Comparing Different Machine Learning Models at Predicting Bitcoin Price Moving Direction at Different Time Intervals from 2017 to 2024*

RNN Can Predict Best at One-Hour Intervals and XGB Trains the Fastest

Justin (Jiazhou) Bi

December 7, 2024

This project aims to build a predictive model for the price moving direction of Bitcoin (BTC) at 1-minute, 1-hour, and 1-day intervals. We explored four popular machine learning algorithms for time-series data prediction: Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Random Forest (RF), and Extreme Gradient Boosting (xGB). We have compared these four models on two different dimensions: model performance and time for training. The data used for this project is gathered from Binance which contains BTC data from late 2017 to present (November 2024).The original variables include high, low, open, close, and volume, as well as the timestamp of BTC, at one-minute, one-hour, and one-day intervals. Ultimately, RNN has the overall best performance in terms of its predictive accuracy, while XGB has the fastest training duration. While most models across different time intervals achieved a better performance than simply guessing at a 50-50 percent chance, their performance varies for each direction of the movemenet. Future steps for improving the model may involve incorporating additional features to enhance model accuracy and reduce errors. Additionally, fine-tuning hyperparameters, experimenting with more sophisticated feature engineering, and handling outliers could further improve performance.

1 Introduction

Cryptocurrencies have drawn much attention from the public recently due to their growing public awareness. Researchers have found that 4,495 pieces of literature related to Bitcoin were published from 2011 to 2020 (Aysan, Demirtaş, and Saraç 2021). More recently, another

*All project related files available at: https://github.com/Jiazhou-Bi/BTC-Price-Prediction

study stated that more than 9,100 research papers were published on cryptocurrencies in the Web of Science (Bâra and Oprea 2024). In 2015, a journal called Ledger was published by the University of Pittsburgh's University Library System, and it is the first journal dedicated to cryptocurrencies (Extance 2015). This reflects the trend of growing interest in digital assets. In December 2019, there were 2,813 different cryptocurrencies existing with 201.47 billion US dollars in their total market capitalization; in particular, bitcoin is the most significant contributor and had a 131.89 billion US dollars market capitalization (Chen 2021). When writing this article on December 4, 2024, the current market capitalization of Bitcoin is 1.95 trillion US dollars, with a 103.22 billion US dollars 24-hour trading volume, according to forbes.com (Forbes, n.d.). As of October 2023, there were a total of 1492 cryptocurrency trading platforms globally, both centralized and decentralized (Oh 2024). Among all the different cryptocurrencies, Bitcoin is the key player in the market and the research field (Aysan, Demirtaş, and Saraç 2021). Therefore, learning more about Bitcoin can benefit both researchers and the public interested in cryptocurrency and its trade.

On the other hand, volatility is a key aspect of Bitcoin due to its non-tangible nature. Much research has focused on the price fluctuations of Bitcoin and compared its volatility to more traditional assets, such as gold. For example, (Kurihara and Fukushima 2018) demonstrated that Bitcoin has higher volatility than gold, indicating its speculative nature. In another study, the authors have concluded that Bitcoin's extreme price volatility may interfere with its use as a stable investment instrument. As a result, speculative investors are often more attracted to Bitcoin (Chen 2021) and are looking for profit by capitalizing on its rapid price changes (Baur, Dimpfl, and Kuck 2018).

Despite its extreme volatility, Bitcoin is still widely accepted as a tradable investment vehicle, as many characteristics are similar to stocks (Chen 2021). In a recently published study, the authors developed a model called TradExpert-NM, which can achieve a 59% accuracy for Bitcoin's daily price movement direction (Ding, Shi, and Liu 2024). In another study, (Mudassir et al. 2020) developed a model that can score up to 65% accuracy for the next-day forecast of Bitcoin's price movement. However, many of these "high-performing" models use various data (social media data in (Ding, Shi, and Liu 2024) and other technical data (Mudassir et al. 2020)) that are usually too complicated for investors with limited knowledge of machine learning and data science. Even for models that used only more accessible data, such as in (Adcock and Gradojevic 2019), they utilized neural networks as their model. However, deep learning models, like deep neural networks, generally require more data than traditional machine learning models and can take longer to train (Janiesch, Zschech, and Heinrich 2021). Therefore, these models pose challenges for individuals who want to start trading Bitcoin with limited knowledge and resources. As a result, in this project, three commonly used models for price prediction are compared on their performance and relative training effort by using only commonly available, easy-to-understand features (i.e., High Price, Low Price, Open Price, Close Price, and Trade Volume). This project aims to provide some knowledge for beginner crypto-traders who want to trade Bitcoin with some help from machine learning algorithms.

This project extracted data from Binance using a free Python (Van Rossum and Drake 2009)

package called ccxt (Kroitor and Kroitor 2024). A total of three datasets were extracted. The three datasets contain the same time horizon, from August 17, 2017, to November 16, 2024. The only difference between them is their time intervals. The datasets were extracted at 1-minute, 1-hour, and 1-day intervals. Binance was chosen because it is the world's largest cryptocurrency exchange (Kowsmann and Ostroff 2021), and it only offers data dating back to August 17, 2017, through its free API. Arguably, data before 2017 may be less important for our models' training because the cryptocurrency market was vastly different before 2017, and its price surged significantly in 2017 (Lim and Gorse 2020). One potential drawback of this approach is that we are relying on the data from Binance as the single point of truth, which may not be the case across the whole time horizon in this case because the quality of the data is not cross-verified with other sources.

The ultimate goal of this project is to compare and contrast various machine learning models in terms of their technical performance and training effort in predicting Bitcoin's price movement direction at three different time intervals. By evaluating models on datasets with 1-minute, 1-hour, and 1-day intervals, this project aims to shed light on the trade-offs between predictive accuracy, computational efficiency, and accessibility for beginner traders. Furthermore, individuals with limited expertise in machine learning can leverage relatively simple and accessible data to make more informed, confident trading decisions by learning more about these models. In short, this project aims to empower crypto-traders by providing practical guidance on selecting and utilizing machine learning models effectively while balancing ease of use and predictive power.

In this analysis, we explored three different models: Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and XGBoost (XGB) to predict Bitcoin's future price movement direction. LSTM and RNN are deep learning models that are excellent at handling sequential data. Thus, they are commonly used in price-prediction models. However, these two models generally need more computational resources, making their training/retraining processes effortful. On the other hand, XGB requires less computational effort thanks to its gradient-boosting framework, making it ideal for individuals with limited resources. By comparing these models on three different time intervals (1 minute, 1 hour, and 1 day), we aim to evaluate their predictive accuracy, training efficiency, and overall practicality for novice traders.

The remainder of this paper is structured as follows: Section 2 introduces the raw dataset, describes the cleaned datasets, and explains our final datasets with additionally engineered features, along with a preliminary analysis through numerical summaries and visualizations. Section 3 describes the three machine learning models being used in our study. Section 4 explores key findings from our analysis. Lastly, Section 5 addresses the limitations of the analysis and offers recommendations for future research projects.

2 Data

As mentioned in Section 1, the datasets were extracted from Binance using a free Python (Van Rossum and Drake 2009) package called ccxt (Kroitor and Kroitor 2024). The extracted raw datasets contain the following variables: Timestamp, High, Low, Open, Close, and Volume. Two new variables were added to the cleaned datasets: direction_t-1 and direction_t+1. Lastly, in our final datasets, we have added four more variables: high_diff, low_diff, open_diff, and close_diff. Furthermore, some missing data were inputed for better data quality. Details of our data cleaning and engineering processes are provided in Section 2.2 and Section 2.3. In the following subsections, we will review all the variables used in this report and provide some basic descriptive statistics.

2.1 Descriptive Data Analysis

In this subsection, we will explore every variable used in our study by examining some of the basic statistics. These preliminary analyses can help us better understand the data we are working with.

2.1.1 Timestamp

The timestamp variable in our datasets is a key component for tracking the temporal aspect of Bitcoin's price movement. All three datasets start on August 17, 2017, and end on November 16, 2024. While ensuring all three datasets cover the same period, each dataset has different granularities. In the 1-minute dataset, the timestamp is recorded for each minute, providing the highest resolution data. In the 1-hour dataset, the timestamp is aggregated at hourly intervals. In the 1-day dataset, the timestamps are recorded at daily intervals. The differences in temporal granularity allow us to examine the trade-offs between datasets containing more information closely but may also increase computational complexity and noise. Table 1 is an example of the timestamp variable from the 1-hour dataset.

Table 1: Example of the timestamp variable from the 1-hour dataset.

timestamp
2017-08-17 04:00:00
2017-08-17 05:00:00
2017-08-17 06:00:00
2017-08-17 07:00:00
2017-08-17 08:00:00

2.1.2 Open

Open refers to the 'open price' of the tradable asset. As defined in (Chen 2021), the open price is the initial price of the cryptocurrency at the beginning of the trading session. For example, in the 1-hour dataset, the open price is the price of Bitcoin at the start of the hour. This is a commonly used indicator in the financial sector, and it shows the perceived value change of the underlying asset over time. Figure 1 reveals the change of the 1-hour open price of Bitcoin over the entire time horizon of the dataset.

Figure 1: Line chart of the open prices from the 1-hour dataset.

2.1.3 Close

Similar to 'open,' 'close' refers to the 'close price' of the tradable asset. As defined in (Chen 2021), the close price is the last recorded price of Bitcoin before the trading session ends. For example, in the 1-minute dataset, the open price is the price of Bitcoin at the end of the minute. This variable is widely used in financial analysis as a key indicator of market sentiment and performance over a given time period. More importantly, close price is often considered the most robust indicator of the assets' perceived price, as suggested in (Hayes, n.d.). Therefore, 'close' is also used as the 'final price' of Bitcoin in this project. Figure 2 shows the change of the 1-minute open price of Bitcoin over the entire time horizon of the dataset.

Figure 2: Line chart of the close prices from the 1-minute dataset.

2.1.4 High and Low

The high and low variables represent the extremes of Bitcoin's price within a given time interval. As suggested by the names, high refers to the highest price of Bitcoin during the time interval, and low refers to the lowest. (Lapitskaya, Eratalay, and Sharma 2024), (Poudel et al. 2023), and many other researchers believe that incorporating basic features such as high and low prices can significantly improve the performance of cryptocurrency price prediction models. In Figure 3, both high and low from the 1-hour dataset are presented in the same graph. It is clear that these two are highly correlated, and this problem will be addressed later in Section 2.3.

Figure 3: Line chart of high and low prices from the 1-hour dataset.

2.1.5 Volume

The last raw variable extracted is volume. Volume represents the total amount of Bitcoin traded during a given time interval and indicates market activity and liquidity. For example, in the 1-hour dataset, the volume indicates the cumulative number of Bitcoins traded within an hour. In (Osina, Adediran, and Babajide 2023), the authors have found that trade volume is an important feature when predicting the price of three different cryptocurrencies. Generally speaking, higher trading volume reflects increased market interest and is often associated with significant price fluctuations. As shown in Figure 4, the trading volume in the 1-day dataset varies significantly over time, illustrating periods of heightened trading activity.

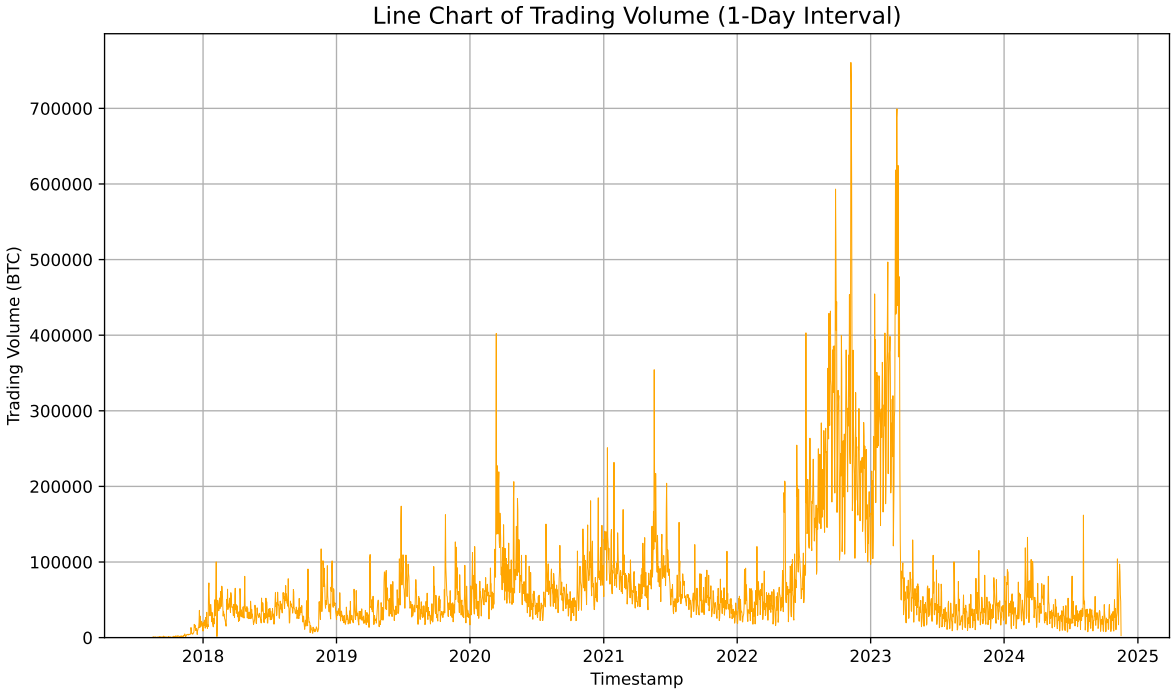

Figure 4: Line chart of trading volume from the 1-day dataset.

2.2 Data Cleaning

There was no "obvious" missing value from the extracted datasets, as all the mentioned variables above are filled with either a timestamp in the "timestamp" column or a numeric value in all the remaining columns. However, upon a closer examination, the timestamps were not exhaustive. For example, in the 1-minute dataset, the timestamp "2017-09-06 16:01:00" does not exist even though it is included within the time horizon of the dataset. 8632 and 128 rows were missing from the 1-minute and 1-hour datasets, respectively. As a result, linear interpolation was used to estimate the missing values because it is a straightforward technique to fill in missing values in time series data (Moritz et al. 2015). Meanwhile, a new column was added to the 1-minute and 1-hour datasets to indicate if the prices of the row were either existing from the raw dataset or input. The new column is named "was_missing." The value of "was_missing" is 0 if the prices were extracted from the original data and 1 if the prices were filled with linear interpolation. Other than interpolating the prices of missing timestamps, no other data were manipulated.

2.3 Feature Engineering

First, two new features were added to indicate the price-moving direction of Bitcoin, namely "direction_t-1" and "direction_t+1". These features provide insights into the direction of Bitcoin's price change relative to the current timestamp. Second, four more new features were engineered to capture Bitcoin's price movements: "open_diff," "high_diff," "low_diff," and "close_diff." These features represent the lagged differences of the respective price variables (open, high, low, and close) from one timestamp to the next.

2.3.1 Direction_t-1

This feature indicates the price movement direction compared to the previous timestamp. It was calculated by calculating the difference between the close price of the current row and the previous row. If the price has decreased from the previous timestamp, "direction_t-1" is marked as -1; if the close price has increased, then "direction_t-1" is marked as 1; if the close price did not change, it is assigned 0—the reason for choosing the close price as the final price indicator was explained in Section 2.1.3.

2.3.2 Direction_t+1

Similar to 'direction_t-1', 'direction_t+1' indicates the price movement direction at the next timestamp relative to the current close price, and it was calculated by calculating the close difference from the next timestamp to the current close price. "direction_t+1" is marked as -1 if the close price is lower in the next timestamp, and is marked as 1 if the close price is going up; if the close price does not change, it is assigned 0—the reason for choosing the close price as the final price indicator was explained in Section 2.1.3. Because 'direction_t+1' indicates the price movement direction of the current timestamp, this variable is used as our models' outcome (i.e., the result the models try to predict).

2.3.3 Open_Diff, High_Diff, Low_Diff, and Close_Diff

An issue was mentioned in Section 2.1.4. That is, the correlation between the price variables (open, high, low, and close) is too high (almost or equal to 1), as illustrated below in Figure 5 and Figure 6.

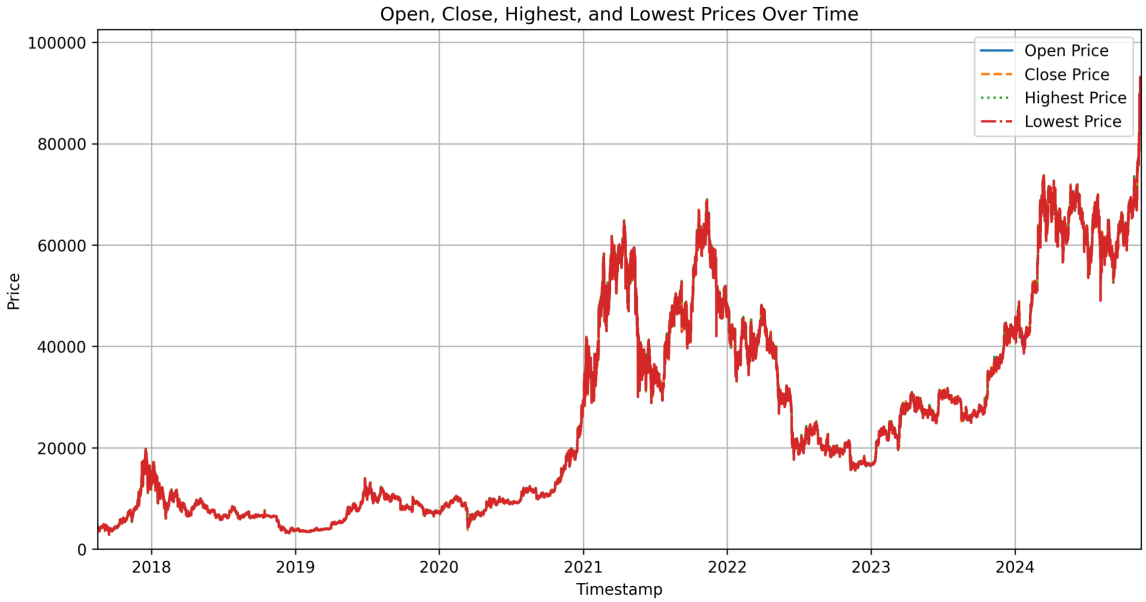

Figure 5: All price variables (open, high, low, close) plotted for the 1-minute dataset.

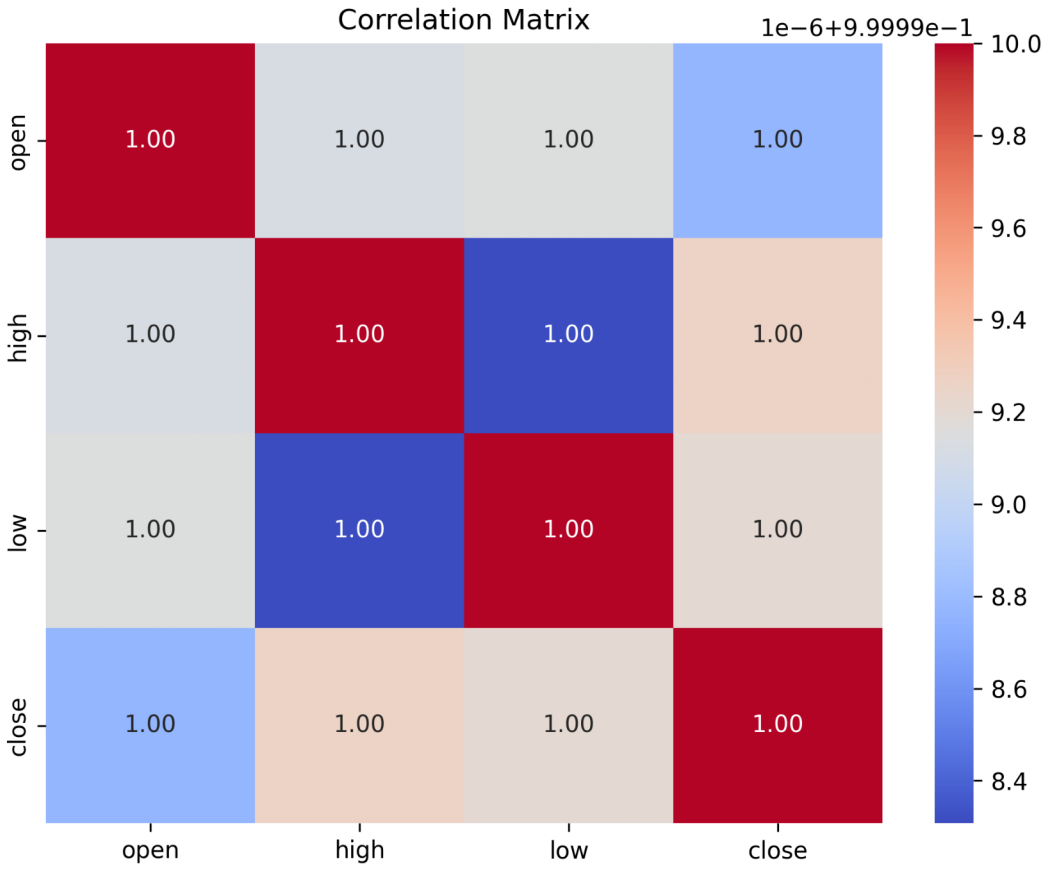

Figure 6: Correlation matrix (open, high, low, close) plotted for the 1-minute dataset.

To solve this issue, "open_diff," "high_diff," "low_diff," and "close_diff" were added to include the lagged differences in the prices. Lagged difference refers to the difference in prices from the current price to the price from the previous timestamp. They are often used in predictive models in financial time series to solve the multicollinearity issue with the original prices, as suggested in (Moews, Herrmann, and Ibikunle 2019). In Figure 7 and Figure 8 below, the correlation between the lagged difference price variables (open_diff, high_diff, low_diff, and close_diff) were displayed.

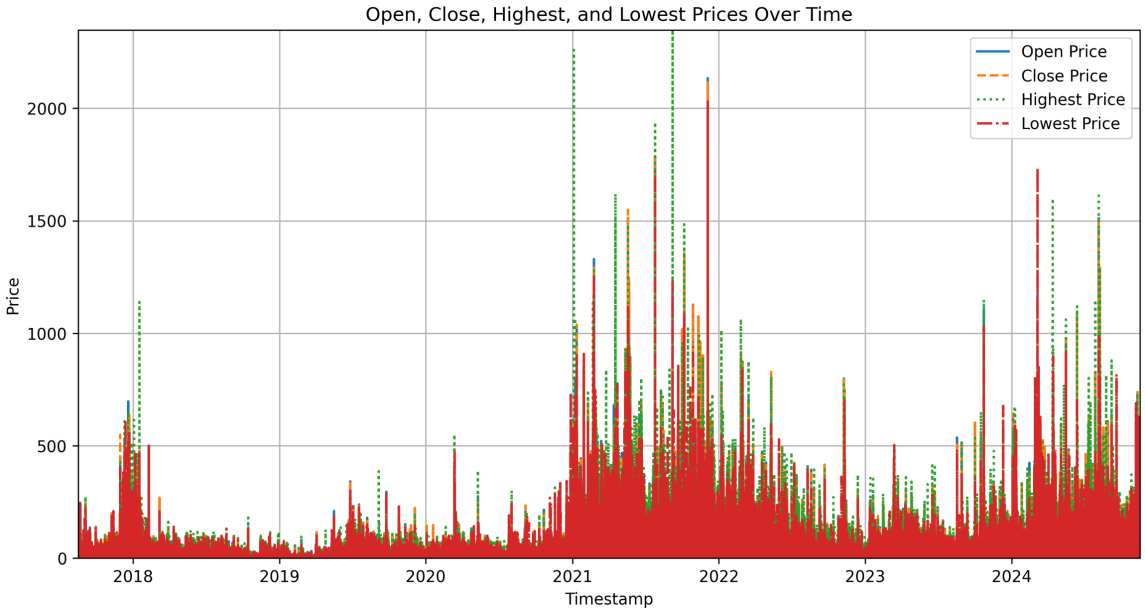

Figure 7: All lagged price differences (open_diff, high_diff, low_diff, and close_diff) plotted for the 1-minute dataset.

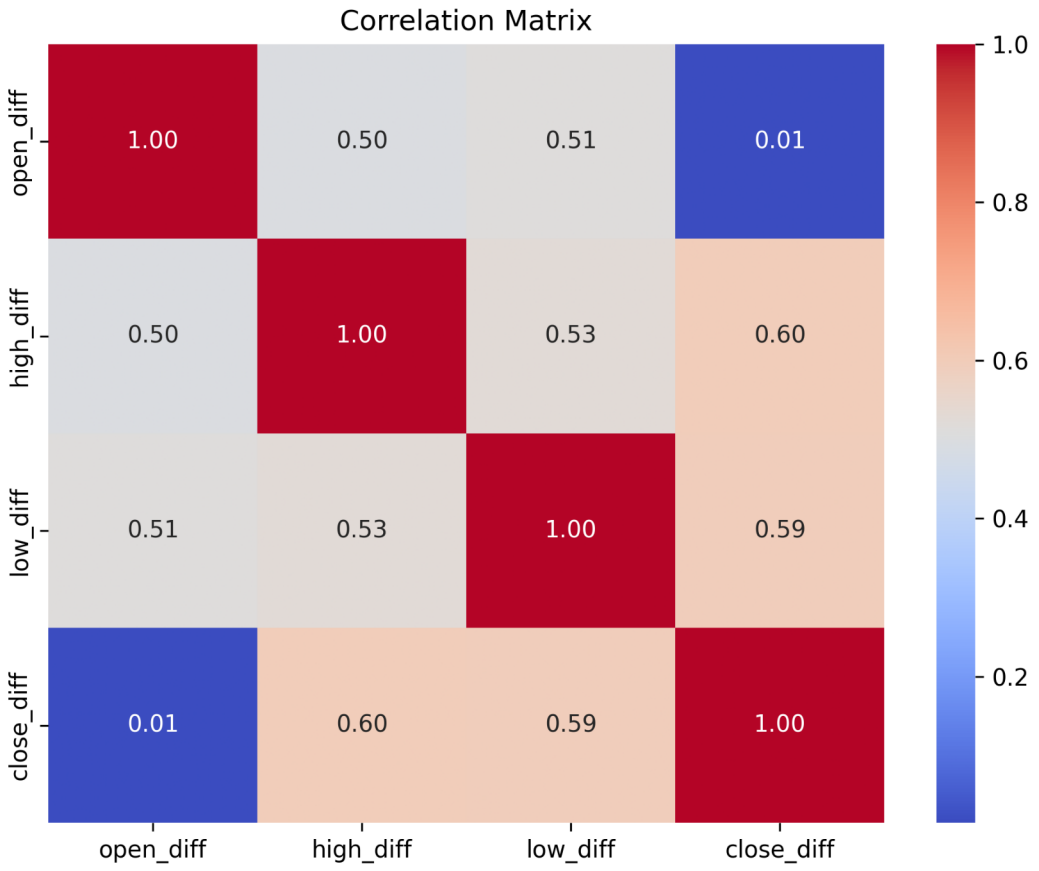

Figure 8: Correlation matrix (open_diff, high_diff, low_diff, and close_diff) plotted for the 1-minute dataset.

3 Models

4 Results

5 Discussion

References

Adcock, Robert, and Nikola Gradojevic. 2019. "Non-Fundamental, Non-Parametric Bitcoin Forecasting." *Physica A: Statistical Mechanics and Its Applications* 531: 121727.

Aysan, Ahmet Faruk, Hüseyin Bedir Demirtaş, and Mustafa Saraç. 2021. "The Ascent of Bitcoin: Bibliometric Analysis of Bitcoin Research." *Journal of Risk and Financial Management* 14 (9): 427.

Bâra, Adela, and Simona-Vasilica Oprea. 2024. "The Impact of Academic Publications over the Last Decade on Historical Bitcoin Prices Using Generative Models." *Journal of Theoretical and Applied Electronic Commerce Research* 19 (1): 538–60.

Baur, Dirk G, Thomas Dimpfl, and Konstantin Kuck. 2018. "Bitcoin, Gold and the US Dollar–a Replication and Extension." *Finance Research Letters* 25: 103–10.

Chen, Yuanyuan. 2021. "Empirical Analysis of Bitcoin Price." *Journal of Economics and Finance* 45 (4): 692–715.

Ding, Qianggang, Haochen Shi, and Bang Liu. 2024. "TradExpert: Revolutionizing Trading with Mixture of Expert LLMs." *arXiv Preprint arXiv:2411.00782*.

Extance, Andy. 2015. "Bitcoin and Beyond." *Nature* 526 (7571): 21.

Forbes. n.d. Forbes Magazine. https://www.forbes.com/digital-assets/crypto-prices/?sh=4d7db3d52478.

Hayes, Adam. n.d. "What Is Closing Price? Definition, How It's Used, and Example." *Investopedia*. Investopedia. https://www.investopedia.com/terms/c/closingprice.asp.

Janiesch, Christian, Patrick Zschech, and Kai Heinrich. 2021. "Machine Learning and Deep Learning." *Electronic Markets* 31 (3): 685–95.

Kowsmann, Patricia, and Caitlin Ostroff. 2021. *Wall Street Journal*.

Kroitor, Igor, and Vitaly Kroitor. 2024. "CCXT: Cryptocurrency Exchange Trading Library." https://github.com/ccxt/ccxt.

Kurihara, Yutaka, and Akio Fukushima. 2018. "How Does Price of Bitcoin Volatility Change?" *International Research in Economics and Finance* 2 (1): 8.

Lapitskaya, Darya, M Hakan Eratalay, and Rajesh Sharma. 2024. "Prediction of Cryptocurrency Prices with the Momentum Indicators and Machine Learning." *Computational Economics*, 1–19.

Lim, Ye-Sheen, and Denise Gorse. 2020. "Deep Recurrent Modelling of Stationary Bitcoin Price Formation Using the Order Flow." In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part i 19*, 170–79. Springer.

Moews, Ben, J Michael Herrmann, and Gbenga Ibikunle. 2019. "Lagged Correlation-Based Deep Learning for Directional Trend Change Prediction in Financial Time Series." *Expert Systems with Applications* 120: 197–206.

Moritz, Steffen, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, and Jörg Stork. 2015. "Comparison of Different Methods for Univariate Time Series Imputation in r." *arXiv Preprint arXiv:1510.03924*.

Mudassir, Mohammed, Shada Bennbaia, Devrim Unal, and Mohammad Hammoudeh. 2020.

"Time-Series Forecasting of Bitcoin Prices Using High-Dimensional Features: A Machine Learning Approach." *Neural Computing and Applications*, 1–15.

Oh, Susan. 2024. "How Many Crypto Exchanges Are There in [Currentyear]? (Updated Data)." *Coinweb*. https://coinweb.com/trends/how-many-crypto-exchanges-are-there/.

Osina, Nataliia, Adeyinka Adediran, and Bola Babajide. 2023. "Exploring the Nexus Between Price and Volume Changes in the Cryptocurrency Market."

Poudel, Samir, Rajendra Paudyal, Burak Cankaya, Naomi Sterlingsdottir, Marissa Murphy, Shital Pandey, Jorge Vargas, and Khem Poudel. 2023. "Cryptocurrency Price and Volatility Predictions with Machine Learning." *Journal of Marketing Analytics* 11 (4): 642–60.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.