Research on prediction of bitcoin price based on machine learning methods

Chenglin Zhao
Shanghai University, Sydney Institute of Language and Commerce, Shanghai, 200444, China

Abstract. Bitcoin, a decentralized digital currency, has gained widespread acceptance and recognition in recent years. The prediction of Bitcoin prices is a challenging task due to its relatively young age and high volatility. Therefore, this study explores the accuracy of price prediction for Bitcoin using machine learning models and makes comparison on the outcome of different models, Linear Regression, Long Short-Term Memory, and Recurrent Neural Network. This study utilizes the closing price of Bitcoin in USD from a Kaggle dataset as the independent variable. The study also adopts Mean Absolute Error (MAE) as the measurement indicators, and comparative performance analysis is conducted under various circumstances. The experimental results demonstrate that LR performs poorly in Bitcoin price prediction, while LSTM and RNN outperform LR. Further analysis reveals that LSTM performs better during price apexes, while RNN performs better during price recessions. Graphical representations illustrate the strengths and weaknesses of each model under different market scenarios. Through comparison, the article provides an insight for other researchers to choose corresponding machine learning models under different circumstances to predict bitcoin price.

1 Introduction

Bitcoin, a fully digital currency, creates a novel payment system as a consensus network [1]. Powered by its users, it peers into a payment network that does not require a central authority to operate. In 2008, Satoshi Nakamoto first introduced the concept of Bitcoin in his research of A completely peer-to-peer version of electronic cash system. So far, bitcoin has become one of the most famous and acceptable cryptocurrencies. By utilizing a digital currency, individuals can exchange value electronically without the need for third-party supervision [2]. Unlike gold, Bitcoin lacks inherent value since it cannot be transformed into tangible items such as valuable jewelry. However, Bitcoin holds value because users believe they can acknowledge it as one way of payment and later utilize it for their desired purchases or necessities [3]. Bitcoin demonstrates a unique prospect for academic research on price forecasting for its relatively recent inception and consequent high instability, surpassing that of traditional fiat currencies.

In contrast to the well-established time series predictions in other financial markets like stocks, Bitcoin's predictions are less developed and receive less research attention due to its current transitional phase. Conventional prediction tools for time series data normally depend on linear regression and demand data that can be divided into visualized trends and periodic patterns [4]. These approaches are more suited for tasks like forecasting sales that involve seasonal effects. However, the Bitcoin market lacks significant seasonality and exhibits high volatility, rendering these methods less effective for predicting its behavior.

Owing to the complicacy of the relative research, deep learning methodology emerges as an intriguing scientific remedy due to the success in associated domains [5].

The paper aims to research on the accuracy of price prediction of bitcoin by varied machine learning methodologies and compare the difference when adopting differentiated machine learning models. To test the effect of different models on time series predictions, three models, respectively Linear Regression, Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN), will be utilized. Rather than focusing on one external factors of bitcoin, the prediction will mainly focus on its historical data, especially its previous price to make related predictions. When comparing the performance of different models, mean absolute error will be imperative index and graphs that contain true value and value predicted will be demonstrated to further compare its performance under differentiated circumstances. The discussion presents a concise framework of three machine learning models and help enhance the accuracy of bitcoin price prediction for investors and researchers.

The remaining of this paper is structured as follows. Section 2 reviews the previous literature. Section 3 introduces how data is collected and which methods are involved. Section 4 demonstrate the outcomes of data analysis. Section 5 makes a conclusion based on the above research. Section 6 involves all the literature cited in this article.

2 Literature review
Related research has been conducted, especially in the field of blockchain, underlying base technology behind bitcoin, and price prediction with different factors concerned. In terms of blockchain, Guo and Liang project that the blockchain technology can be applied into the operation of bank to improve the efficiency of banking [6]. Dujak and Sajter invoke that supply chain management might be easier with the application of the blockchain due to an expanded concept of technologies that use distributed ledgers [7].

As for bitcoin price prediction, it has been attached increasing attention among its breaches in recent years for its market potential. Georgoula and colleagues utilized support vector machines (SVM) in conjunction with the number of online views from Wikipedia and the hash rate to conduct computerized analysis in order to identify the factors that may influence Bitcoin price. [8]. Similar works had also been done by Matta et al, which focus on finding the correlation between Bitcoin price and online popularity of Bitcoin on the social media [9]. Apart from using sentiment from social media, some studies also suggest that transaction graphs can to some extent reflect the trend of bitcoin price. With visualization, Greaves and his coworkers used the bitcoin market trade trends to forecast the Bitcoin price by Linear Regression and other three machine learning methods [10].

Based on the existing research on bitcoin price prediction, this article focuses on comparing the accuracy of different machine learning algorithms and aims to enlighten future studies concerned.

3 Data and Methodology

3.1 Data Collection

Dataset of this article is Bitcoin Historical Dataset from Kaggle website [11]. It contains daily price data of Bitcoin from November 2014 to December 2021, which approximates to 2700 samples, not involving NaN values. The columns collected are unix, date, symbol, prices of four periods, volume BTC, and volume USD. Figure 1 shows the closing price of Bitcoin from 2012 to 2021. According to the figure, after a 3-year plateau phase, the bitcoin price upsurged to approximately 20 thousand dollars in 2018. From 2018 to 2020, low-degree swings had occurred, while 2021 witnessed the rapid growth of bitcoin price, more than 60 thousand dollars. A recorded high occurred in 2020 but ending with drastic swings.

![Fig. 1. Bitcoin Price Trend](image)

3.2 Data Pre-Processing

Several pre-processing procedures are conducted so that this Bitcoin historical dataset is cleaned, embodying feature selection, feature creation and time series train test split.

Although the dataset comprises many features, this article will utilize exactly one column for feature creation, close price. The close price column will be converted into extra seven feature columns that demonstrate previous nine-day price, but not involving the previous two-day price, which aims to predict a longer period price trend.

Due to time-series data, the article applies time-series split to split train and test set, instead of train test split.

3.3 Machine Learning Methods

Based on the literature review, 3 machine learning method will be chosen to help predict bitcoin price, respectively Linear Regression (LR) model, Long Short-term Memory (LSTM) model and Recurrent Neural Network (RNN) model. To prove the point provided by the previous academic papers that LR does not perform well in the prediction of time-series data, LR will be tested and made comparison with other two models.
3.3.1 Linear regression

Linear regression is a scientific mathematics method employed to examine the correlation between a dependent variable and one or multiple independent variables. The main purpose of this model is to determine the most accurate line that minimizes the discrepancy between real values and values forecasted by the linear model [12]. Linear regression can both be a single independent variable or extended to multiple independent variables, known as multiple linear regression. Figure 2 shows linear equation of a basic Linear regression. \( Y \) represents the predicted value. \( B \) reflects the coefficient of factors \( X \).

\[
Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon
\]  

(1)

3.3.2 Long Short-Term Memory

LSTM addresses the challenge of the vanishing gradient problem encountered in standard RNNs. It excels in handling and predicting data sequences due to its specific architectural design. The main advantage of LSTMs is their ability to capture long-term dependencies in sequential data by incorporating memory cells and gates [13]. LSTMs have a complex structure and consist of three key components, memory cell, forget gate and input and output gates. LSTM models undergo training through a technique known as backpropagation through time, enabling the gradients to be propagated from the output layer back to the initial time step. This process facilitates the learning and parameter adjustment of the model based on the sequential data. Figure 2 demonstrates the framework of LSTM.

4 Experimental Results and Analysis

4.1 Prediction results by 3 models

Table 1 represents the prediction results by 3 models utilized. For each predication by models, the first five elements have been listed below for observation and comparison. Due to tedium of prediction set, further analysis will be adopted by indicators in Section 4.2 to provide a more unambiguous comparison.

<table>
<thead>
<tr>
<th>Model</th>
<th>LR</th>
<th>LSTM</th>
<th>RNN</th>
<th>Close</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19327.2</td>
<td>19999.8</td>
<td>18598.8</td>
<td>22814.2</td>
</tr>
<tr>
<td>2</td>
<td>19489.0</td>
<td>19833.5</td>
<td>18753.4</td>
<td>23136.1</td>
</tr>
<tr>
<td>3</td>
<td>21445.0</td>
<td>21061.7</td>
<td>19920.6</td>
<td>23843.0</td>
</tr>
<tr>
<td>4</td>
<td>22806.2</td>
<td>22337.3</td>
<td>21099.0</td>
<td>23470.9</td>
</tr>
<tr>
<td>5</td>
<td>23244.2</td>
<td>23004.9</td>
<td>21752.7</td>
<td>22732.0</td>
</tr>
</tbody>
</table>

4.2 Mean Absolute Error of Three Prediction Models

The article mainly adopts mean absolute error (MAE) as the measurement indicator. Through comparison, it is obvious that the performance of linear regression is the worst, and its mean absolute error is more than 2000 and it also proves one of the perspectives in previous literature that the linear regression does not fit well in time series prediction. Meanwhile, the error of LSTM is the lowest, better than that of RNN, whose error is lower than 300. According to Table 2, the MAE of LSTM is 290.3. The MAE of RNN is around 313, which ranks at the second place and that of LR is 2476.9.
Table 2. MAE of Prediction Models.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>2476.9</td>
<td>3</td>
</tr>
<tr>
<td>LSTM</td>
<td>290.3</td>
<td>1</td>
</tr>
<tr>
<td>RNN</td>
<td>313.4</td>
<td>2</td>
</tr>
</tbody>
</table>

4.3 Comparison on Performance of LSTM and RNN

Some performance details of LSTM and RNN will be demonstrated in the following graphs. As is shown in Figure 4, LSTM does not perform well in recession of bitcoin, especially freefall. When bitcoin price decreases rapidly, the model could not catch up with this trend, resulting in drastic loss and error during prediction.

Fig. 4. Prediction of LSTM

On contrary, RNN performs better at that circumstance. However, as is shown in Figure 5, RNN does not fit well in the rapid growth, especially around its apex. When bitcoin price increases drastically, the model is not able to follow the trend, leading to large gap between predictions and real numbers.

Fig. 5. Prediction of RNN

5 Conclusion

This paper predicts bitcoin price in three machine learning methods and make comparisons. It finds that when predicting time-series data, RNN and LSTM perform better than Linear Regression, which is similar to the results of other studies related to price prediction. However, through observation from two figures, RNN performs better when it is around the recession and LSTM performs better when it is around the apex.

The conclusions of this article include the following implications. Firstly, two algorithms LSTM and RNN should be used to predict the bitcoin price and correspondingly employ these methods depending on whether it is around apex or recession. Additionally, researchers can also focus on other virtual coins that involve financial value to make price prediction based on a clearer framework. Therefore, conclusions
contribute to the research of price prediction and identifying accuracy difference of three models under time-series dataset.

However, there are several limitations. Initially, for future research, more features can be added. It is recommended that hash rate, popularity of Bitcoin, gold price can be added as features in correlational research.

References

1. S. Nakamoto, Decentralized business review, 21260 (2008)