Title: Bitcoin Price Prediction and Analysis

Module Leader: Hisham Ihshaish

Module Title: Machine Learning and Predictive Analytics

Module Code: UFCFMJ-15-M

Student ID: 13040289

Wordcount (1954)

Table of Contents

1. Introduction
2. Data Exploration
3. Analysis Type
4. Learning Algorithm Selection
5. Model Performance
6. Conclusion
7. Reference

1. Introduction(445)

Bitcoin is a decentralized digital cryptocurrency that operates on a global online network, enabling peer-to-peer transactions without the need for a central bank or administrator. It is accepted in more than 40 countries worldwide, including Vietnam, Argentina, and India, and its popularity has led to the emergence of new alternative coins. Bitcoin facilitates the exchange of other cryptocurrencies, products, and services. Since its introduction in 2009, Bitcoin has remained secure against hacking attempts due to its blockchain technology. Each electronic coin is encrypted with a unique digital signature, enhancing traceability and establishing trust in the system.

Despite the multitude of analysts and investors seeking to profit, predicting the price of Bitcoin (BTC) remains a challenging endeavor. The value of Bitcoin is determined continuously, with the market operating around the clock. Being subject to an open market, cryptocurrencies face distinctive challenges associated with volatility, setting them apart from most traditional currencies. Predicting cryptocurrency prices is a constantly shifting task, making market literacy crucial for individuals to optimize their engagement in the crypto economy. Scholars have presented mixed findings on the efficiency of the Bitcoin market. Nonetheless, a majority of researchers observe that the Bitcoin market has exhibited improved efficiency over time.

The cryptocurrency market is known for its high volatility, and among the various cryptocurrencies available, Bitcoin is the one that has garnered the most attention from investors. This can be attributed to its unique combination of anonymity and transparency within the system.(Suneetha, 2022)

It is important to note that the crypto market is highly volatile and cannot be predicted with absolute accuracy. Additionally, the market's behavior is influenced by human sentiment, exemplified by the potential impact of a single individual selling a substantial amount of Bitcoin, causing a significant downturn in the crypto market.

The growing efficiency of the Bitcoin market can be attributed to its continuous expansion since its inception. This widespread adoption has resulted in increased competition among market participants, which in turn has contributed to the improved efficiency of the market.(Jacquard, 2021)

It is important to note that the crypto market is highly volatile and cannot be predicted with absolute accuracy. Additionally, the market's behavior is influenced by human sentiment, exemplified by the potential impact of a single individual selling a substantial amount of Bitcoin, causing a significant downturn in the crypto market.

The primary objective of this research is to explore and examine the effectiveness of a proposed model in predicting the price of Bitcoin. The model incorporates machine learning techniques, including Linear Regression, as well as neural network methods utilising LSTM (Long Short-Term Memory). By employing these techniques, the model aims to accurately forecast and analyse Bitcoin prices.

2. Data Exploration(628)

The Bitcoin dataset was obtained by accessing the Coinbase API and consists of 215,665 rows of data. The dataset covers the period from January 1, 2023, to May 30, 2023. It includes various attributes such as the Bitcoin price in USD, daily opening and closing prices, daily low and high prices, as well as trading volume. The primary aim of data collection is to ensure that the gathered information is both informative and trustworthy for conducting statistical analysis. This enables data-driven decisions to be made with precision and effectiveness, facilitating efficient decision-making processes.

Cryptocurrency prices exhibit significantly greater variability across exchanges compared to traditional financial assets. As a result, this thesis opts to utilise average prices as the chosen metric. Prior to conducting any analysis, it was imperative to ensure the cleanliness of the data by eliminating any null values. This data preparation step was crucial in order to proceed with time series analysis and the application of LSTM (Long Short-Term Memory) techniques.

In the LSTM model, the prediction of the current closing value of Bitcoin is based on utilising the five preceding data points of the closing values. This approach ensures that the model considers recent trends and patterns in the data. Similarly, to eliminate any bias, the Linear Regression model follows the same process of utilising the five previous closing values to predict the current close value. This consistency in the approach helps maintain fairness and comparability between the models.

Before utilizing a supervised learning model like LSTM, it is necessary to transform time series data into a specific format. LSTM, being a supervised learning algorithm, has specific requirements for input data. It needs data to be organized into two components: an input component referred to as X, and an output component known as Y. Therefore, for the historical Bitcoin price data from the beginning of the year, it is essential to convert it into this particular format with X and Y components to effectively feed it into the LSTM model.



Figure 1: Bitcoin price overview from January to May

The figure illustrates the Bitcoin price movement in 2023. Over the course of five months, Bitcoin experienced several notable periods of significant price increase. At the beginning of January, the price started at $16,531 and continued to rise, surpassing $20,000 by mid-January. In March, there was a sharp decrease in price followed by a swift recovery, reaching $30,000 within a short period. Since then, Bitcoin has maintained a price level above $20,000.

Undoubtedly, Bitcoin has brought immense wealth to some individuals while causing significant losses for others. The question that lingers is whether such fluctuations will occur again in the future. To shed light on this, let's delve into a potential model's perspective on the matter.

The data is organized chronologically and captured at regular intervals of one day. This type of sequential data is known as a Time Series, which poses unique challenges and considerations for analysis and prediction.

The dataset was narrowed down to include only the selected attributes.

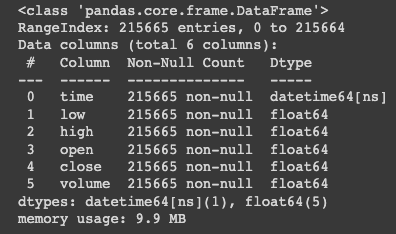
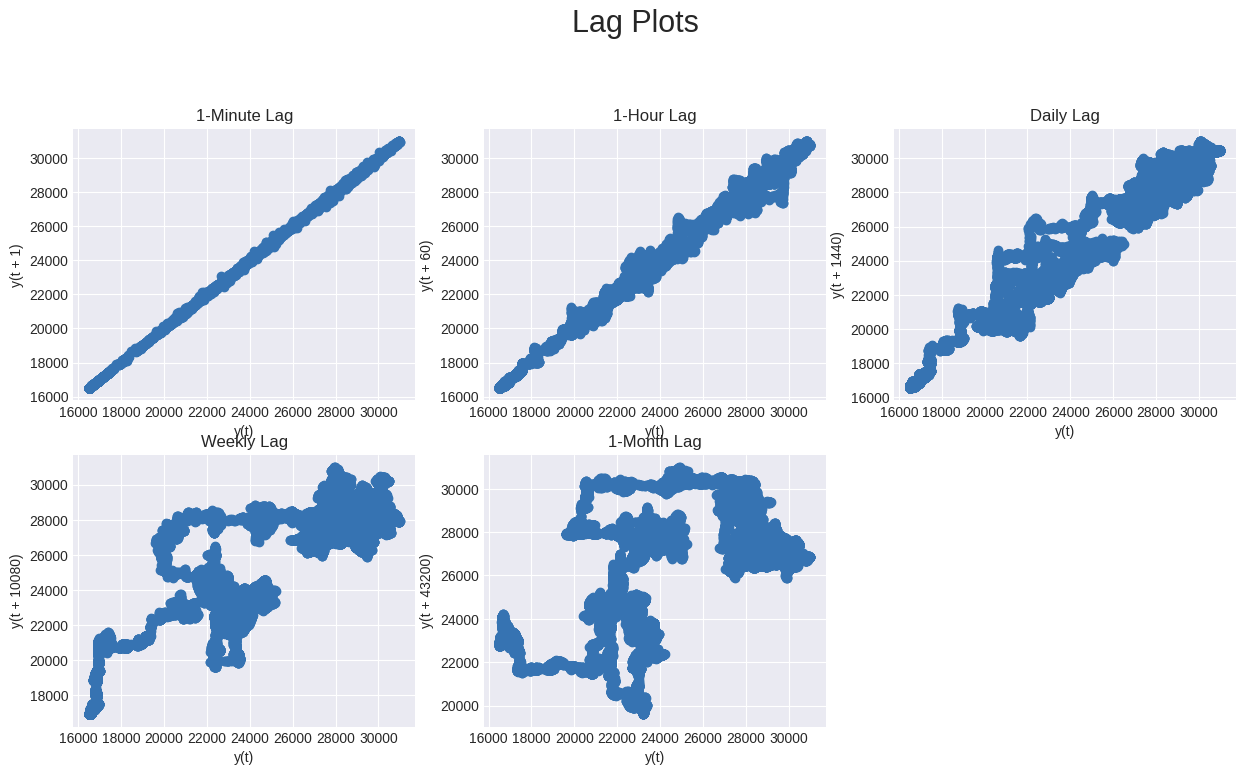


Figure 2: Features represented in the data

Lag plots are employed to examine autocorrelation, particularly when addressing trends, stationarity, and the application of smoothing functions. They play a crucial role in enhancing our comprehension of the data by providing insights and facilitating data analysis.

Figure 3: Lag Plots graph

Upon analysis, positive correlations are observed in the lag plots for minute, hour, and daily intervals. However, the correlation substantially diminishes when considering weekly lag plots, and there is no correlation evident in the monthly lag plots. Consequently, it is advisable to resample the data at a maximum frequency of the daily level. This approach ensures that the autocorrelation is preserved while mitigating the decreasing correlation observed in longer time intervals.

3. Analysis Type(203)

For my research Univariate analysis was the most appropriate because Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own. Descriptive statistics describe and summarize data.

In my research, I determined that conducting univariate analysis was the most suitable approach. Univariate analysis involves examining each variable within a dataset independently. It focuses on analyzing the range of values and the central tendency of each variable, providing insights into the response patterns associated with the variable. This analysis allows for a thorough understanding of each variable individually. Descriptive statistics play a key role in describing and summarising the data during this process.

Utilising LSTM in combination with sentiment data covering broader topics can enhance investment decision-making for Ethereum and Polygon. Including such sentiment data is essential to prevent overfitting and enhance the models' ability to generalize. However, the LSTM models developed for Bitcoin, even with the incorporation of broader sentiment data, were ineffective in accurately predicting price trends. These models suffered from overfitting issues and exhibited poor generalization capabilities on unseen data.

4. Learning Algorithm Selection(268)

Long Short Term Memory (LSTM)

I have chosen LSTM (Long Short Term Memory) as my primary algorithm. LSTM, which stands for Long Short-Term Memory, is a deep learning concept and a type of Recurrent Neural Network (RNN) that addresses the issue of vanishing gradients. It is specifically designed to overcome the problem of backpropagation errors becoming too small or too large as they propagate through the network. The key advantage of using LSTM is its ability to handle long-term dependencies in time series data. It is particularly effective in scenarios where there are significant time delays between events and when the data contains a mixture of low and high-frequency components. LSTM has been widely used in various fields, including stock prediction, where it has demonstrated higher accuracy compared to other methods. In the context of LSTM, researchers have utilized the algorithm to predict time series data such as stock prices. The goal is to accurately predict the value (y) based on the true value (x), taking into account the long-term dependencies and patterns present in the data.

Linear Regression

This method is employed to analyze the correlation between a dependent variable and one or more independent variables, with the aim of making future predictions. When there is only one dependent and one independent variable involved, it is referred to as simple linear regression. However, as the number of independent and dependent variables increases, it becomes known as multiple linear regression. The relationship between the variables is represented by a straight line on a graph, which is determined through the method of least squares to find the best fit.

5. Model Performance(172)

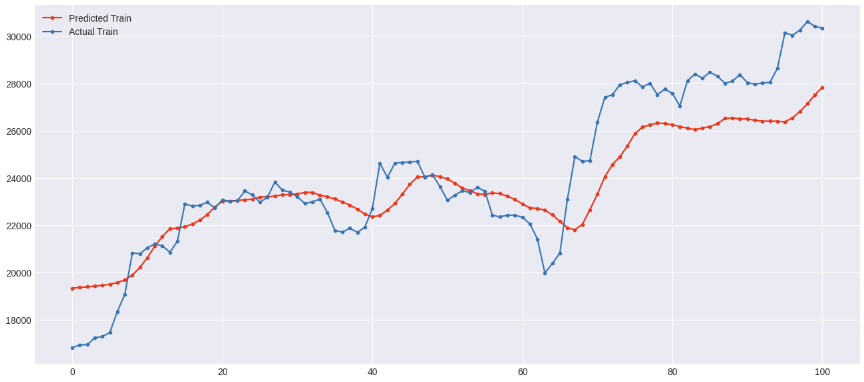


Figure 4: Predicted Train Data

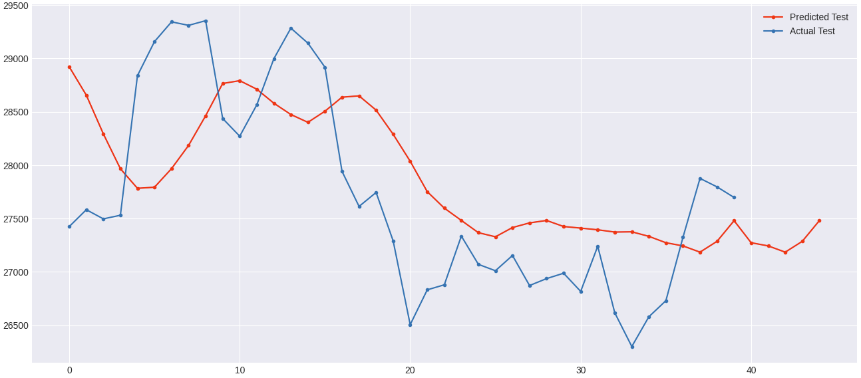


Figure 5: Predicted Test Data

The Test and Train sets have shown a similar pattern in predicting the Bitcoin price.

|  | LSTM | Linear Regression |
| --- | --- | --- |
| Training Data Accuracy | 0.1311 (RMSE) | 0.2333(RMSE) |
| Test Data Accuracy | 0.0506 (RMSE) | 0.1332(RMSE) |

The outputs of two different models are presented: the Machine Learning model, specifically Linear Regression, and the Recurrent Neural Network model, specifically Long Short-Term Memory (LSTM). These models provide distinct outcomes and predictions. The LSTM model's training data accuracy was determined by calculating the Root Mean Squared Error (RMSE) from the model fit. To evaluate the accuracy of the test data, the model predictions were rescaled and compared to the actual values. On the other hand, for the Linear Regression model, the training data accuracy was assessed through prediction validation data. The accuracy of the test data was evaluated by comparing the model predictions to the actual values of the test data. As no rescaling was necessary, the predictions were directly compared to the scaled data. In summary, the accuracy of the LSTM model was determined using RMSE and rescaling, while the accuracy of the Linear Regression model was assessed through direct comparison to the scaled data.

6. Conclusion(238)

The study findings indicated that the LSTM model exhibited higher accuracy compared to the Linear Regression model. This can be attributed to the use of a relatively small dataset in the study. LSTM models are known to perform well when trained on a large volume of data, enabling them to capture long-term dependencies effectively. However, due to the limited dataset used in this study, the Linear Regression model demonstrated poorer performance in predicting the target variable.

The LSTM model demonstrated superior accuracy compared to the Linear Regression model in this study. The LSTM model's effectiveness can be attributed to its ability to capture long-term dependencies, which is especially beneficial for time series data. In contrast, the linear gradient changes in Linear Regression can limit its accuracy when predicting time series data. Therefore, the LSTM model outperformed the Linear Regression model, indicating that it had higher accuracy in this analysis. The review of existing studies revealed several limitations. Firstly, most studies primarily focused on improving performance through the exploration of advanced models and techniques, while neglecting the importance of gathering relevant information that could enhance the results. Secondly, the complexity and non-stationarity of cryptocurrency time series were frequently overlooked in these studies. Additionally, many studies failed to incorporate differentiation or filtering techniques to address the high volatility of cryptocurrency prices. Finally, the majority of studies did not consider the influence of public sentiment, policies, and laws on digital currencies.

7. Reference

Jaquart, P., Dann, D. and Weinhardt, C. (2021) Short-term bitcoin market prediction via machine learning. The Journal of Finance and Data Science. [online]. 7, pp.45–66.

Suneetha Rani R, Sri Vinithri Chowdary D, Uday Kiran C H, Deekshith K and Alekhya V (2022) Bitcoin price prediction based on linear regression and lstm. South Asian Journal of Engineering and Technology. [online]. 12 (3), pp.87–95.

Crypto - BTC : Advanced Analysis & Forecasting’ (no date) kaggle.com. [online]. Available from: https://www.kaggle.com/code/vbmokin/crypto-btc-advanced-analysis-forecasting [Accessed]

Teo, B.G. (2021) Stock Prices Prediction Using Long Short-Term Memory (LSTM) Model in Python The Handbook of Coding in Finance.26 October 2021 [online]. Available from: <https://medium.com/the-handbook-of-coding-in-finance/stock-prices-prediction-using-long-short-term-memory-lstm-model-in-python-734dd1ed6827>.

Maheshwari, R. (2023) Will Bitcoin Rise Again In February 2023? Forbes Advisor INDIA.8 February 2023 [online]. Available from: <https://www.forbes.com/advisor/in/investing/cryptocurrency/bitcoin-prediction/>.

‘Machine Learning in Banking and Finance | Exadel’ (2022) exadel.com.6 July 2022 [online]. Available from: <https://exadel.com/news/how-machine-learning-is-used-in-finance-and-banking/>.

‘univariate statistics’ (no date) home.csulb.edu. [online]. Available from: <https://home.csulb.edu/~msaintg/ppa696/696uni.htm#:~:text=Univariate%20analysis%20explores%20each%20variable>.

Patel, M.M., Tanwar, S., Gupta, R. and Kumar, N. (2020) A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions. Journal of Information Security and Applications. [online]. 55, p.102583.

Huang, X., Zhang, W., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., Liu, Z. and Zhang, J. (2021) LSTM Based Sentiment Analysis for Cryptocurrency Prediction. Database Systems for Advanced Applications. [online]. pp.617–621.

‌Lahmiri, S. and Bekiros, S. (2019) Cryptocurrency forecasting with deep learning chaotic neural networks. Chaos, Solitons & Fractals. [online]. 118, pp.35–40. Available from: https://www.sciencedirect.com/science/article/pii/S0960077918310233 [Accessed 27 February 2019].

‌

‌

‌