoin price and volume to assess whether an asset is being accumulated or distributed, thus identifying divergences between the Bitcoin price and volume flow (Belloca et al., 2022).

3.2 Feature Selection Using Flower Pollination Algorithm

The advantage of feature selection in high-dimensional prediction tasks is that the dimensionality of the feature space is reduced, enabling a faster execution of prediction algorithms and increased accuracy. Feature selection methods are typically divided into wrapper, embedded and filter methods (Chandrashekar and Sahin, 2014). Wrappers use an enumerative algorithm to explore the feature space by testing the predictive performance on each feature subset (Niu et al., 2020). The predictive performance measure is used to assess candidate feature subsets. This approach is computationally intensive and slow because the prediction model must be re-trained for all candidate subsets. To this end, evolutionary algorithms can be used as a search method (Xiong, 2002). In the embedded methods, a feature selection process is conducted during the process of estimating the parameters of the prediction model, leading to a dependence of the predictive performance on the prediction model. Filter methods select features based on the actual properties of the data. By estimating the feature relevance using an assessment function, filters work separately from any prediction algorithm. Traditional filter methods select redundant features without considering the relationships between features. More sophisticated methods, such as the correlation-based filter method, seek to mitigate this issue by eliminating features that are highly correlated with each other. We adopted this method in this study because it proved to be effective in selecting technical indicators for stock market predictions in previous research (Sugunnasil and Somhom, 2010).

In the correlation-based feature selection method, the quality of a feature subset is assessed based on the individual predictive ability of each feature reduced by the degree of redundancy between features. The objective function $f(S)$, representing the merit of the feature subset based on Pearson's correlation coefficient, can be defined as follows:

$$f(S) = \frac{k \times \overline{r_{cf}}}{\sqrt{k + k \times (k-1) \times \overline{r_{ff}}}}, \quad (1)$$

where S is a candidate subset with k features (chosen from the original set of n features), $\overline{r_{cf}}$ is the average feature-class correlation, and $\overline{r_{ff}}$ is the average feature-feature correlation. The numerator in Equation (1) represents the predictive power of selected features, while the denominator denotes their degree of redundancy. As such, irrelevant features exhibit a low correlation with the class or are redundant due to their high correlations with other features. Correlations between numerical and categorical features can be computed using the discretized values of the features. Symmetrical uncertainty (SU) is used as a correlation measure to estimate the degree of association between discrete features (X and Y):

$$SU = 2.0 \times \frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)}, \quad (2)$$

where H stands for the entropy of the given feature. After computing a correlation matrix using Equation (2), a heuristic search strategy is used to find a good subset of features. In metaheuristic search methods, each agent represents a candidate subset S, and the best solution is given by the agent with the highest value of the fitness function $f(S)$ in each cycle. Initially, a random candidate subset is chosen. To select the appropriate heuristic search strategy, we followed the recommendation based on extensive experimentation with swarm search methods (Fong et al., 2018). In that comparative study, the Flower search method provided excellent results, outperforming both traditional metaheuristic methods and state-of-the-art population-based methods in terms of accuracy and execution time.

The Flower Pollination Algorithm (FPA), first introduced by Yang (2012), was inspired by the process of pollinating flowers. The algorithm randomly combines global and local pollination to find the best solution using Equation (3) and Equation (4), respectively. The global pollination can be expressed as follows:

$$x_i^{t+1} = x_i^t + \alpha \times L(\gamma) \times (g^* - x_i^t), \quad (3)$$

$$L(\gamma) = \frac{\gamma \times \Gamma(\gamma) \times \sin(\gamma)}{\pi} \times \frac{1}{s^{1+\gamma}}, \quad (4)$$

where x_i^t denotes an i-th pollen at the t-th iteration, g^* stands for the best solution at the t-th iteration, α is the scaling factor, s denotes the step size ($s > 0$), $L(\gamma)$ is the step size of the Lévy flight, and $\Gamma(\gamma)$ is the gamma function ($1 \leq \gamma \leq 2$). The local pollination is defined as follows:

$$x_i^{t+1} = x_i^t + \varepsilon \times \left(x_j^t - x_k^t\right), \quad (5)$$

where x_j^t and x_k^t are pollens from the *j*-th and *k*-th flowers, $j \neq k$, respectively.

Rodrigues et al. (2015) modified the original version of the FPA to produce a binary version suitable for searching the feature space. In this algorithm, each candidate solution is represented by an *n*-dimensional binary vector indicating whether a feature belongs to the selected set of features or not. The binary vector is obtained from the values of pollens using the sigmoid function.

3.3 Bagged Support Vector Regression

SVR has proven to be a powerful and robust method that is able to account for the multidimensional and dynamic characteristics of financial data. This method is an adaptation of SVM for regression problems. Like SVM, SVR is based on structural risk minimization, the objective of which is to estimate nonlinear data generation processes by using a regularization term. Indeed, by adopting nonlinear interactions, the predictive power of regression models can be enhanced in the financial domain. In this context, SVR and SVM have successfully been applied to the cryptocurrency market (Peng et al., 2018; Mallqui and Fernandes, 2019; Poongodi et al., 2020; Aggarwal et al., 2020; Akyildirim et al., 2021). However, the limited performance of SVR occurs because SVR is implemented using approximation algorithms due to computational complexity. This leads to the possibility that a single SVR model may not accurately learn the parameters reaching the global optimum of the objective function. Therefore, for more complex regression problems, it is recommended to combine several base SVR models into an ensemble model to achieve the desired accuracy and robustness of forecasting. In BSVR, an ensemble of SVR is first trained independently using the bootstrap method and then aggregated, which in turn reduces the variance of base SVR models.

The principle of the BSVR method is presented below.

Let $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be a set of training data, and $f(x)$ denotes the SVR nonlinear estimating function defined as follows:

$$f(x) = \big(w \cdot \Phi(x)\big) + b, \tag{6}$$

where $\Phi(x)$ stands for a nonlinear transformation function, w is the support vector weight, b is a bias term, and $\cdot$ denotes a dot product. To estimate the values of parameters w and b in Equation (6), the regularized risk function $R(f)$ is used as follows:

$$R(f) = C \sum_{i=1}^{n} \Gamma(f(x_i) - y_i) + \frac{1}{2}\|w\|^2, \tag{7}$$

where C is a constant term and $\Gamma(.)$ represents the ε-sensitive loss function:

$$\Gamma(f(x_i) - y_i) = \begin{cases} |f(x) - y| - \varepsilon, & \text{for } |f(x) - y| \geq \varepsilon \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

By solving the quadratic optimization problem and substituting the weight vector w in Equation (6) by $w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i$, $R(f)$ in Equation (7) and the ε-sensitive loss function in Equation (8) can be minimized as follows:

$$\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)k(x_i, x) - \sum_{i=1}^{n}\alpha_i^*(y_i - \varepsilon) - \alpha_i\,(y_i + \varepsilon),$$

s.t.

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \quad \alpha_i, \alpha_i^* \in [0, C]. \quad (9)$$

where $k(x_i, x)$ is a kernel function enabling transformation of the low-dimensional space data into a high-dimensional space and (α_i, α_i^*) is a Lagrange multiplier pair. Note that the Lagrange multipliers represent the solution to the optimization problem, and the constant term C denotes the penalty for the estimation error in Equation (9).

In this study, we use a bagging method (Breiman, 1996) to produce the SVR ensemble. In the bagging method, individual SVRs are first trained independently using a bootstrap procedure and then their outcomes are aggregated into the final prediction. To reduce the variance of the ensemble model and avoid overfitting via aggregation, the bootstrap datasets should be as different as possible. To reach this goal, the bootstrap method repeatedly builds B replicate training datasets (bootstraps) using random resampling with replacement from the original training data. Individual SVR_b, $b = 1, 2, \ldots, B$, are trained on these bootstrap replicates. The aggregated prediction of the bagging ensemble $\{\text{SVR}_b\}$ is computed using a simple average of B predictors. It should be noted that when the diversity of the ensemble is increased at the same time that the mean generalization error of individual predictors is reduced or preserved, a generalization error is reduced.

4. Results and Discussion

4.1 Experimental Setup

First, the Bitcoin price time series were divided into three subsets: (1) The period from March 1, 2018 to March 31, 2020 was used for training (759 transaction dates); (2) the period from April 1, 2020 to November 30, 2021 was used as the first testing period (609 transaction dates); and (3) the period from December 1, 2021 to May 18, 2022 was used as the second testing

period (169 transaction dates). Alternative testing periods were used to incorporate events that might be expected to affect the Bitcoin market. Indeed, bearish sentiment dominated the Bitcoin market in 2020 and 2021, reaching a peak of more than USD 68,000 in November 2021. However, by the end of 2021, the price of Bitcoin began to fall, slipping into a bear market and experiencing one of its largest historical declines in 2022. This was mainly ascribed to the crisis of algorithmic stablecoins.

In-keeping with the settings recommended by Fong et al. (2018), the parameters of the FPA were set as follows: the chaotic coefficient was 4.0, number of iterations was 20, mutation probability was 0.01, pollination was 0.33, and size of the population was 20.

The training parameters of bagging were as follows: the size of each bag (as a percentage of the training data) was 100, and the number of iterations was 10. To set the hyperparameter combination for SVR, the grid search strategy was used; it investigates all combinations by using a 10-fold cross-validation to prevent overfitting and to estimate the predictive performance of BSVR in terms of the mean square error. The hyperparameter combinations were chosen from the candidate values $C = \{2^{-1}, 2^0, 2^1, \ldots, 2^6\}$, and $\varepsilon = \{0.0001, 0.001, 0.01, 0.1\}$, and the polynomial kernel function was used with the exponent of 1.0. The optimized value hyperparameter combination for SVR was $C = 2^2$ and $\varepsilon = 0.001$.

The proposed prediction model was deployed in Weka 3.8.4. More precisely, the MetaphorSearchMethods library was used for the feature selection using the FPA, and experiments with BSVR were carried out in the Ensemble library using the SMOreg algorithm as the base predictor.

In keeping with the guidelines from earlier literature (Altan et al., 2019), the performance of the prediction model was assessed using three standard metrics, namely the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean square error (RMSE). Following previous research (Shen et al., 2020), the predictive performance was assessed for different forecasting horizons, ranging from 1-day-ahead to 30-days-ahead horizons.

4.2 Effect of Feature Selection on Prediction Accuracy

To assess the effectiveness of integrating the feature selection component into our prediction model, we first performed the feature selection process using the correlation-based filter with the FPA. Table 2 lists the selected features. It is important to note that only training data were used in the feature selection process to avoid feature selection bias. The results show that 19

features were selected out of the initial set of 55 features (see Appendix A for the list of features). In other words, the feature set was reduced by approximately 65%, implying that most Bitcoin market and technical indicators are either irrelevant or redundant for predicting the Bitcoin price. The results also suggest that the BMI is a relevant and essential predictor of theBitcoin price.

Insert Table 2 about here

Note that some of the features discarded may seem surprising given their importance in previous studies. For example, the hash rate was identified as an important variable by Hayes (2019), but it did not make the cut in our study because it was strongly correlated with other features, particularly supply-based features (see Appendix B for the correlations between features), and this reduced its overall merit in Equation (1).

The purpose of the following run of experiments was to examine the effect of feature selection on the BSVR forecasting performance. The three indicators of the forecasting performance, MAE, MAPE and RMSE were higher for our model without feature selection than for the model with FPA feature selection (Fig. 1). So, the prediction model using all of the available Bitcoin price determinants was not efficient enough to accurately predict its price over the next 30 days, indicating that relevant predictors must first be identified in order to further improve the accuracy of forecasting.

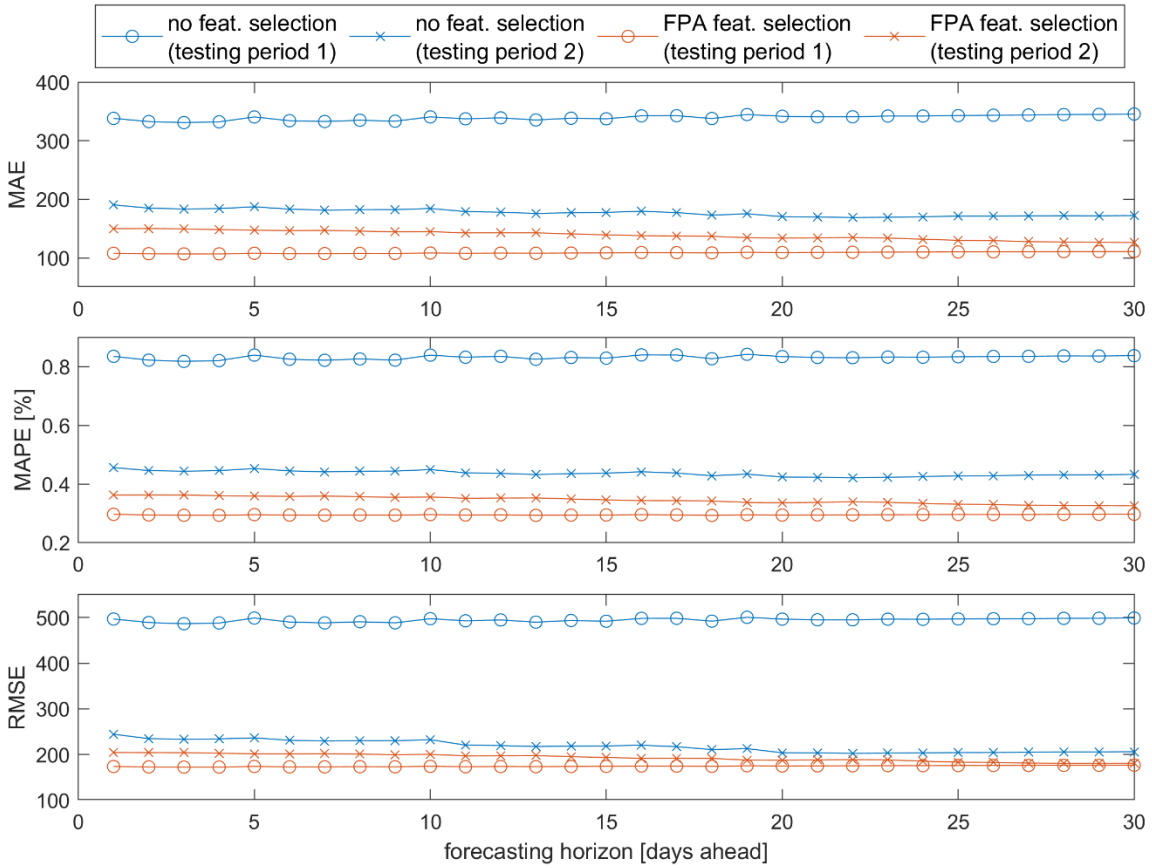

Fig. 1: Effect of feature selection on Bitcoin forecasting performance of BSVR

4.3 Effect of Investor Sentiment on Prediction Accuracy

In the baseline model, only financial indicators, as selected using the correlation-based filter with the FPA (Table 2), were used to predict the Bitcoin price (i.e., without considering investor sentiment). The purpose of this baseline model was to examine whether or not the inclusion of investor sentiment, expressed as the BMI, is effective in terms of prediction accuracy.

On the basis of the MAE, MAPE and RMSE values, Fig. 2 shows that our sentiment indicator, the BMI, improved the efficiency of our forecasting model for both testing periods. More precisely, the effect of investor sentiment was consistent for the MAE and MAPE values for different forecasting horizons and for both testing periods (bull and bear markets). Regarding the performance in terms of the RMSE values, the sentiment effect decreased in the second testing period for the multi-day prediction, indicating the low predictive power of investor sentiment in the case of fluctuations during the period of rapid Bitcoin price decline. First of all, this result extends the work of Gaies et al. (2021), who revealed that the BMI explained Bitcoin returns, without addressing the predictive power of this sentiment variable. Moreover,

it contradicts the results of Aggarwal et al. (2020), and this enable us to state that machine learning methods such as SVR are not adequate alone to predict Bitcoin prices, even if technical indicators are added.

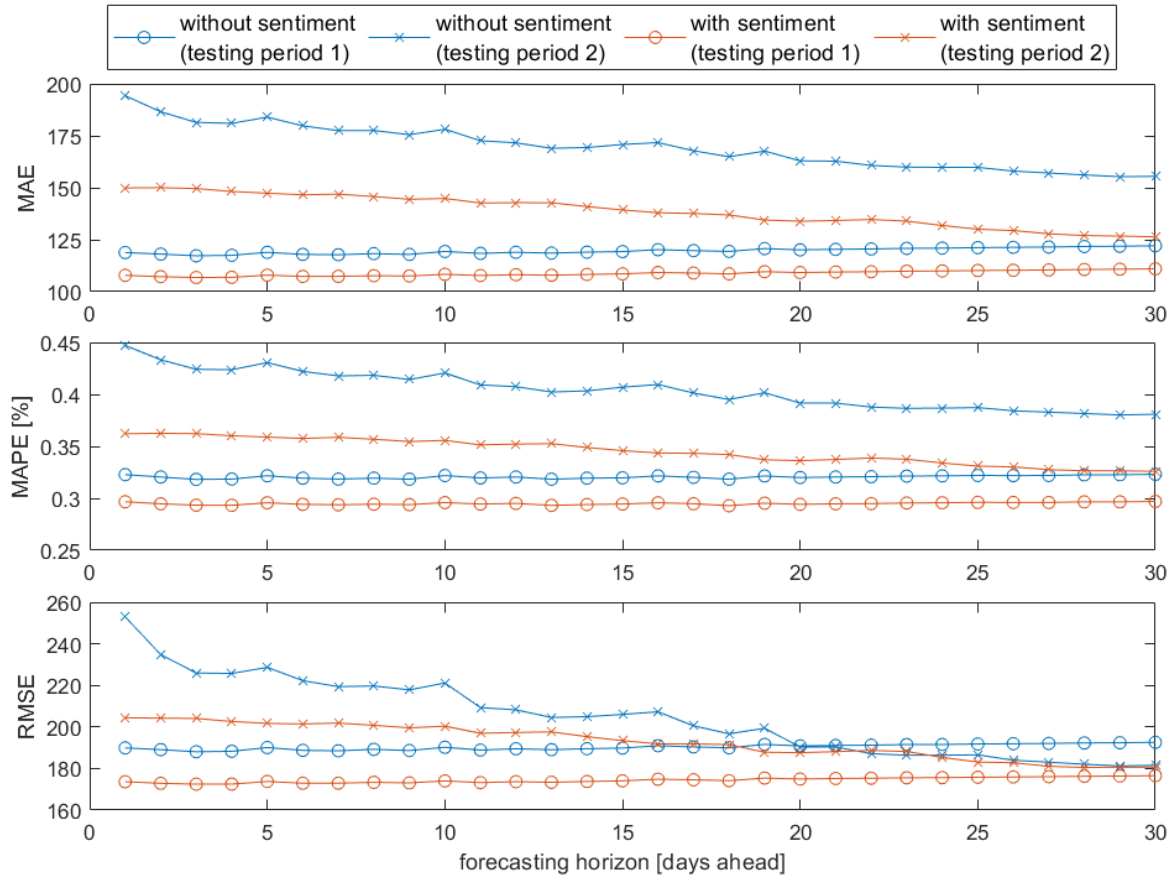

Fig. 2: Bitcoin forecasting performance of BSVR with and without investor sentiment

In addition to this, given the difference between the two curves related to the Bitcoin forecasting performance of BSVR with and without investor sentiment, the explanatory power of the sentiment variable does not seem to depend on the short-term forecasting horizon (up to 30 days in this case). The results show that the prediction model with investor sentiment outperforms the baseline model by 28.37, 0.13%, and 68.16 on average in terms of the MAE, MAPE, and RMSE, respectively. Therefore, we can conclude that integrating the BMI sentiment index into the prediction model could improve the accuracy of Bitcoin market forecasts.

Finally, Fig. 3 and Fig. 4 highlight the good performance of our forecasting model by comparing current and predicted Bitcoin prices for the different forecasting horizons (1 to 30 days), including the highly volatile mid-2020s.

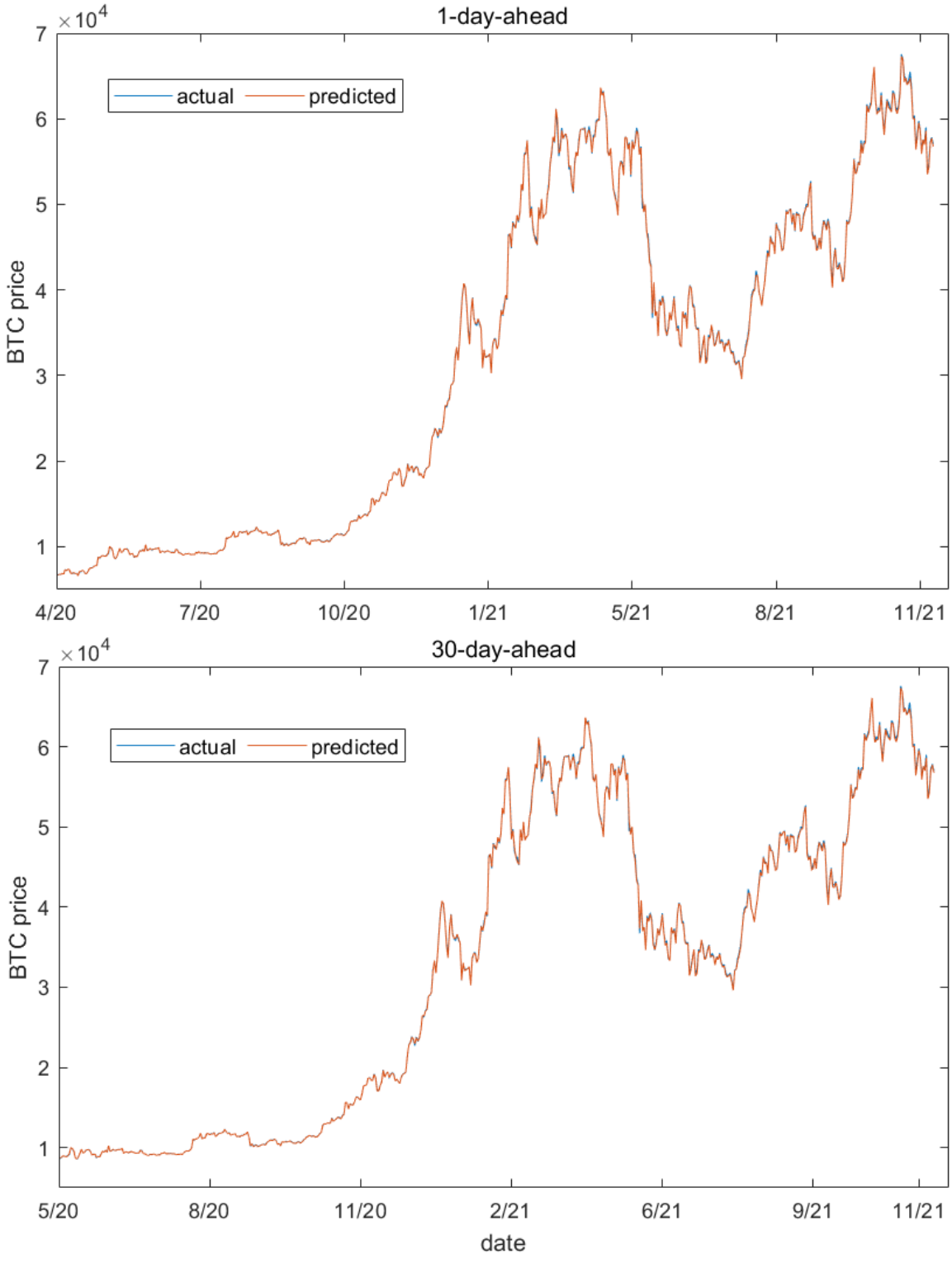

Fig. 3: Actual versus predicted Bitcoin price (BTC) during testing period 1 using BSVR for different forecasting horizons

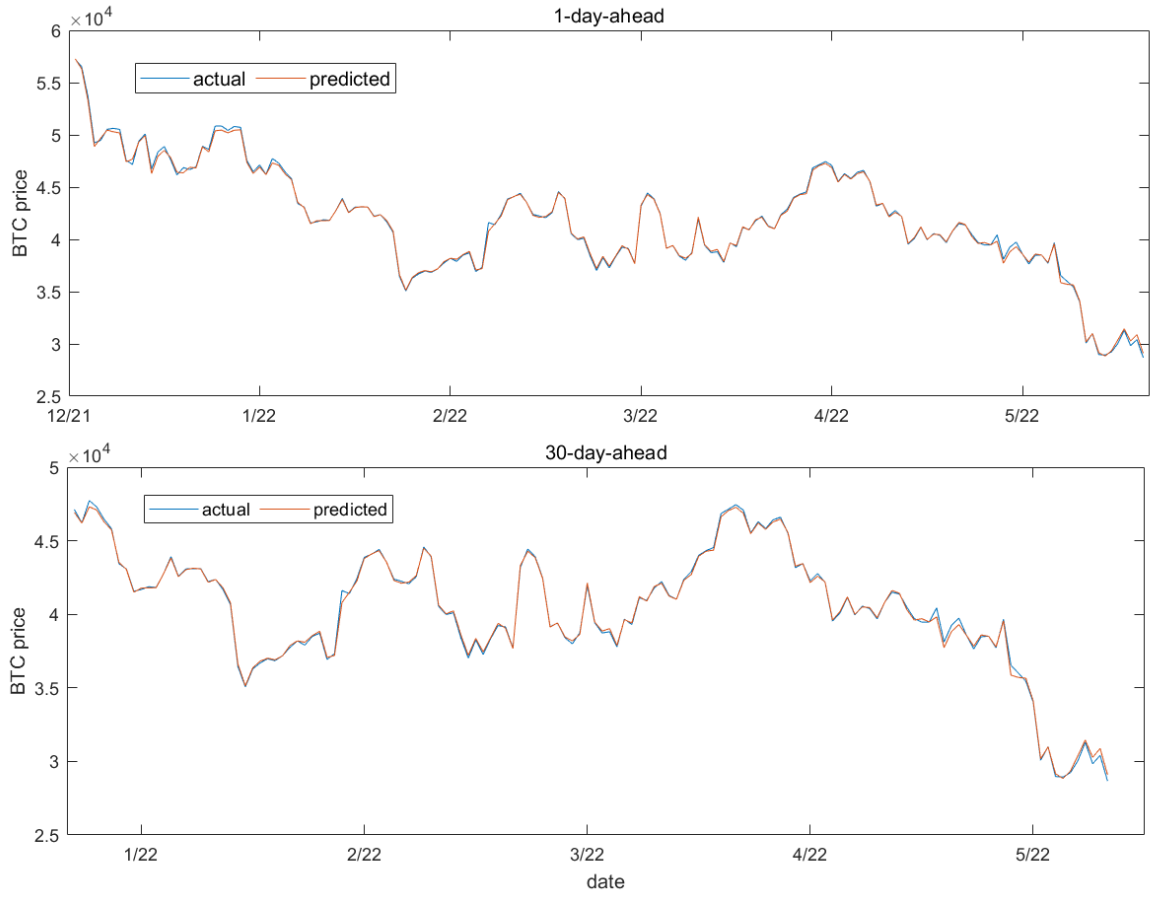

Fig. 4: Actual versus predicted Bitcoin price (BTC) during testing period 2 using BSVR for different forecasting horizons

4.4 Comparison with State-of-the-art Prediction Models

In comparison with the findings of earlier research, the model proposed herein (BSVR) has the benefit of being based on sentiment and using Bitcoin trade data efficiently. Indeed, Table 3 and Table 4 show that it outperforms the baseline random walk martingale model (Zargar and Kumar, 2019) and the following seven leading models:

- SVR (Jana et al., 2021) – a univariate time-series forecasting model using five lagged values of the Bitcoin price as features. In accordance with Jana et al. (2021), SVR was trained using radial basis kernel with $C = 4$, $\varepsilon = 0.02$, and the width of the kernel was $\sigma = 4.0$.
- ARIMA (Yenidoğan et al., 2018) – a multivariate model using foreign exchange rates (GBP, EUR, and JPY) as predictors. Following Yenidoğan et al. (2018), the parameters of the ARIMA model were inferred automatically using the pmdarima Python library.

- Artificial neural network (ANN) (Adcock and Gradojevic, 2019) – a feedforward neural network with 12 hidden layer units using technical indicators and the Chicago Board Options Exchange (CBOE) volatility index as inputs.
- Random Forest (RF) (Gradojevic et al., 2021) – a multivariate model using a wide range of technical indicators derived from the Bitcoin price and volume. In accordance with Gradojevic et al. (2021), the sentiment direction (fear / greed) was also incorporated into this model. We also adopted the RF setup as follows: the feature subset size = $\log_2 k + 1$, where k is the number of features, and 100 trees were generated to construct the RF model.
- Radial basis function neural network (RBFNN) (Lahmiri and Bekiros, 2020) – a univariate model using the previous five observations as inputs and Gaussian functions in the hidden layer units.
- Stacked artificial neural network (SANN) (Mudassir et al., 2020) – a multivariate forecasting model using seven technical indicators; the SANN model consists of five individual ANN models with ReLU (rectified linear unit) activation functions trained using the Adam optimizer.
- Bidirectional long short-term memory (Bi-LSTM) (Wu et al., 2018) – the previous Bitcoin price and volume were used as input features in-keeping with the original study. In accordance with Wu et al. (2018), we used Bi-LSTM with one hidden Bi-LSTM layer containing 1000 ReLUs, followed by one dense layer with linear units. Again, the Adam optimizer was used to train the forecasting model. The values of the remaining hyperparameters, as for the previous models, were determined on a trial and error basis.

Remarkably, the overall performance of the RF and RBFNN models was poor, which can be attributed to model overfitting, despite the fact that we attempted to control for model complexity through the tree depth in the case of the RF and the number of neurons in the hidden layer for the RBFNN. Overfitting then prevented these models from predicting the sharp rise and fall of the price of Bitcoin in the two testing periods.

Another interesting finding came from comparing the results of the prediction models with the baseline random walk martingale model. Most of the machine learning–based models had problems outperforming the baseline model and only achieved better results for longer forecasting horizons. However, it is important to note that the BSVR substantially outperformed the baseline random walk martingale model for all forecasting horizons, and this further strengthened our confidence in the proposed model.

In addition to this, even if the three performance indicators (MAE, MAPE, RMSE) were greatly improved between the SVR and BSVR models, the gap is even greater with other models. This result suggests that (1) SVR-based methods are efficient for multivariate forecasting models of the Bitcoin price; and (2) the bagging-based ensemble learning further improves the predictive performance of SVR by reducing its variance and increasing its robustness to overfitting.

A comparison of the two testing periods in Table 3 and Table 4 reveals that, as might be expected, the accuracy of prediction of all of the models for the later period declined slightly. However, the performance of the proposed BSVR model was still acceptable and the MAPE was below 0.4% even for a 30-day prediction, which can be considered a good result compared with those of other existing studies (Mudassir et al., 2020). Overall, these results emphasize the validity of the proposed model, providing highly accurate predictions in both bullish and bearish sentiment periods.

Insert Table 3 about here

Insert Table 4 about here

5. Conclusion

Recently, an increasing number of studies have shown a correlation between sentiment and Bitcoin prices. However, no consensus on the nature of this connection has yet been reached, nor was there a consensus on the efficiency of a sentiment feature to forecast Bitcoin prices. Indeed, to demonstrate such an effect, it was necessary to devise a multivariate prediction model, because the Bitcoin price is affected by many other factors. This study presented a novel forecasting model to integrate investor sentiment for forecasting the Bitcoin price. The findings from this study made several contributions to the current literature. First, investor sentiment based on the BMI was considered herein for the first time in cryptocurrency forecasting. Second, this was the first research study to show the effectiveness of feature selection in multivariate models for cryptocurrency forecasting. Third, the proposed prediction model provided a state-of-the-art performance for the Bitcoin price compared with existing univariate and multivariate models. Our findings appear to be well substantiated by the results for different Bitcoin price forecasting horizons.

The implications of this research are threefold. First, by demonstrating the capacity of the BMI (based on transaction data, a Bitcoin user sentiment) to forecast Bitcoin prices in the short term, the study's results may assist portfolio managers in developing effective short-term investing

strategies. This result is not trivial when one examines the asymmetric behavior of optimistic and pessimistic investors in the Bitcoin market that has been revealed by Gaies et al. (2021).

Second, because this sentiment factor is an important determinant of Bitcoin's price and thus returns, financial analysts and researchers should incorporate it into their asset pricing models to improve their robustness. Because the most comprehensive asset pricing model, derived from the capital asset–pricing model (CAPM), used by practitioners is Carhart's four-factor model (Sahut and Pasquini-Descomps, 2015), a fifth factor could be added to this model. In this way, some researchers such as Ji et al. (2020), for example, added a sentiment factor to the Fama–French three-factor model to explain the stock performance of Chinese firms with blockchain-centric business.

Third, this paper provides a fresh dimension to the discussion about the nature of Bitcoin. Ever since credit currencies supplanted metallic currencies, it has been commonplace to assert that money is based on trust. In foreign currency markets, national currencies, often known as fiat currencies, are valued based on the trust they inspire among financial operators. If trust is a reflection of a good feeling or attitude, then one could wonder how the value of Bitcoin differs from that of fiat currencies, given that trust is a positive emotion. The main difference is that the mood of Bitcoin users is not based on fundamentals, as is the case with fiat currencies (Chen et al., 2021), but rather on their belief in this innovative technology, as is the case with other Fintech goods.

As with every empirical study, this one has limitations related to the data and features employed. Because the focus of the study was on investor sentiment expressed in terms of the BMI, we were unable to assess the role of other approaches in estimating investor sentiment, including the uncertainty indicators introduced only recently (Aharon et al., 2022). The current model was designed to estimate the Bitcoin price level. Of course, some investors may prefer to forecast Bitcoin price movements (i.e., upward and downward). However, the proposed model could be extended to predict the direction of change by replacing the SVR base predictor with a bagged SVM classification model. The current study was also limited by using only data on Bitcoin and not on other cryptocurrencies. Finally, it is plausible that the specific period analyzed could have influenced the results. These are recommended avenues for further study.

Acknowledgments

This research paper was made possible by a grant from the Czech Sciences Foundation (No. 22-22586S).

References

Adcock, R., Gradojevic, N. 2019. Non-fundamental, non-parametric Bitcoin forecasting. Physica A: Statistical Mechanics and its Applications. 531, 121727.
Aggarwal, D., Chandrasekaran, S., Annamalai, B. 2020. A complete empirical ensemble mode decomposition and support vector machine-based approach to predict Bitcoin prices. Journal of Behavioral and Experimental Finance. 27, 100335.
Aharon, D. Y., Demir, E., Lau, C. K. M., Zaremba, A. 2022. Twitter-based uncertainty and cryptocurrency returns. Research in International Business and Finance. 59, 101546.
Ahn, Y., Kim, D. 2020. Sentiment disagreement and bitcoin price fluctuations: a psycholinguistic approach. Applied Economics Letters. 27(5), 412-416.
Akyildirim, E., Goncu, A., Sensoy, A. 2021. Prediction of cryptocurrency returns using machine learning. Annals of Operations Research. 297(1), 3-36.
Akhtaruzzaman, M., Boubaker, S., Umar, Z. 2022a. COVID–19 media coverage and ESG leader indices. Finance Research Letters. 45, 102170.
Akhtaruzzaman, M., Boubaker, S., Nguyen, D. K., Rahman, M. R. 2022b. Systemic risk-sharing framework of cryptocurrencies in the COVID–19 crisis. Finance Research Letters. 47, 102787.
Akyildirim, E., Goncu, A., Sensoy, A. 2021. Prediction of cryptocurrency returns using machine learning. Annals of Operations Research. 297(1), 3-36.
Alonso-Monsalve, S., Suárez-Cetrulo, A. L., Cervantes, A., Quintana, D. 2020. Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. Expert Systems with Applications. 149, 113250.
Altan, A., Karasu, S., Bekiros, S. 2019. Digital currency forecasting with chaotic meta-heuristic bio-inspired signal processing techniques. Chaos, Solitons & Fractals. 126, 325-336.
Anamika, Chakraborty, M., Subramaniam, S. 2021. Does sentiment impact cryptocurrency?. Journal of Behavioral Finance. 1-17. https://doi.org/10.1080/15427560.2021.1950723.
Baig, A., Blau, B. M., Sabah, N. 2019. Price clustering and sentiment in bitcoin. Finance Research Letters. 29, 111-116.
Balcilar, M., Bouri, E., Gupta, R., Roubaud, D. 2017. Can volume predict Bitcoin returns and volatility? A quantiles-based approach. Economic Modelling. 64, 74-81.
Bellocca, G. P., Attanasio, G., Cagliero, L., Fior, J. 2022. Leveraging the momentum effect in machine learning-based cryptocurrency trading. Machine Learning with Applications. 8, 100310.
Bergsli, L. Ø., Lind, A. F., Molnár, P., Polasik, M. 2022. Forecasting volatility of Bitcoin. Research in International Business and Finance. 59, 101540.
Bouri, E., Shahzad, S. J. H., Roubaud, D. 2019. Co-explosivity in the cryptocurrency market. Finance Research Letters. 29, 178-183.
Breiman, L. 1996. Bagging predictors. Machine learning. 24(2), 123-140.
Brochado, A. 2020. Google search based sentiment indexes. IIMB Management Review. 32(3), 325-335.
Bukovina, J., Marticek, M. 2016. Sentiment and bitcoin volatility. MENDELU Working Papers in Business and Economics 58. University of Brno, 1-12.
Burggraf, T., Huynh, T. L. D., Rudolf, M., Wang, M. 2021. Do FEARS drive Bitcoin?. Review of Behavioral Finance. 13(3), 229-258.

Chandrashekar, G., Sahin, F. 2014. A survey on feature selection methods. Computers & Electrical Engineering. 40(1), 16-28.
Cheikh, N. B., Zaied, Y. B., Chevallier, J. 2020. Asymmetric volatility in cryptocurrency markets: New evidence from smooth transition GARCH models. Finance Research Letters. 35, 101293.
Chen, C., Liu, L., Zhao, N. 2020. Fear sentiment, uncertainty, and bitcoin price dynamics: The case of COVID-19. Emerging Markets Finance and Trade. 56(10), 2298-2309.
Ciaian, P., Rajcaniova, M., Kancs, D. A. 2016. The economics of BitCoin price formation. Applied Economics. 48(19), 1799-1815.
Eom, C., Kaizoji, T., Kang, S. H., Pichl, L. 2019. Bitcoin and investor sentiment: statistical characteristics and predictability. Physica A: Statistical Mechanics and its Applications. 514, 511-521.
Figa-Talamanca, G., Patacca, M. 2019. Does market attention affect Bitcoin returns and volatility?. Decisions in Economics and Finance. 42(1), 135-155.
Flori, A. 2019. News and subjective beliefs: A Bayesian approach to Bitcoin investments. Research in International Business and Finance. 50, 336-356.
Fong, S., Biuk-Aghai, R. P., Millham, R. C. 2018. Swarm search methods in weka for data mining, in: Proceedings of the 2018 10th International Conference on Machine Learning and Computing, pp. 122-127.
Gaies, B., Nakhli, M. S., Sahut, J. M., Guesmi, K. 2021. Is Bitcoin rooted in confidence?–Unraveling the determinants of globalized digital currencies. Technological Forecasting and Social Change. 172, 121038.
Goodell, J. W., Goutte, S. 2021. Diversifying equity with cryptocurrencies during COVID-19. International Review of Financial Analysis. 76, 101781.
Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D., Giaglis, G. M. 2015. Using time-series and sentiment analysis to detect the determinants of bitcoin prices. Available at SSRN 2607167.
Gradojevic, N., Kukolj, D., Adcock, R., Djakovic, V. 2021. Forecasting Bitcoin with technical analysis: A not-so-random forest?. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2021.08.001.
Guégan, D., Renault, T. 2021. Does investor sentiment on social media provide robust information for Bitcoin returns predictability?. Finance Research Letters. 38, 101494.
Guo, H., Zhang, D., Liu, S., Wang, L., Ding, Y. 2021. Bitcoin price forecasting: A perspective of underlying blockchain transactions. Decision Support Systems. 151, 113650.
Hayes, A. S. 2019. Bitcoin price and its marginal cost of production: support for a fundamental value. Applied Economics Letters. 26(7), 554-560.
Hotz-Behofsits, C., Huber, F., Zörner, T. O. 2018. Predicting crypto-currencies using sparse non-Gaussian state space models. Journal of Forecasting. 37(6), 627-640.
Huang, Y., Duan, K., Mishra, T. 2021. Is Bitcoin really more than a diversifier? A pre-and post-COVID-19 analysis. Finance Research Letters. 43, 102016.
Huang, J. Z., Huang, W., Ni, J. 2019. Predicting bitcoin returns using high-dimensional technical indicators. The Journal of Finance and Data Science. 5(3), 140-155.
Jana, R. K., Ghosh, I., Das, D. 2021. A differential evolution-based regression framework for forecasting Bitcoin price. Annals of Operations Research. 306, 295-320.
Ji, Z., Chang, V., Lan, H., Hsu, Robert, C., H., Valverde, R., 2020. Empirical research on the Fama-French three-factor model and a sentiment-related four-factor model in the Chinese blockchain industry. Sustainability 12 (12), 5170.
Jia, B., Goodell, J. W., Shen, D. 2022. Momentum or reversal: which is the appropriate third factor for cryptocurrencies?. Finance Research Letters. 45, 102139.

Jin, J., Yu, J., Hu, Y., Shang, Y. 2019. Which one is more informative in determining price movements of hedging assets? Evidence from Bitcoin, gold and crude oil markets. Physica A: Statistical Mechanics and its Applications. 527, 121121.
Kaabia, O., Abid, I., Guesmi, K., Sahut, J. M. 2020. How do bitcoin price fluctuations affect crude oil markets?. Gestion 2000. 37(1), 47-60.
Karaa, R., Slim, S., Goodell, J. W., Goyal, A., Kallinterakis, V. 2021. Do investors feedback trade in the Bitcoin—and why?. The European Journal of Finance. 1-21.
Kaminski, J., Gloor, P. A. 2014. Nowcasting the bitcoin market with twitter signals. CoRR abs/1406.7577.URL http://arxiv.org/abs/1406.7577.
Kapar, B., Olmo, J. 2021. Analysis of Bitcoin prices using market and sentiment variables. The World Economy. 44(1), 45-63.
Karalevicius, V., Degrande, N., De Weerdt, J. 2018. Using sentiment analysis to predict interday Bitcoin price movements. The Journal of Risk Finance. 19(1), 56-75.
Kearney, C., Liu, S. 2014. Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis. 33, 171-185.
Khaldi, R., El Afia, A., Chiheb, R. 2019. Forecasting of BTC volatility: comparative study between parametric and nonparametric models. Progress in Artificial Intelligence. 8(4), 511-523.
Koki, C., Leonardos, S., Piliouras, G. 2022. Exploring the predictability of cryptocurrencies via Bayesian hidden Markov models. Research in International Business and Finance. 59, 101554.
Kristoufek, L. 2013. BitCoin meets Google trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific Reports. 3(1), 1-7.
Lahmiri, S., Bekiros, S. 2019. Cryptocurrency forecasting with deep learning chaotic neural networks. Chaos, Solitons & Fractals. 118, 35-40.
Lahmiri, S., Bekiros, S. 2020. Intelligent forecasting with machine learning trading systems in chaotic intraday Bitcoin market. Chaos, Solitons & Fractals. 133, 109641.
Lahmiri, S., Bekiros, S. 2021. Deep learning forecasting in cryptocurrency high-frequency trading. Cognitive Computation. 13(2), 485-487.
Li, Y., Goodell, J. W., Shen, D. 2021. Comparing search-engine and social-media attentions in finance research: Evidence from cryptocurrencies. International Review of Economics & Finance. 75, 723-746.
Liebi, L. J. 2022. Is there a value premium in cryptoasset markets?. Economic Modelling. 109, 105777.
Liu, M., Li, G., Li, J., Zhu, X., Yao, Y. 2021. Forecasting the price of Bitcoin using deep learning. Finance Research Letters. 40, 101755.
Livieris, I. E., Stavroyiannis, S., Pintelas, E., Pintelas, P. 2020. A novel validation framework to enhance deep learning models in time-series forecasting. Neural Computing and Applications. 32(23), 17149-17167.
Livieris, I. E., Stavroyiannis, S., Pintelas, E., Kotsilieris, T., Pintelas, P. 2021a. A dropout weight-constrained recurrent neural network model for forecasting the price of major cryptocurrencies and CCi30 index. Evolving Systems. 1-16. https://doi.org/10.1007/s12530-020-09361-2.
Livieris, I. E., Kiriakidou, N., Stavroyiannis, S., Pintelas, P. 2021b). An advanced CNN-LSTM model for cryptocurrency forecasting. Electronics. 10(3), 287.
Loginova, E., Tsang, W. K., van Heijningen, G., Kerkhove, L. P., Benoit, D. F. 2021. Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data. Machine Learning. 1-24. https://doi.org/10.1007/s10994-021-06095-3.

López-Cabarcos, M. Á., Pérez-Pico, A. M., Piñeiro-Chousa, J., Šević, A. 2021. Bitcoin volatility, stock market and investor sentiment. Are they connected?. Finance Research Letters. 38, 101399.
Mallqui, D. C., Fernandes, R. A. 2019. Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques. Applied Soft Computing. 75, 596-606.
Mallqui, D. C., Fernandes, R. A. 2021. Analysis of technical, economic and social information features to predict the bitcoin price direction for day-trade operations. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
Mudassir, M., Bennbaia, S., Unal, D., Hammoudeh, M. 2020. Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. Neural Computing and Applications. 1-15. https://doi.org/10.1007/s00521-020-05129-6.
Munim, Z. H., Shakil, M. H., Alon, I. 2019. Next-day bitcoin price forecast. Journal of Risk and Financial Management. 12(2), 103.
Niu, T., Wang, J., Lu, H., Yang, W., Du, P. 2020. Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting. Expert Systems with Applications. 148, 113237.
Oad Rajput, S. K., Soomro, I. A., Soomro, N. A. 2020. Bitcoin sentiment index, bitcoin performance and US dollar exchange rate. Journal of Behavioral Finance, 1-16. https://doi.org/10.1080/15427560.2020.1864735.
Peng, Y., Albuquerque, P. H. M., de Sá, J. M. C., Padula, A. J. A., Montenegro, M. R. 2018. The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with Support Vector Regression. Expert Systems with Applications. 97, 177-192.
Poongodi, M., Sharma, A., Vijayakumar, V., Bhardwaj, V., Sharma, A. P., Iqbal, R., Kumar, R. 2020. Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system. Computers & Electrical Engineering. 81, 106527.
Rodrigues, D., Yang, X. S., De Souza, A. N., Papa, J. P. 2015. Binary flower pollination algorithm and its application to feature selection, in: Recent advances in swarm intelligence and evolutionary computation, pp. 85-100. Springer, Cham.
Rognone, L., Hyde, S., Zhang, S. S. 2020. News sentiment in the cryptocurrency market: An empirical comparison with Forex. International Review of Financial Analysis. 69, 101462.
Shen, D., Urquhart, A., Wang, P. 2020. Forecasting the volatility of Bitcoin: The importance of jumps and structural breaks. European Financial Management. 26(5), 1294-1323.
Sahut, J.-M., Pasquini-Descomps, H. 2015. ESG impact on market performance of firms: International evidence. Management International / International Management / Gestiòn Internacional, 19(2), 40-63.
Sugunnasil, P., Somhom, S. 2010. Feature selection for neural network based stock prediction, in: International Conference on Advances in Information Technology, pp. 137-146. Springer, Berlin, Heidelberg.
Sun, X., Liu, M., Sima, Z. 2020. A novel cryptocurrency price trend forecasting model based on LightGBM. Finance Research Letters. 32, 101084.
Xiong, N. 2002. A hybrid approach to input selection for complex processes. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 32(4), 532-536.
Yang, X. S. 2012. Flower pollination algorithm for global optimization, in: International conference on unconventional computing and natural computation, pp. 240-249. Springer, Berlin, Heidelberg.
Yenidoğan, I., Çayir, A., Kozan, O., Dağ, T., Arslan, Ç. 2018. Bitcoin forecasting using ARIMA and prophet, in: 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 621-624. IEEE.

White, R., Marinakis, Y., Islam, N., Walsh, S.T. 2020. Is Bitcoin a currency, a technology-based product, or something else? Technological Forecasting and Social Change. 151(C), 119877.
Wu, C. H., Lu, C. C., Ma, Y. F., Lu, R. S. 2018. A new forecasting framework for bitcoin price with LSTM, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 168-175. IEEE.
Zargar, F. N., Kumar, D. 2019. Informational inefficiency of Bitcoin: A study based on high-frequency data. Research in International Business and Finance. 47, 344-353.

Table 1: Summary of previous studies of cryptocurrency price prediction

Study	Approach	Features	Feature selection method	Forecasting method	Performance for BTC
Livieris et al. (2021a)	univariate	XRP, BTC, LTC, ETH, and CCi30 index	no	WCRNN	MAE=108.95, RMSE=151.18
Altan et al. (2019)	univariate	XRP, LTC, Dash, and BTC	no	EWT+LSTM+ CS	MAE=500.16, RMSE=623.41, MAPE=3.55
Livieris et al. (2021b)	univariate	BTC, XRP, and ETH	no	CNN+LSTM	MAE=169.60, RMSE=256.32
Lahmiri and Bekiros (2019)	univariate	BTC, Digital Cash, and XRP	no	LSTM	RMSE=275.00
Guo et al. (2021)	multivariate	8 social interest and inter-exchange transaction features	no	WT+CATCN	MAE=1044.38, RMSE=1204.09
Liu et al. (2021)	multivariate	40 indicators of cryptocurrency market, public attention, and macroeconomic environment	no	SDAE	RMSE=160.63, MAPE=0.10
Livieris et al. (2020)	univariate	BTC	no	CNN+LSTM	MAE=256.53, RMSE=399.66
This study	multivariate	BTC, investor sentiment, technical indicators, and Bitcoin metrics	CFS+FPA	BSVR	

Legend: BSVR – bagged support vector regression, BTC – Bitcoin, CATCN – casual attention temporal convolutional network, CFS+FPA – correlation-based feature selection with Flower Pollination Algorithm, CNN – convolutional neural network, CS – cuckoo search, ETH – Ethereum, EWT – empirical wavelet transform, LSTM – long short-term memory, LTC – Litecoin, MAE – mean absolute error, MAPE – mean absolute percentage error, RMSE – root mean squared error, SDAE – stacked denoising autoencoder, SVR – support vector regression, XRP – Ripple, WCRNN – weight-constrained recurrent neural network, WT – wavelet transform.

Table 2: Features selected using a correlation-based filter with the FPA

Feature	Description	Category
BMI	Bitcoin Misery Index	Investor sentiment
A/D	Accumulate/distribution	Technical indicator
XRP price	Ripple price	Cryptocurrency market / market data
XRP volume	Ripple volume	Cryptocurrency market / market data
LTC price	Litecoin price	Cryptocurrency market / market data
LTC volume	Ripple volume	Cryptocurrency market / market data
BTC price	Bitcoin price	Cryptocurrency market / market data
BTC volume	Bitcoin volume	Cryptocurrency market / market data
ROI30d	Return on investment 30 days prior	Cryptocurrency market / market data
VtyDayRet180d	180-day volatility of daily returns	Cryptocurrency market / market data
CapMVRVCurUSD	Current supply in USD	Cryptocurrency market / market supply
IssContUSD	New native units issued by a protocol-mandated continuous emission schedule	Cryptocurrency market / market supply
IssTotUSD	All new native units issued	Cryptocurrency market / market supply
NVTAdj	Network value (current supply) divided by the adjusted transfer value	Cryptocurrency market / market supply
FeeTotUSD	Fees paid to miners, etc.	Cryptocurrency market / transactions
TxTfrCnt	Count of transfers	Cryptocurrency market / transactions
TxTfrValAdjNtv	Native units transferred between distinct addresses	Cryptocurrency market / transactions
TxTfrValMeanUSD	Mean size of a transfer	Cryptocurrency market / transactions
TxTfrValMedUSD	Median size of a transfer	Cryptocurrency market / transactions

Table 3: Results of Bitcoin price prediction for the testing period from April 2020 to November 2021

		Forecasting horizon (days ahead)								
		1	2	3	4	5	10	15	20	30
RWM	MAE	929.0	1340.1	1664.4	1896.2	2127.7	3106.6	3861.8	4394.3	5792.0
	MAPE	2.74	3.90	4.85	5.57	6.24	9.03	11.39	13.34	17.81
	RMSE	1473.1	2019.3	2456.1	2834.6	3202.4	4675.4	5857.8	6703.5	8226.3
Bi-LSTM	MAE	1401.8	1287.3	1267.2	1453.0	1219.4	1285.3	1211.7	1395.7	1511.3
	MAPE	3.74	3.58	3.54	3.85	3.51	3.64	3.58	3.73	3.93
	RMSE	2173.6	1947.5	1918.0	2260.3	1831.5	1939.1	1855.7	2154.3	2354.8
ARIMA	MAE	1607.0	2782.3	4059.8	5479.7	7212.3	19455	40545	77847	256748
	MAPE	4.54	7.73	11.10	14.82	19.28	50.10	101.32	189.19	595.96
	RMSE	2435.1	4202.8	6144.1	8245.9	10829	29248	61604	119820	408989
ANN	MAE	2295.3	2477.3	2649.2	2808.9	2946.3	3535.4	4014.1	4413.6	4880.1
	MAPE	7.28	7.84	8.36	8.85	9.27	11.05	12.48	13.62	14.83
	RMSE	3096.3	3336.9	3562.8	3772.9	3951.2	4683.3	5251.1	5705.8	6229.3
SANN	MAE	2510.7	2700.7	2875.5	3036.5	3189.9	3853.5	4435.8	4948.5	5580.7
	MAPE	7.91	8.49	9.02	9.50	9.97	11.96	13.66	15.11	16.75
	RMSE	3390.8	3644.3	3875.5	4086.5	4286.4	5116.8	5804.4	6391.1	7117.7
RBFNN	MAE	22642	22675	22709	22743	22777	22953	23129	23309	23688
	MAPE	54.59	54.63	54.67	54.71	54.75	55.01	55.23	55.47	56.11
	RMSE	29624	29648	29673	29697	29721	29845	29969	30095	30353
RF	MAE	20851	20917	20974	21038	21088	21275	21503	21695	22065
	MAPE	47.79	48.12	48.41	48.70	48.93	49.63	50.40	50.95	51.78
	RMSE	27836	27865	27888	27920	27944	28026	28166	28279	28521
SVR	MAE	167.6	164.2	163.7	164.6	167.6	166.1	165.9	167.9	170.4
	MAPE	0.44	0.43	0.43	0.43	0.44	0.44	0.43	0.44	0.44
	RMSE	249.4	244.8	244.1	245.2	249.3	246.5	245.7	247.8	249.7
BSVR	MAE	107.9	107.2	106.8	106.9	108.0	108.4	108.6	109.2	111.0
	MAPE	0.30	0.29	0.29	0.29	0.30	0.30	0.29	0.29	0.30
	RMSE	173.7	172.9	172.5	172.5	173.8	174.0	174.1	175.0	176.6

Legend: ANN – artificial neural network, ARIMA – autoregressive integrated moving average, Bi-LSTM – bidirectional long short-term memory, BSVR – bagged support vector regression, RBFNN – radial basis function neural network, RF – random forest, RWM – random walk martingale, SANN – stacked artificial neural network, SVR – support vector regression.

Table 4: Results of Bitcoin price prediction for the testing period from December 2021 to May 2022

		Forecasting horizon (days ahead)								
		1	2	3	4	5	10	15	20	30
RWM	MAE	1033.3	1532.2	1908.8	2276.2	2550.9	3780.1	4566.8	5152.7	6356.8
	MAPE	2.52	3.71	4.66	5.56	6.22	9.38	11.27	12.62	15.48
	RMSE	1428.7	2047.4	2468.3	2879.1	3233.3	4484.8	5340.6	6168.5	7761.9
Bi-LSTM	MAE	4240.6	4243.5	4154.7	4111.2	4222.6	3905.2	4145.7	4097.5	4070.4
	MAPE	9.86	9.87	9.66	9.55	9.83	9.06	9.63	9.52	9.45
	RMSE	4624.5	4614.4	4539.7	4499.3	4599.4	4302.1	4534.5	4485.6	4466.8
ARIMA	MAE	2401.3	3979.2	5662.7	7424.2	9542.8	22258.7	38931.94	62313.39	136273.7
	MAPE	5.56	9.42	13.57	17.96	23.19	54.97	96.72	155.66	345.94
	RMSE	4720.8	5912.6	7310.5	8860.1	10883.9	23292.7	40203.9	63939.5	140132.2
ANN	MAE	5058.0	5348.0	5607.3	5856.5	6088.7	6793.0	7207.4	7545.7	7638.6
	MAPE	11.47	12.23	12.91	13.54	14.11	15.88	16.98	17.89	18.45
	RMSE	6304.0	6396.0	6499.5	6655.7	6862.8	7472.6	7823.6	8131.5	8154.4
SANN	MAE	5295.9	5601.7	5872.4	6116.4	6358.2	7172.5	7687.5	8107.5	8307.4
	MAPE	11.99	12.79	13.50	14.13	14.71	16.76	18.11	19.23	20.06
	RMSE	6625.9	6732.0	6842.5	6985.3	7193.4	7874.0	8310.9	8698.2	8826.1
RBFNN	MAE	32994	32902	32814	32743	32697	32442	32217	32013	31421
	MAPE	78.38	78.34	78.30	78.27	78.25	78.13	78.02	77.92	77.64
	RMSE	33408	33299	33195	33113	33065	32792	32554	32343	31695
RF	MAE	30929	30828	30741	30648	30588	30137	29940	29668	29372
	MAPE	73.37	73.31	73.27	73.18	73.13	72.92	72.81	72.67	72.51
	RMSE	31372	31248	31140	31035	30972	30483	30279	29982	29655
SVR	MAE	301.2	291.9	288.0	285.3	287.7	278.5	269.3	264.5	255.1
	MAPE	0.70	0.68	0.68	0.67	0.68	0.66	0.64	0.64	0.63
	RMSE	346.5	329.2	322.8	319.2	321.5	308.4	293.4	283.9	273.3
BSVR	MAE	149.9	150.1	149.6	148.2	147.4	144.8	139.3	133.9	126.4
	MAPE	0.36	0.36	0.36	0.36	0.36	0.36	0.35	0.34	0.33
	RMSE	204.4	204.2	204.1	202.6	201.7	200.4	193.4	187.6	180.7

Legend: ANN – artificial neural network, ARIMA – autoregressive integrated moving average, Bi-LSTM – bidirectional long short-term memory, BSVR – bagged support vector regression, RBFNN – radial basis function neural network, RF – random forest, RWM – random walk martingale, SANN – stacked artificial neural network, SVR – support vector regression. For a fair comparison, the same set of selected features were used in all multivariate forecasting models.

Appendix A: Descriptive statistics of features

Feature	Description	Mean	Minimum	Maximum	St. Dev.
BMI	Bitcoin Misery Index	44.73	5.00	95.00	22.64
VltyChk	Chaikin's volatility	10.26	-63.40	374.13	55.96
Psych	Psychological line	50.67	16.67	91.67	13.96
AccDisV	Accumulation / distribution	1409509	1409509	2893198	522783
AdrActCnt	Addresses, active, count	824454.8	414992.0	1366494.0	187584.8
BlkCnt	Block, count	146.49	58.00	197.00	16.04
BlkSizeByte	Block size [bytes]	1.61E+08	71186540	2.62E+08	3.04E+08
BlkSizeMeanByte	Mean block size [bytes]	1103883	442152.4	1527074	205839
CapMVRVCur	Capitalization, current supply	1.773	0.690	3.958	0.631
CapMrktCurUSD	Capitalization, market to current supply, [USD]	3.83E+11	5.55E+10	1.27E+12	3.5E+11
CapRealUSD	Realized capitalization [USD]	1.93E+11	7.62E+10	4.69E+11	1.44E+11
DiffMean	Mean difficulty	1.45E+13	3.01E+12	3.13E+13	7.29E+12
FeeMeanNtv	Mean transaction fees [native units]	0.000172	2.59E-05	0.001125	0.000154
FeeMeanUSD	Mean transaction fees [USD]	3.681	0.161	60.950	6.247
FeeMedNtv	Median transaction fees [native units]	7.8E-05	5.63E-06	0.000583	8.74E-05
FeeMedUSD	Median transaction fees [USD]	1.648	0.026	27.594	3.011
FeeTotNtv	Total fees [native units]	50.29	5.46	302.58	48.02
FeeTotUSD	Total fees [USD]	1055972	42893	17088265	1782188
HashRate	Mean hash rate	1.05E+08	21859737	2.55E+08	53409872
IssContNtv	Continuous issuance [native units]	1396.1	362.5	2362.5	493.9
IssContPctAnn	Continuous issuance, annualized [%]	2.839	0.706	4.897	1.076
IssContUSD	Continuous issuance [USD]	22379403	4593408	71698528	15064715
IssTotNtv	Total issuance [native units]	1396.1	362.5	2362.5	493.9
IssTotUSD	Total issuance [USD]	22379403	4593408	71698528	15064715
NVTAdj	Adjusted NVT, free float	78.87	18.86	252.90	31.47
NVTAdj90	Adjusted NVT, free float, 90-day MA	74.74	29.44	141.51	20.55
ROI1yr	ROI 1 year prior [%]	20615.40	3185.07	67541.76	18361.24
ROI30d	ROI 30 days prior [%]	145.46	-83.73	1064.31	207.54
SplyCur	Current supply	5.23	-51.72	120.67	23.67
SplyExpFut10yrCMBI	Future expected supply, next 10 years	18151837	16894261	19043461	630313
SplyFF	Supply, free float	20500308	20303609	20673897	102179
TxCnt	Count of transactions	14283822	13759056	14647592	235876
TxTfrCnt	Count of transactions	280990	123967	453346	54034
TxTfrValAdjNtv	Adjusted value of transactions [native units]	719002	365031	1146432	138778
TxTfrValAdjUSD	Adjusted value of transactions [USD]	268154	74316	1009539	117155
TxTfrValMeanNtv	Mean value of transactions [native units]	5.96E+09	5.53E+08	3.65E+10	6.81E+09
TxTfrValMeanUSD	Mean value of transactions [USD]	1.023	0.307	10.764	0.709
TxTfrValMedNtv	Median value of transactions [native units]	21365.0	2716.2	506174.4	32822.7
TxTfrValMedUSD	Median value of transactions [USD]	0.0081	0.0018	0.0197	0.0039
TxTfrValNtv	Adjusted value of transactions [native units]	103.25	33.68	285.56	40.47
TxTfrValUSD	Adjusted value of transactions [USD]	728024.2	214079.7	8331638.0	561447.8
VtyDayRet180d	180-day volatility of daily returns	1.65E+10	1.43E+09	3.92E+11	2.63E+10
VtyDayRet30d	30-day volatility of daily returns	0.040	0.025	0.061	0.008
VtyDayRet60d	60-day volatility of daily returns	0.036	0.010	0.106	0.014
WTI price	West Texas Intermediate (WTI) crude oil price	73.27	18.46	118.23	23.05

XRP volume	Ripple volume	51824601	1943778	728520004	60201053
XRP price	Ripple price	0.505	0.141	1.834	0.313
ETH volume	Ethereum volume	44017.9	1286.0	397258.0	41102.1
ETH price	Ethereum price	1111.70	82.91	4811.59	1299.52
BCH volume	Bitcoin Cash volume	9173.7	120.0	115292.0	9957.6
BCH price	Bitcoin Cash price	436.49	76.17	1755.00	271.82
LTC volume	Litecoin volume	50618.2	1894.0	375187.0	49799.1
LTC price	Litecoin price	101.24	22.82	388.32	61.23
BTC volume	Bitcoin volume	14007.4	1176.0	130323.0	10705.0
BTC price	Bitcoin price	20610.6	3183.0	67554.8	18361.7
Direction	Direction of investor sentiment	Extreme Fear (25.24%), Fear (33.77%), Neutral (9.63%), Greed (19.32%), and Extreme Greed (12.04%)			

Appendix B: Correlations between features